Referee's Report

JCGS-20-156

Hole or grain?

A Section Pursuit Index for
Finding Hidden Structure in Multiple Dimensions

Summary

There are two main ideas here.

The first is the notion of comparing the density of an interior section to that outside the section. Differences may be interesting, particular in detecting holes. The idea of density comparison between two datasets comes from Gous and Buja (2004) and is adapted to comparing interior and exterior section projections. This is a novel and interesting research direction.

The second idea idea is that or radial binning. This seems to appear for the first time in this paper (no reference in the paper; search turns up very little). While relatively straightforward, it should be given a formal treatment separate from its application in any projection/section tour.

The review below is lengthy for two reasons. First, the ideas in the paper are interesting and complex enough that they merit more formal treatment. To give a more complete sense of what is meant, Second, the case presented for the methodology is in many places unconvincing. These relate to the curse of dimensionality (which is not compellingly addressed in the paper) and to the informality (including occasional carelessness) of the arguments presented.

The first few sections are critical comments about the presentation and questions about the proposal and reasoning around trimming the data outside a sphere. Towards the end of the review, this spherical trimming is revisited with a sketch discussion of all that I think needs to be said about justifying its use.

The sections which follow, beginning with "Radial binning" include critical review but are largely more formal presentations/development of the ideas underpinning the current manuscript. I could see no other way to make it clear what is intended by "more formal", or how the ideas related to radial binning might be formally separated from their use in sectioning, than to do so by example. To that end, the radial binning is treated in some detail.

The paper's discussion of the generalised index is already formally separated from its use. Some details here could have been critically assessed but have not been given attention due to the length of this review already. The criticisms, if any, would have been minor.

The ideas on radial binning and the generalised index are interesting whether or not they actually perform well in practice for sectioning. I remain somewhat skeptical that they do perform well, or are competitive with other simpler methods. The conclusions section would benefit from an honest assessment of the strengths *and weaknesses* of the proposed methodology.

The reader, and follow-on researchers, would benefit from a re-organization of the paper. As I see it, the organization might be something like this:

1. Introduction
2. Motivation
   (a) Grand tour and slice tour
   (b) Projection pursuit and guided tour
   (c) An index for comparing two sections
   This would essentially be only the current sections 3.1 and 3.2.
3. Radial binning

   This would cover the topics formally without reference to sectioning for the most part, much like, and in much the same order as, the material developed in the review below.

   By, and towards, the end of this section, the effect of sectioning on radial binning and their inter-relationships should have been covered.
4. A generalised index

   Current Section 3.6 and Section 4 combined
5. Application
   (a) Radial binning in practice
   Includes various choices and recommendations based on practice and experience (e.g.resolution, scaling, number of bins, choice of reference distribution, . . .)
   (b) The index in practice
   All choices and recommendations based on practice and experience would be discussed as well as how to use and interpret. E.g. which norm, choosing weights, etc.
   (c) Example (optional): some artificial geometric example,
   perhaps one based on equality boundaries (ideally statistically motivated; physics applications could be referenced then without discussion) or maybe some geometric structure from Laa, Cook, and Valencia (2019)
   (d) Example: classification boundaries.
   The olives data set
   The physics example would NOT appear. There is already significant amount of material in the paper and the physics example distracts from it.
6. Conclusion and discussion

   This should include an honest assessment of the limitations of the approach.

Again, throughout the mathematics should be formal and precise.
The ad hoc recommendations based on experience should appear in the Application section.

Notation and nomenclature

Consistent and, where possible, standard, mathematical and statistical conventions for notation and nomenclature should be followed throughout. For example,

- **Orthogonal projection matrix**: *Projection* matrices are square and idempotent. If they are also symmetric/Hermitian, then they are *orthogonal projection* matrices.

  Page 5, line 9. The $p \times d$ matrix A is therefore *not* a projection matrix, though it has orthonormal columns which provide *one* orthonormal basis for the space being projected into. The orthogonal projection matrix is the $p \times p$ matrix AA^{T} which projects any $p \times 1$ vector orthogonally into the d-dimensional space spanned by the columns of A. In terms of the columns of A as a new axis system (orthonormal basis) for the subspace, the *coordinates* of the points are given by $Y = XA$.

  In the original (p-dimensional) coordinate system, the projected points are the rows of XAA^{T} and the $n \times n$ matrix $H = [h_{ij}]$ defined by

  $$H = X(I_p - AA^{\mathsf{T}})X^{\mathsf{T}} \tag{1}$$

  has as its diagonal element h_{ii} the height $h_i = h_{ii}$ above the d-dimensional column space of A for the ith data point $\mathbf{x}_i$ (i.e. the ith row $\mathbf{x}_i^{\mathsf{T}}$ of X).

  This would seem to be a more careful notation like that found in Laa et al. (2019) (although there too A is incorrectly called a projection matrix).

- **Notation for cdfs and pdfs**: Write upper case Latin letters for a cumulative distribution function (e.g. $F(\cdots)$) and lower case Latin letters for its density function (e.g. $f(\cdots)$).

  For example, the left hand side of **equation (4) on page 7** would be more conventionally (and hence more clearly) written as $F(r; p, R)$, not as $f(r; p, R)$.

- **Notation for unknown parameters**. Convention designates unknown parameters by Greek letters. The radius of the sphere to be used for the uniform calculations is unknown and determined from the data.

  This suggests that the Greek ρ and the Latin R be swapped throughout the manuscript. So in the Appendix A, R would be the variable of integration and ρ would appear in the limit. The cumulative distribution function of Equation (A4) would now be

  $$F(r; p, \rho) = \frac{V_{inside}(r, p, \rho)}{V(p, \rho)} = 1 - \left(1 - \left(\frac{r}{\rho}\right)^2\right)^{p/2}.$$

  The general formula for the volume of a hypersphere appearing in Appendix A could be defined with r in place of R. More importantly, when it came time to determine a value for ρ (the radius of the sphere used to trim the data in the proposed algorithm), the standard notation $\widehat{\rho}$ could be used to make it clear that this was a "statistic" based on the observed data used in place of the unknown parameter ρ.

- **Sections: slices or shells?** It might be best to avoid the term **slice** throughout the paper to define the various sectionings of the space orthogonal to the target d-dimensional space.

  The word "slice" suggests a linear space (as in a slice of cheese or salami) running parallel to the target space. This perfect when $d = p - 1$ but gives the wrong (or at least incomplete) mental picture otherwise.

  When $p > d + 1$, any and all directions orthogonal to the target space are used to define a section. It is the distance r from the target space that defines the section. The sections are hyper-spherical shells within the $p - d$ dimensional space (and centred at its origin) which is orthogonal to the target space.

  This distinction should be pointed out to the reader and, to be geometrically evocative, the word **shell** should replace "slice" (nearly) everywhere in the manuscript. The word **section** should be reserved for the *general* idea of sectioning (whatever the shape might be) and **slice** reserved for those sections which are linear manifolds.

  Also reserve a single notation for section throughout. A suggestion is to use S_i for the ith section, shell, or slice. That would mean a different letter is needed for the bin counts S_k and s_k (page 5 on) – perhaps O_k (o_k) for outer and C_k (c_k) for central?

  "Slice radius", for example, should everywhere be replaced by "shell radius".

  Alternatively, "layer" might be used in place of "shell" (both sandwiches and onions have layers after all). Then L_i could denote the ith layer.

Rotational invariance and spherical trimming

On page 6, first paragraph of Section 3.3, a case is made for using indices that are invariant to rotation in the d-dimensional projected subspace. This makes eminent sense, but I do not see why this implies a trimming of the data to fit within a hypersphere of some radius R.

All points are orthogonally projected into the column space of A, denoted $colsp(A)$. The choice of coordinate system (i.e. basis) for $colsp(A)$ is arbitrary and so should not affect the projection index. That is, if the $p \times d$ matrix A is everywhere replaced by a $p \times d$ matrix $B = AO$ for any $d \times d$ orthogonal matrix O there should be no change in the index, since the columns of B also form an orthonormal basis for $colsp(A)$.

The index depends on identifying which points appear within which bins when projected and on the orthogonal distance of each original point from its projected position. From equation (1) above, it is clear that replacing A by $B = AO$ will not change H and so will not change any of the orthogonal distances. The challenge then lies with the definition of bins in $colsp(A)$.

In the special case $d = 2$, defining bins by polar coordinates seems reasonable. In the limit (i.e. as the number of equi-angular divisions $K_\theta \to \infty$), the bins will not depend upon the basis for $colsp(A)$. For finite K_θ, they will. In practice, one imagines that K_θ is large enough that the rotational invariance is nearly effected. The effect of actual lack of rotational invariance will be felt most, closest to the origin where bins are smaller and points might change bins more readily.

It is not clear what necessitates the original points being trimmed to a sphere; rotational invariance in the projected plane does not seem to require it. A clear case for trimming to a sphere seems not to have been made in the text.

Of course, if the points were within a sphere of known radius, then the mathematics guiding the various choices (based on uniform distribution within a hypersphere) is fairly clear cut. If this convenience is the principal reason, then it should be stated as such. Otherwise some other, more compelling, argument is needed for the trimming.

Spherical trimming: costs and benefits

As indicated above, the most troublesome procedure in the proposed index is the spherical trimming of the data prior to analysis. Even supposing a compelling argument for the doing so, some separate discussion of the cost of the trimming.

On **page 6, lines 43-47**, the spherical trimming is "harmless" either "because it is interior structure that is of interest" or because the data have "reasonably small p so that the vast gap between spheres and cubes in high dimension is not a concern." These points merit elaboration.

For example, consider the relative volume in the corners of a p dimensional hypercube of width $2R$ within which is inscribed a hypersphere of radius R touching the cube on every face. Expressed as a function of p the proportion of the hypercube volume occupied by the hypersphere is

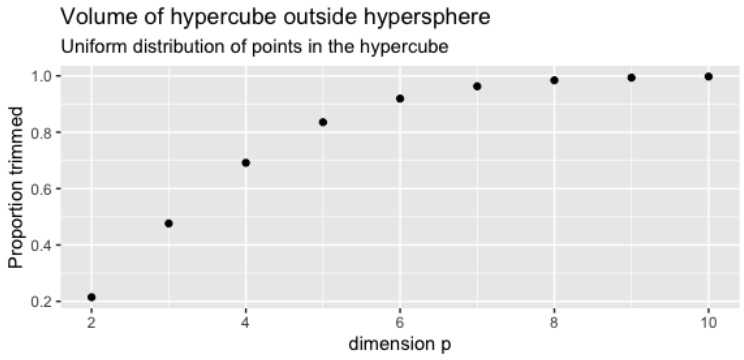
$$\frac{R^p \pi^{p/2}}{\Gamma(1+\frac{p}{2})} \div (2R)^p = \frac{\pi^{\frac{p}{2}}}{2^p \Gamma(1+\frac{p}{2})}$$

giving the proportion lying in the corners of the hypercube outside the sphere to be

$$1 - \frac{\pi^{\frac{p}{2}}}{2^p \Gamma(1 + \frac{p}{2})}.$$

This, then, is also the proportion of points which, when uniformly distributed in the hypercube, is *expected* to lie outside the hypersphere and so *be removed from the dataset by spherical trimming*.

The plot below shows this proportion as a function of the dimension p:



For only $p = 3$, nearly half the points are expected to be trimmed; for $p > 4$ more than 80% are expected to be removed! Spherical trimming seems like it could be fairly radical.

It is not at all clear what would be a "reasonably small p so that the vast gap between spheres and cubes in high dimension is not a concern". Such a claim requires much more discussion to be believed.

The other claim is that this doesn't matter so much because it is the "interior structure that is of interest". For $p = 5$ this is only about than 16% of the data; for $p = 7$ it is less than 4%. If p gets only as high as 10, it is only a quarter of one percent of the data.

When data dimensions must be so low, it is not clear that this projection pursuit method has nearly as much to offer as does interactive brushing, masking, or slicing of the data. Such interactive methods have the further virtue that they are *not restricted to a central slice containing a small proportion of the data*. Any region of the data can be explored for holes, *including the corners*.

Radial binning

Radial binning is an interesting contribution of the manuscript and merits slightly deeper discusstion. Other researchers/JCGS readers could get more out of the idea, if a more general exposition were given in the paper. Below, is an exposition (which could be shortened considerably in manuscript) that separates the several key ideas of the binning and deals with the problem more generally, before introducing the particular choices that might be made in practice.

These choices would be problem specific such as those made in the paper to address the problem of comparing projected densities between central and exterior sections.

In making such choices, the exposition below introduces a reference distribution that guides the choices. In the present manuscript this is always that of spherically uniform data projected onto a 2D space. However, more generally, any distribution could be chosen.

The radial distribution

To that end, let $F_d(r)$ be the marginal cumulative distribution function for radii (i.e. $r = ||\mathbf{x}||$ for $\mathbf{x}$ in the d-dimensional space) a hypersphere in d dimensions.

Note that

- no assumptions are *yet* being made about the distribution F other than it is radial.

  It might have resulted from projecting data uniformly distributed in a hypersphere of dimension $p > d$, or it might not. It doesn't matter *yet*.

- r is not *yet* required to be less than some constant R; so far only $0 < r < \infty$.

- the dimension d is not *yet* required to be $d = 2$.

With the above, we can imagine *radial bins* that are complete *shells* when $d > 2$ or that are complete (annular) *rings* when $d = 2$. Whatever the dimension, choose radii $r_{i-1} < r_i$ with $r_0 = 0$ for $i = 1, \ldots, K_r$, and define the ith **radial bin** to be the shell/ring S_i from r_{i-1} to r_i. The probability associated with the ith radial bin is then simply

$$Pr(S_i) = Pr(\mathbf{x} \in S_i) = Pr(r_{i-1} \leq ||\mathbf{x}|| \leq r_i) = F_d(r_i) - F_d(r_{i-1}) = f_i, \text{ say.}$$

For a sample of N points, f_i is the *expected fraction* of points to fall in S_i and $N \times f_i$ the *expected number*. Trivially straightforward from the Binomial distribution.

One could pause there and simply use radial bins for detecting unexpected structure (holes or otherwise; with or without projection).

One could choose **equal depth shells** $r_i - r_{i-1} = \delta$ to produce a histogram of the data. If $F_d(r)$ is known, the histogram could be compared to the expected counts and even a goodness of fit test conducted. Alternatively, the radii could be chosen to have **constant probability** $Pr(S_i)$ (i.e. being closer together nearer the origin). The resulting histogram would then be comparable to a uniform distribution. Finally, one might even choose radii to have **equal volume** bins (i.e. being closer together as r increases) to match equal bin sizes in the plane rather than equality along a radial ray. Why not?

Maintaining a more general exposition leaves these possibilities open.

The directional distribution

Having the marginal distribution of the radius $r = ||\mathbf{x}||$ in the d-dimensional space, what remains is the conditional distribution of $\mathbf{x}$ given its length is r. If this conditional cumulative distribution function is denoted

$$G_d(\mathbf{x} \mid ||\mathbf{x}|| = r)$$

so that the cdf of $\mathbf{x}$ in the d-dimensional space is

$$G_d(\mathbf{x} \mid ||\mathbf{x}|| = r) \times F_d(r).$$

The key elements determining the distribution of $\mathbf{x}$ are the distribution of the length of $\mathbf{x}$ and the conditional distribution of its *direction* given its length. **Note** the latter *can change with the length* r since $G_d(\mathbf{x} \mid ||\mathbf{x}|| = r)$ is, in general, also a function of r.

If $d = 2$, and all that is of interest is the distribution of the direction of $\mathbf{x}$, then a histogram having K_θ wedge shaped bins A_j for $j = 1, \ldots, K_\theta$ could be used. Here, each A_j would be defined by rays extending from the origin at angles θ_{j-1} and θ_j with $0 = \theta_0 < \theta_1 < \cdots < \theta_{K_\theta} = 2\pi$ radians. The paper takes the difference in angles to be a constant:

$$\theta_j - \theta_{j-1} = \frac{2\pi}{K_\theta} \text{ radians}$$

to produce equiangular bins A_j. (Note that use of the notation A_j here would require a change for the matrix A in the paper to some other letter, e.g. P.) Any loss of rotational invariance due to this binning is mitigated somewhat by choosing equiangular bins and large K_θ.

But it is the conditional distribution of $\mathbf{x}$ for a **particular** $r = ||\mathbf{x}||$ which is of interest. This suggests narrowing consideration to wedges taken only from a ring (or more generally shell) that contains the radius r.

Suppose the radial bin B_i contains r for some $i \in \{1, \ldots, K_r\}$. Then the bins of interest for the conditional distribution of the direction of $\mathbf{x}$ given $||\mathbf{x}|| = r$ are

$$B_{ij} = B_i \cap A_j$$

$j = 1, \ldots, K_\theta$. Again, the A_{ij}, and hence B_{ij}, need not all have the same area/volume for all j.

The joint distribution

The histogram bins appropriate for the joint distribution are **all** bins $B_{ij} = B_i \cap A_j$ for all $i = 1, \ldots, K_r$ and all $j = 1, \ldots, K_\theta$.

When $d = 2$, these bins would produce histograms like those seen in Figure 3 from the paper (but absent any thought of projection or slicing as *yet*). For $d > 2$, instead of a single angle determining the directional distributions, $d - 1$ separate angles would be required.

Choosing a reference distribution

The binning as described above follows the distributions $F_d(\cdot)$ and $G_d(\cdot)$, whatever they may be.

Specified, these distributions become references against which the observed data could be compared. As the paper suggests, a reference distribution would be chosen to be uninteresting so that large differences with the data suggest the data might be interesting.

It is easy to imagine an uninteresting distribution for the conditional distribution G_d of $\mathbf{x}$ given $||\mathbf{x}|| = r$. First, it should not depend on the value of r that is conditioned on. So only the direction of $\mathbf{x}$ matters. Second, there should be no preference in direction. The direction vector should be uniformly distributed on the surface of a d dimensional hypersphere. For $d = 2$, this means that the angle θ is uniformly distributed on $[0, 2\pi)$.

Any spherically symmetric multivariate distribution for $\mathbf{x} \in \mathbb{R}^d$ including the spherical multivariate normal $N_d(\mathbf{0}, \sigma^2\mathbf{I}_d)$ and the uniform distribution within a sphere (of suitably large radius) as that used in the manuscript.

When $d = 2$, the conditional uniform distribution assumption $\theta \sim U[0, 2\pi)$ gives

$$Pr(\mathbf{x} \in B_{ij} \mid \mathbf{x} \in B_i) = \frac{1}{K_\theta}$$

and hence unconditionally that

$$Pr(\mathbf{x} \in B_{ij}) = \frac{f_i}{K_\theta}.$$

54 It follows that the *expected number of points inside* B_{ij} is

$$\frac{N}{K_\theta} f_i.$$

This simple result is what produced equation (9) in the paper.

As to what makes for an uninteresting marginal distribution for $F_d(r)$, there are a multitude of choices available. For example, $F_d(r) = \sigma\sqrt{\chi^2_d}$ (for the uninteresting $\mathbf{x} \sim N_d(\mathbf{0}, \sigma^2\mathbf{I}_d)$). Or if, as in the paper, $\mathbf{x}$ is uniformly distributed within a d-dimensional hypersphere of radius ρ, then $F_d(r) = \left(\frac{r}{\rho}\right)^d$ for $0 \leq r \leq \rho$.

The treatment could be much more general, assuming only that the reference distribution for $\mathbf{x}$ is spherically symmetric and so has some density that is a function only of $r = ||\mathbf{x}||$, say $h(r)$.

Applying the projection

The distributions described above take no account of the fact that a projection is to be made into a lower dimensional space d. In this case, we begin with a $p > d$ dimensional distribution, for $\mathbf{x} \in \mathbb{R}^p$, that is "uninteresting".

For the spherical normal $N_p(\mathbf{0}, \sigma^2\mathbf{I}_p)$, projecting the density onto a d-dimensional subspace yields a d dimensional spherical normal and hence $F_d(r) = \sigma\sqrt{\chi^2_d}$ as before.

For $\mathbf{x}$ distributed uniformly within a p dimensional sphere and then projected onto the d dimensional space, the cdf for r is given in Appendix A when $d = 2$. Why not determine the

result for arbitrary d? It will also be straightforward, involving $V(p-d, \sqrt{R^2-\rho^2})$, in place of $V(p-2, \sqrt{R^2-\rho^2})$, and $d-1$ angles (one imagines with $(2\pi)^{d-1}$ appearing in place of 2π)). The integral should be just as easily determined as that of equation (A2).

This begs a more general question. Why not decompose the density of *any* spherical distribution into the product of its marginal radial and its conditional directional components? Should that really be beyond the scope of this paper?

Effect of sectioning

Two sections are defined based on a radius h. Orthogonal to the d-dimensional projection space, two nested spherical shells S_1 and S_2 are defined as

$$S_1 = \{\mathbf{x} \,:\, 0 \le ||\mathbf{x}|| < h \,,\, \mathbf{x} \in \mathbb{R}^p\} \quad \text{and}$$
$$S_2 = \{\mathbf{x} \,:\, h \le ||\mathbf{x}|| < R \,,\, \mathbf{x} \in \mathbb{R}^p\} \,.$$

These shells are the inner and outer sections, respectively, whose data are to be projected into the d dimensional space.

Presence in either shell is a conditioning event for $\mathbf{x}$, so one would like to know the conditional probability of $\mathbf{x} \mid \mathbf{x} \in S_i$ for $i = 1, 2$ as well as in each case, the marginal d-dimensional probability of the conditional after projecting onto the d-dimensional space.

This can, in principle at least, be defined for any (uninteresting) reference distribution for $\mathbf{x}$ and any value of $h \in [0, R)$. It would add value to the paper if these conditional distributions and their projected d-dimensional marginals, were determined for at least the two "uninteresting" reference distributions specified above (only when $d = 2$?).

(Knowing these exactly would obviate the need for the approximation which follows and allow larger h and hence smaller overall sample sizes to get sufficient points in the central section.)

Instead, for all practical purposes, provided h is small enough, an *approximate* solution is easily had.

Assuming that h is very much smaller than the *range* of the data in any direction in $\mathbb{R}^p$, the inner section will have thickness of only $2h$ in every direction orthogonal to the d-dimensional target space. This section will be sufficiently thin that the distribution of the projected points from the section will not be *that different* from the distribution of the infinitesimally thin section that is the projected space itself. That distribution will be the d-dimensional reference distribution.

For example, when $p = 3$, $d = 2$, and a uniform distribution within the sphere is the reference distribution, one imagines a $2h$ thick equatorial disc as the central section/slice and the infinitesimally thin equatorial disc as the projection target. The distribution in the equatorial disc is uniform and the projected distribution of the equatorial slice is approximately the same, *provided h is small enough*.

Assuming that the central section is so thin that its projected distribution is not that different from that of the d-dimensional space itself, suggests that removing it from the orthogonal $p-d$ dimensional space to which it belongs will have little effect on the distribution in that space. Consequently, the distribution of the outer section projected onto the d-dimensional space is approximately the same as if the mass of the whole $p-d$-dimensional space were projected into the d-dimensional space.

The same argument holds for any spherical reference distribution, not just the uniform hypersphere but also, say, $N_p(\mathbf{0}, \sigma^2\mathbf{I}_p)$. The question is how small must h be to (a) effect these approximations and (b) ensure that enough points remain in the central slice.

To answer the first depends on the reference distributions and knowing the exact conditional distributions of $\mathbf{x} \mid \mathbf{x} \in S_i$ for both $i = 1$ and $i = 2$. This is not done in the paper.

To answer the second, the expected number of points in S_1 is determined as a function of h, using the reference distribution. This is achieved by determining the probability of S_1 given the reference distribution and multiplying by N.

For the spherical uniform, Equation (8) on page 10 purports to give this expectation (and probability after dividing by N. The same could be determined for the spherical normal.

The result given as equation (8) is not obvious (to me at least) and **needs to be derived**:

- it does not appear to be using equation (A4) of the Appendix
- it does not appear to be based on the thin "disk" approximation since it gets the wrong answer when $p = 3$ (should yield approximate probability $\frac{3h}{2R}$ when $p = 3$ and $h << R$)
- it should be described in terms of the probability and/or the **expected** number of points and not solely as "the number of points ... is"
- using the disc approximation with relative volume $disc \div sphere$ of
  - $V(p-2, h) \times V(2, R) \div V(p, R)$ gives $\frac{p}{2}\left(\frac{h}{R}\right)^{p-2}$
  - $Cube(p-2, h) \times V(2, R) \div V(p, R)$ gives $\left(\frac{2h}{R}\right)^{p-2} \frac{p}{2} \frac{p-2}{2} \Gamma(\frac{p-2}{2}) / \pi^{\frac{p-2}{2}}$

  both of which give the correct approximate answer of $\frac{3h}{2R}$ when $p = 3$ but neither is the same as equation (8).

I suspect I have missed or misunderstood something in spite of what I think has been a fairly careful study.

Some care should be taken in denoting sample sizes via N and its various decorated versions. In this review, the text around its use should indicate whether it is referring to the overall sample size (before or after trimming the corners from the data), whether it is referring to the count in a section S_1 or S_2, whether it is referring to the count in a radial bin B_i or an equi-angular bin B_{ij}.

Considering the left side of equation (8) on ms. page 11, there N_S denotes the expected number of points in S_1. As such, it might be better denoted as N_{S_1} to emphasize this fact. Similarly, N_i could denote the count in the radial bin B_i, N_{ij} in the equi-angular bin B_{ij} with the surrounding texts indicating whether these counts are conditional on S_1 or S_2 or neither. The sample size *after spherical trimming* is most convenient as N and to emphasize the number before trimming N_{cube} suggests the count with in the p-dimensional hypercube.

With this notation in mind, and taking equation (8) to be correct, one can calculate N_{S_1} from this equation. The results are worrisome.

A few calculations when $p = 10$ make the point (counts rounded to nearest unit):

p	N_{cube}	N	x	N_{S_1}
10	100,000,000	249,040	0.1	0
”	”	”	0.25	18
10	100,000,000	100,000,000	0.1	5
”	”	”	0.25	7,248
8	572	572	0.1	0
”	”	”	0.25	1
”	”	”	0.50	29
4	572	572	0.1	11
”	”	”	0.25	69
”	”	”	0.50	250

The last six lines have $N = 572$, the size of the olive oil fatty acids data set. First all $p = 8$ variables are used, then only the (judicially chosen?) $p = 4$ used in example of Section 5.2.1. The functions used to get these values were

```
cornerProportion <- function (p) {
    1 - (sqrt(pi)^p) / (2^p * gamma(1 + p/2))
      }

equation8 <- function(Ncube, N, p, res) {
    if (missing(Ncube)) Ncube <- N
    if (missing(N)) N <- (1- cornerProportion(p)) * Ncube
    NS <- (N /2) * res^(p-2) * ( p - (p-2) * res^2)
    list(p = p, Ncube = Ncube, N = round(N), res = res,
         NS = round(NS))
}
```

If these calculations are correct, then it is difficult to understand the recommendation of resolution $x = 0.25$ on page 19 and the working of the example in Section 5.2.1. Values for N_{ij} would necessarily be near zero.

If equation (8) is constructed from the approximation using a thin “disk” as S_1, then one would expect the resolution $x = h/R$ to be very small for the approximation to hold and certainly not as large as that suggested in practice of $x = 0.25$ on **p. 19, line 11** of the ms. or the $x = 0.50$ considered above for the olive data. Of course, if it is calculated exactly without approximation, then any resolution $x \in (0, 1)$ could in principle be considered.

Again, if the dimensionality must be as low as $p = 4$ to get performance, it casts considerable doubt on the methodology compared to interactive sectioning via interactive graphics.

Comparing projected distributions requires matching bin probabilities

The reasoning for reweighting bin counts, as sketched by *Section 3.3.1*, **is not clear** and even confusing in places. For example, equation 7 on page 8 is followed nearly immediately by an apparently contradictory assertion on line 33/34 that $s_i(p, R, K_r) = 1/K_r$. In the text, both statements rely on the uniform hypersphere assumption – the difference is that the former first does a projection, the latter does not. Unfortunately, $s_i(p, R, K_r)$ is *defined* by the former and so the latter is technically wrong (though one might guess at its intended meaning).

What follows is an alternative presentation to justify the reweighting.

The index (defined simply as in equations 1-3 of the manuscript) is based on comparing the estimated densities of two distributions – in this case, the two projected distributions from each section/shell (conditionals given S_1 and S_2)

To compare these, the bins must be identical in shape and location in the projected space for the two distributions. Unfortunately, this means that the probability within any bin will generally differ depending on which section/shell, S_1 or S_2, is being projected (whatever the reference distribution for $\mathbf{x}$).

The reference distribution is chosen to be "uninteresting" so any differences between its two projected distributions should be made negligible. This is done by ensuring that each bin B_{ij} has the same expectation for each of S_1 and S_2.

Suppose that there are N observations to be distributed amongst the K_r radial bins independently with N_i falling in the ith radial bin B_i and $N = \sum_{i=1}^{K_r} N_i$. The probability of an observation falling in B_i has been shown to be

$$f_i = F_d(r_i) - F_d(r_{i-1})$$

where $F_d(r)$ will be a different radial distribution for each section/shell S_1 or S_2 and found after projected the reference distribution of $\mathbf{x}$ onto the d-dimensional space. The observed density in radial bin B_i will simply be the relative frequency

$$n_i = \frac{N_i}{N}$$

(again differing with the section S_1 or S_2) which has *expectation* f_i. So n_i/f_i has expectation 1.

An identical density would assign the same probability, say p_i, to bin B_i regardless of which shell, S_1 or S_2, was being projected ($\sum_{i=1}^{K_r} p_i = 1$). On the density estimate for bin B_i, the relative frequencies n_i would simply be replaced by

$$\frac{n_i}{f_i} \times p_i = n_i w_i$$

a weighted relative frequency with weights $w_i = p_i/f_i$. The weighted relative frequency has expectation p_i, the target probability for B_i. Using the f_i peculiar to each of S_1 and S_2 would make the projected distributions identical for each shell under the reference distribution. Applying the weights to the relative frequencies in the section index (Equations 1-3 in the paper) renders the same bins for S_1 and S_2 comparable on the same scale.

Note that p_i can be any probability. It makes sense to choose p_i to be such that the resulting distribution in the projected d space is an uninteresting type of reference distribution analogous to that used in p dimensions.

For the uniform distribution in the sphere and constant radial difference $r_i - r_{i-1}$, this will be $p_i = 1/K_r$ for all B_i.

For a spherical multivariate normal reference distribution, each p_i will be determined from a $\sigma\sqrt{\chi^2_d}$ random variable; here, the radii could also be adjusted to force $p_i = 1/K_r$ if desired.

Wasn't the preceding page what was meant in Section 3.3.1?

A key ingredient missing in the explication offered in Section 3.3.1 is that the *two* distributions being compared **must be matched** on bin geometry *and* on probability mass in the null (uninteresting) configuration of the reference distribution.

Assuming spherical symmetry of the reference distribution, the relative frequency for the radial equal angular bins B_{ij} will be $n_{ij} = N_{ij}/N$ where N_{ij} is the number falling in bin B_{ij} with probability f_i/K_θ. The corresponding weighted relative frequency

$$n_{ij}w_{ij}$$

with weights

$$w_{ij} = \frac{p_{ij}}{f_i/K_\theta}$$

where p_{ij} is the target reference probability for B_{ij}. Again, for spherically symmetric reference distribution $p_{ij} = p_i/K_\theta$ so that the weighted relative frequencies simplify to

$$n_{ij}w_{ij} = n_{ij}w_i$$

as (implicitly?) proposed in the manuscript.

This gives the expectation of n_{ij} is f_i/K_θ and so that of $w_{ij}n_{ij}$ is $p_{ij} = p_i/K_\theta$ as desired.

For spherically uniform and constant radial difference, $p_i = 1/K_r$ and $p_{ij} = (K_rK_\theta)^{-1}$ so that all bins are treated identically, from the tiniest sliver near the origin to the largest radial arc bin at the outer extremes. On the face of it, this seems a curious choice and one wonders whether, for example the number of equiangular bins shouldn't increase with the radius.

Again, the calculation for a spherical normal would proceed in the same way.

Note that the expectation for the bin frequency, N_{ij} is

$$E(N_{ij}) = \frac{Nf_i}{K_\theta}$$

where N is the number of points to be binned. As in equation (9) of the paper it could be N_S, the number in a section. The latter is given by equation (8) when $d = 2$, the reference distribution is the uniform sphere in $\mathbb{R}^p$ and the section is the central one S.

The above treatment essentially ends the discussion of the radial binning. The subsection below might better appear in the section on the generalised index where ϵ also appears.

The threshold ϵ

In addition to choosing a resolution $x = h/R$ in order to define the central and outer sections and hence expected sample sizes, another value, the threshold ϵ determines whether differences are significant enough to merit counting (according to equation (3) on page 6). This too depends on expected sample sizes. Again, in complete generality this is

$$E(N_{ij}) = \frac{Nf_i}{K_\theta}$$

and is essentially equation (9). Here N is the number of points in the entire data set (within the sphere) or it is the number N_S in a section S.

The random count N_{ij} will is binomial, though for large enough samples, can be treated as Poisson. This is the approach taken in **Section 3.5**. There the coefficient of variation (the ratio of the standard deviation to the mean) is estimated by

$$\delta_S^i = \frac{\sqrt{N_{ij}}}{N_{ij}}.$$

This is essentially what is happening in equation (10) on page 10 of the ms.

As indicated by the notation, there is no dependence of δ_S^i on j, since j only identifies the angle used to determine bin direction and all directions are uniform. It does however depend on i, which determines the outer radius of the bin B_{ij}. **The value of δ_S^i depends on which radial bin B_i contains the data**.

This will be different for each i, though the notation of equation (10) obfuscates this by using r_2 and r_1 instead of r_i and r_{i-1}. The difference $r_2^2 - r_1^2$ is *not constant* even when $r_2 - r_1$ is. It therefore **makes no sense** to imagine that

$$\delta = \delta_S^i/K$$

as in equation (11) and recommended as the value of ϵ on page 11, line 31. It changes with i.

Perhaps something else, like the average or max of the δ_S^is might be sensible. Moreover, since a single ϵ is to used for counts in that bins from both S_1 and S_2, shouldn't the choice depend on counts in the bins projected from each section?

Generalised index

Assuming that the binning and spherical shell sectioning has been delivered to the reader as above, the generalised index of section 3.6 could follow as its own section (not subsection).

It might make more sense to move the above material on the threshold ϵ into the section devoted only to the general index.

Most of the material in this section is inspired by Gous and Buja (2004) and its various normed differences.

This review is already overly long, so I will refrain from discussing the indices and their properties. I will say, that just as above, care should be taken in the mathematical exposition, the language, and the notation.

Application

This section is where various ad hoc recommendations based on experience would go. Recommendations on how to choose the resolution, the threshold, the maximum radius (or standard deviation for a spherical normal), and the index would appear here. These recommendations, and their justifications, would be welcome by the user.

Indeed, here is where the **spherical trimming** might be justified as a practical matter. The argument would be something along the line that in practice, data do not generally fill a hyper cube and if they did, the fact that data from the corners would appear exclusively in bins in the outer section and essentially not at all in the bins of the inner central section would be reason enough to exclude them on the basis of their introducing unwelcome noise or instability in the index.

The single example of the olives data is enough. It makes the point. The pdfSense example does not contribute much beyond what is illustrated by the olives data.

The THDM Higgs-boson example (including Section 5.3 on inequality condition) should be removed entirely (along with its supporting Appendix and references). It adds nothing that helps the reader understand the proposed methodology, and introduces a number of physics concepts and ideas that will be unfamiliar to most JCGS readers. If "physics insights" really are gained from the application of the methodology to the example, then it should be written up as a separate paper and submitted to a physics journal. In the present journal, it is nothing but a distraction for the reader.

Some minor details

- Page 3, line 39 and page 4, Figure 1. The definition of S1 and S2 are unclear.

  Are these defined by defining two different projection planes (as, I think, in the previous paragraph) and the projection of the points within the central shell onto each projection plane? Or, are these two different shells, orthogonal to the same projection plane?

  In either case, it seems strange to talk about a slice through the origin, rather than *about* the origin.

- Page 3, lines 43-46 and page 4, Figure 1.

  How are the two data sets A and B actually constructed? The reference to Meremianin (2009) suggests that methods contained there were used.If so, what distinguishes the construction of A from that of B? It currently sounds like they were different samples of the same mathematical object. Is that correct?

- p. 5, last line. The parenthetical remark "needles in a haystack" does not *necessarily* apply (in spite of Eureka being found by Laa et al. (2019) in the otherwise structure free and infamously unrealistic data set of the ASA 1986 competition; it was also found by simpler rotation and interactive methods back in 1986). It is but an example of concentrated data which *might* be found. The reader should not be led to believe that the proverbial "needle in a haystack" will necessarily be found.

  All that a high I_A^{up} is saying is that there exists a lower, say $q \leq d$, dimensional structure about which the points are concentrated.

  While a "needle" structure could result in positive values of I_A^{up}, it might not result in a large value of I_A^{up}. The dimension q of the structure could be greater than 1 (up to $q = d$) and/or not at all linear, with even a large I_A^{up}. Either proscribes the use of "needle".

  Conversely, were a "needle" to exist in the middle, I_A^{up} could still be very low, possibly zero, depending on the configuration of points inside and outside the section.

  Similarly, while the word "grain" in the title opposite "hole" is the rhetorical licence of the authors, the choice comes at a cost of slight misdirection for the reader (given the concentration could be of any shape in the d-dimensional space).

- p. 7, Figure 5 and similar figures. Use wider lines (especially for legends) and different line types to better distinguish the lines visually.

- p. 19, line 9. h should be called the shell/section *radius* and not the *thickness* since thickness should really refer to the diameter $2h$.

References

Gous, A., & Buja, A. (2004). Visual comparison of datasets using mixture decompositions. *Journal of Computational and Graphical Statistics*, *13*(1), 1-19.

Laa, U., Cook, D., & Valencia, G. (2019, October). A slice tour for finding hollowness in high-dimensional data. *arXiv preprint arXiv:1910.10854v1*. Retrieved from https://arxiv.org/pdf/1910.10854.pdf