Errata & Addenda

Object Oriented Data Analysis

By J. S. Marron and Ian L. Dryden

Acknowledgement: Special thanks to Arthur Pewsey for providing a large number of these.

Page 40, line 6: Add a bullet point which moves the definition of the matrix of ones from (10.1) to here. Give that the equation number (3.6).

Page 52, line 9: Twin Arches 🡪 Tilted Parabolas

Page 64, line 6: from the 🡪 from

Page 97, line -10: shown in Figures 🡪 shown using two different views of the same data set in Figures

Page 106. Lin3 5: research in the 🡪 research by the

Page 106, line 9: of the FDA 🡪 of FDA

Page 106, line 12: However generally better is to 🡪 However, it is generally better to

Page 118, line 26: positions, and 🡪 positions (nor shapes of the circles of points)

Page 118, line -2: (10.1) 🡪 (3.6)

Page 121, line -12: because they represent 🡪 because such directions represent

Page 121, line -11: change, called 🡪 change. The subspace generated by these directions is called

Page 123, line -3: vector-valued data 🡪 vector valued auxiliary data

Page 125, line 4: challenging. In 🡪 challenging and even controversial. In

Page 126, line 20: useful. 🡪 useful, in particular this way of measuring error is the foundation of least squares and the Analysis of Variance. Furthermore this best corresponds to our usual intuitive notion of distance in the three dimensional world in which we live.

Page 126, equation (7.3). Should be:

Note: the wrong formula in the book is the *cosine similarity*, which has the property that it is 1 when **x** and **y** point in the same direction, 0 when they are orthogonal and -1 when the point in antipodal directions. The *cosine distance* given here is 0 when **x** and **y** point in the same direction, 0.5 when they are orthogonal, and 1 when they are antipodal.

Page 126, line 22: radial direction 🡪 radial distance

Page 126, line -10: The cosine 🡪 On the cosine

Page 126, line -9: other. When 🡪 other. However, it is a proper metric when the data objects lie on the sphere . When

Page 127, line 5: Centers In Metric 🡪 Centers in Metric

Page 129, line 3: the middle of the 🡪 the

Page 130, Caption of Figure 7.4: invariance. 🡪 invariance of notions of multivariate median.

Page 132, line 7: Scaling For Object 🡪 Scaling for Object

Page 134, line 22: classification discrimination 🡪 classification/discrimination

Page 134, line -16: e.g. as 🡪 e.g. as done through display of curve modes of variation shown in many places starting with Figure 1.4 for the Spanish Mortality data and as

Page 136, line -1: Missing “dx” in integral.

Page 140, line -7: Bold-faced capital J\_{k,1} should be a boldfaced 1, and reference to (3.6) should be given.

Page 145, line -3: over-rotation 🡪 over rotation

Page 147, line -17: Good recent 🡪 A good recent

Page 164, line -14: deviations along 🡪 deviation along

Page 164, line -15: deviations along 🡪 deviation along

Page 164, line -15: DNA. 🡪 DNA (where ±1 were selected as reasonable representatives of this mode of variation).

Page 166, line 2: deviation along 🡪 deviations along

Page 166, line 6: deviation along 🡪 deviations along

Page 166, line -10: Append a sentence:

Note Figures 8.12 and 8.13 used representatives of the mode of variation at ±2 standard deviations, instead of the ±1 used in Figure 8.11 to better highlight the impact of the outlier on the curvature of the modes of variation in this data set.

Page 173, Caption of Figure 8.14: indicate fibers 🡪 indicate approximately parallel fibers

Page 180, line 19: MDS using 🡪 MDS (recall Multi Dimensional Scaling from Section 7.2) using

Page 185, Caption of Figure 9.7: Data Objects in 🡪 Piece-wise linear data object curves in

Page 188, line 11: panel. 🡪 panel of Figure 9.9.

Page 197, bottom: Append a new paragraph:

A related research area is currently called *connectome analysis*, where the goal is understanding connections in the human brain. When that is studied at the population level, the issue of what should be the data objects becomes central. Many approaches are based on bundles of fibers. Campbell et al. (2021) have proposed a particularly innovative approach based on Riemannian metrics as data objects. Variation is then studied in the space of Riemannian metrics.

Page 198, line -2: Append a new paragraph: Note that the concept of *tree lines* first introduced in Wang and Marron (2007) was an early version of *modes of variation* as discussed in Section 3.1.4. Another early version are the *tree curves* proposed by Aydin et al. (2012).

Page 199, line 5: operator. 🡪 operator, which led to the inclusion of non-anatomical variation in the data set.

Page 199, line 23: objects. The 🡪 objects. In particular an elegant and very widely used mathematical frame work, often called *BHV Space*, was developed by Billera et al. (2001). The

Page 202, line 1: and Nye 🡪 and see Weyenberg (2015) and Nye

Page 202, end of Section 10.1.2: Append a new paragraph:

A promising recent approach to the analysis of phylogenetic trees as data objects has been proposed by Garba et al. (2021). This revolutionary idea steps completely away from the classical BHV space, through representing the data object trees as correlation matrices (motivated by a new probability framework). The resulting *Wald Space* has far different, yet intuitively sensible properties. Statistical analysis in Wald Space may ultimately prove to be more tractable than in BHV Space because the more direct manifold data analysis methods as described in Section 8.7 may be applicable An interesting open problem is analysis of the Brain Artery data set using Wald Space ideas.

Page 204, line -3: After the Bubenick (2015) reference, add a new sentence: Persistence Landscapes have been integrated with the phase shift - alignment ideas of Chapter 9, in a more recent and interesting analysis of the brain artery data, by Matuk et al. (2021).

Page 216, line 15: (10.1) 🡪 (3.6)

Page 229, line -3: maybe 🡪 may be

Page 233, line -8: (10.1) 🡪 (3.6)

Page 237, line -13: (2014) DWD 🡪 (2014), DWD

Page 241, line 12: for good overview 🡪 for a good overview

Page 241, line 19: (1984) The 🡪 (1984). The

Page 241, line -17: properties make classification an 🡪 properties provide the justification of the statement made at the beginning of this chapter that classification is an

Page 244, line 4: Within, Cluster 🡪 Within Cluster

Page 245, line 4: while second 🡪 while the second

Page 247, line 16: seems be 🡪 seems to be

Page 249, line 15: Finally because of its length the 🡪 Finally, because of its length, the

Page 252, Figure 12.7: Would be interesting to also consider the distance of 1 – Correlation in these comparisons, as that is commonly used in bioinformatics.

Page 259, line 17: as noted in Section 11.4 DWD 🡪 Section 11.4, DWD

Page 260, line 13: the 2 sample t statistic 🡪 the two-sample *t* statistic

Page 260, line -13: the t statistic 🡪 the *t* statistic

Page 260, line -2: -999 🡪 $-999$ (in LaTeX)

Page 264, line 21: using the definition in terms of coming from a single Gaussian this is taken 🡪 using the single Gaussian definition, this is taken

Page 265, line -14: the fit density 🡪 the fitted density

Page 283: Figure 14.4 should be a little smaller, so it fits on this page.

Page 285, line -6: Insert a new paragraph:

Ricardo Cao has pointed out that the unfamiliarity and non-intuitive nature of high dimensional space that is clear from the geometric representation can also be seen in purely deterministic ways. For example, consider the symmetric (about the origin) unit cube . Note that there are orthants (whose boundaries are the hyperplanes whose normal vectors are the unit coordinate vectors). Each of these orthant cubes (the intersection of the orthant and ) has an inscribed sphere, which is centered at a point of the form , which is distance from the origin. Furthermore, each inscribed sphere has radius . Next let , denote the *Diagonal Line* through the origin connecting the far vertex of the cube in the positive orthant with its antipodal point. For each value of , the distance of the point on the diagonal line to the origin is . Now the diagonal line intersects the inscribed sphere at the two points which are distances from the origin. Next consider the sub-sphere centered at the origin, with radius . That sub-sphere is tangent to each of the orthant inscribed spheres, at their point closest to the origin. Now for that sub-sphere has radius 1 (hence has points tangent to the unit cube, ). Furthermore, for that sub-sphere actually extends *outside* the unit cube, , showing the terminology “sub-sphere” is not actually appropriate. Beyond that, note that the orthant inscribed spheres all lie within the range of distances of from the origin. Note that is a purely deterministic way of understanding how the stochastic geometric representation is actually quite natural in high dimensions. It also gives an indication of the parallel result to (14.4), where the Gaussian distribution there is replaced by the uniform distribution on the interval , with a different coefficient of . This also reveals that although the unit cube, , is compactly supported in , it perhaps surprisingly “puts most of its mass in the corners”.

Page 294, line 17: node numbers 🡪 mode numbers

Page 296, line 12: 0.005,showing 🡪 0.005, showing

Page 296, line 13: 0.0006,with 🡪 0.0006, with

Page 296, line 14: 0.0016,with 🡪 0.0016, with

Page 297, line 1: After the Marron and Nolan (1988) reference, add the new sentences:

The main idea is to treat the kernel function as an equivalence class indexed by the bandwidth, i. e. to work with the quotient space with respect to bandwidth rescaling. This approach is a perhaps simpler variation on the quotient space ideas that underlie the curve registration in Section 9.1.3.

Page 302, line 7: discovered that surprisingly the speed 🡪 discovered, surprisingly, that the speed

Page 305, line 4: For example one 🡪 For example, one

Page 305, line -4: As noted in Section 15.1 that 🡪 Section 15.1, that

Page 307, line 9: approximations, that 🡪 approximations that

Page 307, line 14: dashed 🡪 dotted

Page 309, line 7: -20 🡪 $-20$ (in LaTeX)

Page 311, line -9: Delete sentence starting “SiZer ideas …”

Page 315, line -17: large values 🡪 large or small values

Page 315, line -2: above the 🡪 above, the

Page 323, line 22: Figure 16.6. 🡪 Figure 16.6).

Page 327, line 8: Furthermore pairwise 🡪 Furthermore, pairwise

Page 333, line 16: (10.1) 🡪 (3.6)

Page 350, line 19: studied those 🡪 studied in those

Page 351, line 15: Good overview 🡪 A good overview

Page 351, line -10: In , the matrix should be .

Page 352, line 11: .Because 🡪 . Because

Page 358, line 6: of residuals 🡪 of errors (or perturbations).

Page 361, line -5: In contrast statisticians 🡪 In contrast, statisticians

Page 361, line -4: conservative, about 🡪 conservative about

Page 361, line -2: this writing *sparsity* 🡪 this writing, *sparsity*

Page 372, line -15: Marron, J. S. should be a sixth co-author.