Reassessing PCA of Acceptability Rating Data for Japanese (ARDJ) data using kernel Multivariate Analysis

Kow Kuroda
kow.k@ks.kyorin-u.ac.jp
Medical School, Kyorin University

1. Introduction

Acceptability Rating Database of Japanese (ARDJ) [4] is a research project that aims at providing normative data that prepares for "evidence-based linguistics" [12]. The project released a dataset called `s2u` (for "survey 2 unified") that consists of 1,880 responses to 300 stimuli sentences by 877 individuals, as well as answers to 11 questions for social attributes like age, gender, lived places.[1]

Recent work [5] presented a preliminary analysis of the data thus acquired, but it was likely that it suffered from a flaw. Four kinds of data were prepared: 1) raw count data, 2) normalized data[2], 3) scaled by sentence without centering, and 4) scaled by sentence with centering. Application of PCA to these data yields the results in Figure 1, where colors are added to identify clusters membership[3] and contours are added to indicate density.

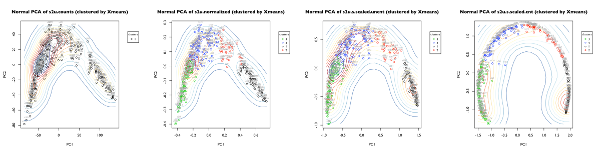

Figure 1: PCA of s2u data, raw count, normalized, and s-wise scaled (uncentered and centered)

The horseshoe-like shape of data points, common to all plots here, is called the "horseshoe" effect, or Guttman effect [2, 8]. It is not peculiar to PCA. Results of Correspondence Analysis (CA) and Multi-dimensional Scaling (MDS) of the same data take the same shape. In other words, normal methods for MVA suffer from this.

This effect is considered an artifact arising from the response design (e.g., Likert or Guttman scales). Quite a number of techniques [7, 8] were proposed to remove or at least attenuate this effect, but most of them require reimplementation of basic functions used to calculate variance. I wondered if there was an easier way to make a shortcut and came up with the idea of applying *kernel-based multivariate analysis (kMVA)* to this end.

The most obvious possibility to explore is to use kernel PCA [10]. For example, `kernlab` package [3] for `R` provides Gaussian and Laplacian kernels among others.[4]

Another possibility worth exploration was unsupervised metric learning such as *Isometric Mapping (Isomap)* [11] and *Locally Linear Embedding (LLE)* [9], which serves as kernel-based versions of MDS. `RDRToolbox` [1] package for `R` provides them.

In what follows, I present results of kernel based PCA in 2. (ones using Gaussian RBF in §2.1 and ones using Laplacian RBF in §2.3), ones using Isomap in §3.1, and ones using LLE in §3.2. Due to space limitation, presentation of results is selective.

2. Applying Kernel PCA

2.1 RBF kernels

Gaussian kernel uses $\exp(-\sigma\|x-x'\|^2)$, and Laplace kernel uses $\exp(-\sigma\|x-x'\|)$ for radial base functions (RBFs), as plotted with varying σs in Figure 2.

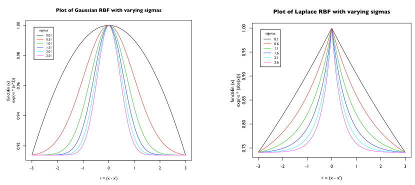

Figure 2: Gaussian and Laplacian RBFs with varying σ's

2.2 Gaussian kernel PCA

2.2.1 Gaussian PCA on raw count data

Gaussian PCA applied to the raw count data yielded the results in Figure 3, where isolation seems successful.

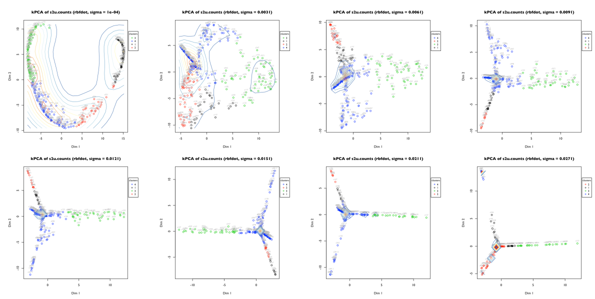

Figure 3: Gaussian PCA of s2u raw count data

2.2.2 Gaussian PCA on s-wise scaled, uncentered data

Gaussian PCA applied to the the sentence-wise scaled, uncentered data yielded the results in Figure 4.

[1] This dataset, along with some others, is available at https://kow-k.github.io/Acceptability-Rating-Data-of-Japanese/

[2] Normalization means conversion of range-wise raw counts into discretized density distribution.

[3] Recognition of clusters inside the dimensionally reduced data is always done with X-means provided by `clusternor` package [6].

[4] Only results from Gaussian and Laplacian kernels are discussed in this paper, though other options like polynomial, ANOVA RBF, hyperbolic, Bessel kernels were tested. `R` version 3.5.3 was used for all runs.

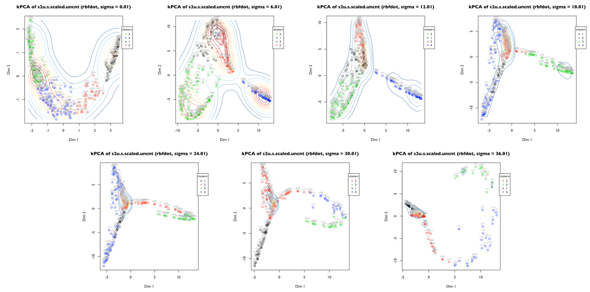

Figure 4: Gaussian PCA of s2u s-scaled, uncentered data

2.2.3 Gaussian PCA on s-wise scaled, centered data

Gaussian PCA applied to the sentence-wise scaled, centered data yielded the results in Figure 5.

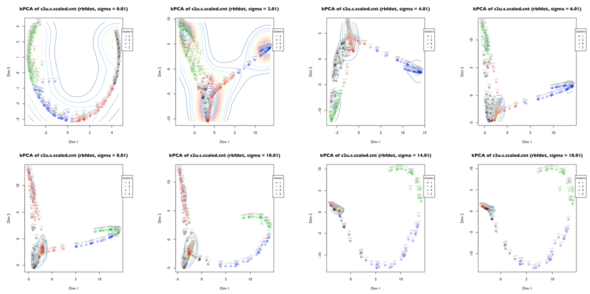

Figure 5: Gaussian PCA of s2u s-scaled, centered data

2.2.4 Gaussian PCA on normalized data

Gaussian PCA applied to the normalized data yielded the results in Figure 6. Effective values for σ are very large, compared to other cases, suggesting that normalization is harmful.

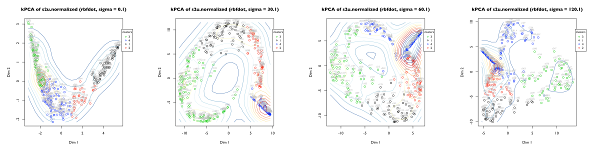

Figure 6: Gaussian PCA of s2u normalized data

2.2.5 Summary of Gaussian PCA results

In terms of σ optimization, raw count data and normalized turned out to be harder to handle with Gaussian RBF kernel: raw data requires extremely small values, and normalized data required rather large values. Scaled data, with or without centering, seem to give us better results.

2.3 Laplacian kernel PCA

2.3.1 Laplacian PCA on raw count data

Laplacian PCA applied to the raw count data yielded the results in Figure 7.

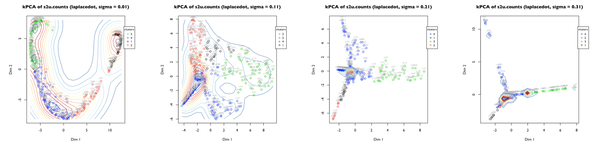

Figure 7: Laplacian PCA of s2u raw count data

2.3.2 Laplacian PCA on s-wise scaled, uncentered data

Laplacian PCA applied to the scaled, uncentered data yielded the results in Figure 8.

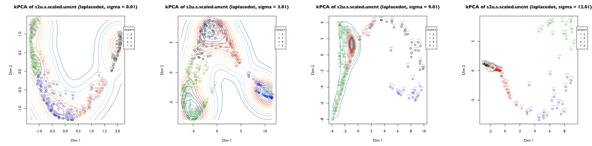

Figure 8: Laplacian PCA of s2u s-scaled, uncentered data

2.3.3 Laplacian PCA on s-wise scaled, centered data

Laplacian PCA applied to the scaled, centered data yielded the results in Figure 9.

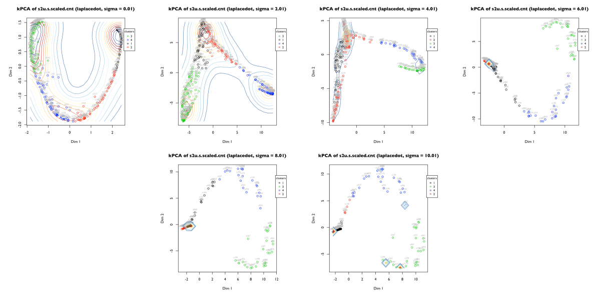

Figure 9: Laplacian PCA of s2u s-scaled, centered data

2.3.4 Laplacian PCA on normalized data

Laplacian PCA applied to the normalized data yielded the results in Figure 10.

2.3.5 Summary of Laplacian RBF

In terms of σ optimization, Laplacian kernel turned out to be easier to handle than Gaussian kernel. But it can be more selective in terms of data: the normalized data looks incompatible with Laplacian kernel.

2.4 Gaussian vs Laplacian kernels

On comparing the results from Gaussian and Laplacian kernels, it would be safe to say that Laplacian kernel performs better. In addition, selection of effective values for σ is harder with Gaussian kernel, probably because it responds to a squared term $\|x - x'\|^2$.

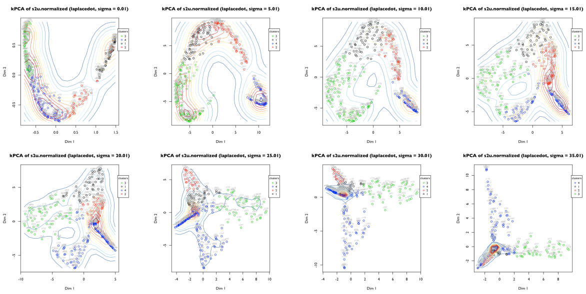

Figure 10: Laplacian PCA of s2u normalized data

3. Applying Isomap and LLE

Isometric Mapping (Isomap) [11] and Locally Linear Embedding (LLE) [9] form another class of kernel-based MVA, serving as kernel-based versions of MDS.

3.1 Isometric Mapping (Isomap)

3.1.1 Isomap on raw count data

Isomap applied to the raw count data yields the results (k = 5 was accepted) in Figure 11.

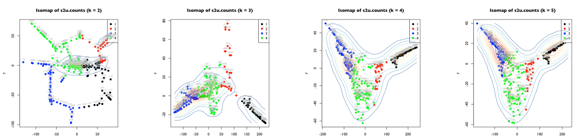

Figure 11: Isomap of s2u raw count data

3.1.2 Isomap on s-wise scaled, uncentered data

Isomap applied to the sentence-wise scaled, uncentered data gave us the results in Figure 12.

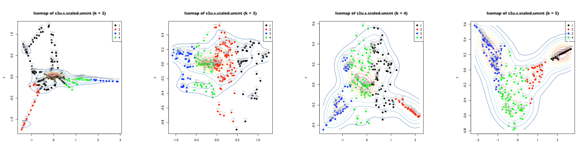

Figure 12: Isomap of s2u s-scaled, uncentered data

3.1.3 Isomap on s-wise scaled, centered data

Isomap applied to the sentence-wise scaled, centered data gave us the results in Figure 13.

3.1.4 Isomap on normalized data

Isomap applied to the normalized data gave us the results (k = 5 was accepted) in Figure 14.

3.1.5 Summary of Isomap results

Isomap brings stable and coherent analysis in that it produces roughly the same results on raw count, s-wise scaled with and without centering, with larger k.

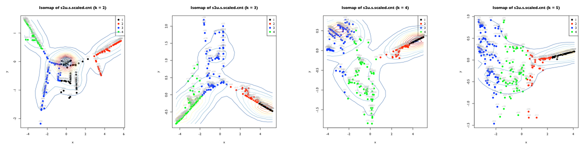

Figure 13: Isomap of s2u s-scaled, centered data

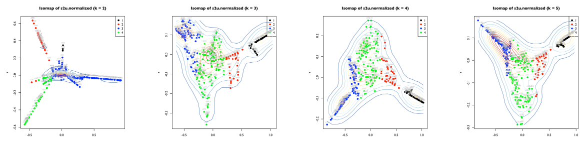

Figure 14: Isomap of s2u normalized data

3.2 Locally Linear Embedding (LLE)

Like Isomap, LLE [9] has a parameter k that encodes the number of nearest neighbors to consider. With our data, k cannot exceed 3 on whichever data.

3.2.1 LLE on raw count data

LLE applied to the raw count data gave the results (k = 4 was not accepted) in Figure 15.

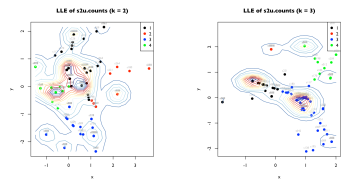

Figure 15: LLE of s2u raw count data

3.2.2 LLE on s-wise scaled, uncentered data

LLE applied to the sentence-wise scaled, uncentered data gave us the results (k = 4 was accepted, but k = 5 was not) in Figure 16.

3.2.3 LLE on s-wise scaled, centered data

LLE applied to the sentence-wise scaled, centered data gave us the results in Figure 17.

3.2.4 LLE on normal data

LLE applied to the normalized data gave the results (k = 4 was not accepted) in Figure 18.

3.2.5 Summary of LLE results

LLE produces less stable and coherent results than Isomap in that it gives different results to raw count, s-wise scaled with and without centering, and only with smaller values like as 2, 3.

3.3 Isomap vs LLE

Based on comparing results from Isomap and LLE, it would be safe to say that Isomap performs significantly better. Isomap is easier to use and provides better analyses.

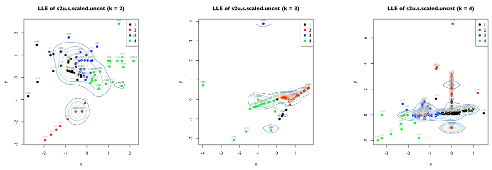

Figure 16: LLE of s2u s-scaled, uncentered data

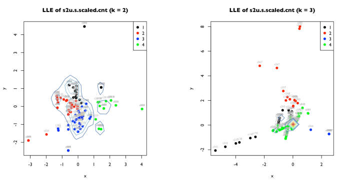

Figure 17: LLE of s2u s-scaled, centered data

4. Discussion

My motivation for the investigation reported here was to check if any of kernel MVA's can remove the Guttman effect [2, 8]. I tested kernel PCA [10] with various kernels (Gaussian, Laplacian, ANOVA RBF, polynomial, Bessel), and two methods for metric learning, i.e., Isomap [11] and LLE [9]. Based on the results reported and some others omitted here, I conclude that the attempt was successful if proper methods are chosen and hyper-parameters (e.g,. σ, k) are properly set.

Which ones are successful, however? By comparing the results, the most promising remedy for the Guttman effect seems to be achieved by Isomap for two reasons. First, Isomap seems to yield results similar enough to the results of normal PCA: no drastic transformations are detected. This is a good property. Second, Isomap gives us more robust results in comparison with results from kernel PCA, which tend to be "fragile" in that they are highly sensitive to different values for hyper-parameters (especially σ): a lot of tweaking was required to obtain reasonable results in kernel PCA. Metric-based methods like Isomap and LLE are free from this sort of annoyance, though LLE turned out to have less descriptive power.

The performance of kernel PCA would be best characterized as a mixture of wheat and chaff. If Laplacian and Gaussian kernels are compared, the former seems to work better with my data, especially in exploration for effective values for σ.

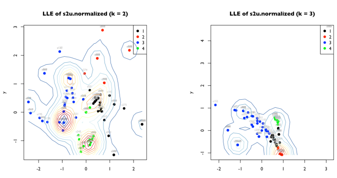

Figure 18: LLE of s2u normalized data

5. Conclusion

The Guttman effect shown by PCA, CA and MDS of ARDJ `s2u` data was effectively removed by Isometric Mapping (Isomap). Kernel PCA was shown to be also effective, but to a limited extent, due to messiness in optimization for hyper-parameters such as σ.

Acknowledgement

This research was supported by JSPS through Grant-in-Aid (16K13223).

References

[1] Ch. Bartenhagen. `RDRToolBox`*: A package for nonlinear dimension reduction with Isomap and LLE*, 2014. R package version 1.30.0.

[2] D. W. Goodall. Objective methods in the classification of vegetation. III. An essay on the use of factor analysis. *Australian Journal of Botany*, 2:304–324, 1954.

[3] A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis. `kernlab` — An S4 package for kernel methods in R. *J. of Statistical Software*, 11(9):1–20, 2004.

[4] K. Kuroda, H. Yokono, K. Abe, T. Tsuchiya, Y. Asao, Y. Kobayashi, T. Kanamaru, and T. Tagawa. Development of Acceptability Rating Data of Japanese (ARDJ): An initial report. In *Proc. of the 24th Ann. Meeting of the Assoc. for NLP*, pp. 65–68, 2018.

[5] K. Kuroda, H. Yokono, K. Abe, T. Tsuchiya, Y. Asao, Y. Kobayashi, T. Kanamaru, and T. Tagawa. Insights from a large scale web survey for Acceptability Rating Data for Japanese (ARDJ) project. In *Proc. of the 25th Ann. Meeting for the Assoc. of NLP*, pp. 253–256, 2019.

[6] D. Mhembere, D. Zheng, C. E. Priebe, J. T. Vogelstein, and R. Burns. `knor`: A NUMA-optimized in-memory, distributed and semi-external-memory k-means library. *High-Performance Parallel and Distributed Computing*, 26, 2017.

[7] J. Podani and I. Miklós. Resemblance coefficients and the horseshoe effect in Principal Coordinates Analysis. *Ecology*, 83(12):3331–3343, 2002.

[8] J. Rijckesvorsel. *The Application of Fuzzy Coding and Horseshoes in Multiple Correspondence Analysis*. DSWO Press, 1987.

[9] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[10] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10, 1998.

[11] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

[12] 黒田 航. 証拠に基づく医療 (EBM) との比較を通じて理論言語学の方法論を見直す. In **第** *16* **回日本認知言語学会発表論文集**, pp. 580–585, 2016.