Additional materials for "Principal curves" presentation

Stas Kolenikov

This handout present the graphs that are (to be) reproduced during the presentation along with additional comments.

The first example shown is that of circular data, which is the example both Hastie and Stuetzle (1989) and Tibshirani (1992) were using for their demonstrations. Fig. 1 shows the example of raw data (circle + 2D Gaussian noise), and Fig. 2 and 3 shows how iteration of the algorithm proceeded.

Another example of simulated data is the parabola data: the points approximately lie on a parabola, and the first principal component roughly connects the ends of the curve. Raw data are shown at Fig. 4, and the two examples of application of unmodified algorithm are shown on the next two pictures. As we see, the algorithm failed to converge to what we would expect to be the principal curve. The proposed remedies are: identify and drop out the intervals to which no points are projected (Fig. 7); change the starting point (random start — Fig. 8 and 9; starting with parametrization given by one of the variables — 10); simply apply more smoothing (Fig. 11).

Finally, the application that we are most interested in is that of functional data. The raw data simulated for this example are shown at Fig. 12. I tried to incorporate several nice features of the data that should be difficult for the principal components: nonlinearity, bimodality at the center of the graph, skewness at the right end.

The first few principal components of the (centered) data explain 81.15%, 18.45%, 0.15%, and then no greater than 0.02%; see Fig. 13. They seem to capture the main features of the data, but note that the first component at Fig.r̃effdapc1 swings too far away and above the data, which we would hardly want. Also, it is quite likely that the first two components are not independent, as they both demonstrate a lot of action in the middle — the second component is shown on Fig. 15. Other components do not seem to contribute much, and the difference from the mean is probably noise or sampling variation (Fig. 16 and 17). Lack of independence is evidenced by Fig. 18 for the first two components; for othes, Fig. 19 suggests that the problem is not too awful.

The iterations of the principal curve algorithm are shown on Fig. 20 – 23. The resulting principal curve captures 97.86% of variance. Alternative options of the algorithm produce quite similar results (Fig. 24 – 28). The quality of approximation is thus very high, as evidenced by the movie to be shown in class. Also, the quantiles of the distribution are reproduced closely enough — see Fig. 29.

Conclusion: principal curve might be a sensible thing to try along with all other method we are touching.

Figure 1: Raw circular data

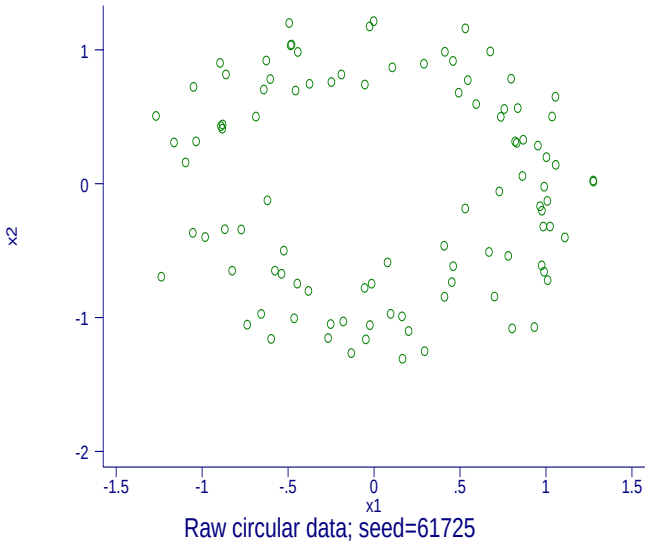

Figure 2: Iterations of principal curve algorithm.

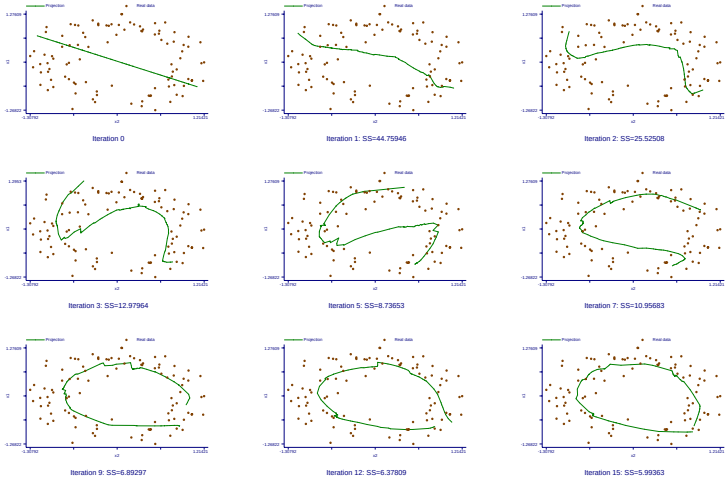

Figure 3: Iterations of principal curve algorithm — another random sample.

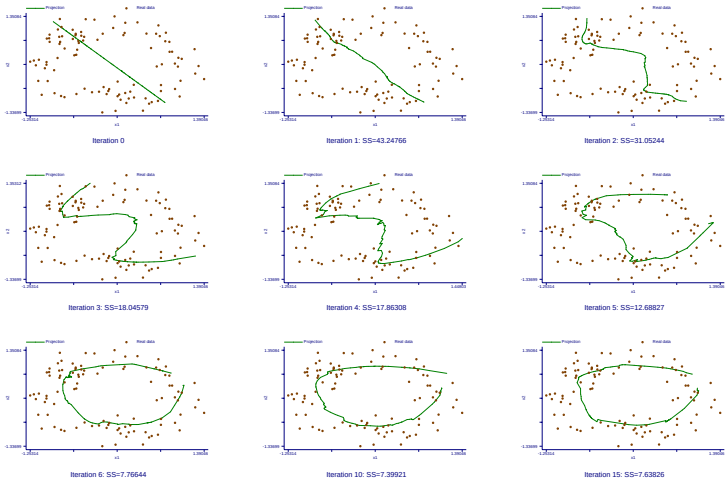

Figure 4: Raw parabola data

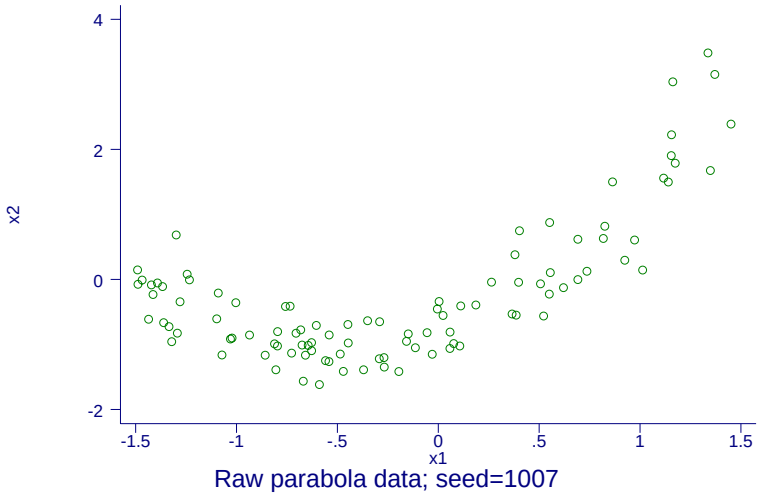

Figure 5: Iterations for parabola data: difficult case?

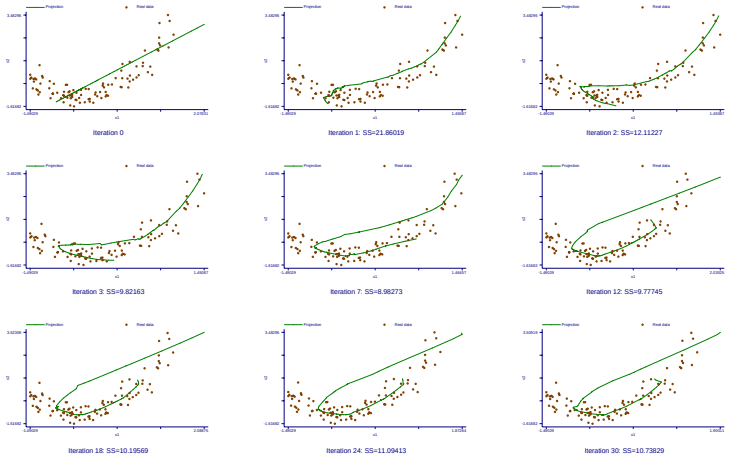

Figure 6: Iterations for parabola data: difficult case again!

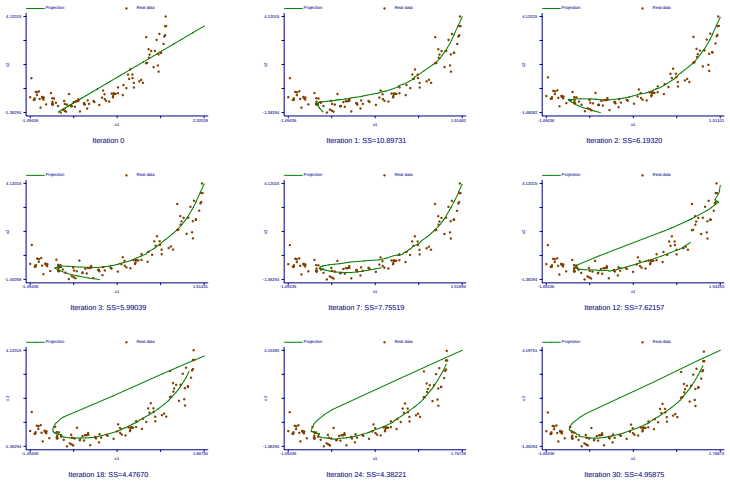

Figure 7: Iterations for parabola data: the algorithm tracks the intervals with zero density of the data points.

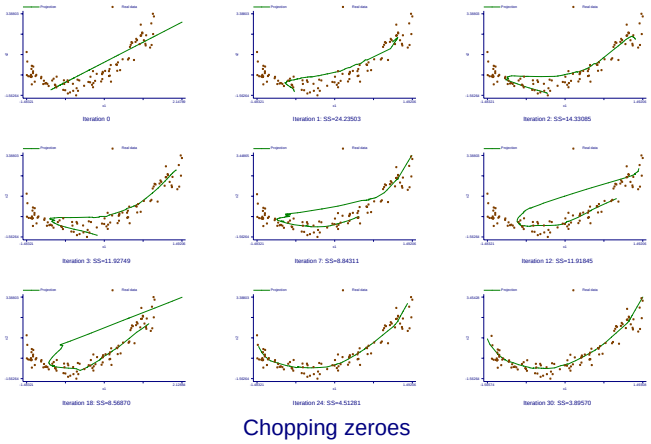

Figure 8: Iterations for parabola data: points are randomly connected at start.

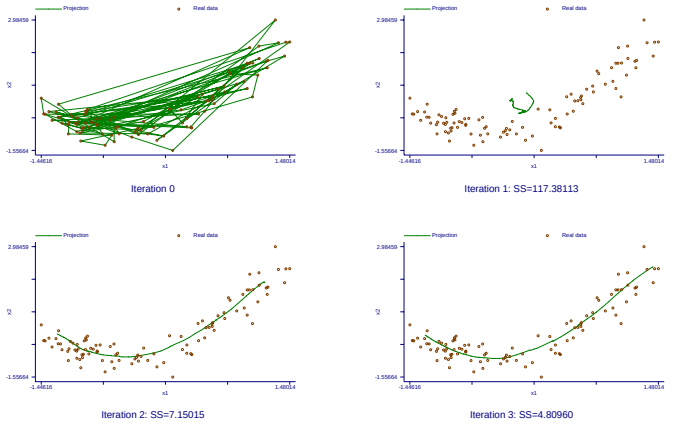

Figure 9: Iterations for parabola data: another random start.

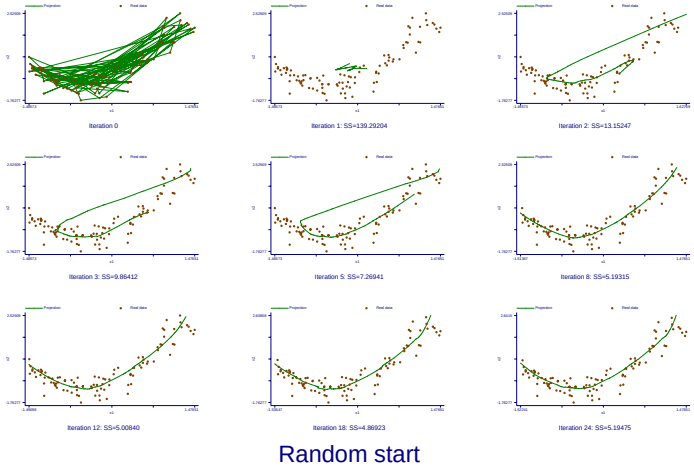

Figure 10: Iterations for parabola data: starting from one of the variables.

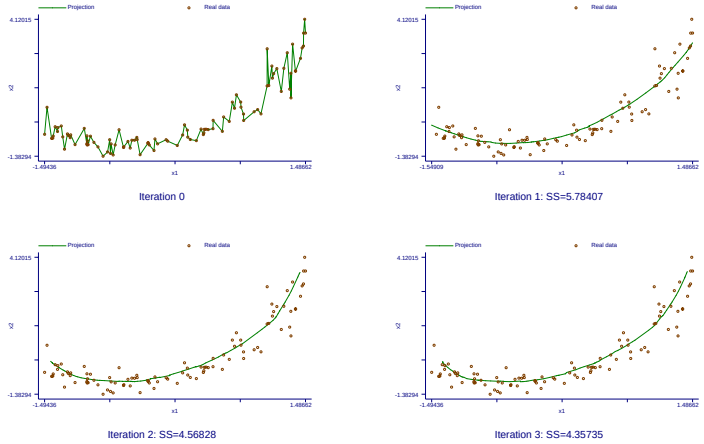

Figure 11: Iterations for parabola data: larger bandwidth

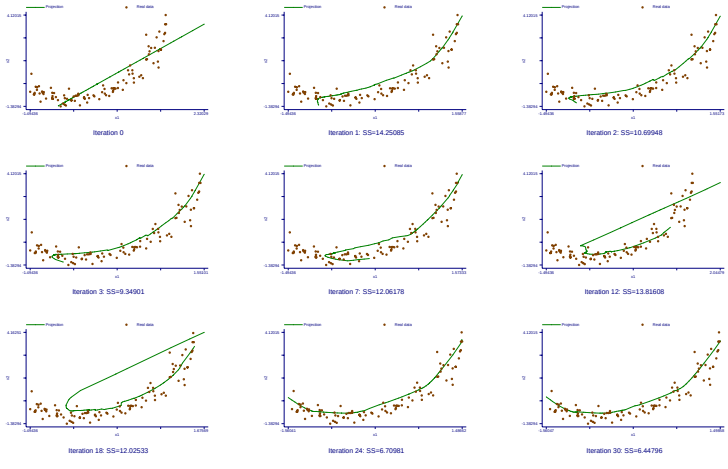

Figure 12: Raw functional data.

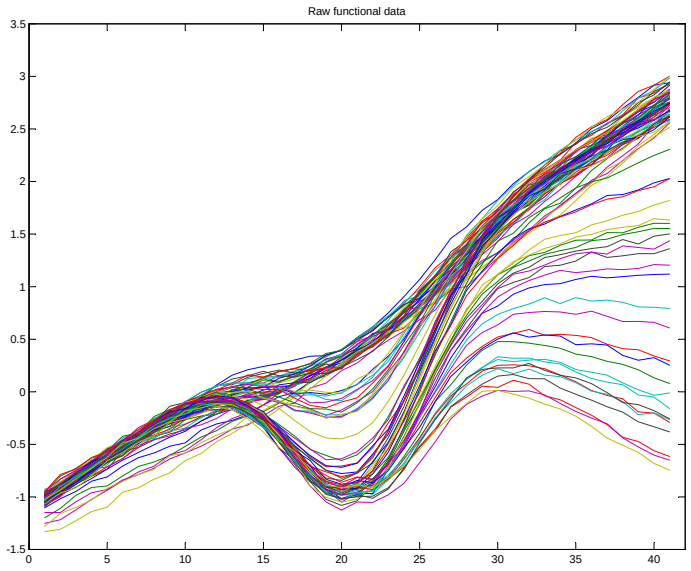

Figure 13: Principal components of the centered data.

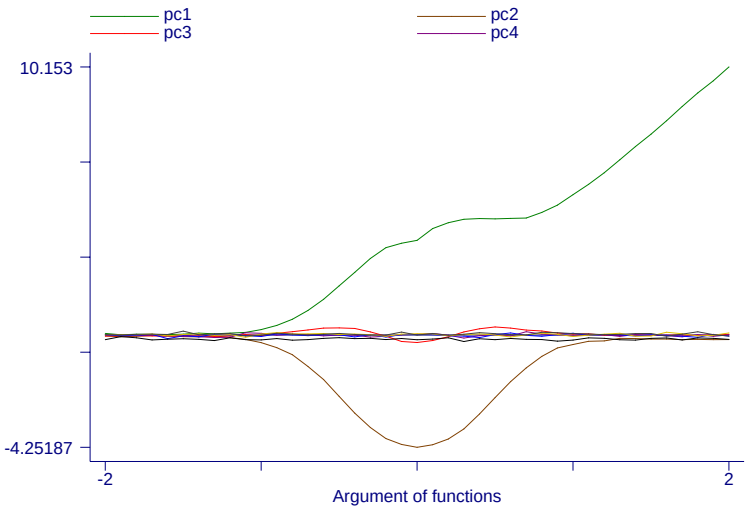

Figure 14: Variation in PC 1: major shape differences.

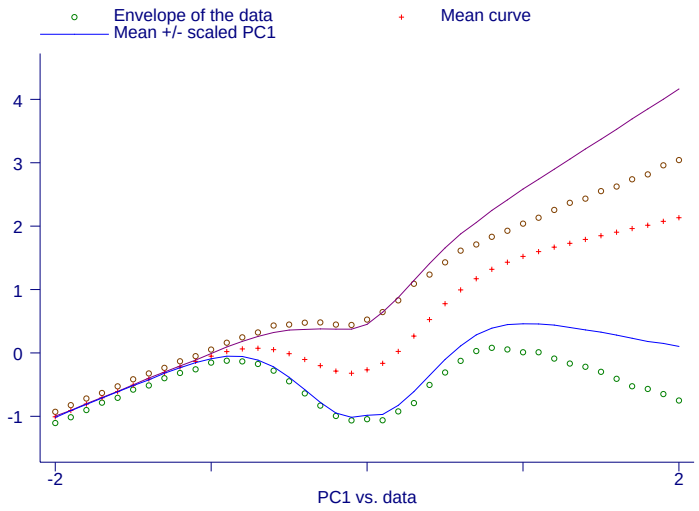

Figure 15: Variation in PC 2: change in the valley.

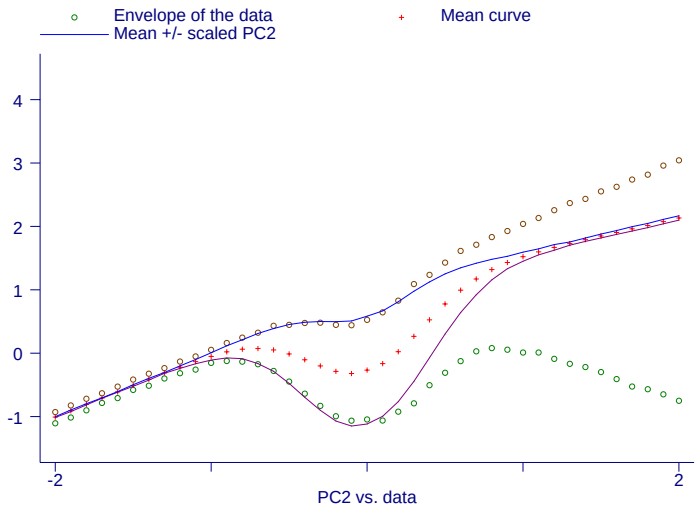

Figure 16: Variation in PC 3: some tilting?

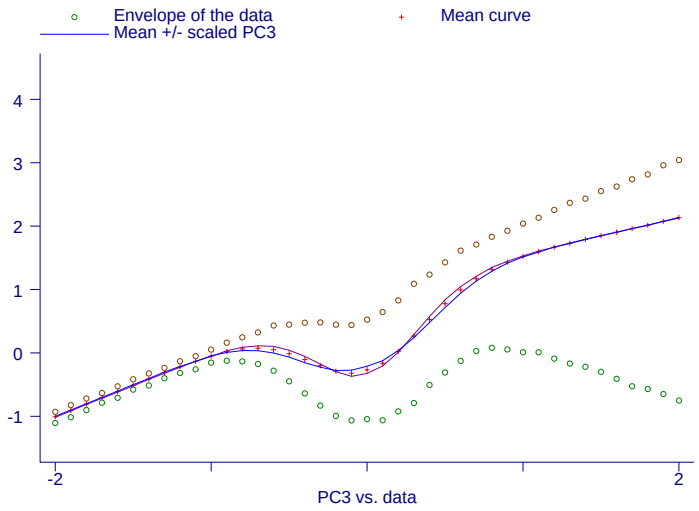

Figure 17: Variation in PC 4: noise?

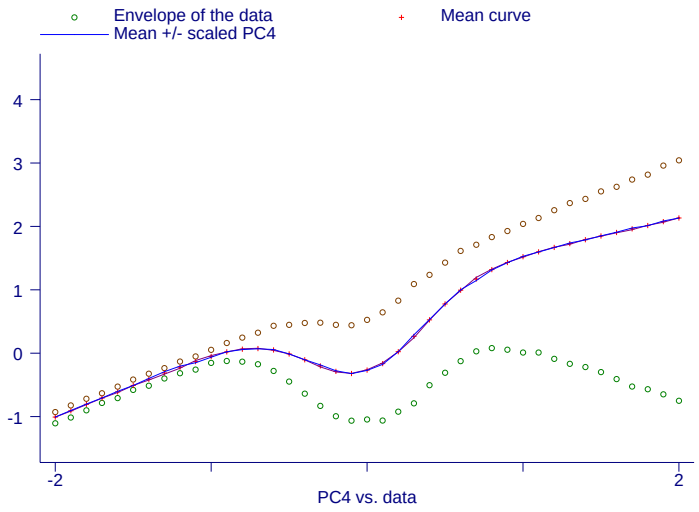

Figure 18: The first two components are not independent!

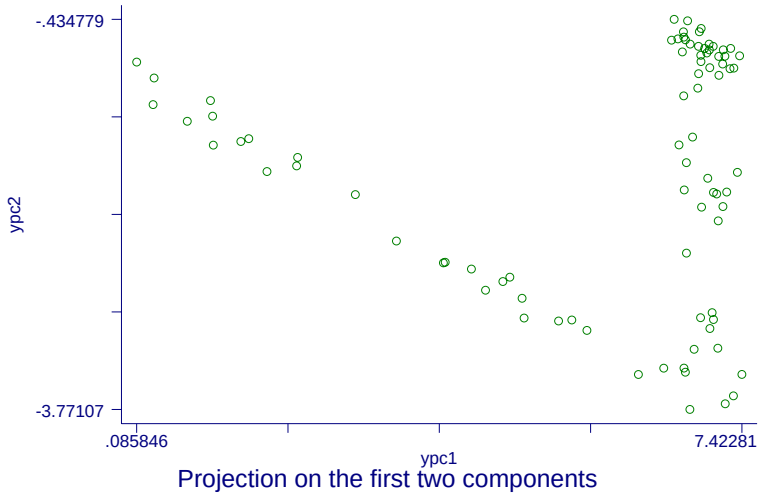

Figure 19: The draftsman plot for the first components.

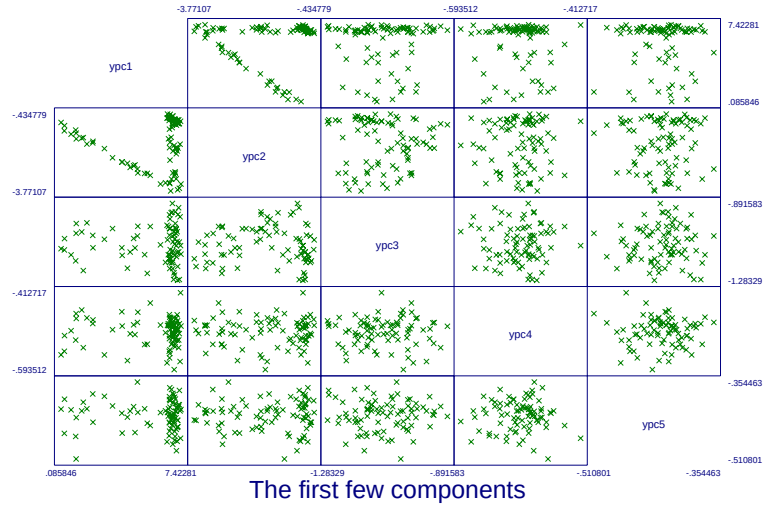

Figure 20: Several projections of the PC 1: starting point.

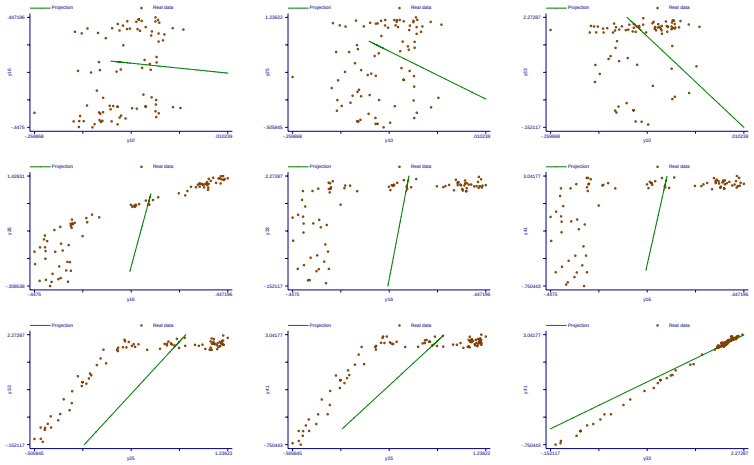

Figure 21: The first iteration.

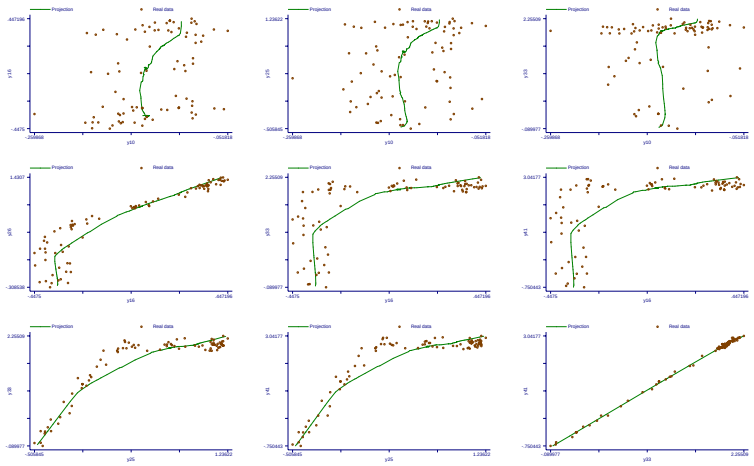

Figure 22: The second iteration.

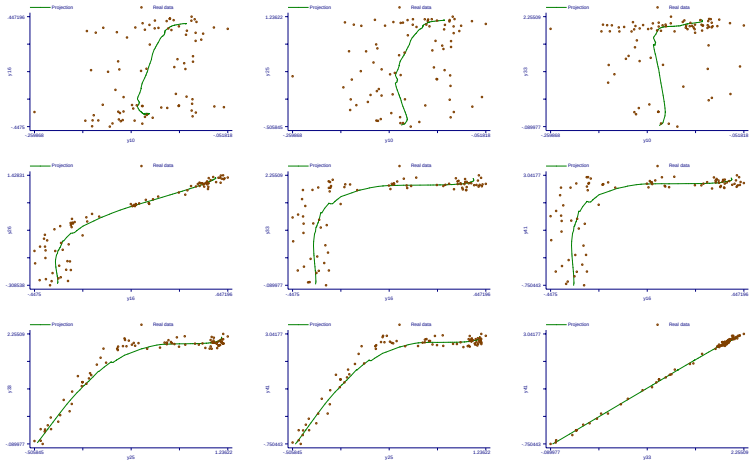

Figure 23: The last iteration (starting from the PC).

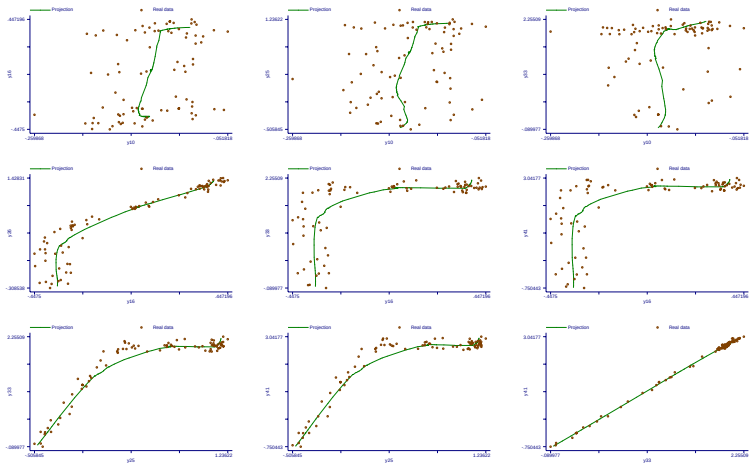

Figure 24: Random starting point: the first iteration.

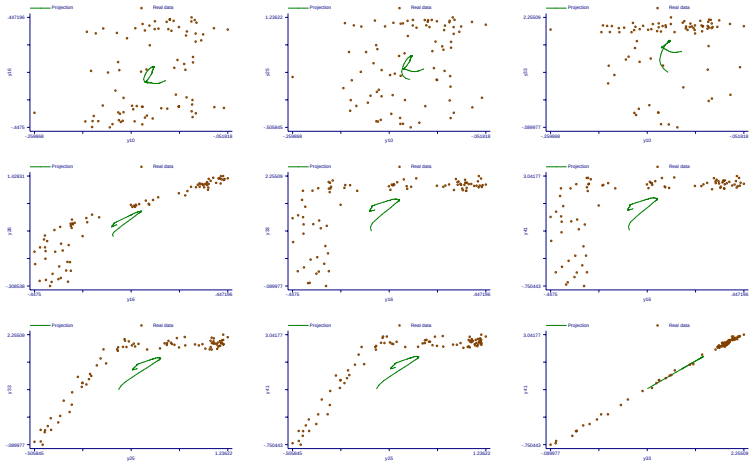

Figure 25: Random starting point: the second iteration.

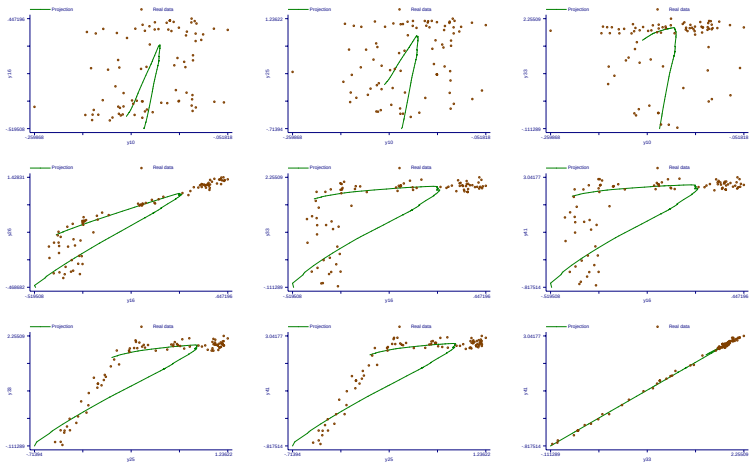

Figure 26: Random starting point: the third iteration.

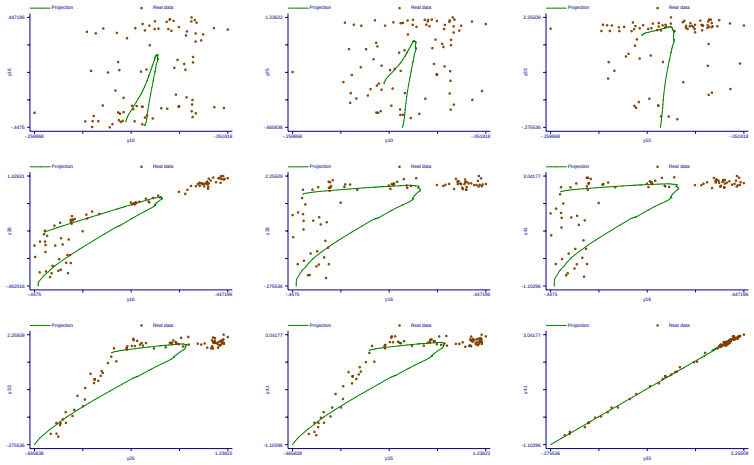

Figure 27: Random starting point: somewhere in the middle.

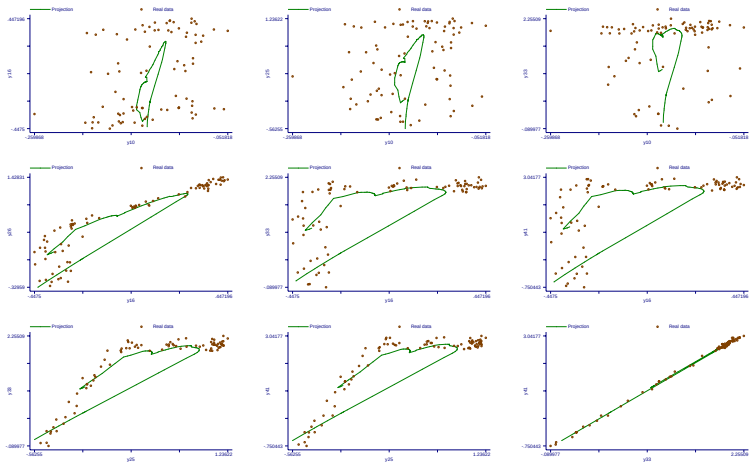

Figure 28: Random starting point: last iteration.

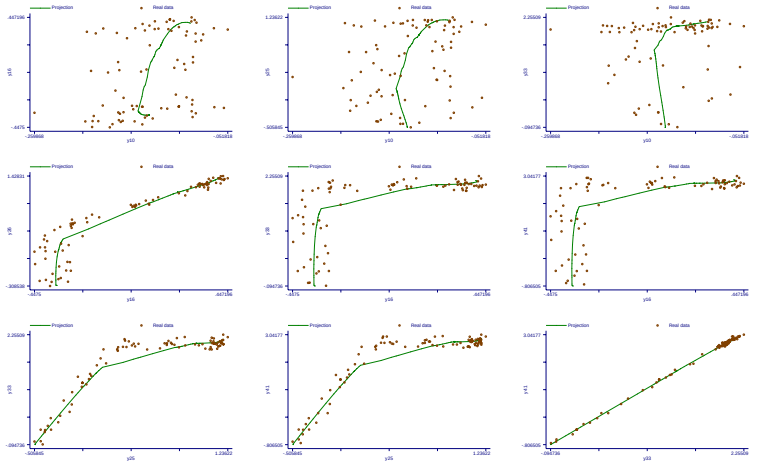

Figure 29: Principal curves got quantile approximately right