Estimation of Level Set Trees Using Adaptive Partitions: Supplementary Material

Lasse Holmström, Kyösti Karttunen, and Jussi Klemelä

November 5, 2016

Abstract

We provide supplementary material for the article "Estimation of Level Set Trees Using Adaptive Partitions"

1 Introduction

Section 2 studies how the mean integrated squared error (MISE) of the adaptive grid kernel estimator increases when the minimum allowed observation number is increased. Section 3 studies the MISE of the adaptive grid kernel estimator when the partition is such that the split points are always the mid points of the intervals. Section 4 shows the complete scatter plot matrix for the simulation example of Section 4.2 of the article. Section 5 shows the complete scatter plot matrix and the barycenter plots for the simulation example of Section 4.3 of the article. In addition, the barycenter plots of the regular grid kernel estimator are shown in order to demonstrate that the computational complexity of the regular grid kernel estimator is prohibitive for the detection of the modes. Section 6 gives additional figures for Section 5 of the article, where flow cytometry data was analyzed.

2 The Minimum Observation Number

Figure 1 shows how the mean integrated squared error (MISE) of the adaptive grid kernel estimator increases when the minimum allowed number of observations m is increased. Number m is such that the cells with this number of observations are not split anymore. We plot the MISE as a function of the sample size. In panel (a) $d = 2$ and in panel (b) $d = 3$. In panel (a) the curves with labels $1 - 4$ correspond to the minimal observation numbers

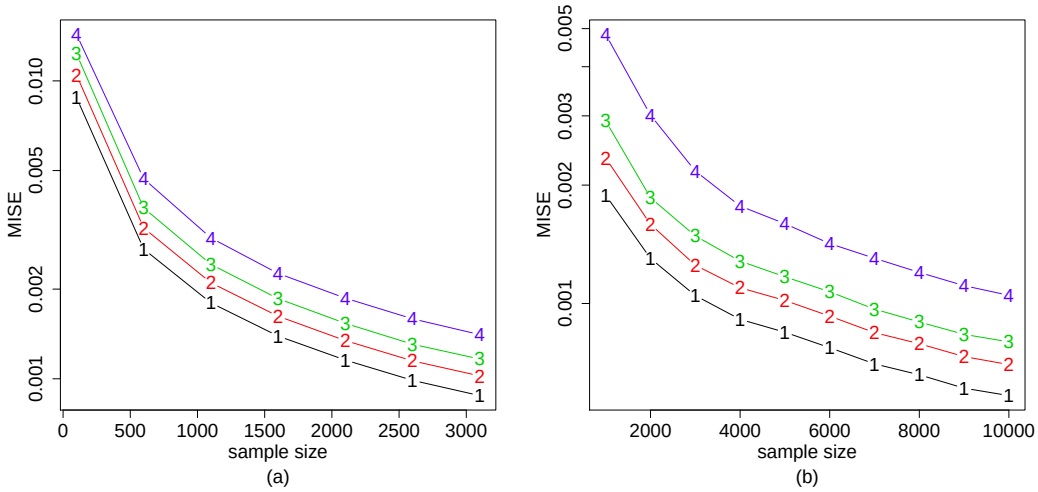

Figure 1: *Minimum allowed observation number.* (a) $d = 2$; (b) $d = 3$. The curves show the MISE as a function of the sample size for the adaptive grid kernel estimator. In panel (a) the curves with labels $1-4$ correspond to the minimal observation numbers $1, 5, 10$, and 20. In panel (b) the curves with labels $1-4$ correspond to the minimal observation numbers $1, 10, 20$, and 50.

$1, 5, 10$, and 20. In panel (b) the curves with labels $1-4$ correspond to the minimal observation numbers $1, 10, 20$, and 50. The data are generated from the standard normal distribution. The smoothing parameter is chosen by the normal reference rule and the kernel function is the standard normal density.

3 Dyadic Partitions

We study the MISE of the discretized kernel estimator when the partition is adaptive but the split points are always the mid points of the intervals: the best split is searched only over the directions (variables).

Figure 2 shows the MISE of discretized kernel estimators as a function of sample size when the true density is the standard normal density. In panel (a) $d = 2$ and in panel (b) $d = 3$. The figure is otherwise similar to the Figure 3 of the article, but we have added black lines with symbols "x", which show the MISE of the dyadic partition kernel estimator. We see that the restriction to dyadic splits increases the MISE considerably.

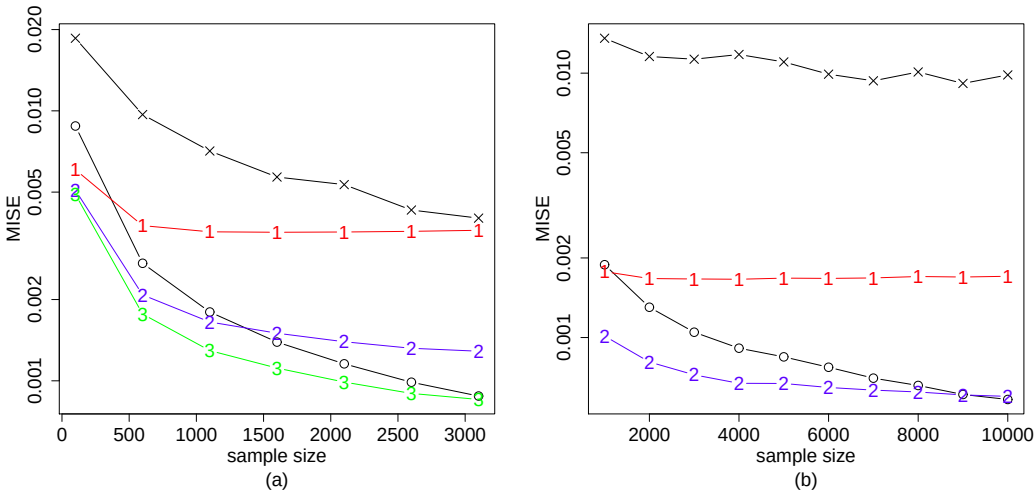

Figure 2: *MISE as a function of the sample size for the standard normal density.* (a) $d = 2$; (b) $d = 3$. The black curves with symbol "x" show the MISE of the adaptive dyadic grid kernel estimator. The black curves with symbol "o" show the MISE of the adaptive grid kernel estimator. The red curves show the regular grid kernel estimator with 10^d cells, the blue curves show the case of 20^d cells, and the green curve shows the case of 30^d cells.

4 7D Simplex: Scatter Plot Matrix

Figures 3–5 show all pairwise scatter plots when the data are simulated from a mixture of 8 normal distributions in 7D. The data are described in Section 4.2 of the article and it is generated from a mixture of 8 normal distributions with marginal standard deviations 0.25 and modes at the corners of the unit simplex. Figure 5(b) of the article shows the scatter plot of the 1st and the 2nd coordinates but now we show all 21 pairwise scatter plots.

5 4D Pentahedron: Scatter Plot Matrix and Barycenter Plots

5.1 Adaptive Grid Kernel Estimator

Figure 6 shows all pairwise scatter plots when the data are simulated from a mixture of 5 peaked distributions in 4D. The data are described in Section 4.3 of the article and it is generated from a mixture of 5 peaked distributions whose modes are at the corners of a simplex. Figure 6(b) of the article shows the scatter plot of the 1st and the 2nd coordinates but now we show all five pairwise scatter plots.

Figure 7 shows the barycenter plots for the adaptive grid kernel estimator whose volume function is shown in Figure 6 of the article. Panels (a)–(d) show the coordinates 1–4. The five modes can be detected, as was already shown with the volume function. A barycenter plot is defined at the end of Section 5 of the article.

5.2 Regular Grid Kernel Estimator

Figure 8 shows the barycenter plots for the regular grid kernel estimator when the grid has 20^d points. Panels (a)–(d) show the coordinates 1–4. We see that this grid is too small in order that the five modes could be detected with the regular grid kernel estimator, even when the computational complexity is very large for the grid of this size. Thus, even when the dimension is only $d = 4$, the computational complexity of the regular grid kernel estimator is so large that the modes cannot be detected. The grid size of the regular grid would have to be very large in order that the modes could be detected because in this example there is five thin modes that are close to each other.

Note that there is a large number of spurious modes at the tail areas of the density. The normal reference rule is used to select the smoothing parameter.

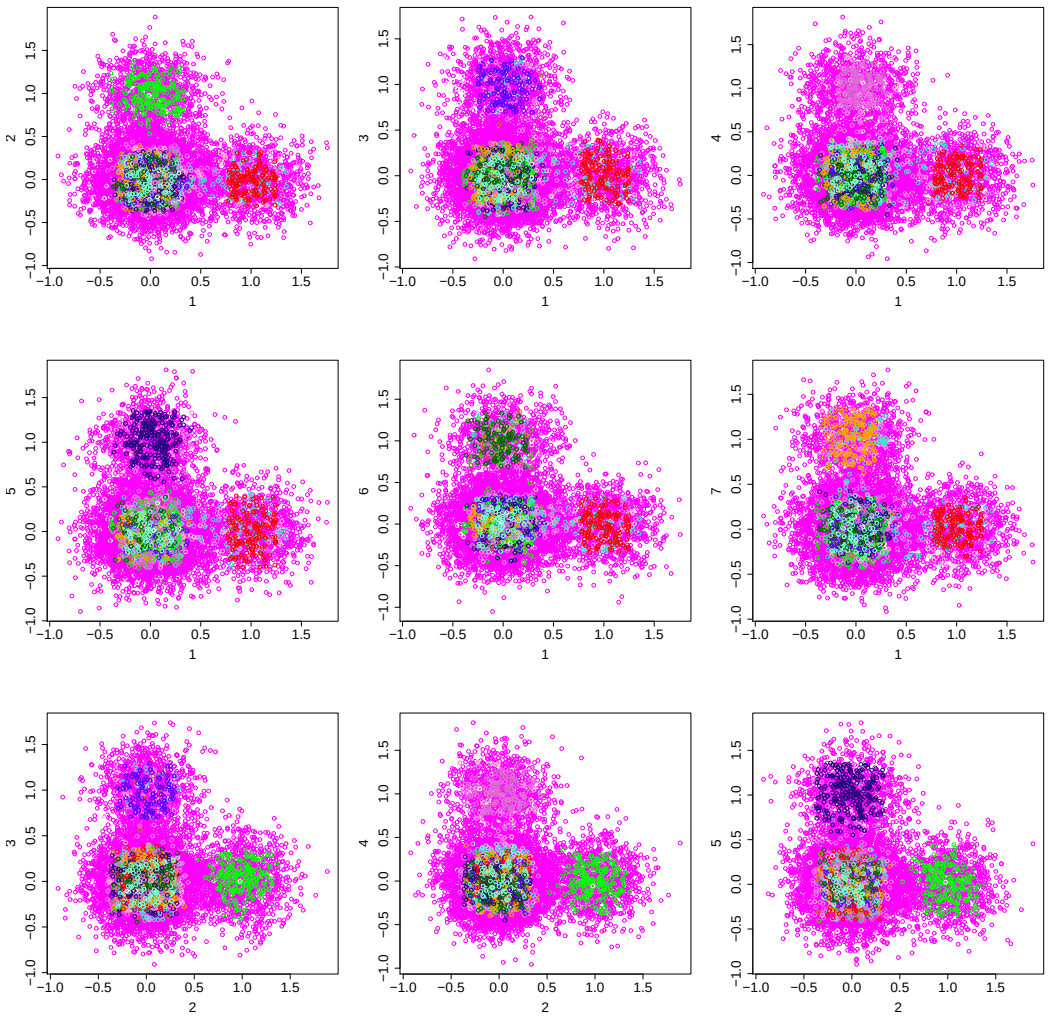

Figure 3: *Scatter plots for the 7D simplex data.* Scatter plots $(X_1, X_2), (X_1, X_3), \ldots, (X_2, X_5)$ are shown.

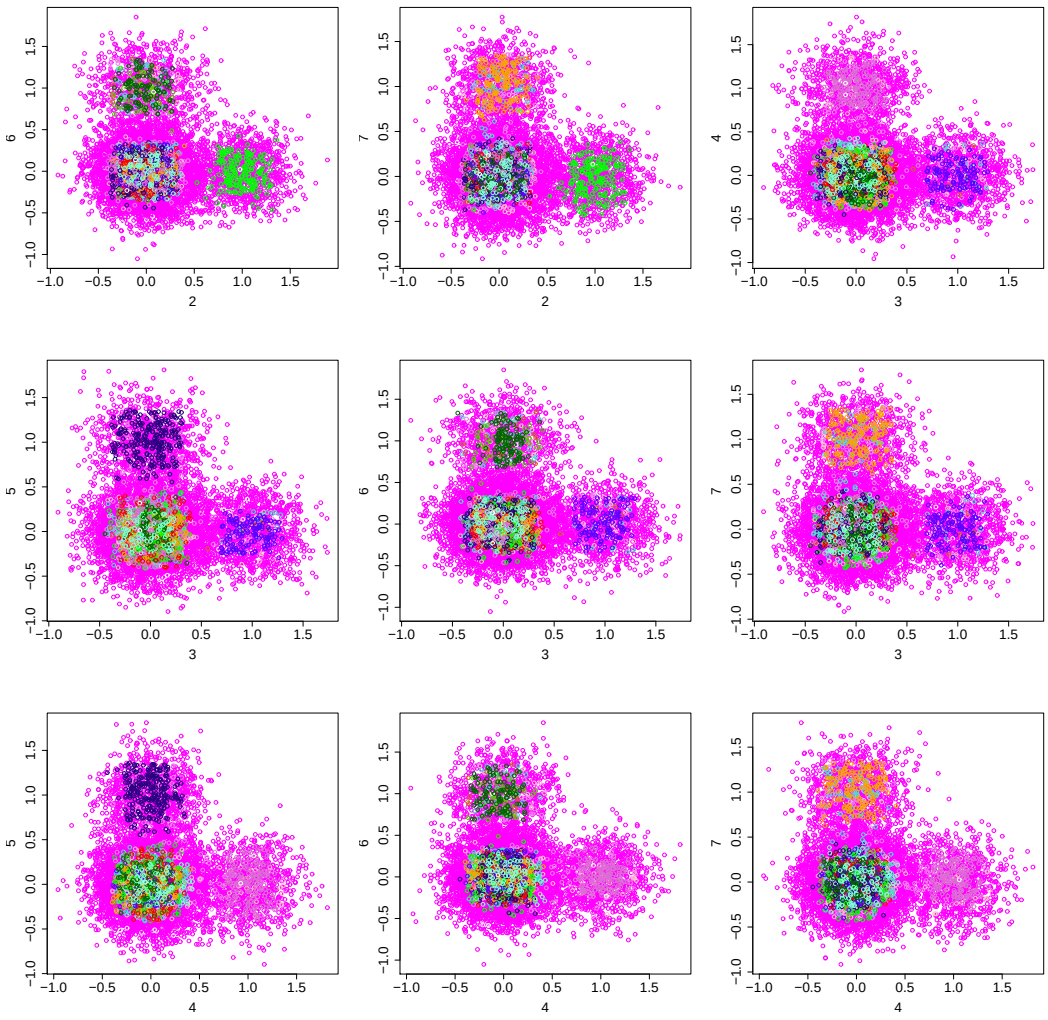

Figure 4: *Scatter plots for the 7D simplex data.* Scatter plots $(X_2, X_6), (X_2, X_7), \ldots, (X_4, X_7)$ are shown.

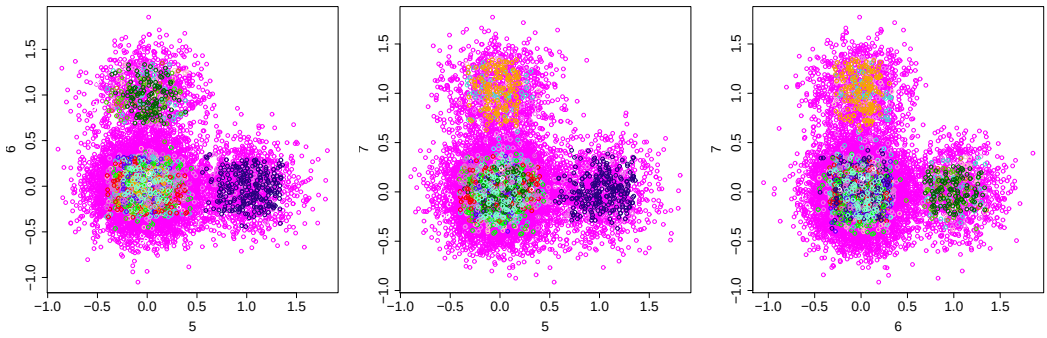

Figure 5: *Scatter plots for the 7D simplex data.* Scatter plots $(X_5, X_6), (X_5, X_7), (X_6, X_7)$ are shown.

6 Flow Cytometry Data: Scatter Plot Matrix and Barycenter Plots

Figure 9 shows a complete 6×6 scatter plot matrix of FCM data, whereas Figure 8 of the article shows only a 4×4 scatter plot matrix.

Figure 10 shows the barycenter plots for the FCM data. The corresponding volume function is shown in Figure 7 of the article. The barycenter plots show six modes, and the coloring of the modes is the same as in the volume function.

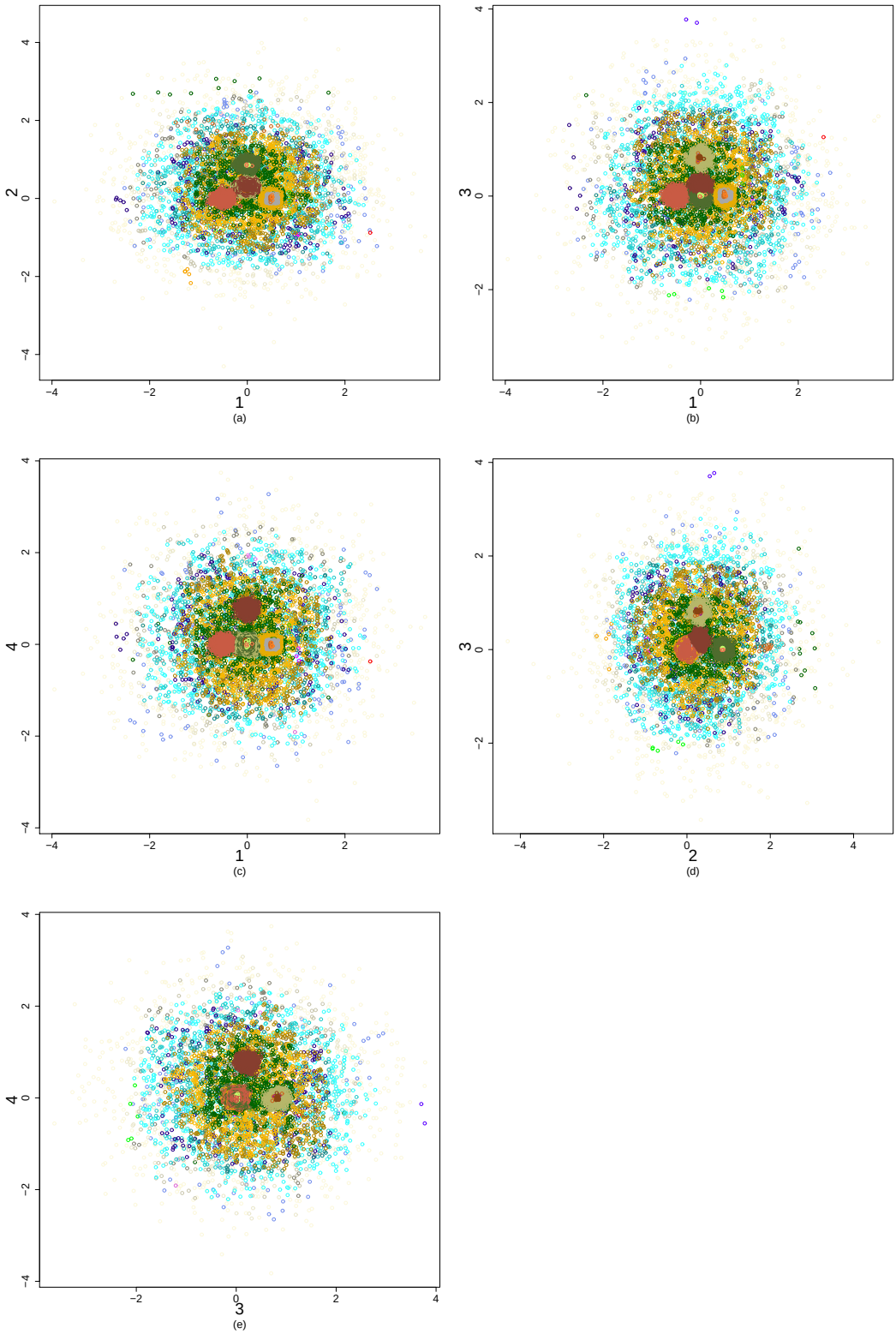

Figure 6: *Scatter plots for the 4D pentahedron data.* All pairwise scatter plots are shown.

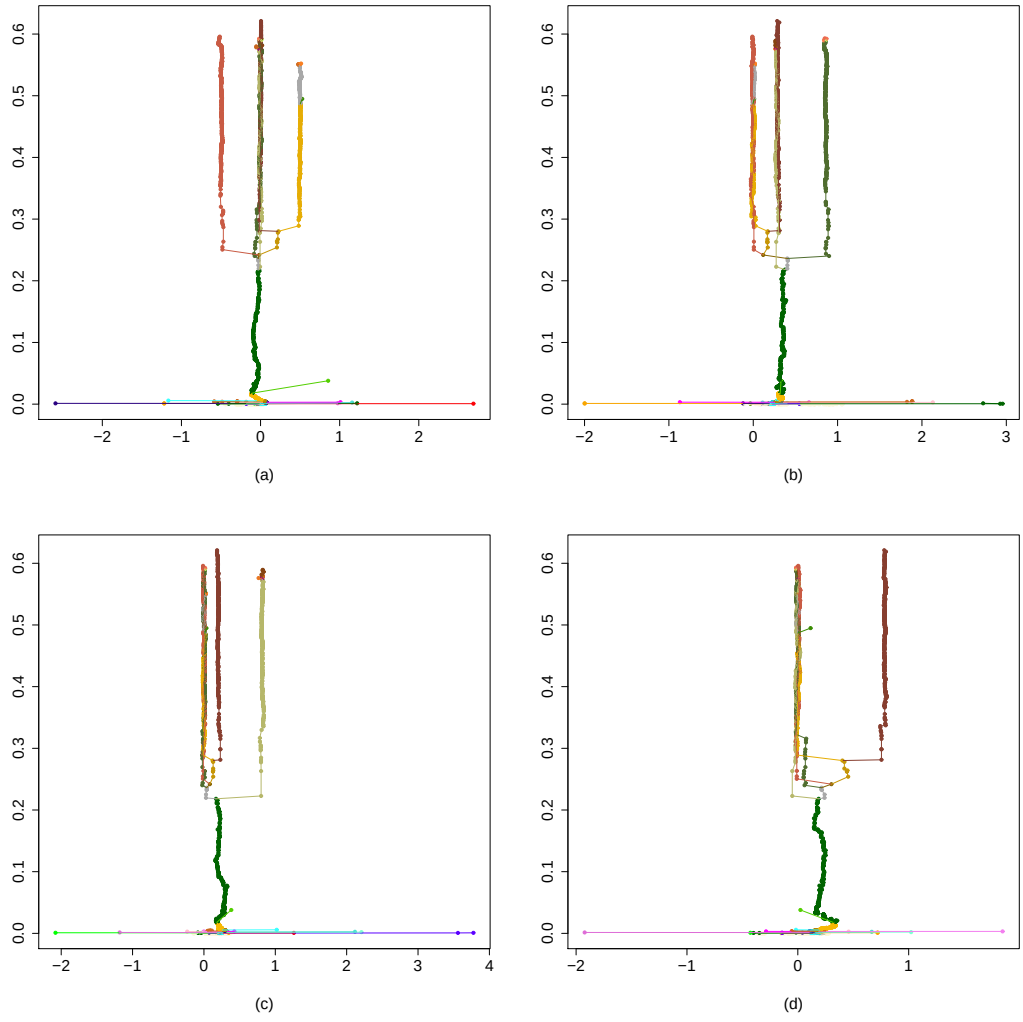

Figure 7: *Barycenter plots for the 4D pentahedron data.* The barycenter plots of the adaptive grid kernel estimator. Panels (a)–(d) show the coordinates 1–4. The estimate is the same as in Figure 6 of the article.

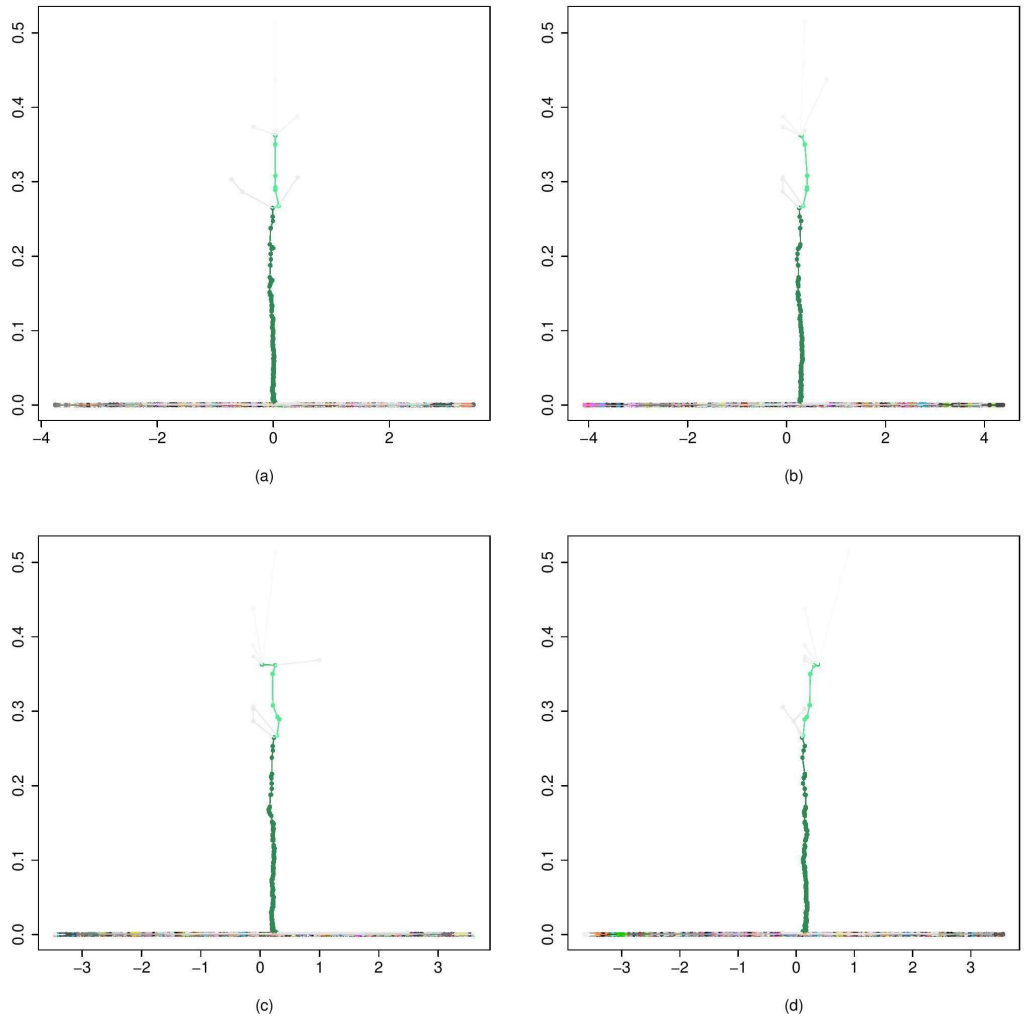

Figure 8: *Barycenter plots for the 4D pentahedron data.* The barycenter plots of the regular grid kernel estimator. Panels (a)–(d) show the coordinates 1–4.

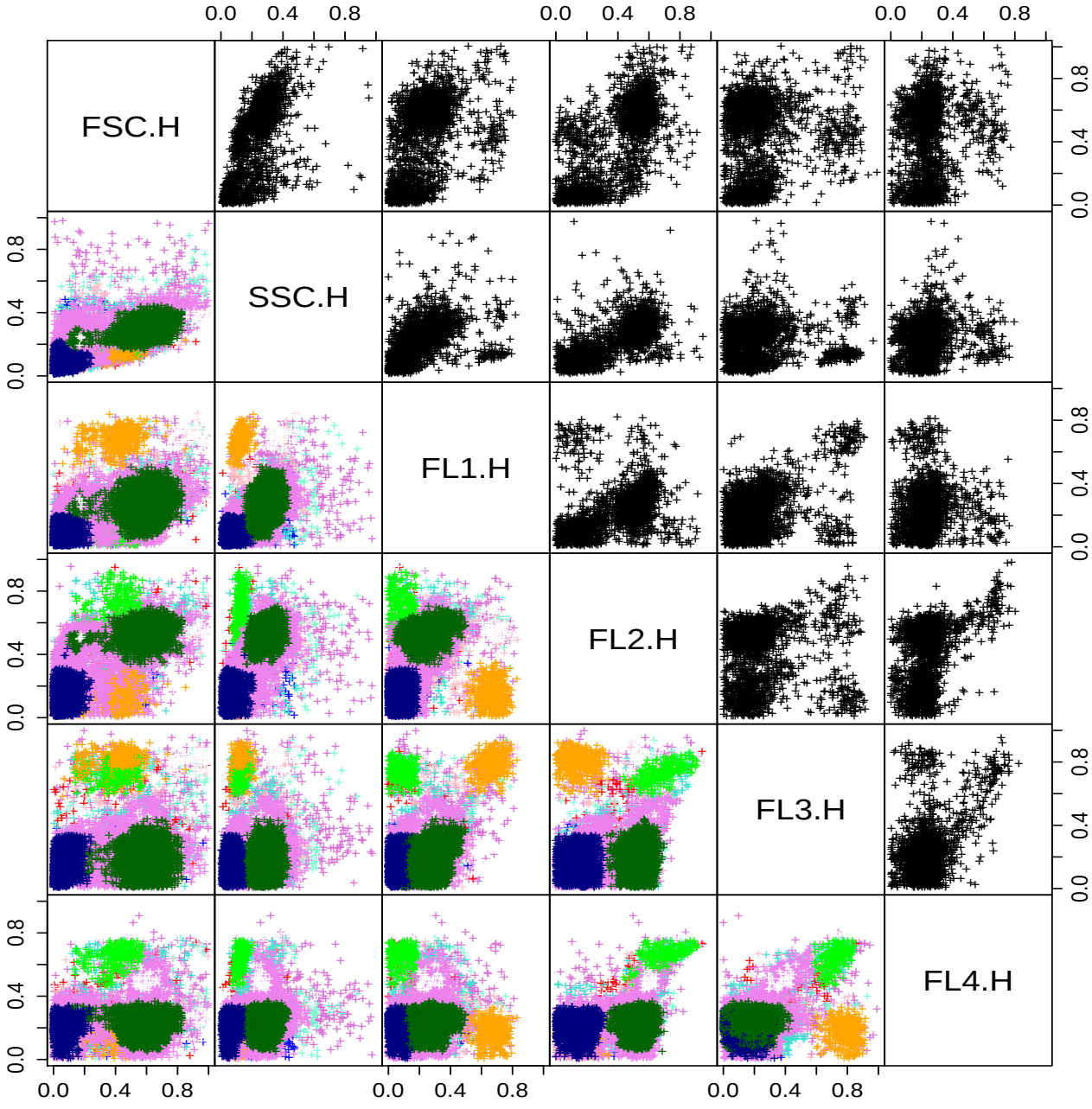

Figure 9: *Scatter plots of FCM data.* The colouring scheme is the same as in Figure 8 of the article. To emphasize the clusters, the densest regions are plotted last, possibly covering sparser cluster regions. Lower left panels show all data, but in the upper right plots a subset of 2000 observations is plotted.

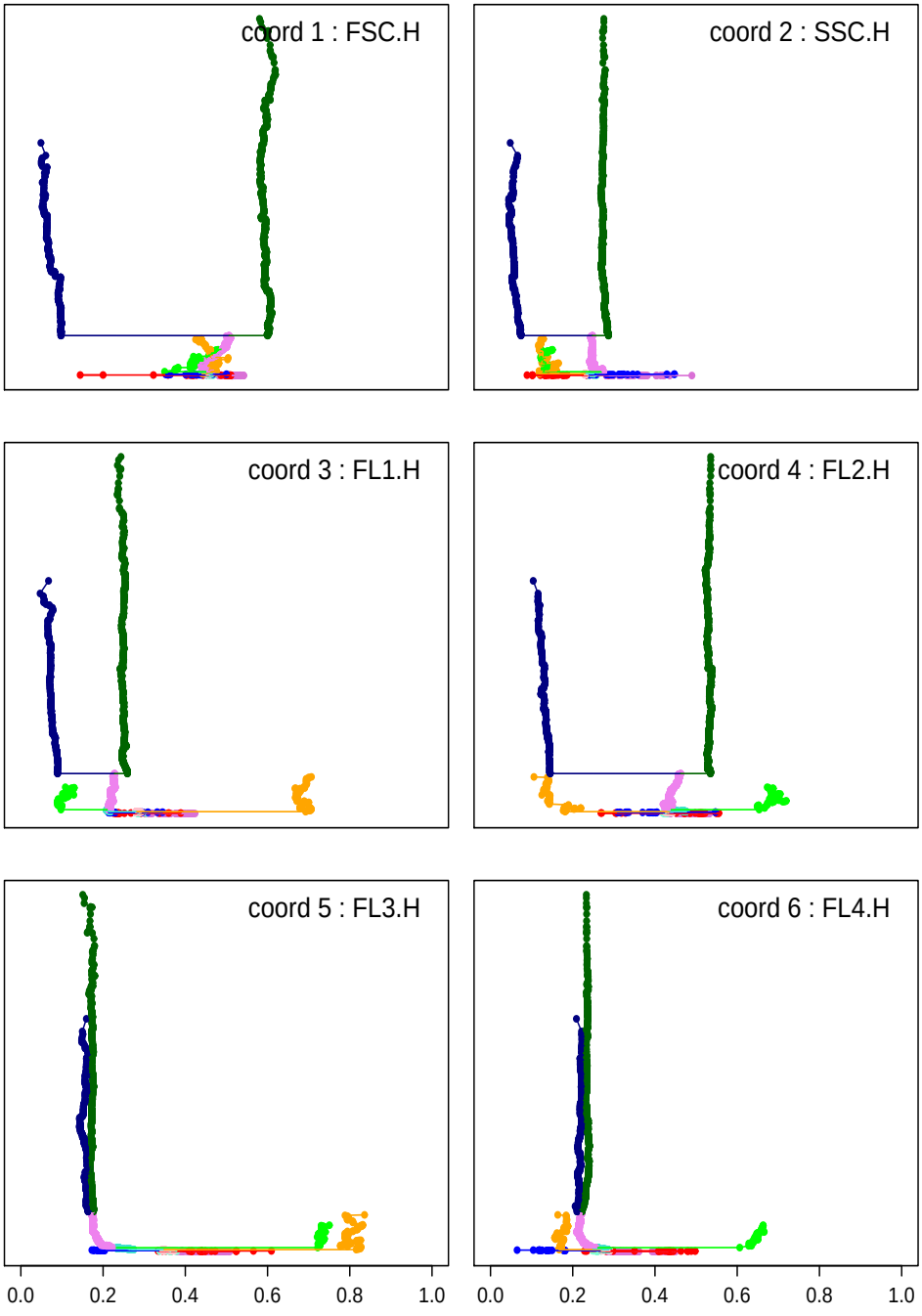

Figure 10: *Barycenter plots for the FCM data.* The six panels show the coordinates 1-6. The density estimate is the same as in Figure 7 of the article.