Feature Selection, Clustering, and Prototype Placement for Turbulence Data Sets

Matthew Barone, Jaideep Ray and Stefan Domino

Sandia National Laboratories[1], Albuquerque, NM, 87185

Online Supplementary Material

In this document, we provide ancilliary information that supports the conclusions of the paper.

Clustering and Prototypes for Bump in Channel

Figure 1 shows the channel flow clusters and wavy wall clusters for the bump flow, with the cluster colors consistent with previous figures. We might expect this situation, given the similarity in geometric configuration between the wavy wall and bump flows. Figure 2 shows the six additional clusters identified by the FSS algorithm. Member points lie in a region above the bump but relatively distant from the surface. There is no obvious physical significance one can attach to these clusters.

Figure 1: Left: Bump in channel data that is well-classified by the channel flow clusters, colored by cluster. Right: Bump in channel data that is well-classified by the wavy wall clusters, colored by cluster.

We have also applied the prototype placement algorithm to the bump dataset. The feature set used is $\xi = f_b = \{C_1, C_2, C_3, \cos\theta_{S-\tau}, \lambda_3\}$. As seen in Figure 3 (left), there are 11 clusters, whose sizes vary by two orders of magnitude (cluster 16 versus cluster 6). In addition, many of the physics (clusters) seen in the previous flows (channel flow and wavy wall) are not seen here; cluster IDs 3, 5, 8, 9 and 10 are missing. The ratio of the number of

[1]Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the 'views of the U.S. Department of Energy or the United States Government.

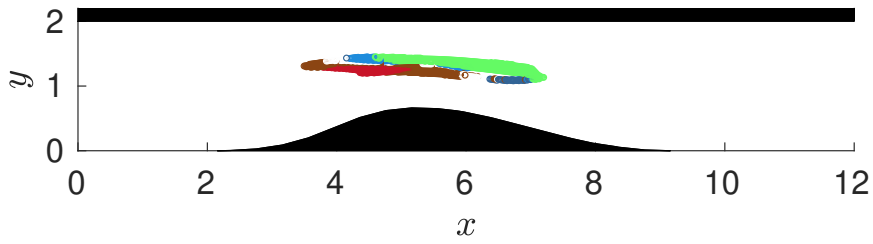

Figure 2: Bump in channel data clusters.

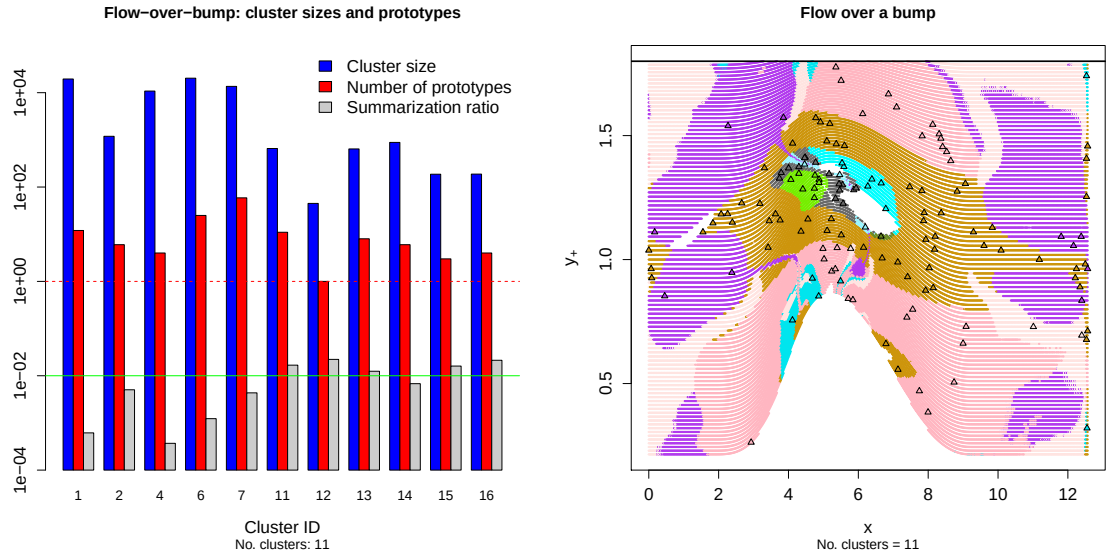

Figure 3: Left: Cluster sizes, prototypes and the degree of summarization obtained. The solid horizontal line corresponds to 0.01 whereas the dashed line corresponds to 1. We see that prototypes summarize the clusters by a factor of about 100x or more. Right: The flow-field segregated into clusters. There are 11 clusters in all.

prototypes placed in a cluster to the cluster size are plotted with gray bars; the prototypes vary between 10% to 0.1% of the cluster. The quality of the cover the prototypes provide is given by $(u, \bar{i}) = (18\%, 15\%)$; $N_{proto} = 139$.

Figure 3 (right) plots the clustering on either side of the bump and the placement of the prototypes. Turbulent flows in regions with favorable and unfavorable pressure gradients are clearly seen, as well at the region above the tip of the bump, where the flow quickly changes character. We see that the prototypes are *not* uniformly distributed in physical space, indicating severe contortions of the clusters as they mapped between physical and ξ-space where clustering and prototype-pacement is performed.