A Metric For Assessing The Quality of Low-Rank Models

PRESENTER: Joshua Cook

Background

A low-rank model (LRM, e.g. PCA) is frequently used as part of the greater pipeline in preparing a cluster model.

Raw Data → Low Rank Model → Cluster Model

An unsupervised learning task, a LRM can not be assessed against a label vector, but must be measured against some intrinsic quality of the data.

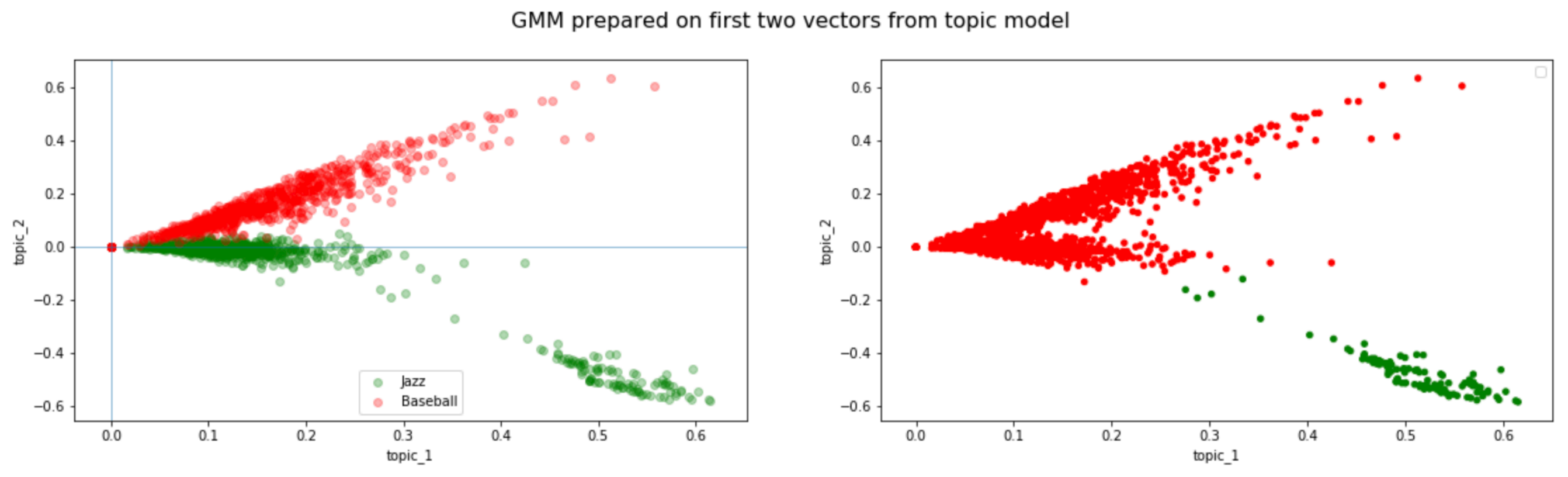
Semantic Coherence

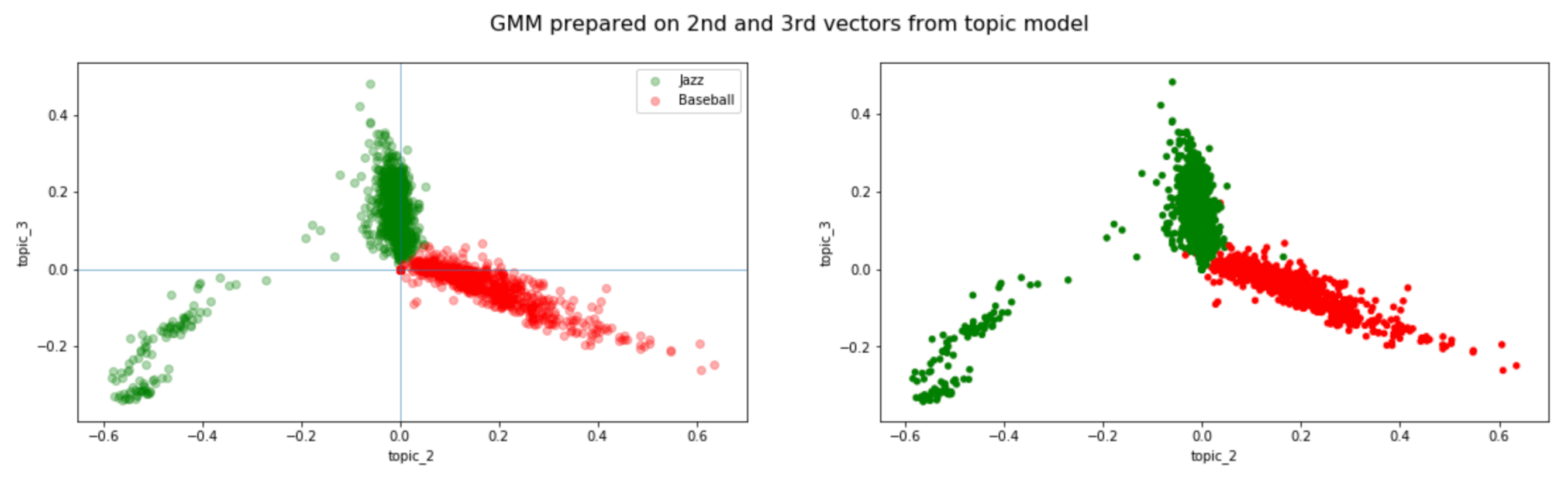
Newman, et al (2010) and Mimno, et al (2011) introduced the metric **Semantic Coherence** for assessing the quality of LRM topic models as part of the preparation of a document clustering model.

Applications of Semantic Coherence

Consider the following cluster models prepared using article content from two Wikipedia categories: Jazz and Baseball. The first is prepared with the first two vectors of the LRM. The second is prepared with the second and third vectors of the LRM.

In each, the plot on the left is a plot of the articles labeled by actual category. The plot on the right is labeled with the results of a Gaussian Mixture Model (GMM).


GMM prepared on first two vectors from topic model


GMM prepared on 2nd and 3rd vectors from topic model

Note that the first model fails to capture the correct clustering, where the second succeeds.

Failure Due To Lack of Coherence

The first model fails due to a lack of semantic coherence in the first vector of the low rank representation. This can be seen in looking at ten most important words in each topic. Intuitively, we can see that Topics 2 and 3 are more semantically coherent.

Topic 1		Topic 2		Topic 3	
semantic coherence: 5.0211		**semantic coherence: 5.8259**		**semantic coherence: 6.9165**	
american	born	baseball	batter	jazz	new
died	pianist	ball	pitcher	records	released
jazz	league	base	team	label	labels
baseball	ball	league	runner	music	recordings
composer	saxophonist	game	home	record	musicians

The purpose of the Semantic Coherence metric is to identify this lack of coherence automatically, without leveraging human understanding of the topics. The Semantic Coherence metric has captured this, showing Topics 2 and 3 to be more semantically coherent than Topic 1.

A generalized approach to computing the **Coherence** of a low-rank model provides an objective, scale-invariant measurement its quality.

Take a picture to download the full paper

Coherence

Semantic Coherence is a metric specific to the domain of topic modeling. We propose a metric **Coherence**, more general than Semantic Coherence, and capable of being applied to a more general category of LRMs.

The Metric: Coherence

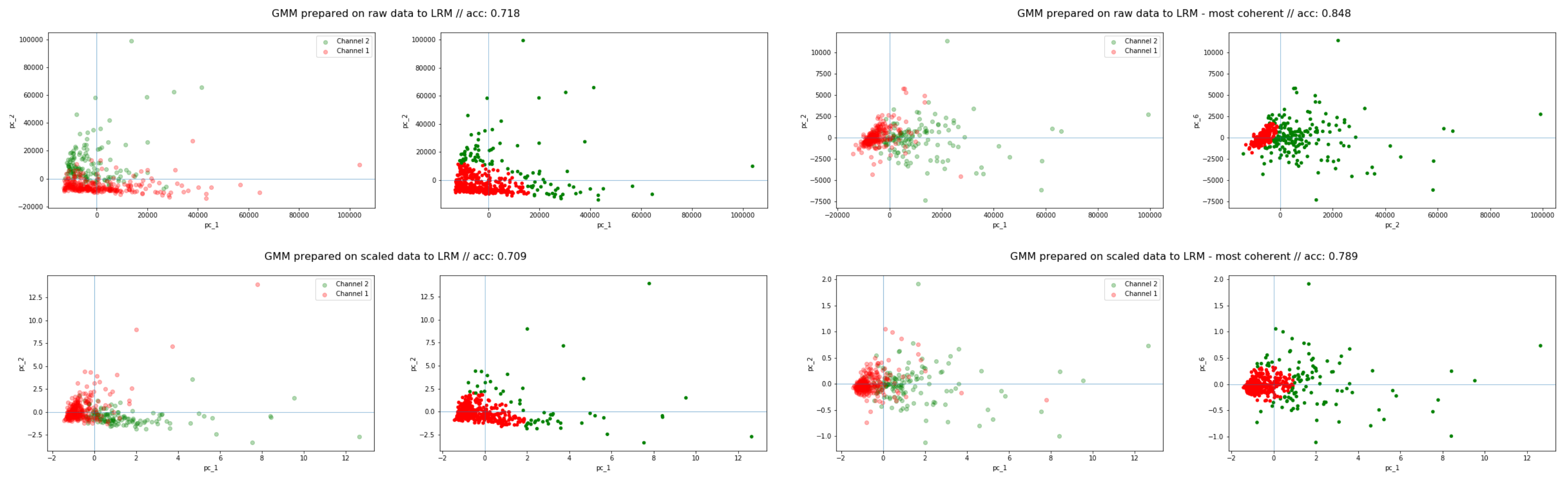
1. Low-rank models typically generate a loadings matrix, L, as a by-product of the model fitting process.
2. The values of a loading matrix column represents the expression of each original feature vector in the corresponding low-rank model column.
3. These values can be used to compute a scaled mutual information for each pair of vectors.
4. A sum taken over all pairs is an intrinsic measurement of the mutual information in the given column vector. We call this sum **coherence**.

$$C_i = \sum_{f_j \in F} \sum_{\substack{f_k \in F;\\ j \neq k}} L_{ji} \cdot L_{ki} \mathrm{MI}(f_j, f_k)$$

Applications of Coherence

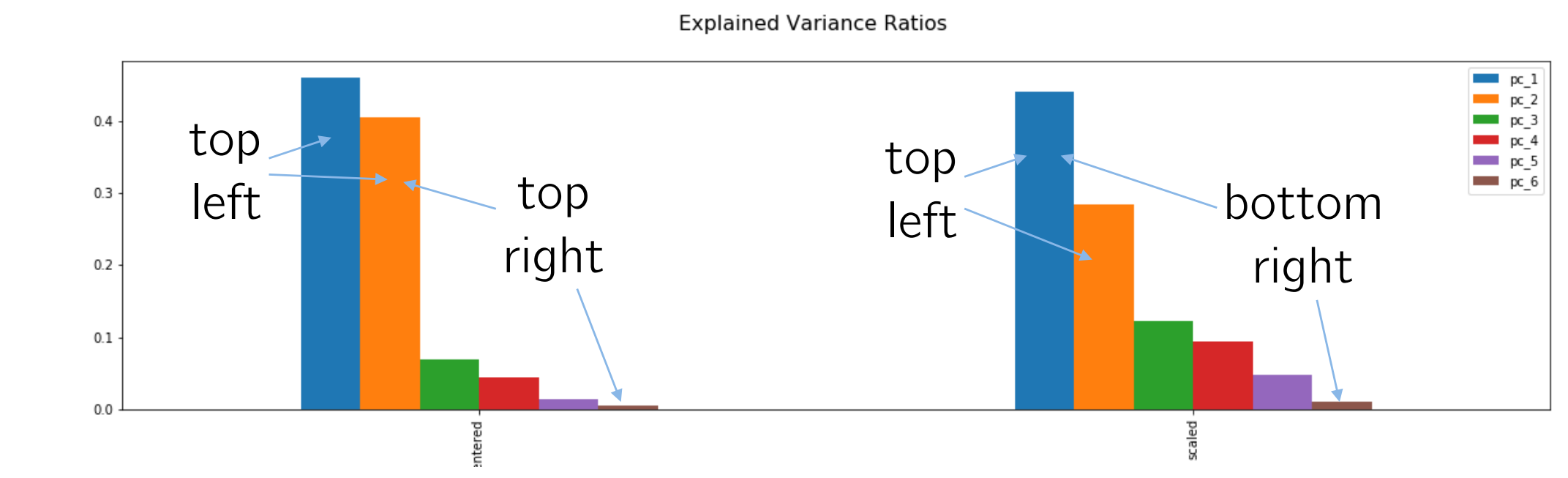
Consider the following cluster models prepared using the Wholesale Customers data set made available by the UCI Machine Learning Repository. Each of the four plots shows actual labels on the left and labels generated by preparing a cluster model using a GMM and two features from a PCA model. The top row are models prepared using centered data passed to a PCA model. The bottom row are models prepared with an additional preprocessing step: the data are scaled by the standard deviation of each feature prior to being passed to a PCA model.

The labels generated by the GMM are compared to the actual labels. This results is reported as accuracy (acc).



Results

The typical process for selecting vectors for an LRM is to use the explained variance of each of the vectors. This process was followed for the generation of the two models on the left. For each of these models, the first two principal components corresponding to those with the highest explained variance were selected.



An alternative process was used for the two models on the right. For each of these, the principal components with the highest coherence were selected. For the model on the top right, this was principal components 2 and 6. For the model on the bottom right, this was principal components 1 and 6.

Preliminary results on this small data set show that cluster models using LRM vectors selected using coherence to be more performant than models generated by maximizing explained variance.

Joshua Cook
@joshuacook
http://joshuacook.me