Accepted Manuscript

ISSN 1063-5203
ELSEVIER
Applied and Computational Harmonic Analysis
Time–Frequency and Time–Scale Analysis, Wavelets, Numerical Algorithms, and Applications
Available online at www.sciencedirect.com
ScienceDirect

A variational approach to the consistency of spectral clustering

Nicolás García Trillos, Dejan Slepčev

PII: S1063-5203(16)30063-X
DOI: http://dx.doi.org/10.1016/j.acha.2016.09.003
Reference: YACHA 1164

To appear in: *Applied and Computational Harmonic Analysis*

Received date: 7 August 2015
Accepted date: 12 September 2016

Please cite this article in press as: N.G. Trillos, D. Slepčev, A variational approach to the consistency of spectral clustering, *Appl. Comput. Harmon. Anal.* (2016), http://dx.doi.org/10.1016/j.acha.2016.09.003

A VARIATIONAL APPROACH TO THE CONSISTENCY OF SPECTRAL CLUSTERING

NICOLÁS GARCÍA TRILLOS AND DEJAN SLEPČEV

Abstract. This paper establishes the consistency of spectral approaches to data clustering. We consider clustering of point clouds obtained as samples of a ground-truth measure. A graph representing the point cloud is obtained by assigning weights to edges based on the distance between the points they connect. We investigate the spectral convergence of both unnormalized and normalized graph Laplacians towards the appropriate operators in the continuum domain. We obtain sharp conditions on how the connectivity radius can be scaled with respect to the number of sample points for the spectral convergence to hold. We also show that the discrete clusters obtained via spectral clustering converge towards a continuum partition of the ground truth measure. Such continuum partition minimizes a functional describing the continuum analogue of the graph-based spectral partitioning. Our approach, based on variational convergence, is general and flexible.

1. Introduction

Clustering is one of the basic problems of statistics and machine learning: having a collection of n data points and a measure of their pairwise similarity the task is to partition the data into k meaningful groups. There is a variety of criteria for the quality of partitioning and a plethora of clustering algorithms, overviewed in [14, 36, 52, 53]. Among most widely used are centroid based (for example the k-means algorithm), agglomeration based (or hierarchical) and graph based ones. Many graph partitioning approaches are based on dividing the graph representing the data into clusters of balanced sizes which have as few as possible edges between them [4, 5, 24, 37, 41, 42, 51]. Spectral clustering is a relaxation of minimizing graph cuts, which in any of its variants, [29, 37, 49], consists of two steps. The first step is the embedding step where data points are mapped to a euclidean space by using the spectrum of a *graph Laplacian*. In the second step, the actual clustering is obtained by applying a clustering algorithm like k-means to the transformed points.

The input of a spectral clustering algorithm is a weight matrix W which captures the similarity relation between the data points. Typically, the choice of edge weights depends on the distance between the data points and a parameter ε which determines the length scale over which points are connected. We assume that the data set is a random sample of an underlying ground-truth measure. We investigate the convergence of spectral clustering as the number of available data points goes to infinity.

For any given clustering procedure, a natural and important question to consider is whether the procedure is consistent. That is, if it is true that as more data is collected, the partitioning of the data into groups obtained converges to some meaningful partitioning in the limit. Despite the abundance of clustering procedures in the literature, not many results establish their consistency in the nonparametric setting, where the data is assumed to be obtained from an unknown general distribution. Consistency of k-means clustering was established by Pollard [32]. Consistency of

Date: September 21, 2016.

1991 *Mathematics Subject Classification.* 49J55, 49J45, 60D05, 68R10, 62G20.

Key words and phrases. spectral clustering, graph Laplacian, point cloud, discrete to continuum limit, Gamma-convergence, Dirichlet energy, random geometric graph .