Contributions to multiscale statistical analysis on random geometric graphs

Franziska Göbel
Universität Potsdam

Summary

The aim of statistical data analysis is to describe and extract and discover (useful) information and conclusions from collected data by applying statistical methods. These methods are diverse and they depend on and are adapted to the task and the type of data.

The initial task of interest in this thesis is the following (fixed design) regression problem, known as denoising problem: informally speaking, having observed the noisy values of a function at a finite set of points, the aim is to recover the true values of the function at these points. To be more precise, the model $Y = f(x) + \epsilon$ with additive noise ϵ (independent and identically distributed) is considered. Given a set of points $(x_1, \ldots, x_n)$ with $x_i \in \mathbb{R}^K$ the values $y_i = f(x_i) + \epsilon_i \in \mathbb{R}$ at point x_i for $i = 1, \ldots, n$ are observed. The goal is to recover $f(x_i)$ for $i = 1, \ldots, n$ given $(x_i, y_i)_{i=1,\ldots,n}$ where f may be of inhomogeneous regularity. In the classical setting with $x_i \in \mathbb{R}$, wavelet bases are widely used, especially orthonormal wavelets with compact support. Due to their properties like multi-resolution and localization in time and space they are in particular well-suited for the (sparse) representation and estimation of functions with inhomogeneous regularity (for example a smooth function with jumps). In this thesis the focus is on a special kind of data, namely data with intrinsic geometrical structure. The interest in data with intrinsic geometrical structure is motivated by high-dimensional data analysis. Many classical methods suffer from the curse of dimensionality. A way out was presented by the observation that data sets in applications often exhibit an intrinsic structure of low dimensionality. The common approach is that the structure can be described as submanifold $\mathcal{M}$ of $\mathbb{R}^K$. Methods that exploit this structure depend on its low dimension and not on the dimension of the ambient space. But in general the structure is unknown. A fundamental tool introduced to recover, implicitly or explicitly, this unknown structure is the construction of a neighborhood graph based on the observed sample. A neighborhood graph is a geometric graph whose vertices are the sample points themselves, and neighbor points which are defined in a suitable sense based on the ambient Euclidean distance in $\mathbb{R}^K$ are joined by an edge. Neighbors may be defined as points whose distance is smaller than a fixed $\varepsilon > 0$ resulting in the ϵ-graph. Another possibility are the k-nearest neighbor graphs where each point is connected to its k closest neighbors. In spectral graph theory it is known that the graph Laplace operator $\mathbf{L}$ (which can be defined in various ways) is a very useful tool in analyzing properties of graphs.

The main question addressed in this thesis is: Can a dictionary adapted to the intrinsic structure of observed data be constructed that exhibits some properties

of wavelets, making it in particular well suited for denoising functions of inhomogeneous regularity? In this thesis a positive answer is given to this question. Given a neighborhood graph representation of a finite set of points $x_i \in \mathbb{R}^K, i = 1, \ldots, n$, the construction of a frame (redundant dictionary) $\{\Psi_{k\ell}\}_{(k,\ell)\in\mathcal{I}_\mathcal{F}}$ for the space of real-valued functions defined on the graph is presented.

The construction proposed here, based on a transform of the spectral decomposition of the graph Laplacian, follows the ideas of a work from Hammond, Vandergheynst und Gribonval (2011). One key point is that the construction in this thesis ends up with a tight (or Parseval) frame: $\|f\|^2 = \sum_{(k,\ell)\in\mathcal{I}_\mathcal{F}} |\langle f, \Psi_{k\ell}\rangle|^2$. This means a very simple, explicit reconstruction formula holds for every function f defined on the graph: $f = \sum_{(k,\ell)\in\mathcal{I}_\mathcal{F}} \langle \Psi_{k\ell}, f\rangle \, \Psi_{k\ell}$. The choice of the multiscale bandpass filter functions in the construction is inspired by a work of Coulhon, Kerkyacharian und Petrushev (2012), This frame is adapted to the underlying geometrical structure of the x_i, has finitely many elements, and these elements are localized in frequency as well as in space. Most of these properties are reminiscent of wavelets. That is why the frame elements are called "Graph Wavelets"(even though this name is already used in literature for other sometimes similar objects) in this thesis. The reconstruction formula is used in the setting of denoising where noisy observations of a function f defined on the graph are given. By applying a thresholding method to the coefficients in the reconstruction formula, an estimator of f is proposed whose risk satisfies a tight oracle inequality. Furthermore the application of the frame in semi-supervised prediction is investigated. In contrast to the denoising setting, only a (random) part of the data values y_i are observed, while all the covariate points x_i are known. The aim is to determine the missing function values.

The „Graph Wavelet“ octave/matlab library offers an implementation of the proposed data-dependent Parseval frame construction which adapts to random data points having an unknown low-dimensional structure (e.g. lying on a low-dimensional manifold). It provides tools for signal denoising based on thresholding of corresponding signal coefficients. In experiments based on the package the spatial localization is visualized, and the performance of the frame thresholding estimator in the denoising setting compared to reference methods is investigated, in particular good results are obtained when using scale-dependent thresholds.

Strongly related to the spatial localization of the frame elements is the existence of a (subgaussian) heat kernel bound associated to the graph Laplace operator. In certain cases this is in turn equivalent to two fundamental measure-metric properties of the associated graph: namely a volume doubling condition and a local Poincaré inequality. The central question is therefore: do (random) neighborhood graphs satisfy these two properties? The focus in this thesis is on random ε-neighborhood graphs whose vertices are drawn independently and identically distributed from a common distribution defined on a regular submanifold of $\mathbb{R}^K$. For this class of neighborhood graphs it is shown that a volume doubling condition and local Poincaré inequality hold (with high probability, and uniformly over all shortest path distance balls in a certain radius range) under suitable regularity conditions of the underlying submanifold and the sampling distribution.