Multi-resolution Regression, Divide and Conquer Risk Estimation, and the Large-scale Universe

Collin Eubanks
Advisors: Larry Wasserman, Jessi Cisewski, Rupert Croft
Committee: Peter Freeman, Ryan Tibshirani

May 31, 2017

Abstract

In this proposal, we extend our previous work on mapping the large-scale intergalactic medium by developing relevant areas of statistical methodology and subsequently applying them to the twelfth data release of the Baryon Oscillation Spectroscopic Survey (BOSS DR12). Specifically, our two main statistical contributions, as outlined below, are the development of a novel multi-resolution spatio-temporal model on the sphere and a highly generalizable, computationally efficient algorithm for predictive risk estimation of nonparametric regression estimators.

In the first part of this proposal, we develop a multi-resolution regression model for predicting spatio-temporal fields on the sphere embedded in three-dimensional Euclidean space. The multi-resolution feature of the model stems from the organization of wavelet functions on lattices of increasing resolution in $S^2 \times \mathbb{R}^+$. Several potential configurations are discussed, however, the proposed model for mapping the IGM exploits irregular lattices constructed from a combination of a HEALPix hierarchical tessellation and k-centers clustering of the sphere. In particular, this organization is designed to create efficient local adaptivity in the time domain by mimicking the highly irregular redshift distribution of BOSS DR12. In this setting, the temporal dimension of the model is regarded as perpendicular to the surface of the sphere, signifying the lookback time to an observed location in intergalactic space. Furthermore, by defining the wavelet coefficient vector to be the solution of an $L_1$-penalized weighted regression, we allow for further efficient local adaptivity at fixed redshifts, namely in relatively noisy and/or sparsely sampled regions of the two-dimensional sky.

In the second part of this proposal, we develop a spatial divide and conquer (SDAC) algorithm for efficient estimation of predictive risk curves for a very general class of nonparametric regression estimators. The generalizability is achieved by sampling compact, disjoint subsets of the data instead of random (non-compact) subsets, as is typically done in divide and conquer algorithms. A predictive risk curve is then estimated by cross validation, or another criterion, for each compact subset and the global risk curve is then taken to be a function of the local risk curves. In particular, we consider both a weighted average of the local risk curves and a weighted geometric median. A natural implication of this algorithm is a procedure for (global) tuning parameter selection. We have used this algorithm as a heuristic but, to this point, a theoretical foundation has not been developed. Our primary goals in laying this foundation
are: (1) to establish convergence rates, asymptotic distributions, and approximate confidence intervals for the SDAC risk estimators and their respective minimum risk tuning parameters in various nonparametric regression settings, and (2) to explore functional data analysis procedures for testing whether the modeling could be improved by utilizing a spatially adaptive tuning parameter.

1 Introduction

Pictured in Figure 1.1 is a cross section of the largest scale three-dimensional map of the Universe currently available (Eubanks et al. [In preparation], Eubanks et al. [In preparation]). As part of a collaboration between the Department of Statistics, Machine Learning Department and McWilliams Center for Cosmology at Carnegie Mellon University, the Department of Statistics at Yale University, and the National Center for Atmospheric Research (NCAR), we constructed this map from approximately 100 million Lyman-α absorptions in the quasar spectra of the twelfth data release of the Baryon Oscillation Spectroscopic Survey (BOSS DR12) of SDSS-III (Alam et al. [2015]). The map therefore displays the relative over-densities and under-densities of H I in a region of the intergalactic medium (IGM) intervening Earth and the targeted background quasars, with a sky coverage of 10,400 deg$^2$ ($\sim$25% sky coverage) and spanning the redshift range $1.94 < z < 3$. The map will soon become available for interactive viewing on the Microsoft Research WorldWide Telescope.

In particular, we constructed this map using a multi-resolution Gaussian random field (GRF) model developed for spatial modeling in up to three dimensions by our collaborator at NCAR (Nychka et al. [2014]). The GRF model is a modified kriging (Krige [1951], Matheron [1963]) algorithm designed specifically for very large, irregularly sampled, heteroskedastic spatial data sets—all of which describe BOSS DR12. Despite the notable strengths of the GRF model in the context of this problem, we believe there is room to improve upon the map presented in Eubanks et al. [In preparation]. In this proposal we therefore develop a novel model specifically designed to address the shortcomings of the GRF model while still preserving its strengths. The proposed model makes use of a multi-resolution wavelet basis organized on irregular lattices in $S^2 \times \mathbb{R}^+$ according to a combination of a HEALPix hierarchal tessellation (Górski et al. [2005]) and $k$-centers clustering of the sphere embedded in three-dimensional Euclidean space, accompanied by a temporal dimension (redshift). This organization is specifically designed to create local adaptivity in redshift space by mimicking the highly irregular redshift distribution of BOSS DR12 observations. Such a construction will thus provide a much more efficient mapping than the GRF model, which utilized a regular three-dimensional Euclidean lattice. Furthermore, we define the wavelet coefficient vector to be the solution of an $L_1$-penalized weighted regression, which allows for further efficient local adaptivity at fixed redshifts, namely in relatively noisy and/or sparsely sampled regions of the two-dimensional sky. We will refer to this model as a three-dimensional spherical wavelet (SW) model.

The second major component of this proposal centers around a spatial divide and conquer (SDAC) algorithm we developed in Eubanks et al. [In preparation] for efficient predictive risk estimation in order to make tuning parameter estimation feasible when modeling a data set as large as BOSS DR12. The idea of the algorithm is simple:
Figure 1.1: A cross section of the three-dimensional map of the intergalactic medium presented in Eubanks et al. [In preparationa] at $z \approx 2.1$. The map has a sky coverage of approximately 10,400 deg$^2$ ($\sim$25% sky coverage) and spans the redshift range $1.94 < z < 3$. The negative and positive $\delta$-fluctuations signify relative over-densities and under-densities of H I, respectively.

1. Sample compact, disjoint subsets $S_1, \ldots, S_m$ of the feature space, for some $m$.

2. Compute the respective local risk functions $\hat{R}_1(\hat{f}_\theta), \ldots, \hat{R}_m(\hat{f}_\theta)$, where $\hat{f}_\theta$ is the predicted function depending on a tuning parameter $\theta$.

3. Combine them in some way to obtain a global risk function estimate. That is,

$$\hat{R}(\hat{f}_\theta) = h(\hat{R}_1(\hat{f}_\theta), \ldots, \hat{R}_m(\hat{f}_\theta)), \quad (1.1)$$

for some function $h$. 

In the case of the GRF model, each $\hat{R}_i$ took the form of the (negative) log-likelihood inherent to the covariance of the data induced by the GRF assumption of the model. Nevertheless, the algorithm can generalize to any notion of risk. Furthermore, although $h$ can also be arbitrary, our initial investigation will consider the cases where it is a weighted average or a weighted geometric (spatial) median. As the presentation of the algorithm provided in Eubanks et al. [In preparationa] was heuristic and ad hoc, the goal in the thesis is not only to present it in full generality, but also establish a foundation of theoretical guarantees for both the risk estimates of the algorithm.
and the corresponding tuning parameter estimates. Furthermore, the spatial nature of the SDAC algorithm affords the interesting opportunity to establish testing procedures for the use of spatially adaptive nonparametric smoothers. That is, given the local risk functions \( \hat{R}_1(f), \ldots, \hat{R}_m(f) \in \mathbb{R}^r \) computed over the grid of tuning parameters \( \{\theta_1, \ldots, \theta_r\} \), how can we combine them in such a way to test whether a single fixed tuning parameter \( \theta \) is indeed suitable for modeling the data? Given the observations in this case are functions, this investigation will lie within the realm of functional data analysis.

The remainder of this proposal is organized as follows. In Section 2 we provide a background on the IGM and BOSS DR12—the data set that motivated our proposed research. In Section 3 we give an overview of our previous work in applying the GRF model to BOSS DR12 to construct a three-dimensional map of the IGM. Lastly, in Section 4 we will map out our proposed research plan for both the SW model and the SDAC algorithm.

2 Background

2.1 The intergalactic medium

The intergalactic medium (IGM) is a diffuse gas that permeates intergalactic space and hosts a majority of the Universe’s baryonic matter. Since the evolution of the Universe is intrinsically tied to the distribution of the matter in it, conclusions about the structure of the IGM are therefore of paramount cosmological significance. The IGM itself does not emit light but its presence and structure are revealed by the absorption and emission features in the spectra of luminous background quasars (Lynds [1971]). As light travels to Earth from a quasar it occasionally intersects intergalactic gas and is partially absorbed. Absorptions occur at multiple wavelengths because of the various chemical elements in the IGM; however, the absorptions at the Lyman-\( \alpha \) spectral line are of particular interest because they can be used to trace the density of H I, which constitutes a majority of the baryonic matter in the Universe. The series of absorptions originating from the Lyman-\( \alpha \) spectral line—referred to as the Lyman-\( \alpha \) forest—thus provides a means for probing the H I density fluctuations along one-dimensional lines of sight (LOSs) to quasars. Given an appropriate choice of model, a sufficiently dense set of these LOSs, such as that provided by the recently completed Baryon Oscillation Spectroscopic Survey (BOSS) of the Sloan Digital Sky Survey’s third phase (SDSS-III), can be used to construct a continuous map of the IGM in three dimensions.

Quasars are distant, extremely luminous cosmological objects powered by supermassive black holes. The immense gravity of these black holes and the friction of the matter falling inward result in an outpouring of thermal radiation from outside the event horizon that shines anywhere between 10 and 10,000 times brighter than the entire Milky Way. This thermal radiation spans the electromagnetic spectrum and allows astronomers to study the IGM by analyzing the absorption patterns in the observed spectra left by the chemical elements present in intergalactic space. Quasar spectra are particularly dominated by the absorptions at the Lyman-\( \alpha \) spectral line (rest wavelength of 1215.67 Å) because of the prominence of H I in the IGM. A Lyman-\( \alpha \) absorption occurs when a photon with wavelength 1215.67 Å hits a H I atom in intergalactic space, is absorbed, and sends the H I electron from the ground state (\( n = 1 \)) to the first excited state (\( n = 2 \)). This absorption results in a decrease in the flux of the quasar continuum at the Lyman-\( \alpha \) spectral line, where flux
is a measure of the amount of energy traversing a two-dimensional surface per unit time per unit area. Mathematically,

\[ f(\lambda) = \frac{\partial^2 Q_e(\lambda)}{\partial t \partial A} \quad (2.1) \]

where \( Q_e(\lambda) \) is the radiant energy received at wavelength \( \lambda \), \( t \) is time, and \( A \) is area. An observed quasar spectrum, as seen in Figure 2.1, is therefore characterized by the measured flux \( f \) of the quasar at each wavelength \( \lambda \). Here, due to the expansion of the Universe, the Lyman-\( \alpha \) forest is seen originating from \( \lambda \approx 4296 \ \text{Å} \) and extending blueward (to the left).\(^1\)

### 2.2 BOSS DR12

In order for the Lyman-\( \alpha \) forest to be observed by ground-based spectrographs such as those used by BOSS, the Lyman-\( \alpha \) spectral line, which rests in the ultraviolet (UV) portion of the spectrum, must be stretched by a factor of \( \gtrsim 2.9 \) since most UV light is absorbed in Earth’s stratosphere. Thus, BOSS DR12 does not contain any Lyman-\( \alpha \) data from sources at redshift \( z \lesssim 1.9 \). Furthermore, due to a lack of photons from high-redshift sources reaching Earth, BOSS DR12 contains relatively few spectra for quasars at \( z > 3 \) (see the top right panel of Figure 2.2). Consequently, we focus on mapping the BOSS coverage in the redshift range \( 1.94 < z < 3 \). Because looking into the depths of space is akin to looking back in time, such a map gives an imprint of the IGM as it was between 10.3 and 11.5 billion years ago.

The Lyman-\( \alpha \) absorptions alone do not convey information about the H I density along the LOS, without some notion of the flux of the quasar continuum. The bottom panel of Figure 2.1 displays the Lyman-\( \alpha \) absorptions extracted from the spectrum of a BOSS DR12 quasar, with a very smooth local polynomial fit shown in red. Here the local polynomial fit \( m(\lambda) \) approximates the shape of the quasar continuum and allows us to estimate \( \delta \)-flux measurements

\[ \delta^j_F(\lambda_i) = \frac{f^j(\lambda_i)}{m^j(\lambda_i)} - 1, \quad (2.2) \]

for this spectrum \( j \) at each wavelength \( \lambda_i \). By construction, \( \delta_F > 0 \) corresponds to a higher-than-average transmitted flux\(^2\) and \( \delta_F < 0 \) corresponds to a lower-than-average transmitted flux. Applying this transformation to each individual spectrum therefore results in a response variable that is inversely proportional to H I density and can be directly compared across spectra. Inevitably, the choice of \( m(\lambda) \) can have an effect on the eventual density map. In order to calibrate how significant this effect may be, we also compute \( \delta \)-flux measurements using the stacked mean spectrum approach of Harris et al. [2016] and use this data set to produce a second three-dimensional map.

Altogether we utilize Lyman-\( \alpha \) data from 168,953 BOSS DR12 quasar spectra. This figure accounts for the removal of 1) spectra with no Lyman-\( \alpha \) data in the range \( 1.94 < z < 3 \), 2) broad absorption line quasar spectra, and 3) a small number of excessively noisy spectra. The locations of the remaining quasars are indicated by the top-left and bottom panels of Figure 2.2. The data therefore cover two compact regions of the sky: one in the Northern Galactic cap (larger region) and

---

\(^1\)This allows us to infer the quasar is located at redshift \( z = 2.538 \) (with a standard error of 0.00013).

\(^2\)Defined to be the ratio of observed flux to the flux of the unabsorbed continuum.
Figure 2.1: Spectrum collected by BOSS of a quasar located at RA = 12.02527, Dec = −1.05598, z = 2.5338 ± 0.00013. The top panel displays the full spectrum and the bottom displays only the Lyman-α absorptions, with a smooth local polynomial fit to approximate the shape of the quasar continuum.

one in the Southern Galactic cap (smaller region). Altogether, this represents a ~ 25% sky coverage. The sampling design of the data as a function of redshift is highly heterogeneous, as the collective data set is densest at z ~ 2.1 and sharply declines at both lower and higher redshifts. Additionally, although to a much lesser extent, there is sampling heterogeneity on the two-dimensional sky. In particular, the high density of spectra in Stripe 82 can be seen in the Southern Galactic cap at Dec ≈ 0°. The most significant heterogeneity is of course a byproduct of the LOS sampling (top-right panel of Figure 2.2), with data points along an individual LOS having an average separation of ~ 0.6 Mpc/h and the average separation between LOSs being anywhere between ~ 12.5 Mpc/h and ~ 32.5 Mpc/h, depending on the redshift. This irregularity poses the most significant modeling challenge, as most models that are locally adaptive enough to preserve the disproportionate wealth of information at certain redshifts and regions of the sky will also overadapt to the LOS sampling in a way that effectively reduces the resulting map to a set of high-order one-dimensional maps.
Figure 2.2: **Left**: Sample of lines of sight collected in BOSS DR12 (Alam et al. [2015]), displayed in equatorial coordinates. **Right**: Locations of observed BOSS DR12 quasars in RA and $z$.

3 Multi-resolution Gaussian Random Field Model

In Eubanks et al. [In preparationa] we collaborated with Doug Nychka, a senior scientist at the National Center for Atmospheric Research, to adapt and apply the multi-resolution Gaussian Random Field (GRF) model introduced in Nychka et al. [2014] to BOSS DR12. The desirability of the GRF model in this context stems from its ability to model highly irregularly sampled and heteroskedastic spatial data, all the while smoothing adaptively with respect to the sampling density and measurement variance to maximize local signal recovery. That is, the smoothness of the map predicted by the GRF model inherently adapts to the number and quality\(^3\) of locally available data points. Holding measurement errors fixed, regions where data are sparse naturally result in relatively low resolution predictions while dense regions are mapped in higher resolution. Likewise, holding sampling density fixed, the resolution of the map is relatively low in regions with large measurement errors and high in regions with small measurement errors. Moreover, predictions overadapting to closely-spaced observations along LOSs is avoided by utilizing a truncated basis expansion, as outlined in Section 3.2.1. Because of the degree to which the number and quality of locally available data points in BOSS DR12 vary with redshift, our map is best viewed as a continuous set of two-dimensional sky maps in the range $z = 1.94$ to $z = 3$, with the resolution of each cross section varying with the local sampling density and measurement errors characteristic of the redshift. In particular, due to a relatively high density of data at $z \sim 2.1$, the structure of the IGM is recovered in the greatest detail at this redshift.

\(^3\)In terms of measurement variance
3.1 Transformation to Euclidean coordinates

Although the locations of the data are sampled in equatorial space, the most suitable spatial geometry currently available for the GRF model is three-dimensional Euclidean space, thus requiring a transformation. Specifically, for each \( \delta \)-fluctuation observed at location \((RA_i, Dec_i, z_i)\), the Euclidean coordinates \( x_i = (x^1_i, x^2_i, x^3_i) \) are calculated according to the following equations:

\[
\begin{align*}
  x^1_i &= D_C \cdot \sin(\pi/2 - Dec_i) \cdot \cos(RA_i) \\
  x^2_i &= D_C \cdot \sin(\pi/2 - Dec_i) \cdot \sin(RA_i) \\
  x^3_i &= D_C \cdot \cos(\pi/2 - Dec_i)
\end{align*}
\]  

(3.1)  
(3.2)  
(3.3)

where

\[
D_C = D_H \int_0^{z_i} \frac{dz'}{\sqrt{\Omega_M(1 + z')^3 + \Omega_k(1 + z')^2 + \Omega_\Lambda}}
\]

(3.4)

with \( D_H = 3000 \text{ Mpc}/h, \ \Omega_M = 0.3, \ \Omega_k = 0, \) and \( \Omega_\Lambda = 0.7 \).

3.2 Overview

Let \( y_1, \ldots, y_n \) designate the observed \( \delta \)-fluctuations of the IGM given by (2.2) and made at spatial locations \( x_1, \ldots, x_n \in \mathbb{R}^3 \), where each \( x_i \) is a point in Euclidean space given by equations (3.1)-(3.4). We assume the observations are generated according to the model

\[
y_i = g_0(x_i) + \epsilon_i
\]

(3.5)

where \( g_0 \) is a realization of a smooth Gaussian random field (GRF) \( g \) and \( \epsilon_1, \ldots, \epsilon_n \) are mean zero measurement errors. Since \( \delta \)-flux intrinsically has mean zero it is advantageous to also impose this property on \( g \). Furthermore, it is convenient to regard \( g \) as a sum of \( L \) independent GRFs \( g_1, \ldots, g_L \) with marginal variances \( \rho_\alpha_1, \ldots, \rho_\alpha_L \),

\[
g(x) = \sum_{l=1}^L g_l(x),
\]

(3.6)

where \( \rho > 0 \) is useful as a leading scaling parameter for the covariance matrix and \( \alpha_1, \ldots, \alpha_L > 0 \) sum to one. This representation allows \( g \) to exhibit much more complex spatial dependence than each of its individual components. Each component \( g_l \) is then defined through a basis expansion

\[
g_l(x) = \sum_{j=1}^{m(l)} \beta^l_j \phi^l_{j,l}(x),
\]

(3.7)

where \( \phi^l_{j,l}, j = 1, \ldots, m(l), \) is a sequence of compactly supported radial basis functions organized on regular three-dimensional lattices in \( \mathbb{R}^3 \) of increasing resolution and \( \beta^l \) is a vector of coefficients such that

\[
\beta^l \sim MN(0, \rho \Sigma_l), \quad l = 1, \ldots, L.
\]

(3.8)
Thus, the model for \( g \) is constructed as a sum of fixed basis functions coupled with stochastic coefficients. These two key elements of the model are further detailed in the sections below.

### 3.2.1 Radial basis functions

Each level of resolution in the GRF model is provided by a sequence of compactly supported radial basis functions (RBFs) organized on a regular lattice in three-dimensional Euclidean space. The node spacing of the lattices is given by

\[
\delta_l = 2^{-(l-1)}\delta_1, \quad l = 1, \ldots, L, \tag{3.9}
\]

where \( \delta_1 \) and \( L \) are regarded as tuning parameters of the model. At each level \( l = 1, \ldots, L \), the RBFs are given by a real-valued radial function

\[
\phi_{j,l}(x) = \phi \left( \frac{\|x - u_j\|}{\theta_l} \right), \quad j = 1, \ldots, m(l), \tag{3.10}
\]

where \( u_1, \ldots, u_{m(l)} \in \mathbb{R}^3 \) are the nodes of the lattice and \( \theta_l \) is a scaling parameter to adjust the amount of overlap in the RBFs at each level. In particular, we choose \( \phi \) to be the three-dimensional Wendland covariance function [Wendland, 1995] given by

\[
\phi(x) = \begin{cases} 
(1 - x)^6(35x^2 + 18x + 3)/3 & 0 \leq x \leq 1, \\
0 & \text{otherwise}.
\end{cases} \tag{3.11}
\]

Following Nychka et al. [2014], we fix \( \theta_l \) to be 2.5 times the node spacing at level \( l \) so the RBFs have a sufficient amount of overlap to avoid artifacts in the covariance function from the organization of the RBFs on a lattice. Moreover, we add five extra layers of nodes beyond the edges of each lattice to mitigate edge effects.

### 3.2.2 Gaussian Markov random fields

To simplify notation we combine (3.6) and (3.7) so

\[
g(x) = \sum_{j=1}^{m} \beta_j \phi_j(x), \tag{3.12}
\]

where we have combined the multi-resolution coefficients into a single vector \( \beta \) and the multi-resolution bases into a single basis \( \{\phi_i\}_{i=1}^{m} \). From (3.8) we have

\[
\beta \sim MN(0, \rho \Sigma), \tag{3.13}
\]

for some matrix \( \Sigma \).

The foundation of the computational efficiency of this methodology is the enforcement of sparsity in matrix computations in a way that does not sacrifice covariance models with many degrees of
freedom and multi-scale correlations. In addition to the use of basis functions with compact support, we accomplish this by directly computing the precision matrix $\frac{1}{\rho} \Sigma^{-1}$ of the basis coefficients instead of the covariance matrix $\rho \Sigma$ and restricting $\Sigma^{-1}$ to be sparse. This approach allows us to utilize efficient sparse matrix algorithms and still permits $\Sigma$ to be dense. Specifically, we first assume that coefficients between levels are independent, which gives the precision matrix a convenient block diagonal structure

$$\frac{1}{\rho} \Sigma^{-1} = \frac{1}{\rho} \begin{bmatrix} \frac{1}{\alpha_1} \Sigma_1^{-1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\alpha_2} \Sigma_2^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \frac{1}{\alpha_L} \Sigma_L^{-1} \end{bmatrix}. \tag{3.14}$$

We then model the distribution of the basis coefficients at each level of resolution as a Gaussian Markov random field (GMRF) organized on the nodes of each lattice. The Markov property of each GMRF can be described by an undirected graph. In particular, given the set of lattice nodes $V_l$ at level $l$ we define an edge set $E_l$ such that

$$(\Sigma_l^{-1})_{i,j} = 0 \quad \text{if} \quad \{i,j\} \notin E_l. \tag{3.15}$$

We assume the special case that each $\beta_i$ follows a first-order spatial autoregression with respect to the nodes of the lattice at level $l$. Specifically, this means that for every $i = 1, \ldots, m(l)$ the off-diagonal nonzero elements of the $i$th row of $\Sigma_l^{-1}$ correspond to the first-order neighbors of node $i$. Given an autoregression matrix $B_l$ for level $l$, we construct the distribution of $\beta_i$ according to $\beta_i = B_l^{-1}e$, where $e \sim N(0, \rho I)$. Following Lindgren et al. [2011], we let

$$B_{i,j} = \begin{cases} 6 + \kappa^2 & i = j, \\ -1 & \{i,j\} \in E_l \text{ and } i \neq j, \\ 0 & \text{otherwise}, \end{cases} \tag{3.16}$$

where $\kappa \geq 0$. It follows that $\beta_i \sim N(0, \rho B_l^{-1}B_l^{-T})$. Moreover, the covariance matrix of the full coefficient basis is given by

$$\rho \Sigma = \rho \begin{bmatrix} \alpha_1 B_1^{-1}B_1^{-T} & 0 & \cdots & 0 \\ 0 & \alpha_2 B_2^{-1}B_2^{-T} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \alpha_L B_L^{-1}B_L^{-T} \end{bmatrix}. \tag{3.17}$$

### 3.3 Parameter estimation

Based on the introduced model setup, $g$ is a mean zero Gaussian random field with covariance function

$$\text{Cov}(g(x), g(x')) = \sum_{j=1}^{m} \sum_{k=1}^{m} \rho \Sigma_{j,k} \phi_j(x) \phi_k(x'). \tag{3.18}$$
We assume $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ are independent with
$$\epsilon \sim MN(0, \sigma^2 W)$$
and $\text{Cov}(\beta, \epsilon) = 0$. Here, $W$ is diagonal with elements proportional to the measurement variances of the observations and $\sigma^2$ is a free parameter of the measurement error distribution. Now letting $\Phi$ be the regression matrix with $\Phi_{i,j} = \phi_j(x_i)$ we can rewrite (4.2) in matrix vector notation
$$y = \Phi \beta + \epsilon. \quad (3.20)$$
Thus, the assumed observational model is
$$y \sim MN(0, \rho \Pi), \quad (3.21)$$
where $\Pi = (\Phi \Sigma \Phi^T + \lambda W)$ and $\lambda = \sigma^2$. Here, $\lambda$ is the “noise to signal ratio” and is used as a reparametrization of $\sigma^2$ to tune the smoothness of the maps. From (3.21) we have the likelihood
$$L(\rho, \Sigma, \lambda \mid y) = \frac{1}{(2\pi)^{n/2} |\rho \Pi|^{1/2}} e^{-\frac{1}{2} y^T (\rho \Pi)^{-1} y} \quad (3.22)$$
and log likelihood
$$\ell(\rho, \Sigma, \lambda \mid y) = -\frac{1}{2} y^T (\rho \Pi)^{-1} y - \frac{1}{2} \log |\rho \Pi| - \frac{n}{2} \log(2\pi). \quad (3.23)$$
In principle, we can use (3.23) to compute maximum likelihood estimates (MLEs) for $\rho$, $\lambda$, and all parameters associated with $\Sigma$. Nevertheless, such an approach is not computationally feasible nor necessary given that several parameters do not significantly affect the predicted map or can be determined a priori. In particular, we can simplify computations by specifying the covariance $\Sigma$ and focusing on MLEs for $\lambda$ and $\rho$. Details are provided in sections 3.3.1, 3.3.2, and 3.3.3.

### 3.3.1 Lattice parameters
From Figure 3.1 we see that the minimum mean LOS separation over all redshifts is approximately 12.5 Mpc/h for both the Northern and Southern Galactic caps and occurs at $z \sim 2.1$. For this reason, we fix the finest level of basis functions to have a node spacing of 12.5 Mpc/h. This level therefore models the smallest scale structures of the IGM that are detectable in three dimensions given the average LOS density of BOSS DR12. Moreover we add three coarser levels with node spacings of 25, 50, and 100 Mpc/h. The coarsest level models the largest scale structures of the IGM. Although the basis functions themselves have a regular spacing of 100 Mpc/h at this level, the structures that they approximate can be much larger since the basis coefficients exhibit longer range unconditional correlations.

Given the significantly higher density of LOSs in Stripe 82, the mapping of this region could benefit from a finer level of basis functions than that which we use here. Nevertheless, these basis functions would contribute added variance to the rest of the map. Therefore, it is preferable
to exclude them here and possibly follow up with a separate map of Stripe 82, with parameters independently optimized to this subset of the data.

3.3.2 $\kappa$ and $\{\alpha_l\}_{l=1}^4$

The precision matrix $\Sigma^{-1}$ is determined by the scale parameter $\kappa$ and a smoothness parameter $\nu$, which specifies the relative variances of the levels according to $\alpha_l \sim e^{l \nu}$. The predicted map is relatively insensitive to the choice of $\kappa$ and $\nu$ so we follow the practice of fixing these parameters and optimizing $\lambda$ and $\rho$.

3.3.3 $\lambda$ and $\rho$

Let $\hat{\Sigma}$ denote the covariance matrix determined by the parameter choices in sections 3.3.1 and 3.3.2. Substituting $\hat{\Sigma}$ into (3.23), we first maximize (3.23) over $\rho$ analytically, yielding

$$\hat{\rho} = y^T \Pi^{-1} y / n.$$  

(3.24)

This estimate is then substituted back into (3.23) to give a profile log likelihood,

$$\ell(\hat{\rho}, \hat{\Sigma}, \lambda \mid y) = -\frac{1}{2} y^T (\hat{\rho} \Pi)^{-1} y - \frac{1}{2} \log |\hat{\rho} \Pi| - \frac{n}{2} \log(2\pi),$$  

(3.25)

which depends only on $\lambda$. The maximum likelihood estimate for $\lambda$ is then computed via optimization, specifically a combination of golden section search and successive parabolic interpolation [Brent, 1973].
3.3.4 Divide and conquer optimization of $\lambda$

The evaluation of (3.24) and (3.25) can be done much more efficiently by using sparse matrix decompositions and matrix identities. Nevertheless, the matrices associated with a problem of this scale are simply too large to fit in memory. Therefore, we implement a divide and conquer (DAC) strategy for efficiently estimating $\hat{\lambda}$. The DAC algorithm is similar to that which we propose in Section 4.2, modified to account for the fact that our objective function (3.25) is optimized iteratively instead of being calculated in full. Since the point estimate (3.35) is independent of the choice of $\rho$, we postpone discussion of estimating $\rho$ to Eubanks et al. [In preparationb], in which we will release pointwise confidence bands for the map.

The idea of the DAC strategy is to compute the MLE of $\lambda$ on a large number of compact topological subsets of the DR12 coverage, estimate a density $\hat{p}(\lambda)$ from the sample of “sub-MLEs”, and then choose a global estimate for $\lambda$ from $\hat{p}(\lambda)$. Specifically, we consider subsets with a sky coverage of $(1 \text{ deg})^2$ and spanning the redshift range $1.94 < z < 3$. The selected subsets are shown in Figure 3.2. We uniformly sampled 400 subsets in total—300 from the Northern Galactic cap and 100 from the Southern Galactic cap—from a partition of the footprint. A sufficient enough disparity between the distributions of sub-MLEs in each cap exists to warrant the use of a separate parameter estimate for each—$\hat{\lambda}_N$ and $\hat{\lambda}_S$.

We describe the procedure for computing the estimate $\hat{\lambda}_N$ for the Northern Galactic cap. The computation of $\hat{\lambda}_S$ is done analogously. Let $S_j \subset \mathbb{R}^3$ denote the jth subset sampled from the Northern Galactic cap,

$$\mathcal{D}_j = \{(x_i, y_i) : x_i \in S_j\},$$

and $n_j = |\mathcal{D}_j|$. Furthermore, let $\hat{\lambda}_j$ be the sub-MLE obtained by optimizing the profile log likelihood (3.25) of the subsample $\mathcal{D}_j$. By definition, $\lambda > 0$ so we consider the truncated and scaled kernel

\[\hat{\lambda}_j = \frac{1}{n_j} \sum_{(x_i, y_i) \in \mathcal{D}_j} \frac{1}{1 + e^{-\lambda (x_i, y_i)}} \]

Figure 3.2: Subsets of coverage selected for maximum likelihood estimation of parameters $\lambda$. Each subset extends through the redshift range $1.94 < z < 3$. 
density estimate
\[ \hat{p}(\lambda) = \begin{cases} \left( \int_0^\infty \tilde{p}(\lambda)d\lambda \right)^{-1} \cdot \tilde{p}(\lambda) & \lambda \geq 0 \\ 0 & \lambda < 0 \end{cases} \] (3.27)

where
\[ \tilde{p}(\lambda) = \frac{1}{\sum_{i=1}^{300} n_i \sum_{j=1}^{300} n_{ij} h K \left( \frac{\| \lambda - \hat{\lambda}_j \|^2}{h} \right)} \] (3.28)

We select \( h \) by 10-fold cross validation and define the global parameter estimate to be the mode
\[ \hat{\lambda}_N = \arg\max_{\lambda} \hat{p}(\lambda). \] (3.29)

Since the profile log likelihood of each subsample is a smooth function of \( \lambda \), \( \hat{\lambda}_N \) will be nearly optimal for any subset \( S_j \) with MLE \( \hat{\lambda}_j \) lying in a neighborhood of \( \hat{\lambda}_N \). The left panel of Figure 3.3 shows \( \hat{p}(\lambda) \) for each of the two sets of \( \delta \)-flux estimates and the right panel is the analogous plot for the Southern Galactic cap. The DAC approach for computing \( \hat{\lambda}_N \) is summarized in the algorithm table below.

**Algorithm 1** Divide and conquer estimation of \( \lambda \) for NG cap

Require: \( \mathcal{D}_1, \ldots, \mathcal{D}_{300}, \hat{\Sigma} \)

1: for all \( j \) do
2: Compute \( \hat{\lambda}_j \) via optimization of the profile log likelihood for \( \mathcal{D}_j \)
3: end for
4: Let
\[ \hat{p}(\lambda) = \begin{cases} \left( \int_0^\infty \tilde{p}(\lambda)d\lambda \right)^{-1} \cdot \tilde{p}(\lambda) & \lambda \geq 0 \\ 0 & \lambda < 0 \end{cases} \] (3.30)

where
\[ \tilde{p}(\lambda) = \frac{1}{\sum_{i=1}^{300} n_i \sum_{j=1}^{300} n_{ij} h K \left( \frac{\| \lambda - \hat{\lambda}_j \|^2}{h} \right)} \] (3.31)

and \( h \) is chosen by 10-fold cross validation
5: Define \( \hat{\lambda}_N = \arg\max_{\lambda} \hat{p}(\lambda) \)

---

4 Corresponding to the two methods we consider for approximating the shape of continua: local polynomial regression and utilizing the stacked mean spectrum approach of Harris et al. [2016].
Figure 3.3: **Left:** Kernel density estimates fit to sample of 300 sub-MLEs in the Northern Galactic cap for each of the two data sets corresponding to the two continuum fitting methods. We take the global parameter estimates to be the modes $\hat{\lambda}_{N,1} \approx 0.815$ and $\hat{\lambda}_{N,2} \approx 0.747$. **Right:** Kernel density estimates fit to sample of 100 sub-MLEs in the Southern Galactic cap for each of the two data sets corresponding to the two continuum fitting methods. We take the global parameter estimates to be the modes $\hat{\lambda}_{S,1} \approx 1.105$ and $\hat{\lambda}_{S,2} \approx 2.737$. 
3.4 Point estimate

For prediction of $g_0$, we consider the conditional distribution of the basis coefficients given the data and basis covariance. Fixing the covariance parameters at their true values, it follows that

$$
(\beta, y) \mid \{\lambda, \rho, \Sigma\} \sim MN\left(0, \rho \begin{bmatrix}
\Sigma & \Sigma \Phi^T \\
\Phi \Sigma & \Pi
\end{bmatrix}\right).
$$

By basic normal theory, the conditional distribution of $\beta$ given $y$ and all covariance parameters is

$$
\beta \mid \{y, \lambda, \rho, \Sigma\} \sim MN(\beta_0, \rho \Sigma - \rho \Sigma \Phi^T \Pi^{-1} \Phi \Sigma),
$$

where

$$
\beta_0 = \Sigma \Phi^T \Pi^{-1} y.
$$

The true covariance parameters are not known in practice so we replace them with the estimates in Section 3.3 and take our point estimate to be

$$
\hat{g}_0(x) = \Phi_x^T \hat{\beta},
$$

where

$$
\Phi_x = (\phi_1(x), \ldots, \phi_m(x)),
$$

$$
\hat{\beta} = \Sigma \Phi^T \Pi^{-1} y,
$$

$$
\hat{\Pi} = (\Phi \Sigma \Phi^T + \hat{\lambda} W).
$$

3.4.1 Assembling an aggregate map

As with the evaluation of the likelihood, the size of the matrices associated with prediction, if performed all at once, is too large to fit in memory. As noted in Eubanks et al. [In preparation a], this issue can be avoided by partitioning the volume into subsets, computing the predicted “submaps” individually, and combining them into an aggregate map. The only necessary provision is that sufficiently large margins are included when fitting each submap. Specifically, we partition the volume into $(400 \text{ Mpc/h})^3$ cubes and consider all observations within $(200 \sqrt{3} + 125) \text{ Mpc/h}$ of the center of each cube when constructing the corresponding submap. Since smoothing is done locally this stipulation ensures that the submaps fit together seamlessly. Altogether, the aggregate map is a montage of 932 submaps of the Northern Galactic cap and 422 submaps of the Southern Galactic cap.

4 Proposed Research

Ultimately motivated by the further improvement of the three-dimensional map of the intergalactic medium (IGM) we presented in Eubanks et al. [In preparation a], our proposed research pertains to the development of:
1. a novel three-dimensional spherical wavelet (SW) model capable of producing a more efficient mapping from BOSS DR12

2. a generalization of the previously employed heuristic DAC algorithm for computationally efficient risk and tuning parameter estimation, accompanied by a solid theoretical foundation of large sample properties

3. functional data analysis procedures for testing whether a fixed tuning parameter is suitable in a supervised nonparametric analysis, or should be rejected in favor of a spatially-adaptive tuning parameter.

These three components are outlined in the sections below.

4.1 The Three-dimensional Spherical Wavelet Model

Recall that, due to the constraints of the GRF model, a transformation of the feature space to Euclidean space preluded the modeling done in Section 3. An implication of this transformation is that the regular lattices were constructed in $\mathbb{R}^3$ and thus did not exploit the design of the data lying on lines of sight (LOSs) originating from a common vertex. In particular, the finest lattice was chosen to allow for an optimal reconstruction of the IGM at $z \sim 2.1$, where LOS sampling is most dense. However, because the lattices are organized in $\mathbb{R}^3$ this specification results in added variance at other redshifts, which would be mapped more efficiently with coarser lattices. This imperfection is summarized by Figure 4.1, where the left panel shows a three-dimensional Euclidean lattice (as was utilized by the GRF model) and the right panel shows the corresponding basis function separation versus redshift (drawn in red), compared to the average LOS separation of BOSS DR12. Ideally we would like these two curves to be close, insuring that we are mapping the IGM with the greatest level of three-dimensional resolution allowed by BOSS DR12 while also not using too many basis functions at redshifts where fewer would suffice.

As an early effort to develop an optimal organization of basis functions, we considered an organization on regular grids in $\mathbb{S}^2 \times \mathbb{R}^+$, as shown in Figure 4.2, where the basis functions are organized on the sphere according to a HEALPix hierarchal tessellation. This configuration naturally results in an increased basis function separation at higher redshifts. However, as shown in the right panel of Figure 4.1 (in green), the node spacing still does not approximate the average LOS separation of BOSS DR12 very well, due to LOSs appearing and then gradually disappearing as redshift increases. This organization would, however, be ideally suited for modeling data sets where each LOS extends through the full radial range.

Our proposed basis function organization can be viewed as a continuous version of a HEALPix hierarchical tessellation of the sphere. As shown in Figure 4.3, a HEALPix hierarchy partitions the surface of the sphere into subsets of equal area (each centered at a node), beginning with a partition of size 12, and successively further partitioning each subset into four subsets of equal area. Consequently, the number of nodes increases by a factor of 4 at each level of resolution. These hierarchical increases in resolution are too abrupt to alone construct a lattice organization in $\mathbb{S}^2 \times \mathbb{R}^+$ suitable for modeling BOSS DR12, so intermediate spherical lattice layers are filled in using a $k$-centers clustering of the sphere. That is, letting $k$ be the desired number of nodes on the
sphere, the organization is determined by minimizing

$$\max_{(RA, \text{Dec}) \in S^2} \min_{(c_1, c_2) \in C} (RA - c_1)^2 + (\text{Dec} - c_2)^2$$

(4.1)

with respect to the node set $C \subset S^2$ where $|C| = k$. Given such a sequence of lattices, we can construct multi-resolution levels using irregular lattices in $S^2 \times \mathbb{R}^+$ with wavelets fixed at the nodes in such a way so the average node separation of the finest lattice approximates the average LOS separation of BOSS DR12 through the full redshift range $1.94 < z < 3$.

Figure 4.1: Left: Regular three-dimensional Euclidean lattice. Right: Mean LOS separation of BOSS DR12 versus redshift, compared with mean node separation of regular three-dimensional Euclidean lattice and regular sequence of spherical lattices.

Given this new multi-resolution wavelet basis $\psi_1, \ldots, \psi_m \in S^2 \times \mathbb{R}^+$, we propose our new model for mapping the IGM from BOSS DR12. Let $y_1, \ldots, y_n$ designate the observed $\delta$-fluctuations of the IGM given by (2.2) and made at spatio-temporal locations $(RA_1, Dec_1, z_1), \ldots, (RA_n, Dec_n, z_n) \in S^2 \times \mathbb{R}^+$ according to the model

$$y_i = f_0(RA_i, Dec_i, z_i) + \epsilon_i.$$  

(4.2)

The idea of the multi-resolution spherical wavelet (SW) model is then to estimate $f_0$ with a basis expansion

$$\hat{f}_0(RA, Dec, z) = \sum_{j=1}^m \psi_j(RA, Dec, z) \beta_j$$

(4.3)

for some $\beta \in \mathbb{R}^m$. In particular, we will carry this out with a penalized weighted regression

$$\hat{\beta} = \argmin_{\beta \in \mathbb{R}^m} \sum_{i=1}^n w_i(y_i - \Phi_i^T \beta)^2 + \lambda \| \beta \|_1$$

(4.4)
where $w_i = \frac{1}{\sigma_i}$ and $\Psi$ is the regression matrix of wavelet functions. The penalty in (4.4) results in superfluous basis functions being regularized—namely in relatively noisy and/or sparsely sampled regions of the two-dimensional sky. Moreover, the penalty parameter $\lambda$ can be efficiently estimated using the SDAC algorithm proposed in the following section.

4.2 Spatial Divide and Conquer Risk Estimation

In a general data analysis setting, suppose we have a sample $(x_1, Y_1), \ldots, (x_n, Y_n)$ generated according to the model

$$Y_i = f_0(x_i) + \epsilon_i$$

(4.5)

where $x_1, \ldots, x_n \in S \subset \mathbb{R}^d$ and $\mathbb{E}[\epsilon_i] = 0$. Let $\hat{f}_\theta$ be a nonparametric estimator of $f_0$ depending on a vector of tuning parameters $\theta$ and suppose $\hat{f}_\theta$ satisfies the following local smoothing condition: for every $x \in S$, there is a nonempty set of observations satisfying

$$\hat{f}_\theta(x) \perp Y_i$$

(4.6)

for all $i$ such that $\|x - x_i\|_2 > \epsilon$, for some $\epsilon > 0$. Define

$$\delta_\theta(x) = \inf \{\epsilon : \hat{f}_\theta(x) \perp Y_i \text{ for all } \|x - x_i\|_2 > \epsilon\}$$

(4.7)

and

$$B_\theta(x) = \{i : \|x - X_i\|_2 \leq \delta_\theta(x)\}.$$
We are particularly interested in the case $|B_{\theta_0}(x)| \ll n$ for all $x \in S$, where $\theta_0$ is the optimal choice of $\theta$. Let $Q$ be a partition of $S$ into compact topological subspaces $S_1, \ldots, S_m$ and let $\hat{f}_\theta^j$ be the nonparametric estimator trained on the data

$$D_\theta^j = \{(x_i, Y_i) : i \in \bigcup_{x_k \in S_j} B_\theta(x_k)\}.$$  \quad (4.9)

Then, for each $j = 1, \ldots, m$, we have

$$\hat{f}_\theta(x_i) = \hat{f}_\theta^j(x_i)$$ \quad (4.10)

for all $x_1, \ldots, x_{n_j} \in S_j$.

Now let $\hat{R}(\hat{f}) \in \mathbb{R}^r$ be some estimated risk function of $\hat{f}$ evaluated over a grid of tuning parameters $(\theta_1, \ldots, \theta_r)$. For example, data splitting, leave-one-out cross validation, generalized cross validation, $K$-fold cross validation, Mallows’ $C_p$ (Mallows [1973]), and Stein’s unbiased risk estimate (Stein [1981]) are all valid. From (4.10) some simple algebra shows

$$\hat{R}(\hat{f}) = \frac{1}{n} \sum_{j=1}^m n_j \cdot \hat{R}_j(\hat{f}^j)$$ \quad (4.11)

where $\hat{R}_j(\hat{f}^j) \in \mathbb{R}^r$ is the estimated risk of $\hat{f}^j$ computed over the grid of tuning parameters $(\theta_1, \ldots, \theta_r)$ on the subsample $D_\theta^j \cap S_j$. That is, the global risk function $\hat{R}(\hat{f}) \in \mathbb{R}^r$ can be exactly computed as a linear combination of the local risk functions $\hat{R}_1(\hat{f}^1), \ldots, \hat{R}_m(\hat{f}^m) \in \mathbb{R}^r$. This distributed risk estimator in itself can be very useful for computational speedups via parallelization or to reduce the task to subproblems requiring less memory. Nevertheless, when $n$ is very large, (4.11) can also take an unacceptably long time to compute. We therefore propose the following two spatial divide and conquer (SDAC) risk estimators.

**Definition 1.** Let $Q$ be a partition of $S$ into compact topological subsets $S_1, \ldots, S_m$ where $|S_j| = n_j$ and let $Q' = \{S'_1, \ldots, S'_\ell, \ell < m\} \subset Q$. The weighted average (WA) SDAC risk estimator of $\hat{f}$ over
the grid \( \{\theta_1, \ldots, \theta_r\} \) is given by
\[
\hat{R}^{WA}(\hat{f}) = \frac{1}{n'} \sum_{j=1}^{\ell} n_j \cdot \hat{R}_j(\hat{f}^j)
\] (4.12)
where
\[
n' = \sum_{j=1}^{\ell} n_j
\] (4.13)
and \( \hat{R}_j(\hat{f}^j) \) is the estimated risk of \( \hat{f}^j \) computed over the grid of tuning parameters \( \{\theta_1, \ldots, \theta_r\} \) on the subsample \( D_\theta \cap S_j' \).

The chosen subset \( Q' \) of the partition \( Q \) may be sampled randomly or chosen by the data analyst. In either case, the motivation is for \( Q' \) to be a good representative of the full space \( S \) in some sense. The WA estimator is of course highly sensitive to extreme local risk functions so in some situations it may be preferable to use the following spatially robust risk estimator.

**Definition 2.** Let \( Q \) be a partition of \( S \) into compact topological subsets \( S_1, \ldots, S_m \) where \( |S_j| = n_j \) and let \( \{S_1', \ldots, S_{\ell}', \ell < m\} \subset Q \). The weighted geometric median (GM) SDAC risk estimator of \( \hat{f} \) over the grid \( \{\theta_1, \ldots, \theta_r\} \) is given by
\[
\hat{R}^{GM}(\hat{f}) = \arg\min_{z \in \mathbb{R}^r} \sum_{j=1}^{\ell} n_j \| \hat{R}_j(\hat{f}^j) - z \|_2
\] (4.14)
where \( \hat{R}_j(\hat{f}^j) \) is the estimated risk of \( \hat{f}^j \) computed over the grid of tuning parameters \( \{\theta_1, \ldots, \theta_r\} \) on the subsample \( D_\theta \cap S_j' \).

Naturally, (4.12) and (4.14) can each be used to select a global tuning parameter estimate. Namely,
\[
\hat{\theta}^{WA} = \arg\min_{\theta \in \{\theta_1, \ldots, \theta_r\}} \hat{R}^{WA}(\hat{f})
\] (4.15)
and
\[
\hat{\theta}^{GM} = \arg\min_{\theta \in \{\theta_1, \ldots, \theta_r\}} \hat{R}^{GM}(\hat{f}).
\] (4.16)

The goal of our proposed research is to establish a theoretical foundation for the proposed estimators (4.12), (4.14), (4.15), and (4.16) in various nonparametric regression settings. Specifically, we aim to establish convergence rates and asymptotic distributions for each of the estimators. Doing so will also allow us to derive approximate confidence intervals for the less computationally efficient estimators they are approximating.

### 4.3 Hypothesis Testing for Spatially Adaptive Tuning Parameters

The purpose of the SDAC algorithm is to efficiently compute a global tuning parameter estimate in a very general nonparametric regression setting. However, it is not always reasonable to assume a
data set can be appropriately modeled by a single fixed tuning parameter. The goal of this section of the thesis will therefore be to exploit the spatial nature of the SDAC algorithm to simultaneously provide hypothesis tests for the validity of the fixed tuning parameter assumption. That is, given a sample \((x_1, Y_1), \ldots, (x_n, Y_n) \sim F\) and a nonparametric regression estimator \(\hat{f}\), we would like to establish a test statistic

\[ T(\hat{R}_1(\hat{f}^1), \ldots, \hat{R}_m(\hat{f}^m)) \sim P \]

with some distribution \(P\), from the local risk functions computed by the SDAC algorithm, for testing the hypothesis

\[ H_0 : \mathbb{E} \int_S (f_0 - \hat{f}_\theta)^2 dF \leq \frac{1}{m} \sum_{j=1}^{m} \mathbb{E} \int_{S_j} (f_0 - \hat{f}_{\theta_j})^2 dF \]

against the alternative

\[ H_1 : \mathbb{E} \int_S (f_0 - \hat{f}_\theta)^2 dF > \frac{1}{m} \sum_{j=1}^{m} \mathbb{E} \int_{S_j} (f_0 - \hat{f}_{\theta_j})^2 dF. \]

Here, the alternative hypothesis (4.19) states that each topological subset \(S_j\) of the partition \(Q = \{S_1, \ldots, S_m\}\) should have an independent tuning parameter estimate. Note in (4.17) that \(\hat{R}_1(\hat{f}^1), \ldots, \hat{R}_m(\hat{f}^m)\) are functional data. This investigation will therefore lie within the realm of functional data analysis.
References


D. Krige. A statistical approach to some mine valuations and allied problems at the witwatersrand. Master’s thesis of the University of Witwatersrand, 1951.


