Carnegie Mellon University

MELLON COLLEGE OF SCIENCE

THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF Doctor of Philosophy

TITLE Models and Methods for One-Dimensional Approximations to Point Cloud Data

PRESENTED BY Slav Kirov

ACCEPTED BY THE DEPARTMENT OF Mathematical Sciences

Dejan Slepcev	August 2017
MAJOR PROFESSOR	DATE
Thomas Bohman	August 2017
DEPARTMENT HEAD	DATE

APPROVED BY THE COLLEGE COUNCIL

Rebecca Doerge	August 2017
DEAN	DATE

Models and Methods for One-Dimensional Approximations to Point Cloud Data

BY

SLAV KIROV

DISSERTATION

Submitted in Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in

MATHEMATICAL SCIENCES

at

CARNEGIE MELLON UNIVERSITY

Department of Mathematical Sciences

Advised by: Prof. Dejan Slepčev

Approved by:

Prof. Dejan Slepčev (Chair)

Prof. Hayden Schaeffer

Prof. Shlomo Ta'asan

Prof. Ryan Tibshirani

Pittsburgh, PA, August 2017

Abstract

In this thesis we investigate the problem of approximating point cloud data, or more generally, measures, by one-dimensional objects. Our approach is variational, as we will study certain functionals and the extent to which their minimizers (of finite length) can provide adequate approximations to data. In the first part of this thesis, the approximating objects we consider are curves, and our main goals are to understand their behavior and provide a robust and efficient algorithm for computing them. Aside from data analysis applications in which we assume data to have a one-dimensional structure, we are also motivated by settings in which the data approximation problem has a physical meaning. Such is the case in urban planning, and in particular the problem of finding optimal networks for transportation. In the second part of the thesis we will propose a new set of functionals that model this problem and establish basic existence properties. We then develop an algorithm for computing local minimizers, and investigate the suitability of the approach through a set of numerical examples that also give a glimpse into how the complexity of low-energy configurations increases with the total mass of the data.

Acknowledgments

First and foremost, I would like to thank my advisor Dejan Slepčev for his guidance, patience, and support in working with me over the years. Dejan has been a remarkable mentor, and the rewarding experience of working with him is one that I will always be grateful for. I am thankful to my committee members Hayden Schaeffer, Shlomo Ta'asan, and Ryan Tibshirani for contributing thoughtful discussions and valuable feedback to this thesis. I am furthermore thankful to the Center for Nonlinear Anlysis for its support, and to NSF grants CIF 1421502 and DMS 1516677.

There are many people who helped inspire my curiosity and admiration for mathematics, and who helped shape my maturity and development also as a person. Spanning my undergraduate and graduate years at CMU I give many thanks to professors Tom Bohman, Irene Fonseca, David Kinderlehrer, Dmitry Kramkov, Giovanni Leoni, John Mackey, Dana Mihai, Bob Pego, Dejan Slepčev, Shlomo Ta'asan, and Noel Walkington. I also thank Deborah Brandon, Bill Hrusa, and Stella Andreoletti for their assistance throughout my time in the PhD program. My experience during the program would not have been the same if it were not for the incredible people I met and friends I made; in particular, Matteo Rinaldi, Daniel Rodríguez, Nicolás García Trillos, Ryan Murray, and everyone who played on the Energy Minimizers soccer team (and any of its past incarnations). Outside of CMU, I thank Alex Vladimirsky for providing one of my first experiences in applied math research during a summer REU at Cornell, and people I met there including Elisabeth Cai and Zach Clawson.

Finally I would like to thank my parents Milka and Assen, and my brother Stefan for their unwavering and unconditional support throughout my studies. I am extremely fortunate to have them throughout my life, along with the rest of my family in Bulgaria – my grandmother Didi, my aunt Nina, my countless cousins, and my grandparents who are no longer here but whose memories remain clear – Elka, Stefan, and Slavcho.

Contents

Chapter 1

Introduction

The focus of this thesis is on approximating a given measure by a one-dimensional object. We will call the measure μ and think of it as a distribution of data in $\mathbb{R}^d$. The motivation for studying this problem arrives from at least two different channels.

One is from the perspective of data analysis and machine learning, where data is often given as high-dimensional due to of a large number of raw features or variables observed. In such cases, it is common that the high-dimensionality of the data combined with noise obscures a simpler intrinsic structure that can be efficiently described by a lower-dimensional object. To put it simply, if we have data in $\mathbb{R}^d$ (for $d > 1$) that we have reason to believe is supported close to a one-dimensional subset, our goal is to find a one-dimensional set that best represents the data.

From another perspective, even if the distribution of data is not supported near a one-dimensional subset, we may still want to approximate it with such. One example of this is in urban planning for the purpose of designing networks for transportation or irrigation systems. In this setting, the measure μ is regarded as a population distribution, and the goal is to find a network (a one-dimensional subset of $\mathbb{R}^d$) that both provides a good coverage of the population and is feasible to build and maintain (i.e., it is not too costly).

Our approach to this problem is variational – we investigate functionals defined over one-dimensional subsets of $\mathbb{R}^d$, and whose values aim to measure how well a given set approximates the data μ, with lower values being preferred. Occasionally we may refer to a given one-dimensional set as a configuration, and its value in the functional as its energy. We will therefore seek configurations that minimize the considered energy.

In this thesis we explore a number of functionals for approximating measures with one-dimensional objects. Where we propose new functionals, we will prove that minimizers exist, and investigate some their basic properties. A major component of this thesis is to provide efficient numerical algorithms for approximating minimizers of the functionals. It will become evident that the functionals have a complicated energy landscape, with many local minima, which we will pay particular attention when designing algorithms.

This thesis consists of two main parts. The first part deals with approximating the measure μ by a single or multiple curves. The relevance and applicability of this approach

lie mostly in the data analysis and machine learning domains where one seeks to recover the one-dimensional structure of μ, if it has such. The second part of the thesis deals with general one-dimensional sets and functionals that are particularly relevant to designing networks for the transportation needs of a given population represented by μ.

We begin by introducing the average-distance problem, which ties together the data analysis and urban planning perspectives, and served as one starting point for this work. Throughout, μ will be a finite and positive compactly supported Borel measure on $\mathbb{R}^d$ for $d \geq 2$ unless otherwise noted.

1.1 Average-distance problem

The average-distance problem was introduced by Buttazzo, Oudet, and Stepanov in [12], and was provided in the following constrained form.

Problem 1.1.1 (ADP, constrained). Given $\ell > 0$,

$$\text{minimize} \quad \int_{\mathbb{R}^d} d(x, \Sigma) d\mu(x)$$

over

$$\Sigma \in \mathcal{A}_\ell := \{\Sigma \subseteq \mathbb{R}^d : \Sigma \text{ compact and connected}, \mathcal{H}^1(\Sigma) \leq \ell\}.$$

In the above $d(x, \Sigma)$ represents the Euclidean distance to the set Σ, while $\mathcal{H}^1(\Sigma)$ represents the one-dimensional Hausdorff measure, which measures the length of the set Σ [1]. The problem has been studied having in mind applications to transportation network design. In this context, a network Σ is sought which minimizes total distances to the passengers (distributed according to μ), with the constraint representing a budget to build and/or maintain the network. A related problem, which was the focus of Buttazzo and Stepanov in [14], involves finding a network along which one measure can be optimally transported to a second given measure. There one may think a distribution of resources (e.g. workers) that are to be transported where they are needed (e.g. workplaces), and traveling along the network is free. It was shown that this latter problem can be reduced to the average-distance problem 1.1.1 with an appropriate choice of μ [61].

Existence of minimizers of the average-distance problem follows from the Blaschke and Gołąb theorems (2.2.2, 2.2.4), and in recent years much progress has been made in understanding the properties of the minimizers. An overview article summarizing many of the findings has been written by Lamenant [43]. We highlight some of them below.

Perhaps some of the most important results are regarding the topology of the minimizers. For dimension $d = 2$ it was shown by Buttazzo and Stepanov in [14] that any minimizer is topologically equivalent to a tree, i.e., it is a finite union of Lipschitz curves containing no loops. Furthermore, Buttazzo and Stepanov showed that curves will only meet at triple junctions, that is, no more than three curves will meet at a point.

[1] see Chapter 2 for definitions for these and other notation to follow

As is common in optimization and calculus and variations, a problem in a penalized form can be eaiser to study than its constrained form. The penalized form of the average distance problem is as follows.

Problem 1.1.2 (ADP, penalized). Given $\lambda > 0$,

$$\text{minimize} \quad E_\mu^\lambda(\Sigma) := \int_{\mathbb{R}^d} d(x,\Sigma) d\mu(x) + \lambda \mathcal{H}^1(\Sigma)$$

over

$$\Sigma \in \mathcal{A} := \{\Sigma \subseteq \mathbb{R}^d : \Sigma \text{ compact and connected}\}.$$

The applications and interpretations of the penalized problem are very similar. The difference is that instead of having a budget for building the network, it is incorporated with the cost of traveling to the network, and one seeks to minimize this aggregate cost. From a data analysis perspective, the first term can be seen as a fidelity term that measures how well the network approximates the data, while the $\mathcal{H}^1$ term as a regularization that penalizes the complexity of the approximation.

A topic of significant interest has been the regularity of minimizers. In [59], Slepčev showed that minimizers can have corners, even if the measure μ has a smooth density with convex support. The construction involved approximating a counterexample measure μ by a sequence discrete measures μ_n, and used the fact that the sequence of corresponding minimizers converge (up to a subsequence) in Hausdorff distance to the minimizer corresponding to μ, which is implied by $\Gamma-$convergence of the functional with respect to convergence of measures in the weak-$*$ topology.

Later in [45], Lu and Slepčev proved that minimizers satisfy the following total curvature bound

$$\sum_{i=1}^k |\gamma_i'|_{TV} \leq \frac{1}{\lambda} \mu(\mathbb{R}^d)$$

where $\{\gamma_i\}_{i=1}^k$ are Lipschitz curves whose union gives the minimizing network. The total variation (TV) above allows to treat the curvature as a measure, with delta masses at locations of corners, which is necessary in light of the possible lack of regularity. On the way to showing the above bound, the authors proved a desirable topological lower semi-continuity result. More precisely it states that if $\mu_n \overset{\star}{\rightharpoonup} \mu$ and $\Sigma_n \in \operatorname{argmin} E_{\mu_n}^\lambda$, then as along a subsequence $\Sigma_n \overset{H}{\to} \Sigma \in \operatorname{argmin} E_\mu^\lambda$, it holds that for sufficiently large n, Σ is homeomorphic to a subset of Σ_n.

For applications, it is often the case that the measure μ is discrete, consisting of finitely many masses, in which case more can be said about minimizers of the average-distance functional 1.1.2. For the moment, let us consider the empirical measure

$$\mu = \sum_{i=1}^n m_i \delta_{x_i}.$$

Suppose that Σ is a minimizing network, and let v_i denote the projection of x_i onto Σ. Then Σ is a Steiner graph for the points $\{v_i\}_{i=1}^n$, i.e., Σ is the geometric graph of minimum

length that contains $\{v_i\}_{i=1}^n$ (see Figure 1.1 for an illustration). Unlike a minimum spanning tree, the vertices of the Steiner tree do not need to be a subset of $\{v_i\}_{i=1}^n$, which makes the problem significantly more difficult. Indeed, the Steiner tree problem is NP-hard in general. Together with the fact that the points $\{v_i\}_{i=1}^n$ are not known a priori makes the problem very challenging from an optimization point of view.

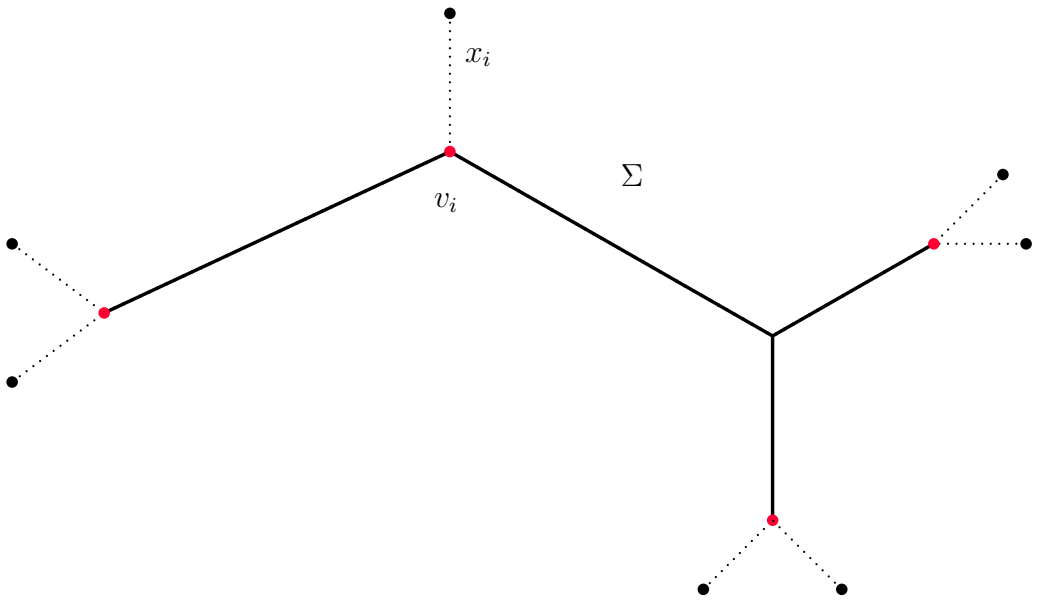


Figure 1.1.1: Given data points x_i the minimizing network Σ for the average-distance problem will be a Steiner tree for the points v_i (in red) – the projections of x_i onto Σ.

On a different note, minimizers inherit an interesting property of Steiner trees, which is that all angles are at least 120 degrees, and at vertices of degree 3 all angles are exactly 120 degrees. These facts, proven by Gilbert and Pollack in [33], hold for any dimension d, and thus imply that the edges at triple junctions lie in a common plane.

1.2 Dimensionality reduction

Let us temporarily shift our perspective to what from machine learning and data mining is known as the dimensionality reduction problem. Given a distribution of data μ in $\mathbb{R}^d$, the dimensionality reduction problem is to find a lower-dimensional representation of μ. There have been many approaches to this problem from the machine learning and statistics community. One of the most well-known is principal component analysis (PCA), which finds a lower dimensional subspace spanned by lines (referred to as *principal components*), along which the variance of the data is maximized. Mathematically, the principal components can be obtained via a singular value decomposition of the mean-zero centered data

matrix. As this reduction is linear, a number of non-linear approaches have been proposed. Among them are methods including multidimensional scaling [41], locally linear embedding [55], Isomap [62], Laplacian eigenmaps [5], diffusion maps [18], which all find lower-dimensional representations of the data by aiming to preserve local distance information. All of these methods apply to one-dimensional, as well as higher-dimensional, reductions of the data.

Our focus here will be on the problem of reducing the data to one dimension, where more work has been done. An interesting recent approach is based on estimating the density function of the data and finding its ridges [16, 24, 30], which we further discuss in Chapter 3. In the 80's, Hastie and Stuetzle [35] studied the problem in a more classical setting where the approximating object is a curve. The curves they sought are called *principal curves*, and can seen as a nonlinear generalization of the first principal component. Principal curves have the mean projection property, which is that every point on the curve is the mean of data that project there. As we later note, the original principal curves are unstable and prone to overfitting, and many modifications followed [21, 23, 32, 39, 60, 63]. A number of works treat the problem by adding a regularization term to the objective functional. Among them are works of Tibshirani [63] (square curvature penalization), Kegl, Krzyzak, Linder, and Zeger [39] (length constraint), Biau and Fischer [6] (length constraint), and Smola, Mika, Schölkopf, and Williamson [60] (a variety of penalizations including penalizing length). The following formulation with a length penalty regularization plays a signifcant role in much of this thesis.

1.3 Penalized principal curves

Problem 1.3.1 (PPC). Given $\lambda > 0$ and $p \geq 1$, the *penalized principal curves* are minimizers of

$$E_\mu^\lambda(\gamma) := \int_{\mathbb{R}^d} d(x, \Gamma)^p d\mu(x) + \lambda \, \mathrm{L}(\gamma) \tag{PPC}$$

over

$$\gamma \in \mathcal{C} := \{\gamma : [0, a] \to \mathbb{R}^d \,\big|\, a \geq 0, \gamma \text{ is Lipschitz with } |\gamma'| \leq 1, \ \mathcal{L}^1 - \text{ a.e.}\},$$

and where $\mathrm{L}(\gamma) := |\gamma|_{TV}$ is the length of γ, and $\Gamma := \gamma([0, a])$ is its image.

The penalized principal curve problem is an average-distance problem for parametrized curves, with note that the general exponent p can also be included in the ADP. We remark that in machine learning and statistics, $p = 2$ is most often used. That is the case for the original principal curves, which were initially sought as critical points of the above energy without the length term [35]. It easy to see that without the length term or other regularization Problem 1.3.1 is ill-posed, as one can approximate μ arbitrary well with a long enough curve. The length term can therefore be seen to penalize the complexity of the approximation, whose quality is represented by the fidelity or error term $\int_{\mathbb{R}^d} d(x, \Gamma)^p$.

Penalized principal curves were studied in [44] by Lu and Slepčev. There, the authors showed their existence and investigated their regularity. In particular, they proved that minimizing curves are injective (i.e. do not self-intersect) in dimension $d = 2$ if $p \geq 2$ or if μ has bounded density. They also showed that any minimizer $\gamma_{\min}$ has the following total curvature bound

$$|\gamma'_{\min}|_{TV} \leq \frac{p}{\lambda} \text{diam}(\text{supp}(\mu))^{p-1} \mu(\mathbb{R}^d).$$

We will investigate penalized principal curves in more detail in Chapter 3. Our objective will be to understand the relationship between the data and the minimizers, and illustrate how the length scales present in the data and the parameters of the functional dictate the length scales seen in the minimizers. In particular, we will provide the critical length scale below which variations in the input data are treated as noise and establish the typical error (bias) when the input curve is smooth. We emphasize that the former has direct implications for when penalized principal curves begin to overfit, and can furthermore assist with parameter selection when one has knowledge of level of noise present in the data.

We will also propose a fast numerical algorithm for computing (approximate) penalized principal curves. As we further discuss in Chapter 3, several of the approaches for computing regularized principal curves suffer from poor local minima arising from the non-convexity of the considered functionals. Our strategy will be to enlarge the space over which the (PPC) functional is considered, allowing for multiple curves.

1.4 Multiple penalized principal curves

We introduce an extension of (PPC) which allows for configurations to consist of more than one curve. Since (PPC) can be made arbitrarily small by considering γ with many components, a penalty on the number of curves is needed. We will refer to minimizers for the problem that follows as *multiple penalized principal curves*.

Problem 1.4.1 (MPPC). Given $\lambda_1 > 0, \lambda_2 > 0$, minimize

$$E_\mu^{\lambda_1,\lambda_2}(\gamma) := \int_{\mathbb{R}^d} d(x,\Gamma)^p d\mu(x) + \lambda_1 \left(\mathrm{L}(\gamma) + \lambda_2 \left(\mathrm{k}(\gamma) - 1\right)\right) \tag{MPPC}$$

over

$$\gamma \in \mathcal{A} := \left\{\gamma = \{\gamma^i\}_{i=1}^k : k \in \mathbb{N},\, \gamma^i \in \mathcal{C},\;\; i = 1, ..., k\right\},$$

and for $\gamma \in \mathcal{A}$ we define $\mathrm{k}(\gamma) := |\gamma|$, the cardinality of the set γ.

We may view this functional as penalizing both zero- and one-dimensional complexities of approximations to μ. On the one hand, we can recover the (PPC) functional by taking λ_2 large enough. On the other hand, taking λ_2 small enough leads to a k-means clustering problem which penalizes the number of clusters, and has been encountered in [11, 42].

The main motivation for introducing (MPPC), even if only one curve is sought, has to do with the non-convexity of the (PPC). We will see that numerically minimizing (MPPC) often helps evade undesirable (high-energy) local minima of the (PPC) functional. As

(MPPC) is a relaxation of (PPC) to a larger configuration space, the energy descent for (MPPC) allow for curve splitting and reconnecting. It will be this mechanism that enables us to evade the local minima of (PPC).

In Chapter 3 we will prove the existence of multiple penalized principal curves, and will investigate the connectedness. We do the latter by identifying a critical *linear density*, which is the density of the projected data onto the curve with respect to its length. The critical linear density together with the scale over which it is recognized give us insight as to when minimizers can recover one-dimensional components of the data.

We will also provide a detailed algorithm for computing approximate minimizers. The algorithm has a desirable computational complexity, that is both linear in the number of data points, and their dimension d. We will demonstrate the algorithm on a number of datasets, and compare results with popular approaches (subspace-constrained mean shift algorithm, and diffusion maps).

1.5 Optimal networks for selective-transport

We turn our attention to what may seem like a substantially different problem. Let us revisit the setting where μ represents a distribution of agents, and suppose that every agent has a distribution of locations they desire or need to visit. In this setting, instead of only being given the measure μ, we are given a joint distribution π on $\mathbb{R}^d \times \mathbb{R}^d$ that describes the coupling between transport origins and destinations. The first marginal of π is the distribution of origins μ, and the second marginal gives the aggregated distribution of destinations. For example, if $\pi = \mu \otimes \mu$, then every agent wishes to visit every other agent. The goal is then to find a network that minimizes a total cost consisting of the cumulative transportation cost together with the cost of building the network. If we again represent the latter cost by $\mathcal{H}^1(\Sigma)$ and we let $c_\Sigma(x, y)$ denote cost of reaching y from x with the available network Σ, then we have the following problem.

Problem 1.5.1 (Optimal network for selective-transport). Given $\lambda > 0$ and a Borel measure π on $\mathbb{R}^d \times \mathbb{R}^d$, minimize

$$ONT_\pi(\Sigma) := \int_{\mathbb{R}^d \times \mathbb{R}^d} c_\Sigma(x, y) d\pi(x, y) + \lambda \mathcal{H}^1(\Sigma) \tag{1.5.1}$$

over $\Sigma \in \mathcal{A}$.

Let us now discuss the choice of the cost function c_Σ. Given the motivation for the problem, one may assume that every agent pays the distance they travel outside the network, while the distance they travel inside the network is discounted by some factor $\alpha \in [0, 1)$. In other words, if we let

$$\mathcal{G}_{x,y} := \{\gamma \text{ closed and connected: } \{x, y\} \subseteq \gamma \subseteq \mathbb{R}^d\}$$

denote the set of paths containing x and y, then we may define the cost

$$c_\Sigma(x, y) = \inf\{\mathcal{H}^1(\gamma \cap \Sigma^C) + \alpha \mathcal{H}^1(\gamma \cap \Sigma) : \gamma \in \mathcal{G}_{x,y}\}.$$

This cost function will be of particular interest for transportation-inspired applications, and will be the emphasis in Chapter 4.

We note a connection to the average-distance problem, which can be recovered with a different choice of the cost function. Indeed, let

$$c_\Sigma(x,y) = d(x,\Sigma)^p + d(y,\Sigma)^p + \alpha \tilde{d}_\Sigma(x,y)$$

where $\tilde{d}_\Sigma(x,y)$ denotes the intrinsic distance between the projections of x and y onto Σ. That is, let

$$\tilde{d}_\Sigma(x,y) := \inf_{\tilde{x}\in\Pi_\Sigma(x),\tilde{y}\in\Pi_\Sigma(y)} d_\Sigma(\tilde{x},\tilde{y})$$

where

$$d_\Sigma(\tilde{x},\tilde{y}) := \inf_{\tilde{x},\tilde{y}\in\Sigma'\subseteq\Sigma,\ \Sigma' \text{connected}} \mathcal{H}^1(\Sigma')$$

and $\Pi_\Sigma(x) := \arg\min_{\tilde{x}\in\Sigma} |\tilde{x} - x|$ denotes the projection set. Then (1.5.1) reduces to

$$2\int_{\mathbb{R}^d} d(x,\Sigma)^p d\mu(x) + \alpha \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \tilde{d}_\Sigma(x,y) d\mu(x) d\mu(y) + \lambda \mathcal{H}^1(\Sigma).$$

Taking $\alpha = 0$ then gives the average-distance problem.

In Chapter 4, we will study this problem and some close variants. One of the variants we consider includes a scenario in which the distribution of destinations for each agent changes with the available network, with less costly destinations being preferred. We also propose a model for planning an optimal city, which includes a joint search over the transportation network Σ and the population distribution μ. The problems are closely related to optimal transportation, including the works of [12, 14], which we will discuss in more detail. We will prove existence of minimizers for the problems, and provide a numerical algorithm for approximating minimizers. Finally, we will show applications to both detecting one-dimensional structure of data, and computing optimal transportation networks for populations of varying sizes.

Chapter 2

Preliminaries

2.1 Notation

- $|\cdot|$: Euclidean norm
- $|\cdot|_{TV}$: total variation (defined in 3.2.1)
- $d(x, A)$: distance from point $x \in \mathbb{R}^d$ to set $A \subseteq \mathbb{R}^d$, i.e. $d(x, A) := \inf\{|x - y| : y \in A\}$
- $\xrightarrow{H}$: Hausdorff convergence of sets (Def. 2.2.1)
- $\mathcal{H}^1$: one-dimensional Hausdorff measure (Def. 2.2.3)
- $\mathcal{L}^d$: Lebesgue measure on $\mathbb{R}^d$
- $\mathcal{B}(X)$: the set of Borel subsets of X
- $\mathcal{P}(X)$: the set of Borel probability measures on X
- $\mathrm{supp}(\mu)$: the support of the measure μ (Def. 2.3.2)
- $\overset{*}{\rightharpoonup}$: weak-$\star$ convergence of measures (Def. 2.3.1)
- $f_{\#}\mu$: push-forward measure of μ by f (Def. 2.3.4)
- $\xrightarrow{\Gamma}$: $\Gamma-$convergence of functionals (Def. 2.4.1)

2.2 Hausdorff distance and measure, Blaschke and Gołąb theorems

We briefly state some results regarding compactness of the Hausdorff metric (Blaschke's theorem), and lower semi-continuity of the one-dimensional Hausdorff measure (Gołąb's

theorem), that will be important in establishing existence of minimizers. For convenience of the reader, we first recall the Hausdorff distance between sets.

Let (X, d) be a metric space. Let E be a subset of X, and let $\mathcal{C}_E$ denote the family of all nonempty closed subsets of E.

Definition 2.2.1 (Hausdorff distance). For $C, D \in \mathcal{C}_E$ the *Hausdorff distance* is defined as

$$d_H(C, D) := \min\{1, h(C, D)\}$$

where

$$h(C, D) := \inf\{r \in [0, +\infty] : C_r \subseteq D \text{ and } D \subseteq C_r\}$$

and for any set $A \subseteq X$, A_ϵ is the $\epsilon-$neighborhood of A, i.e.

$$A_\epsilon := \{x \in X : d(x, A) < \epsilon\}.$$

The Hausdorff distance is a metric, and we will write $\xrightarrow{H}$ to denote convergence in Hausdorff distance. That is, $C_n \xrightarrow{H} C$ if $\lim_{n\to\infty} d_H(C, C_n) = 0$. The following compactness theorem holds.

Theorem 2.2.2 (Blaschke)**.** *Let (X, d) be a metric space. If E is a compact subset of X, then $(\mathcal{C}_E, d_H)$ is a compact metric space.*

We now recall the Hausdorff measure.

Definition 2.2.3 (Hausdorff measure). Let A be any subset of X, and let $\delta > 0$. Define

$$\mathcal{H}^k_\delta(S) := \inf\left\{\sum_{i=1}^{\infty}(\operatorname{diam}(A_i))^k \,\middle|\, \operatorname{diam}(A_i) < \delta, A \subseteq \bigcup_{i=1}^{\infty} A_i\right\}$$

where $\operatorname{diam}(A_i) := \sup\{d(x, y)|x, y \in A_i\}$ denotes the diameter. We then define the *$k-$dimensional Hausdorff measure*

$$\mathcal{H}^k(A) := \sup_{\delta>0} \mathcal{H}^k_\delta(A).$$

We note that it is common for the Hausdorff measure to include a scaling factor so that $\mathcal{H}^d = \mathcal{L}^d$ (on $\mathbb{R}^d$) for $d > 1$. Throughout we will restrict our attention to $\mathcal{H}^1$, which as defined coincides with $\mathcal{L}^1$ (on $\mathbb{R}$). The use of $\mathcal{H}^1$ is to provide a measure of the length of sets that lie in $\mathbb{R}^d$, $d \geq 1$.

We have the following lower semicontinuity result for the one-dimensional Hausdorff measure.

Theorem 2.2.4 (Gołąb)**.** *Let (E, d) be a complete metric space. Suppose $\{C_n\}_{n\in\mathbb{N}} \subseteq \mathcal{C}_E$ is such that each C_n is connected, and $C_n \xrightarrow{H} C$ for some C. Then C is connected and*

$$\mathcal{H}^1(C) \leq \liminf_{n\to\infty} \mathcal{H}^1(C_n).$$

For more on Hausdorff distance and measures, and proofs of these theorems, see for example [3].

2.3 Measure theory

We start by recalling the definition of weak-$\star$ convergence (sometimes also known as narrow or weak convergence, or convergence in distribution) and Prokhorov's theorem. We then define the push-forward measure, and state the disintegration theorem, which will be used in computing the second variation of the (MPPC) functional. For further reference on these, one may for instance see [2]. Throughout, we let X be a metric space and let $\mathcal{P}(X)$ denote the set of Borel probability measures on X.

Definition 2.3.1 (Weak-$\star$ convergence of measures). A sequence of measures $\{\mu_n\} \subseteq \mathcal{P}(X)$ converges *weakly-$\star$* to $\mu \in \mathcal{P}(X)$ if

$$\lim_{n\to\infty} \int_X f(x) d\mu_n(x) = \int_X f(x) d\mu(x)$$

for every continuous and bounded function f on X. In this case we write $\mu_n \overset{\star}{\rightharpoonup} \mu$.

Definition 2.3.2 (Support). The *support* of a measure $\mu \in \mathcal{P}(X)$ is the set

$$\operatorname{supp}(\mu) := \{x \in X : \mu(U) > 0 \quad \text{for each neighborhood } U \text{ of } x\}.$$

Theorem 2.3.3 (Prokhorov)**.** *If a set of measures $\mathcal{M} \subseteq \mathcal{P}(X)$ is tight, i.e.,*

$$\forall \epsilon > 0 \quad \exists K_\epsilon \textit{ compact in } X \textit{ such that} \quad \mu(X \setminus K_\epsilon) \leq \epsilon \quad \forall \mu \in \mathcal{M}$$

then $\mathcal{M}$ is relatively compact in $\mathcal{P}(X)$ with respect to weak-$\star$ convergence.

An immediate consequence that we will later use is that if $\{\mu_n\}$ is a sequence of measures with $\operatorname{supp}(\mu_n) \subseteq K$ for all n with K compact, then along a subsequence $\mu_n \overset{\star}{\rightharpoonup} \mu$.

Definition 2.3.4 (Push-forward measure). Let X_1, X_2 be separable metric spaces, $\mu \in \mathcal{P}(X_1)$, and $f : X_1 \to X_2$ be a Borel function. Then the *push-forward of μ by f* is the measure on X_2 defined by

$$f_{\#}\mu(B) := \mu(f^{-1}(B)) \quad \forall B \in \mathcal{B}(X_2),$$

where $\mathcal{B}(X_2)$ denotes the set of Borel subsets of X_2.

Theorem 2.3.5 (Disintegration)**.** *Let $\mu \in \mathcal{P}(\mathbb{R}^d)$, $\pi : \mathbb{R}^d \to \mathbb{R}^d$ be a Borel function, and $\nu := \pi_{\#}\mu$. Then there exists a ν–a.e uniquely determined family of Borel probability measures $\{\mu_x\}_{x\in\mathbb{R}^d} \subseteq \mathcal{P}(\mathbb{R}^d)$ such that*

$$\mu_x(\mathbb{R}^d \setminus \pi^{-1}(x)) = 0 \quad \textit{for } \nu - \textit{a.e. } x \in \mathbb{R}^d$$

and

$$\int_{\mathbb{R}^d} f(x) d\mu(x) = \int_{\mathbb{R}^d} \left(\int_{\pi^{-1}(x)} f(y) d\mu_x(y) \right) d\nu(x)$$

for every Borel function $f : \mathbb{R}^d \to [0, +\infty]$.

We note that we have stated a simplified version of the disintegration theorem, which applies more generally when initial and target spaces are any separable Radon metric spaces.

Let us now state two more results that help us establish existence of minimizers for optimal network models in Chapter 4. The first is a generalized Fatou lemma, provided somewhat recently in [26] (Theorem 1.1).

Proposition 2.3.6 (Generalized Fatou)**.** *Let S be a metric space, $\{\pi_n\}$ be a sequence of finite Borel measures on S such that $\pi_n \overset{\star}{\rightharpoonup} \pi$, and let $\{f_n\}$ be a sequence of non-negative measurable functions on S. Then*

$$\int_S \liminf_{n\to\infty,\, s'\to s} f_n(s')d\pi(s) \leq \liminf_{n\to\infty} \int_S f_n(s)d\pi_n(s).$$

We also have the following lower semi-continuity result that can be found in [28] (Theorem 5.19).

Theorem 2.3.7. *Let E be a compact subset of $\mathbb{R}^n$ and let $f : \mathbb{R}^m \to (-\infty, \infty]$ be a convex, lower semi-continuous function. Then the functional*

$$v \mapsto \int_E f(v(x))dx$$

is sequentially lower semi-continuous with respect to weak-$\star$ convergence, for integrable functions $v : E \to \mathbb{R}^m$.

2.4 Γ-convergence

We now review a tool that allows us to formalize convergence of a sequence of functionals. Γ-convergence provides one such notion, and we recall its definition.

Definition 2.4.1 (Γ-convergence). Let X be a metric space, and $F_n : X \to [-\infty, \infty]$ be a sequence of functionals on X. We say that the sequence $\{F_n\}$ Γ-*converges* to F, and we write $F_n \overset{\Gamma}{\to} F$, if the following two properties hold:

1. (Liminf inequality) For every $x \in X$ and every sequence $\{x_n\} \subset X$ such that $x_n \to x$, it holds
$$F(x) \leq \liminf_{n\to+\infty} F_n(x_n).$$

2. (Limsup inequality) For every $x \in X$ there exists a sequence $\{x_n\} \subset X$ such that $x_n \to x$ and
$$F(x) \geq \limsup_{n\to+\infty} F_n(x_n).$$

Γ-convergence has the important property that minimizers of converging functionals will converge to a minimizer of the limiting functional.

Proposition 2.4.2. *Suppose $\{F_n\}$ is a sequence of functionals bounded from below with corresponding minimizers x_n, i.e. $F_n(x_n) = \min_{x\in X} F_n(x)$. If $x_n \to x$ and $F_n \xrightarrow{\Gamma} F$, then x is a minimizer of F, and*

$$\lim_{n\to\infty} F_n(x_n) = F(x).$$

$\Gamma-$ convergence has several other nice properties, such as stability under continuous perturbations, and we refer the reader to [8] for further information.

2.5 ADMM algorithm

Here we describe the alternating direction method of multipliers (ADMM), which we will later use as a sub-step in computing approximate minimizers of the considered functionals. The algorithm is a special case of the split Bregman algorithm of Goldstein and Osher [34] when the constraints are linear, and is closely related to a number of other splitting or alternating algorithms for convex problems [25]. For more information we refer the reader to a review article by Boyd et al. [7].

The ADMM algorithm solves problems of the form

$$\begin{aligned} &\text{minimize} && f(x) + g(z) \\ &\text{subject to} && Ax + Bz = c \end{aligned}$$

where f and g are convex functions on $\mathbb{R}^m$ and $\mathbb{R}^l$ respectively, and $A \in \mathbb{R}^{p\times m}$, $B \in \mathbb{R}^{p\times l}$. The augmented Lagrangian for the problem is

$$L_\rho(x, z, y) = f(x) + g(z) + y^T(Ax + Bz - c) + \frac{\rho}{2}\big|Ax + Bz - c\big|^2$$

where y represents a dual variable, and $\rho > 0$ is a parameter. ADMM then consists of the following iterations

$$\begin{aligned} x^{k+1} &= \arg\min_x L_\rho(x, z^k, y^k) \\ z^{k+1} &= \arg\min_z L_\rho(x^{k+1}, z, y^k) \\ y^{k+1} &= y^k + \rho(Ax^{k+1} + Bz^{k+1} - c) \end{aligned}$$

The updates can be seen a gradient ascent on the dual problem with the augmented Lagrangian. The only difference is that the updates for x and z are split into separate, alternating steps. By rescaling the dual variable via $u = \frac{1}{\rho}y$, one can rewrite the steps into the following simpler form

$$x^{k+1} = \arg\min_x \left(f(x) + \frac{\rho}{2}\big|Ax + Bz^k - c + u^k\big|^2\right) \tag{2.5.1}$$

$$z^{k+1} = \arg\min_z \left(g(z) + \frac{\rho}{2}\big|Ax^{k+1} + Bz - c + u^k\big|^2\right) \tag{2.5.2}$$

$$u^{k+1} = u^k + Ax^{k+1} + Bz^{k+1} - c. \tag{2.5.3}$$

One of the main appeals of ADMM is that it allows one to split the problem into two separate minimizations that decouple the roles of the functions f and g. This is particularly useful when one of the functions is not smooth, such as a term involving the ℓ_1-norm (which we will be our case). In such problems, ADMM leads to updates that can be computed very quickly and which are simple to implement.

The ADMM algorithm is known to converge for any value of $\rho > 0$ under mild assumptions. Two sufficient assumptions provided by Boyd et al. [7] are

1. the epigraphs of the function f and g are closed convex sets, where the epigraph of f is
$$\mathbf{epi} f = \{(x, t) \in \mathbb{R}^m \times \mathbb{R} : f(x) \leq t\},$$
and

2. the unaugmented Lagrangian L_0 has a saddle point.

These somewhat basic assumptions will hold for the problems we consider, and they guarantee convergence of the objective to an optimal value, of the residual to zero (which implies feasibility of iterates in the limit), and of the dual variable to an optimal value. In practice, the algorithm can sometimes be slow to converge to high accuracy, and linear convergence has been proven in special cases (e.g. [36, 51]). However, ADMM does usually convergence to a reasonable degree of accuracy in less than 50 iterations, which together with its fast updates and simplicity make it very suitable for our purposes.

Chapter 3

Multiple Penalized Principal Curves

3.1 Introduction

The goal of this chapter is to more closely investigate the penalized principle curves and multiple penalized principal curves that were briefly described in the introduction. The contents here are mostly contained in paper written by Slepčev and the author [40]. The contributions of this chapter break down into two main components. The first concerns obtaining a better understanding of the behavior of penalized principal curves, while the second involves introducing their extension through the (MPPC) functional to multiple curves, largely for the purpose of more robust computation.

As we touched upon, one of the shortcomings of the original principal curves is that they tend to overfit noisy data. Several variants, and in particular, regularizations, of principal curves have thus been proposed. However, a closer understanding of the behavior of the objects studied has been lacking. Adding a regularization term to the objective functional is a common way to address overfitting, but doing so also introduces bias: when the data lie on a smooth curve the minimizer will only approximate them, and will not exactly recover the curve. Our objective will be to understand the relationship between the data and the minimizers, and illustrate how the length scales present in the data and the parameters of the functional dictate the length scales seen in the minimizers. In particular, we will provide the critical length scale below which variations in the input data are treated as noise and establish the typical error (bias) when the input curve is smooth. We emphasize that the former has direct implications for when penalized principal curves begin to overfit, and can furthermore assist with parameter selection when one has knowledge of the level of noise present in the data.

Our second contribution will involve introducing and further investigating the (MPPC) functional, permitting its minimizers to consist of more than one curve (*multiple penalized principal curves*). The motivation is twofold. The data itself may have one-dimensional structure that consists of more than one component, and the relaxed setting would allow it to be appropriately represented. The less immediate appeal of the new functional is that it will guide the design of an improved scheme for computing penalized principal curves.

Namely, for many datasets the penalized principal curves functional has a complicated energy landscape with many local minima. This is a typical situation and an issue for virtually all present approaches to regularized principal curves. As we explain below, enlarging the set over which the functional is considered (from a single curve to multiple curves) and appropriately penalizing the number of components leads to significantly better behavior of energy descent methods (they more often converge to low-energy local minima).

We will find that topological changes of multiple penalized principal curves are governed by a critical *linear density*. The linear density of a curve is the density of the projected data on the curve with respect to its length. If the linear density of a single curve drops below the critical value over a large enough length scale, a lower-energy configuration consisting of two curves can be obtained by removing the corresponding curve segment. Such steps are the means by which configurations following energy descent stay in higher-density regions of the data, and avoid local minima that penalized principal curves are vulnerable to. Identification of the critical linear density and the length scale over which it is recognized by the functional further provide insight as to the conditions under and the resolution to which minimizers can recover one-dimensional components of the data.

Finally, we will provide a detailed algorithm for computing approximate minimizers. We will apply modern optimization algorithms based on alternating direction method of multipliers (ADMM) [7] and closely related Bregman iterations [34, 52] (see Section 2.5) for local curve fitting, and we outline routines for executing topological changes, curve re-parametrization, and initialization. The resulting algorithm has favorable computational complexity that is linear in both the number of data points and the dimension of the space they lie in. At the end, we will present numerical examples that both illustrate the theoretical findings and support the viability of the approach for point clouds with substantial noise and in high dimensions.

3.1.1 Related work.

The original principal curves are prone to overfitting, as carefully explained in [32], and are difficult to compute numerically. A number of works treat the problem by adding a regularization term to the objective functional, in a similar fashion to penalized principal curves. Among them are works of Tibshirani [63] (square curvature penalization), Kegl, Krzyzak, Linder, and Zeger [39] (length constraint), Biau and Fischer [6] (length constraint), and Smola, Mika, Schölkopf, and Williamson [60] (a variety of penalizations including penalizing length). The work of Biau and Fischer [6] also discusses model-selection based automated ways to choose parameters of the given functional for the specific data set. Wang and Lee [67] also use model selection to select parameters, but ensure the regularity of the minimizer in a different way. Namely they model the points along the curve as an autoregressive series.

Regarding methods for computation of regularized principal curves, Kegl, Krzyzak, Linder, and Zeger [39] proposed a polygonal-line algorithm that penalizes sharp angles.

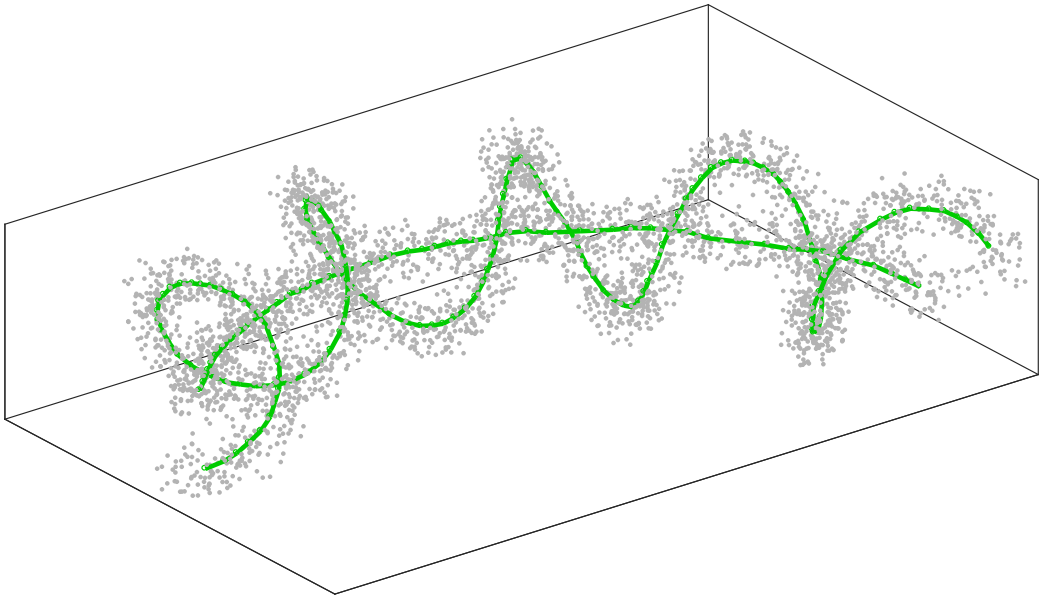


Figure 3.1.1: Example of a point cloud generated by noisy samples of two curves (not shown): a section of a circle and a curved helix wrapping around it. The green curves shown represent the one-dimensional approximation of the data cloud obtained by minimizing the proposed functional (MPPC) using the algorithm of Section 3.4.

Feuersänger and Griebel employ sparse grids to minimize a functional with length squared regularization [27] for manifolds up to dimension three. While these approaches take measures against overfitting data, they do not address the problem of local minima, resulting in performance that is very sensitive to the initialization of the algorithms. Verbeek, Vlassis and Kröse [65] approach this issue by iteratively inserting, fitting, and connecting line segments in the data. This approach is effective in some situations where others exhibit poor performance (e.g. spiral in 2-d, some self-intersecting curves, and curvy data with little noise). However, in cases of higher noise the algorithm overfits if the number of segments is not significantly limited. A better understanding of the impact of the number of segments on the final configuration is still needed, despite some efforts to automate selection of this parameter [67].

Gerber and Whitaker [32] offer an interesting approach to principal curves without regularization. Their approach recognizes that the difficulty in the principal curve problem lies in the unknown intrinsic ordering of the data, and they therefore minimize a suitable functional over coordinate mappings – functions that induce an ordering on the ambient space. The authors show that for any sufficiently smooth coordinate mapping, there is a unique corresponding differentiable curve that satisfies the self-consistency property. In [31], Gerber, Tasdizen, and Whitaker minimize a functional over coordinate mappings that penalizes the corresponding total squared projection distance. However, critical points of

the functional are only saddle points, as was later noted in [32], where Gerber and Whitaker instead minimize non-orthogonality of projections. While critical points of the functional in [32] are global minima, the functional values no longer indicate fit quality, and obtaining a desirable curve strongly depends on the coordinate mapping initialization. The authors note that initialization can be provided by spectral methods such as Isomap [62], Locally Linear Embedding [55], and Laplacian Eigenmaps [5]. In Example 3.4.2 with data on a noisy spiral, one can see in Figure 3.4.7 that the ordering obtained by Diffusion Maps [18] (a robust spectral method) can be incorrect.

A different class of approaches to finding one-dimensional structures is based on estimating the probability density function of the point cloud and then finding its ridges [24]. Estimation of density ridges has been substantially developed and studied – see works of Chen, Genovese, and Wasserman [16], Genovese, Perone-Pacifico, Verdinelli, and Wasserman [30], and Pulkkinen [54]. Of existing methods, these approaches seem to have the best performance in consistently locating one-dimensional structure. The Subspace Constrained Mean Shift (SCMS) algorithm of Ozertem and Erdogmus [53] is widely used for this approach, and is based on the Mean Shift algorithm of Comaniciu and Meer [19]. It is important to note that the SCMS algorithm does not parameterize the found one-dimensional structure, which consists of an (unordered) set of points. Another significant difference between our approach and SCMS is that we seek a one-dimensional structure with low approximation error, measured as part of the functional we consider, while SCMS does not require the found ridges to approximate the data well. See Example 3.4.2 and Figure 3.4.3 for an example where not all ridges are close to the data. We also note that in high dimensions SCMS faces a combination of computational difficulties, the primary of which is accurately estimating the Hessian of the density function (found using a kernel density estimator), as is discussed in Section 3 of [53].

Finally we mention the work of Arias-Castro, Donoho, and Huo [4] who studied optimal conditions and algorithms for detecting (sufficiently smooth) one-dimensional structures with uniform noise in the background.

3.1.2 Outline

We first restate the objective functionals both for single and multiple penalized principal curves, and recall some of the related approaches in Section 3.2. We then establish simple properties of the functionals, including the existence of minimizers and their basic regularity. Under assumption of smoothness we derive the Euler-Lagrange equation for critical points of the functional. We conclude Section 3.2 by computing the second variation of the functional. In Section 3.3 we provide a number of illustrative examples through which we investigate the relation between the length scales present in the data, the parameters of the functional, and the length scales present in the minimizers. At the end of Section 3.3 we discuss parameter selection for the functional when one has some estimates of quantitative properties of the data. In Section 3.4 we describe the algorithm for computing approximate minimizers of the (MPPC) functional. In Section 3.4.7 we provide some further numerical

examples that illustrate the applicability of the functionals and algorithm, including comparisons to the SCMS and Diffusion Maps algorithms. Section 3.5 contains the conclusion. Appendix 3.3.4 contains some technical details of an analysis of a minimizer considered in Section 3.3.

3.2 The functionals and basic properties

Let us we restate the functionals for single and multiple penalized principal curves. We will also recall and prove some of their basic properties. Let $\mathcal{M}$ be the set of finite, compactly supported measures on $\mathbb{R}^d$, for $d \geq 2$ and $\mu(\mathbb{R}^d) > 0$.

3.2.1 Penalized principal curves

Given a measure (distribution of data) $\mu \in \mathcal{M}$, $\lambda > 0$, and $p \geq 1$, the *penalized principal curves* are minimizers of

$$E_\mu^\lambda(\gamma) := \int_{\mathbb{R}^d} d(x,\Gamma)^p d\mu(x) + \lambda \, \mathrm{L}(\gamma) \tag{PPC}$$

over $\gamma \in \mathcal{C} := \{\gamma : [0,a] \to \mathbb{R}^d : a \geq 0, \gamma \text{ is Lipschitz with } |\gamma'| \leq 1, \ \mathcal{L}^1 - \text{ a.e.}\}$, and where $\Gamma := \gamma([0,a])$, $d(x,\Gamma)$ is the distance from x to set Γ and $\mathrm{L}(\gamma)$ is the length of γ:

$$\mathrm{L}(\gamma) := |\gamma|_{TV} := \sup\left\{ \sum_{i=2}^{n} |\gamma(x_i) - \gamma(x_{i-1})| \ : \ 0 \leq x_1 < x_2 < ... < x_n \leq a, \, n \in \mathbb{N} \right\}. \tag{3.2.1}$$

We recall that the functional is closely related to the average-distance problem introduced by Buttazzo, Oudet, and Stepanov [12] having in mind applications to optimal transportation networks [14]. The (PPC) functional can be viewed as a restriction of the penalized average distance problem to the set of curves, instead of general connected one-dimensional sets. One notable difference is also that in general the length term depends on the curve γ as a function, not through its image. However, in [44] Lu and Slepčev showed that minimizing curves are injective (i.e. do not self-intersect) in dimension $d = 2$ if $p \geq 2$ or if μ has bounded density, and therefore in such cases the problem has a geometric interpretation, allowing one to minimize the energy over embedded curves.

The first term in (PPC) is a fidelity term that measures how well the curve γ approximates the data, while the second term serves as a regularization that penalizes the complexity of the approximation. As mentioned earlier, the functional is similar to a number of others in the statistics and machine learning literature as regularizations of the principal curves problem [63], [39], [6], [60].

We recall that (PPC) has been studied in [44], where existence of minimizers was shown, along with the following total curvature bound

$$|\gamma'_{\min}|_{TV} \leq \frac{p}{\lambda} \mathrm{diam}(\mathrm{supp}(\mu))^{p-1} \mu(\mathbb{R}^d).$$

The total variation (TV) above treats the curvature as a measure, with delta masses at locations of corners, which may exist even if μ has smooth density supported in a convex set [59].

3.2.2 Multiple penalized principal curves

We now restate the extension of (PPC) which allows for configurations to consist of more than one component. Since (PPC) can be made arbitrarily small by considering γ with many components, we penalize the number of components through a second parameter λ_2 with the following functional

$$E_\mu^{\lambda_1,\lambda_2}(\gamma) := \int_{\mathbb{R}^d} d(x,\Gamma)^p d\mu(x) + \lambda_1 \left(\mathrm{L}(\gamma) + \lambda_2 \left(\mathrm{k}(\gamma) - 1\right)\right), \tag{MPPC}$$

where γ is now a set consisting of $\mathrm{k}(\gamma)$ curves. More precisely, we aim to minimize (MPPC) over the admissible set

$$\mathcal{A} := \left\{\gamma = \{\gamma^i\}_{i=1}^k : k \in \mathbb{N},\, \gamma^i \in \mathcal{C},\;\; i = 1, ..., k\right\},$$

and for $\gamma \in \mathcal{A}$ we define $\mathrm{k}(\gamma) := |\gamma|$, the cardinality of the set γ. With an abuse of notation we also denote the length of $\gamma \in \mathcal{A}$ by $\mathrm{L}(\gamma) = \sum_{i=1}^{\mathrm{k}(\gamma)} \mathrm{L}(\gamma^i)$.

We call elements of $\mathcal{A}$ *multiple curves*, and the minimizers to (MPPC) *multiple penalized principal curves*. The functional can be seen to penalize a combination of the zero- and one-dimensional complexities of approximations to μ. Namely, at one end when λ_2 is large enough, the minimizer will be a connected curve, and this problem recovers the (PPC) problem. At the other end, if λ_2 is small enough, the problem becomes a k-means clustering problem which penalizes the number of clusters. The connection to clustering will be made more evident when we further investigate the role of λ_2, and discuss a numerical algorithm for finding minimizers to (MPPC).

As we will see, our main motivation for considering (MPPC), is due to the difficulty of minimizing (PPC) over its considered space. The energy landscape of (PPC) has many local minima, and the ability of energy-descent methods to find acceptable configurations is often very sensitive to initialization. Enlarging the configuration space to allow for multiple curves provides energy-descent methods new directions in which to go, specifically, the disconnecting and reconnecting of curves, which is the mechanism that enables one to evade many local minima of (PPC).

3.2.3 Existence of minimizers of (MPPC)

We show that minimizers of (MPPC) exist in $\mathcal{A}$. We follow the approach of [44], where existence of minimizers was shown for (PPC). We first cover some preliminaries, including defining the distance between curves. If $\gamma_1, \gamma_2 \in \mathcal{C}$ with respective domains $[0, a_1], [0, a_2]$, where $a_1 \leq a_2$, we define the extension of γ_1 to $[0, a_2]$ as

$$\tilde{\gamma}_1(t) = \begin{cases} \gamma_1(t) & \text{if } t \in [0, a_1] \\ \gamma_1(a_1) & \text{if } t \in (a_1, a_2]. \end{cases}$$

We let

$$d_{\mathcal{C}}(\gamma_1, \gamma_2) = \max_{t \in [0, a_2]} |\tilde{\gamma}_1(t) - \gamma_2(t)|.$$

We have the following lemma, and the subsequent existence of minimizers.

Lemma 3.2.1. *Consider a measure $\mu \in \mathcal{M}$ and $\lambda_1, \lambda_2 > 0, p \geq 1$.*

(i) For any minimizing sequence $\{\gamma_n\}$ of (MPPC)

(a) $\limsup_{n\to\infty} \mathrm{k}(\gamma_n) \leq 1 + \frac{1}{\lambda_1\lambda_2}(\mathrm{diam}(\mathrm{supp}(\mu)))^p$*, and*

(b) $\limsup_{n\to\infty} \mathrm{L}(\gamma_n) \leq \frac{1}{\lambda_1}(\mathrm{diam}(\mathrm{supp}(\mu)))^p$

(ii) There exists a minimizing sequence $\{\gamma_n\}$ of (MPPC) such that $\forall n,\ \Gamma_n$ is contained in $\mathrm{Conv}(\mu)$, the convex hull of the support of μ.

Proof. The first property follows by taking a singleton as a competitor. The second follows from projecting any minimizing sequence onto $\mathrm{Conv}(\mu)$. Doing so can only decrease the energy, as shown in [14, 44]. The argument relies on the fact that projecting onto a convex set decreases length. □

Lemma 3.2.2. *Given a positive measure $\mu \in \mathcal{M}$ and $\lambda_1, \lambda_2 > 0, p \geq 1$, the functional (MPPC) has a minimizer in $\mathcal{A}$. Moreover, the image of any minimizer is contained in the convex hull of the support of μ.*

Proof. The proof is an extension of the one found in [44] for (PPC). Let $\{\gamma_n\}_{n\in\mathbb{N}}$ be a minimizing sequence in $\mathcal{A}$. Since the number of curves $\mathrm{k}(\gamma_n)$ is bounded, we can find a subsequence (which we take to be the whole sequence) with each member having the same number of curves k. We enumerate the curves in each member of the sequence as $\gamma_n = \{\gamma_n^i\}_{i=1}^k$. We assume that each curve γ_n^i is arc-length parametrized for all $n \in \mathbb{N}$, $i \leq k$. Since the lengths of the curves are uniformly bounded, let $L = \sup_{n,i} \mathrm{L}(\gamma_n^i)$, and extend the parametrization for each curve in the way defined above. Then for each $i \leq k$, the curves $\{\gamma_n^i\}_{n\in\mathbb{N}}$ satisfy the hypotheses of the Arzelà-Ascoli Theorem. Hence for each $i \leq k$, up to a subsequence γ_n^i converge uniformly to a curve $\gamma^i : [0, L] \to \mathbb{R}^d$. Diagonalizing, we find a subsequence (which we take to be the whole sequence) for which the aforementioned convergence holds for all $i \leq k$. Moreover, the limiting object is a collection of curves which are 1-Lipschitz since all of the curves in the sequence are. Thus $\gamma := \{\gamma^i\}_{i=1}^k \in \mathcal{A}$.

The mapping $\Gamma \mapsto \int_{\mathbb{R}^d} d(x, \Gamma)^p d\mu(x)$ is continuous and $\Gamma \mapsto \mathrm{L}(\Gamma)$ is lower-semicontinuous with respect to convergence in $\mathcal{C}$. Thus $\liminf_{n\to\infty} E_\mu^{\lambda_1,\lambda_2,p}(\gamma_n) \geq E_\mu^{\lambda_1,\lambda_2,p}(\gamma)$, and so γ is a minimizer. □

3.2.4 First variation

In this section, we compute the first interior variation of the (PPC) functional considering $\overline{\gamma}$ to be a smooth curve and μ to be a measure with density and supported near $\overline{\gamma}$. In the case of multiple curves, one can apply the following analysis to each curve separately.

Given a curve $\overline{\gamma} : [a, b] \to \mathbb{R}^d$, we consider variations of the form $\gamma(s, t) = \overline{\gamma}(s)+tv(s)$, where $v \in C^2([a, b], \mathbb{R}^d)$, $v(a) = v(b) = 0$. That is, we only perturb the interior of the curve, and not its endpoints. We note that one could allow for $v(a)$ and $v(b)$ to be nonzero (as has been considered for example in [59]), but it is not needed for our purposes. Letting γ_s denote the partial derivative in s, we can furthermore assume (by re-parameterizing the curves if necessary) that $|\gamma_s| = 1$ and that v is orthogonal to the curve: $v(s) \cdot \gamma_s(s) = 0 \ \forall s \in [a, b]$.

We make a few simplifying assumptions on γ and the underlying measure μ. Namely, we assume that the compactly supported measure μ is absolutely continuous with respect to the Lebesgue measure $\mathcal{L}^d$, and that γ is C^2. Although minimizers may have corners as mentioned earlier, we generally expect that minimizers to be C^2 except at finitely many points, and that our analysis therefore applies to intervals between such points. In addition, we assume that the projection of data onto γ is unique. That is, letting $\Gamma_t := \gamma([a, b], t)$, we assume that

$$\Pi_t(x) = \arg\min \{|x - y| : y \in \Gamma_t\}$$

is unique for $x \in \mathrm{supp}(\mu)$. In other words μ is supported within the reach of γ. We note that the assumption is inconsequential since the absolutely continuity of μ implies that the set of points where Π_t is non-unique has μ-measure zero [49].

In order to determine how the energy (PPC) is changing when $\overline{\gamma}$ is perturbed, we first compute how the distance of points to Γ_t is changing with t. For $x \in \mathrm{supp}(\mu)$ let $g(x, t) := d(x, \Gamma_t)^2 = |x - \Pi_t(x)|^2$. Then

$$\frac{\partial g}{\partial t} = -2(x - \Pi_t(x)) \cdot \gamma_t \tag{3.2.2}$$

and further computations give

$$\frac{\partial^2 g}{\partial t^2} = 2\left(|\gamma_t|^2 - (x - \Pi_t(x)) \cdot \gamma_{tt} - \frac{(\gamma_t \cdot \gamma_s - (x - \Pi_t(x)) \cdot \gamma_{st})^2}{|\gamma_s|^2 - (x - \Pi_t(x)) \cdot \gamma_{ss}}\right). \tag{3.2.3}$$

In the above, γ and its derivatives are evaluated at $(s^*(x, t), t)$, where $s^*(\cdot, t) := \gamma(\,\cdot\,, t)^{-1} \circ \Pi_t$ maps points in $\mathrm{supp}(\mu)$ to $[a, b]$. Equivalently, $s^*(x, t) = \arg\min_{s\in[a,b]} d(x, \gamma(s, t))$. We postpone use of (3.2.3) until the second variation in the next section.

In what follows we will, somewhat selectively, suppress dependence on s and t for readability. Taking the derivative in t of $\mathrm{L}(\gamma) = \int_a^b |\gamma_s|\, ds$, combining it with (3.2.2), and changing coordinates so that the approximation-error term is written as double integral, we obtain

$$\frac{dE}{dt} =$$

$$\int_a^b \left(\lambda_1 \frac{\gamma_s}{|\gamma_s|} \cdot \gamma_{st} - 2\,\alpha(s) \int_{\Pi_t^{-1}(\gamma)} (x - \Pi_t(x)) \cdot \gamma_t \, |1 - \vec{\mathcal{K}} \cdot (x - \Pi_t(x))| \, d\mu_s(x) \right) d\mathcal{L}^1(s),$$

where $|1+\vec{\mathcal{K}}\cdot(\Pi_t(x)-x)|$ is the Jacobian for change of coordinates, and $\vec{\mathcal{K}}$ is the curvature vector of γ. Here we have used the disintegration theorem (Theorem 2.3.5) to rewrite an integral for μ over $\mathbb{R}^n$ as an iterated integral along slices orthogonal to the curve (which contain the set of points that project to a given point on the curve). The probability measure supported on the slice $\Pi_t^{-1}(\gamma(s,t))$ is denoted by μ_s, while α is the linear density of the projection of μ to Γ_t, pulled back to the parameterization of $\gamma(\,\cdot\,,t)$. More precisely, $\alpha := d(s^*(\cdot,t)_{\#}\mu)/d\mathcal{L}^1$ is the Radon-Nikodym derivative of $s^*(\cdot,t)_{\#}\mu$ with respect to the Lebesgue measure on the line. The measure $s^*(\cdot,t)_{\#}\mu$ is the push-forward of μ by the mapping $s^*(\cdot,t)$, that is the measure defined by $s^*(\cdot,t)_{\#}\mu(A) = \mu((s^*(\cdot,t))^{-1}(A)) = \mu(\Pi_t^{-1}(\gamma(A,t)))$ for any Borel set $A \subset [a,b]$. We note that although $s^*(\cdot,t)_{\#}\mu$ may have atoms at the endpoints a, b, γ_t is zero there so it does not affect the above expression.

Integrating by parts we obtain

$$\left.\frac{dE}{dt}\right|_{t=0} =$$

$$\int_a^b \left(-\lambda_1 \vec{\mathcal{K}} \cdot \gamma_t - 2\,\alpha(s) \int_{\Pi^{-1}(\overline{\gamma})} (x - \Pi(x)) \cdot \gamma_t \, |1 - \vec{\mathcal{K}} \cdot (x - \Pi(x))| \, d\mu_s(x) \right) d\mathcal{L}^1(s).$$

We conclude that $\overline{\gamma}$ is a stationary configuration if and only if

$$\lambda_1 \vec{\mathcal{K}}(s) = -2\,\alpha(s) \int_{\Pi^{-1}(\overline{\gamma})} (x - \Pi(x)) \, |1 - \vec{\mathcal{K}}(s) \cdot (x - \Pi(x))| \, d\mu_s(x) \tag{3.2.4}$$

for $\mathcal{L}^1-$ a.e. $s \in (a,b)$.

3.2.5 Second variation.

In this section we compute the second variation of (PPC) for the purpose of providing conditions for linear stability. That is, we focus on the case that a straight line segment is a stationary configuration (critical point), and find when it is stable under the considered perturbations (when the second variation is greater than zero). This has important implications for determining when the penalized principal curves start to overfit the data, and is further investigated in the next section.

If $\overline{\gamma}$ is a straight line segment, $\vec{\mathcal{K}} = 0$, and (3.2.4) simplifies to

$$\overline{\gamma}(s) = \bar{x}(s) := \int_{\Pi^{-1}(\overline{\gamma}(s))} x \, d\mu_s(x)$$

for $\mathcal{L}^1-$ a.e. $s \in (a,b)$ such that $\alpha(s) \neq 0$. This simply states that a straight line is a critical point of the functional if and only if almost every point on the line is the mean of points projecting there. In other words, the condition is equivalent to $\overline{\gamma}$ being a principal curve (in the original sense).

The second variation of the length term is

$$\begin{aligned}\frac{d^2}{dt^2}\,\mathrm{L}(\gamma) &= \int_a^b \left(\frac{\gamma_{st}}{|\gamma_s|} - \frac{\gamma_s}{|\gamma_s|^2}\left(\frac{\gamma_s}{|\gamma_s|}\cdot\gamma_{st}\right)\right)\cdot\gamma_{st} + \frac{\gamma_s}{|\gamma_s|}\cdot\gamma_{stt}\,ds \\ &= \int_a^b \frac{1}{|\gamma_s|}\left(|\gamma_{st}|^2 - \left(\frac{\gamma_s}{|\gamma_s|}\cdot\gamma_{st}\right)^2\right) + \frac{\gamma_s}{|\gamma_s|}\cdot\gamma_{stt}\,ds.\end{aligned} \tag{3.2.5}$$

We note that $0 = (\gamma_s\cdot\gamma_t)_s = \gamma_{ss}\cdot\gamma_t + \gamma_s\cdot\gamma_{st}$, and therefore $\gamma_s\cdot\gamma_{st} = 0$, so that the second variation of the length term becomes just $|\gamma_{st}|^2$. Using (3.2.3) we again change coordinates, and evaluating at $t = 0$ we obtain

$$\left.\frac{d^2E}{dt^2}\right|_{t=0} = \int_a^b \left(\lambda_1|\gamma_{st}|^2 + 2\,\alpha(s)\int_{\Pi^{-1}(\overline{\gamma})}\left(|\gamma_t|^2 - \left((x-\Pi(x))\cdot\gamma_{st}\right)^2\right)d\mu_s(x)\right)ds. \tag{3.2.6}$$

We will use this second variation in the next section to determine when straight lines are linearly stable (local minimizers of the functional).

3.3 Relation between the minimizers and the data

In this section, our goal is to relate the parameters of the functional, the length-scales present in the data, and the length-scales seen in the minimizers. To do so we consider examples of data and corresponding minimizers, use the characterization of critical points of (MPPC), and perform linear stability analysis.

3.3.1 Examples and properties of minimizers

Here we provide some insight as to how minimizers of (MPPC) behave. We start by characterizing minimizers in some simple yet instructive cases. In the first couple of cases we focus on the behavior of single curves, and then investigate when minimizers develop multiple components.

Data on a curve.

Here we study the bias of penalized principal curves when the data lie on a curve without noise. If μ is supported on the image of a smooth curve, and a local minimizer γ of (PPC) is sufficiently close to μ, one can obtain an exact expression for the projection distance. More precisely, suppose that for each $s \in (a,b)$, $\Pi^{-1}(\gamma(s))$ contains one element x_s in $\mathrm{supp}(\mu)$. That is, x_s is the only point in $\mathrm{supp}(\mu)$ projecting to $\gamma(s)$. Then (3.2.4) implies

$$\lambda_1\mathcal{K}(s) = 2\alpha(s)h(s)(1+\mathcal{K}(s)h(s))$$

where $h(s) := |\gamma(s) - x_s|$, $\mathcal{K}$ denotes the unsigned scaler curvature of γ, and α is the projected linear density. Suppressing dependence on s, we have

$$h = \frac{1}{2\mathcal{K}}\left(\sqrt{1+2\frac{\lambda_1\mathcal{K}^2}{\alpha}}-1\right) \approx \begin{cases} \sqrt{\frac{\lambda_1}{2\alpha}} & \text{if } \frac{1}{\mathcal{K}} \ll \sqrt{\frac{\lambda_1}{\alpha}} \\ \frac{\lambda_1\mathcal{K}}{2\alpha} & \text{if } \frac{1}{\mathcal{K}} \gg \sqrt{\frac{\lambda_1}{\alpha}}. \end{cases} \tag{3.3.1}$$

Note that always $h \leq \sqrt{\frac{\lambda_1}{2\alpha}}$. We illustrate the transition of the projection distance h indicated in (3.3.1) with the example below.

Example 3.3.1. *Curve with decaying oscillations.* We consider data uniformly spaced on the image of the function $\frac{x}{5}\sin(-4\pi\log(x))$, which ensures that the amplitude and period are decreasing with the same rate, as $x \to 0^+$. In Figure 3.3.1, the linear density of the data is constant (with respect to arc length) with total mass 1, and solution curves are shown for two different values of λ_1. For x small enough the minimizing curve is flat, as it is not influenced by oscillations whose amplitude is less than $\sqrt{\frac{\lambda_1}{\alpha}}$. As the amplitude of oscillations grows beyond the smoothing length scale the minimizing curves start to follow them. As x gets larger and $\mathcal{K}$ becomes smaller, the projection distances at the peaks start to scale linearly with λ_1, as predicted by (3.3.1).

Indeed, as $\mathcal{K}$ decreases to zero the ratio of the curvature of the minimizer to that of the data curve approaches one and α converges to a constant. Hence from (3.3.1) it follows that the ratio of the projection distances at the peaks converges to the ratio of the λ_1 values.

Linear stability.

In this section we establish conditions for the linear stability of penalized principal curves. For simplicity we consider the case when $\operatorname{supp}(\mu) \subset \mathbb{R}^2$. Suppose that $\gamma : [0, L] \to \mathbb{R}^2$ is arc-length parametrized and a stationary configuration of (PPC), and that for some $0 \leq a < b \leq L$, $\gamma([a, b])$ is a line segment. As previously, we let α denote the projected linear density of μ onto γ.

We evaluate the second variation (3.2.6) over the interval $[a, b]$, where the considered variations of γ are $\gamma_t(s) = v(s) = (v_1(s), v_2(s))$, where $\gamma_s \cdot \gamma_t = 0$. Since γ is a line segment on $[a, b]$, we can consider coordinates where $v_1(s) = 0$. We then have

$$\left.\frac{d^2E}{dt^2}\right|_{t=0} = \int_a^b \lambda_1(v_2')^2 + 2\alpha(s)\int_{\Pi^{-1}(\gamma)}\left(v_2^2 - (v_2'(x-\Pi(x)))^2\right)d\mu_s(x)\,ds.$$

We define the *mean squared projection distance*

$$H(s) := \left(\int_{\Pi^{-1}(\gamma)}(x-\Pi(x))^2 d\mu_s(x)\right)^{\frac{1}{2}} \tag{3.3.2}$$

and obtain

$$\left.\frac{d^2E}{dt^2}\right|_{t=0} = \int_a^b \left(\lambda_1 - 2\alpha(s)H(s)^2\right)(v_2')^2 + 2\alpha(s)v_2^2\,ds. \tag{3.3.3}$$

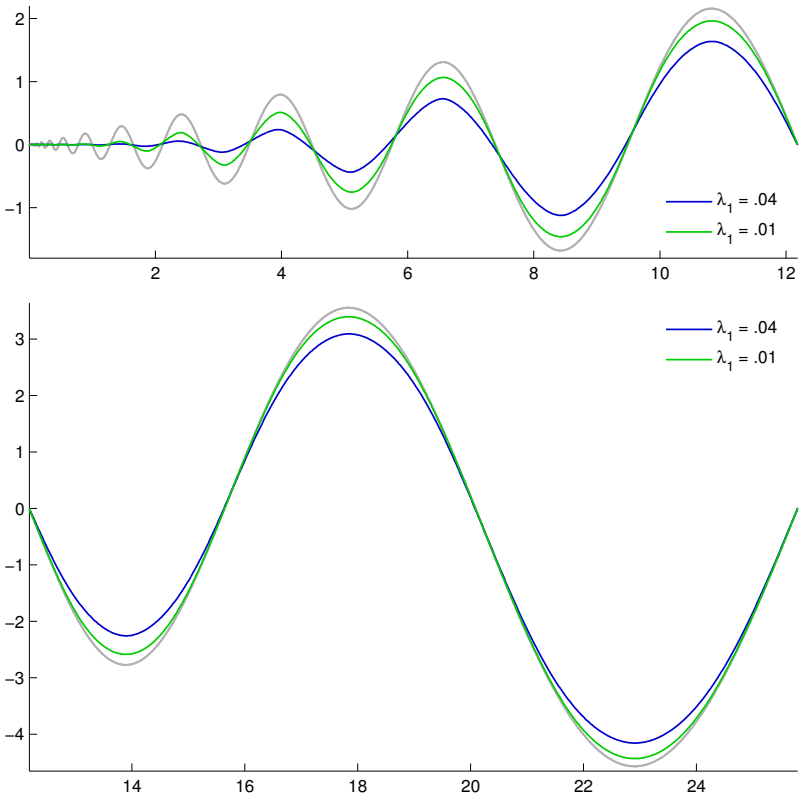


Figure 3.3.1: Numerical results shown for $n = 3000$ uniformly spaced data points (in gray) on the image of $\frac{x}{5}\sin(-4\pi\log(x))$ for $x \in [.001, e^{3.25}]$, and two different values of λ_1. (The bottom plot is a continuation of the top.)

We see that if $\lambda_1 \geq 2\alpha(s)H(s)^2$ for almost every $s \in (a,b)$, then $\frac{d^2E}{dt^2}\Big|_{t=0} > 0$ and so γ is linearly stable.

On the other hand, suppose that $\lambda_1 < 2\alpha(s)H(s)^2$ on some subinterval – without loss of generality we take it to be the entire interval (a,b). Consider the perturbation given by $v_2(s) = \sin(ns)$. Then the RHS of (3.3.3) becomes

$$n^2 \int_a^b \left(\lambda_1 - 2\alpha(s)H(s)^2\right) \cos^2(ns)ds + 2\int_a^b \alpha(s)\sin^2(ns)ds$$

and we see the first term dominates (in absolute value) the second for n large enough. Hence

$$\gamma \text{ is linearly unstable if} \quad \lambda_1 < 2\alpha(s)H^2(s) \quad \text{for all } s \in (a',b') \tag{3.3.4}$$

for some $a \leq a' < b' \leq b$.

In the following examples we examine linear stability for some special cases of the data μ.

Example 3.3.2. *Parallel lines.* We start with a simple case in which data, μ, lie uniformly on two parallel lines. In Figure 3.3.2 we show computed local minimizers starting with a slight perturbation of the initial straight line configuration, using the algorithm later described in Section 3.4. The data lines are of length 2, so that $\alpha = 0.5$ for the straight line configuration. Using $\lambda_1 = 0.16$ the condition for linear instability (3.3.4) of the straight line steady state becomes $0.4 < H$. The numerical results show that the straight line steady state does indeed become unstable when H becomes slightly larger than 0.4.

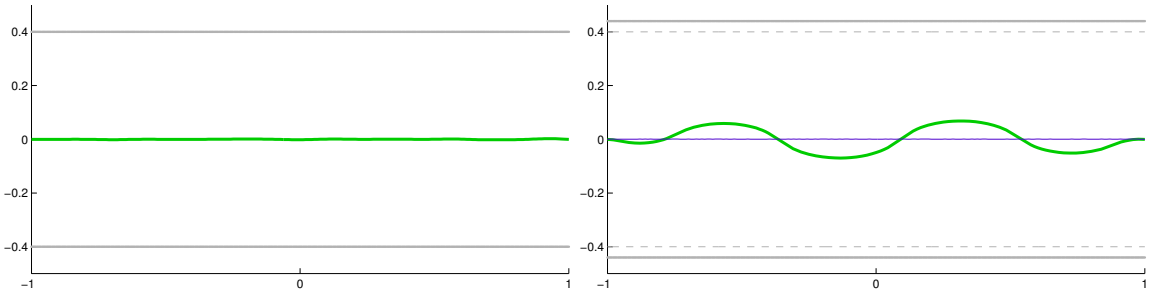


Figure 3.3.2: The data are gray line segments at height $H = \pm 0.4$ on the left image and $H = \pm 0.44$ on the right image. We numerically computed the local minimizers (green) of (MPPC) among curves with fixed endpoints at $(-1,0)$ and $(1,0)$, starting with slight perturbation of the line segment $[-1,1] \times \{0\}$.

Example 3.3.3. *Uniform density in rectangle.* Consider a probability measure, μ, with uniform density over $[0,L] \times [0,2h]$ with $L \gg h$. Linear instability of the line segment $[0,L] \times \{h\}$ (which is a critical point of (PPC)) can be seen as indication of when a local minimizer starts to overfit the data. It follows from (3.3.2) that $H^2 = \frac{1}{3}h^2$, and from (3.3.4) that $\lambda_1^* = \frac{2}{3L}h^2$ is the critical value for linear stability.

In Figure 3.3.3, we show the resulting local minimizers of (PPC) when starting from a small perturbation of the straight line, for several values of λ_1, for $h = \frac{1}{2}$ and $L = 4$. The results from the numerical experiment appear to agree with the predicted critical value of $\lambda_1^* = 1/24$, as the computed minimizer corresponding to $\lambda_1 = 1/27$ has visible oscillations, while that of $\lambda_1 = 1/23$ does not.

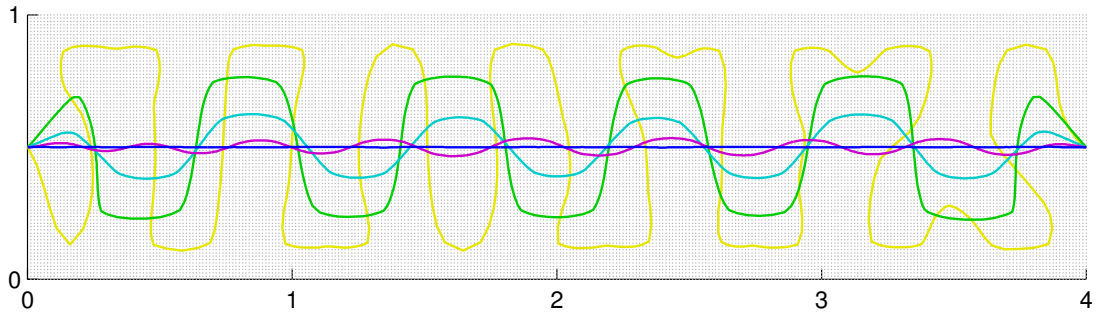


Figure 3.3.3: Numerical results showing local minimizers of (PPC) for various values of λ_1. The data are a grid of $n = 361 \times 81$ uniformly spaced points with total mass equal to 1. Curves with decreasing amplitude correspond to $\lambda_1 = 1/1000, 1/150, 1/50, 1/27, 1/23$. Recall that the critical value for linear stability is $\lambda_1^* = 1/24$. The initial curve used for all results was a randomly perturbed straight line segment $[0, 4] \times \{\frac{1}{2}\}$. The endpoints were kept fixed at $(0, 0.5), (4, 0.5)$ to avoid boundary effects.

To illustrate how closely the curves approximate that data we consider the average mean projection distance, H, for various values of λ_1. We expect that the condition for linear stability of straight-line critical points (3.3.4) applies, approximately, to curved minimizers. In particular, we expect that curves where H is larger than approximately $\sqrt{\frac{\lambda_1}{2\alpha}}$ will not be minimizers and will be evolved further by the algorithm. Here we investigate numerically if for minimizers $H \approx \sqrt{\frac{\lambda_1}{2\alpha}}$, as is the case in one regime of (3.3.1). Our findings are presented in Figure 3.3.4.

Example 3.3.4. *Vertical Gaussian noise.* Here we briefly remark on the case that μ has Gaussian noise with variance σ^2 orthogonal to a straight line. We note that the mean squared projection distance H is just the standard deviation σ. Therefore linear instability (overfitting) occurs if and only if $\lambda_1 < 2\alpha\sigma^2$.

Role of λ_2.

We now turn our attention to the role of λ_2 in (MPPC). Our goal is to understand when transitions in the number of curves in minimizers occur.

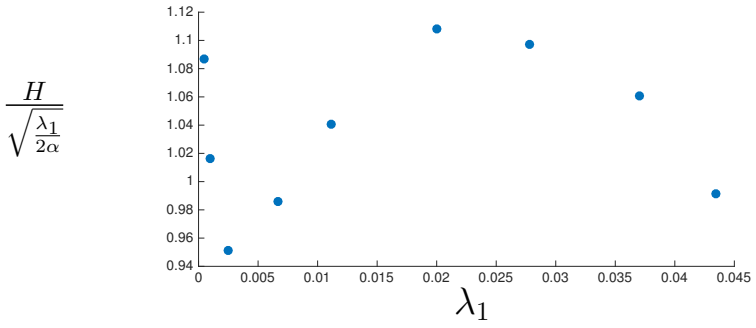


Figure 3.3.4: We compare the average mean projection distance H (defined in (3.3.2)) to $\sqrt{\frac{\lambda_1}{2\alpha}}$ (smoothing length scale) for the Experiment 3.3.3. We consider a somewhat broader set of λ_1 values than in Figure 3.3.3. We observe good agreement with the expectation, partly motivated by (3.3.4), that $H \sim \sqrt{\frac{\lambda_1}{2\alpha}}$.

By direct inspection of (MPPC), it is always energetically advantageous to connect endpoints of distinct curves if the distance between them is less than λ_2. Similarly, it is never advantageous to disconnect a curve by removing a segment which has length less than λ_2. Thus λ_2 represents the smallest scale at which distinct components can be detected by the (MPPC) functional. When distances are larger than λ_2, connectedness is governed by the projected linear density α of the curves, as we investigate with the following simple example.

Example 3.3.5. *Uniform density on line.* In this example, we consider the measure μ to have uniform density α on the line segment $[0, L] \subset \mathbb{R}$. We defer the technical details of the analysis to Section 3.3.4; here we report the main conclusions. By (3.3.11) there is a critical density

$$\alpha^* = \left(\frac{4}{3}\right)^2 \frac{\lambda_1}{\lambda_2^2}$$

such that if $\alpha > \alpha^*$ then the minimizer γ has one component and is itself a line segment contained in $[0, L]$. It is straightforward to check that γ will be shorter than L by a length of $h = \sqrt{\lambda_1/\alpha}$ on each side. Note that at the endpoints $H^2 = h^2/3$, which is less than the upper bound at interior points predicted by (3.3.1).

On the other hand, if $\alpha < \alpha^*$ and L is long enough then the minimizer consists of regularly spaced points on $[0, L]$ with space between them approximately (because of finite size effects)

$$\text{gap} \approx 2\left(\frac{3\lambda_1\lambda_2}{4\alpha}\right)^{\frac{1}{3}}. \tag{3.3.5}$$

An example of this scenario is provided in Figure 3.3.5.

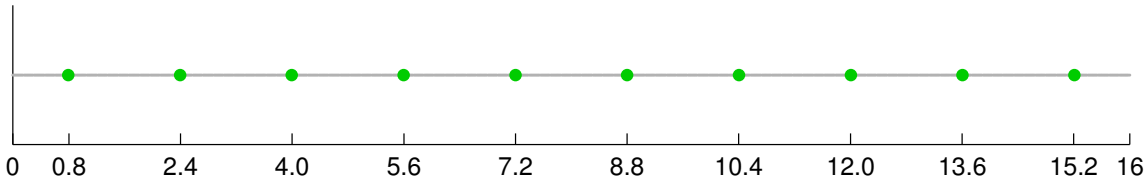


Figure 3.3.5: A minimizer for $n = 1000$ uniformly spaced points on a line segment, with total mass 1. Here $\lambda_1 = 1/16,\ \lambda_2 = .6$ and the critical value for connectedness is $\lambda_2^* = 4/3$. The optimal gap between the points is 1.6, compared to the approximation of ≈ 1.53 given by (3.3.5). The discrepancy is due to the finite length of the line segment considered in the example.

3.3.2 Summary of important quantities and length scales.

Here we provide an overview of how length scales present in the minimizers are affected by the parameters λ_1 and λ_2, and the geometric properties of data. We identify key quantities and length scales that govern the behavior of minimizers to (MPPC). We start with those that dictate the local geometry of penalized principal curves.

α — linear density (defined in Section 3.2.4). Given data as an absolutely continuous measure with compact support, and a smooth curve $\gamma : [a, b] \to \mathbb{R}^d$, the linear density $\alpha : (a, b) \to \mathbb{R}$ is the density of data projected onto the curve. Since α can vary over the curve, one may consider its values locally to help facilitate the following discussion.

$\sqrt{\frac{\lambda_1}{2\alpha}}$ — smoothing length scale (discussed in Sections 3.3.1 and 3.3.1 and illustrated in Example 3.3.3). This scale represents the resolution at which data will be approximated by curves. Consider data generated by a smooth curve with data density per length α and added noise (high-frequency oscillations, uniform distribution in a neighborhood of the curve, etc.). Noise centered around the curve will be ignored as long as its mean squared projection distance is less than $\sqrt{\frac{\lambda_1}{2\alpha}}$. In other words, $\sqrt{\frac{\lambda_1}{2\alpha}}$ is the length scale over which the noise is averaged out. Noise below this scale is neglected by the minimizer, while noise above is interpreted as signal that needs to be approximated. For example, if we take as data a line drawn by a pen, then $2\sqrt{\frac{\lambda_1}{2\alpha}}$ is the widest the pen tip can be, for the line to be considered as such by a minimizer of (PPC).

$\frac{\lambda_1 \mathcal{K}}{\alpha}$ — bias or approximation-error length scale (discussed in Section 3.3.1). Consider again data generated by smooth curve with data density per length α and curvature $\mathcal{K}$. If the curvature of the curve is small (compared to $\sqrt{\frac{\lambda_1}{\alpha}}$) and its reach is comparable

to $1/\mathcal{K}$, then the distance from the curve to the minimizer is going to scale like $\frac{\lambda_1 \mathcal{K}}{\alpha}$. That is, the typical error in reconstruction of a smooth curve that a minimizer makes (due to the presence of the length penalty term) scales like $\frac{\lambda_1 \mathcal{K}}{\alpha}$.

In addition to the above length scales, the following quantities govern the topology of multiple penalized principal curves:

λ_2 — connectivity threshold (discussed in Section 3.3.1). This length scale sets the minimum distance between distinct components of the solution. Gaps in the data of size λ_2 or less are not detected by the minimizer. Furthermore, this quantity provides the scale over which the following critical density is recognized.

$\frac{\lambda_1}{\lambda_2^2}$ — linear density threshold (discussed in Example 3.3.5 and Appendix 3.3.11). Consider again data generated (possibly with noise) by a smooth curve (with curvature small compared to $\sqrt{\frac{\lambda_1}{\alpha}}$) with data density per length α. If α is smaller than $\alpha^* = \left(\frac{4}{3}\right)^2 \frac{\lambda_1}{\lambda_2^2} + O(\mathcal{K})$, then it is cheaper for the data to be approximated by a series of points than by a continuous curve. That is if there are too few data points the functional no longer sees them as a continuous curve. If $\alpha > \alpha^*$, then the minimizers of (PPC) and (MPPC) are expected to coincide, while if $\alpha < \alpha^*$, then the minimizer of (MPPC) will consist of points spaced at distance about $\left(\frac{\lambda_1 \lambda_2}{\alpha}\right)^{\frac{1}{3}}$. Note that the condition $\alpha < \alpha^*$ can also be written as $\sqrt{\frac{\lambda_1}{\alpha}} < \frac{3}{4}\lambda_2$, and thus the minimizer can be expected to consist of more than component if the connectivity threshold is greater than the smoothing length scale.

We also remark the following scaling properties of the functionals. Note that $E_{a\mu}^{\lambda_1,\lambda_2} = aE_{\mu}^{\lambda_1/a,\lambda_2}$ for any $a > 0$. Thus, when the total mass of data points is changed, λ_1 should scale like $|\mu|$ to preserve minimizers. Alternatively, if $\mu_L(A) := \mu(\frac{A}{L})$ for every $A \subseteq \mathbb{R}^d$ and some $L > 0$, one easily obtains that $E_{\mu_L}^{\lambda_1,\lambda_2}(L\gamma) = L^2 E_{\mu}^{\lambda_1/L,\lambda_2/L}(\gamma)$.

3.3.3 Parameter selection

Understanding the length scales above can guide one in choosing the parameters λ_1, λ_2. Here we present a couple of approaches for selecting parameters when one has some estimate of quantitative properties of the data, including the linear density of the one-dimensional structures, and the level of noise or the distance between distinct components. In what follows, we assume that the data measure μ has been normalized, so that it is a probability measure.

A natural quantity to specify is a critical density α^*, which ensures that the linear density of any found curve will be at least α^*. From Section 3.3.2 it follows that setting α^* imposes the following constraint on the parameters: $\frac{16}{9}\frac{\lambda_1}{\lambda_2^2} = \alpha^*$. Alternatively, one can set

α^* if provided a bound on the desired curve length – if one is seeking a single curve with approximately constant linear density and length l or less, then set $\alpha^* = l^{-1}$.

There are a couple of ways of obtaining a second constraint, which in conjunction with the first determine values for λ_1, λ_2.

Specifying critical density α^* and desired resolution H^*.

One can set a desired resolution for minimizers by bounding the mean squared projection distance H. If α^* is set to equal the minimum of α along the curves then, the spatial resolution H from the data to minimizing curves is at most $\sqrt{\frac{\lambda_1}{2\alpha^*}}$. Consequently, if one specifies α^* and desires spatial resolution H^*, or better, the desired parameters are:

$$\lambda_1 = 2\alpha^* H^{*2} \quad \text{and} \quad \lambda_2 = \frac{4\sqrt{2}}{3} H^*.$$

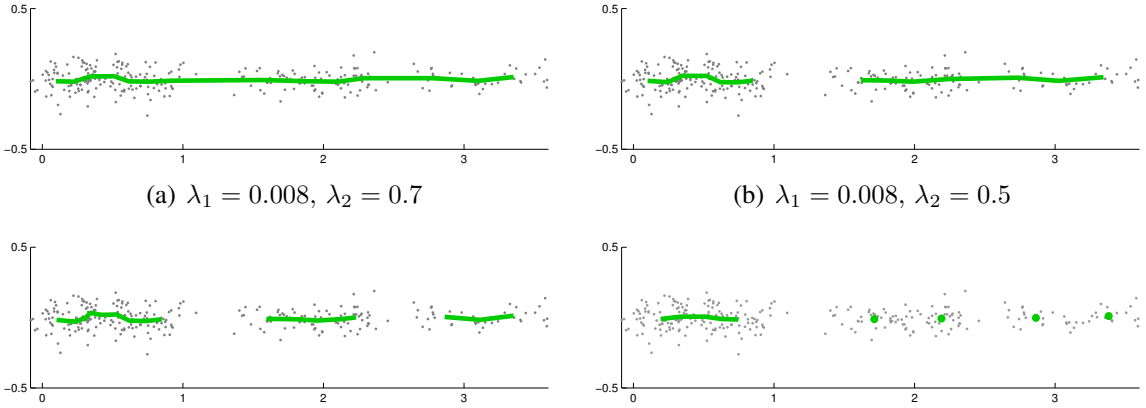
Choosing proper H^* depends on the level of noise present in the data. In particular, H^* needs to be at least the mean squared height of vertical noise in order to prevent overfitting.

Specifying critical density α^* and λ_2.

One may be able to choose λ_2 directly, as it specifies the resolution for detecting distinct components. In particular, there needs to be a distance of at least λ_2 between components, in order for them to detected as separate. Once set, $\lambda_1 = \frac{9}{16}\alpha^* \lambda_2^2$.

Typically one desires the smallest (best) resolution λ_2, that does not lead to α^* larger than desired. Even if a single curve is sought, taking a smaller value for λ_2 can ensure less frequent undesirable local minima. One case of this is later illustrated in Example 3.4.1, where local minimizers can oscillate within the parabola.

Example 3.3.6. *Line segments.* Here we provide a simple illustration of the role of parameters, using data generated by three line segments with noise. The line segments are of the same length, and the ratio of the linear density of data over the segments is approximately 4:2:1 (left to right). In addition, the first gap is larger than the second gap. Figure 3.3.6 shows how the minimizers of (MPPC) computed depend on parameters used. In the Subfigures 3.6(a), 3.6(b) 3.6(c) we keep λ_1 fixed while decreasing λ_2. As the critical gap length is decreased, and equivalently having more components in the minimizer becomes cheaper, the gaps in the minimizer begin to appear. It no longer sees the data representing one line but two or three separate lines. the only difference between functionals in Subfigures 3.6(c) and Subfigures 3.6(d) is that λ_1 is increased from 0.008 to 0.024. This results in length of the curve becoming more expensive. In Subfigure 3.6(d) we see that, due to low data density per length (α), the minimizer approximates the two data patches to the right by singletons rather than curves.



(a) $\lambda_1 = 0.008$, $\lambda_2 = 0.7$ (b) $\lambda_1 = 0.008$, $\lambda_2 = 0.5$

(c) $\lambda_1 = 0.008$, $\lambda_2 = 0.35$ (d) $\lambda_1 = 0.024$, $\lambda_2 = 0.35$

Figure 3.3.6: Minimizer of (MPPC) shown for different parameter settings. λ_1 and λ_2.

3.3.4 Analysis of the uniformly distributed data on a line segment

In this subsection we provide the derivation of the critical density α^* referenced in Example 3.3.5. Consider data uniformly distributed with density α on a line segment $[0, L]$. The functional (MPPC) takes the form

$$E_\mu^{\lambda_1,\lambda_2}(\gamma) := \int_0^L d(x,\gamma)^p \alpha dx + \lambda_1(\mathrm{L}(\gamma) + \lambda_2\,\mathrm{k}(\gamma)) \tag{3.3.6}$$

where for convenience we have let $\mathrm{k}(\gamma)$ denote the number of components of γ *minus 1*. We restrict ourselves to γ such that $\{0, L\} \subset \mathrm{range}(\gamma)$, so that γ takes the form $\gamma = \bigcup_{i=1}^{\mathrm{k}+1}[a_i, b_i]$, where $a_1 = 0$, and $b_{\mathrm{k}+1} = L$. Define $\tau := \sum_{i=1}^{\mathrm{k}+1}\tau_i$, $g := \sum_{i=1}^{\mathrm{k}} g_i$, where $\tau_i := b_i - a_i$ and $g_i := a_{i+1} - b_i$. We make the following observations:

Lemma 3.3.7. *The energy $E_\mu^{\lambda_1,\lambda_2}$ is invariant under redistribution of total length of γ, assuming that the number of components is $\mathrm{k}+1$, and that the gap sizes remain constant. More precisely, if $\bar\gamma = \bigcup_{i=1}^{\mathrm{k}+1}[\bar a_i, \bar b_i]$, $\tilde\gamma = \bigcup_{i=1}^{\mathrm{k}+1}[\tilde a_i, \tilde b_i]$ and there exists a permutation σ of $\{1, \ldots, \mathrm{k}\}$ such that $\bar g_i = \tilde g_{\sigma(i)}$ for $i = 1, ..., \mathrm{k}$, then $E_\mu^{\lambda_1,\lambda_2}(\bar\gamma) = E_\mu^{\lambda_1,\lambda_2}(\tilde\gamma)$.*

Lemma 3.3.8. *For $\mathrm{k} > 0$ fixed, the energy $E_\mu^{\lambda_1,\lambda_2}$ is minimized when the length of the gaps between components are uniform. More precisely, consider an arbitrary $\gamma = \bigcup_{i=1}^{\mathrm{k}+1}[a_i, b_i]$, with total gap g defined as above. Let $\tilde\gamma$ have $\mathrm{k}+1$ components such that $\tilde g_i = g/\mathrm{k}$, implying that $\tilde g = g$. Then $E_\mu^{\lambda_1,\lambda_2}(\tilde\gamma) \le E_\mu^{\lambda_1,\lambda_2}(\gamma)$, with equality only if $g_i = \tilde g_i$.*

Proof. The result is trivial for $\mathrm{k} = 0$. We prove the result for $\mathrm{k} = 1$. Consider γ with $g = g_1 + g_2$. The fidelity part of the energy $E_\mu^{\lambda_1,\lambda_2}$, as a function of g_1 is

$$F(g_1) = 2\int_0^{g_1/2} x^p \alpha dx + 2\int_0^{(g-g_1)/2} x^p \alpha dx.$$

Thus

$$\frac{dF}{dg_1} = \frac{1}{2^{p-1}}\alpha\left(g_1^p - (g-g_1)^p\right)$$

and

$$\frac{d^2F}{dg_1^2} = \frac{p}{2^{p-1}}\alpha\left(g_1^{p-1} + (g-g_1)^{p-1}\right) \geq 0.$$

By these we see that g_1 minimizes the energy if and only if $g_1 = g/2 = g_2$. The result for $\mathrm{k} \geq 2$ follows since one can consider the above situation by looking at the gaps formed by three consecutive components. □

Using Lemma 3.3.7 we may assume that each component not containing the endpoints 0 or L has the same length l, and that the two components containing the endpoints are of length $l/2$. By Lemma 3.3.8, the gaps between the components are $\frac{L-\mathrm{k}\,l}{\mathrm{k}}$. We first consider $\mathrm{k} > 0$ fixed, and minimize the energy w.r.t. l in the range $l \in [0, \frac{L}{\mathrm{k}})$. The energy

$$\begin{aligned} E &= \lambda_1 \mathrm{k}\, l + 2\,\mathrm{k}\int_0^{\frac{L-\mathrm{k}\,l}{2\,\mathrm{k}}} x^p \alpha dx + \lambda_1\lambda_2\,\mathrm{k} \\ &= \lambda_1 \mathrm{k}\, l + \frac{2}{p+1}\,\mathrm{k}\left(\frac{L-\mathrm{k}\,l}{2\,\mathrm{k}}\right)^{p+1}\alpha + \lambda_1\lambda_2\,\mathrm{k} \end{aligned} \tag{3.3.7}$$

is convex on $[0, \frac{L}{\mathrm{k}})$. Taking a derivative in l we obtain

$$\frac{dE}{dl} = \lambda_1\,\mathrm{k} - \mathrm{k}\left(\frac{L-\mathrm{k}\,l}{2\,\mathrm{k}}\right)^p \alpha.$$

Setting the derivative to zero and solving for l, and by noting that if there is no solution on $[0, \frac{L}{\mathrm{k}})$ then E is a nondecreasing function of l, we get that the energy is minimized at

$$l_{\mathrm{k}}^* = \begin{cases} \frac{L}{\mathrm{k}} - 2\left(\frac{\lambda_1}{\alpha}\right)^{1/p} & \text{if} \quad \mathrm{k} \leq \frac{L}{2\left(\frac{\lambda_1}{\alpha}\right)^{1/p}} =: \bar{\mathrm{k}} \\ 0 & \text{else.} \end{cases} \tag{3.3.8}$$

As we indicate above let $\bar{\mathrm{k}} = \frac{L}{2\left(\frac{\lambda_1}{\alpha}\right)^{1/p}}$. For k between 1 and $\bar{\mathrm{k}}$, plugging back into (3.3.7) we get that the minimal energy is

$$E_{min}(\mathrm{k}) = \lambda_1 L + \lambda_1\lambda_2\,\mathrm{k} - \frac{2p}{p+1}\left(\frac{\lambda_1}{\alpha}\right)^{1/p}\lambda_1\,\mathrm{k} \tag{3.3.9}$$

By direct inspection we verify that (3.3.9) is the (minimal) energy in the case that there is only one component (no breaks in the line). We note that (3.3.9) is linear in k, and hence for k between 0 and $\bar{\mathrm{k}}$, the minimizing value is at a boundary:

$$\mathrm{k}^* = \begin{cases} 0 & \text{if} \quad \lambda_2 \geq \frac{2p}{p+1}\left(\frac{\lambda_1}{\alpha}\right)^{1/p} \\ \lfloor \bar{\mathrm{k}} \rfloor & \text{otherwise} \end{cases}$$

We now consider $\mathrm{k} > \bar{\mathrm{k}}$ when all components have length zero ($l^*_{\mathrm{k}} = 0$). The energy in this case is

$$E_{l=0}(\mathrm{k}) := \frac{2}{p+1}\left(\frac{L}{2}\right)^{p+1}\frac{1}{\mathrm{k}^p}\alpha + \lambda_1\lambda_2\,\mathrm{k}$$

Considering k as a real variable we note that $E_{l=0}(\mathrm{k})$ is a convex function. Taking a derivative in k gives

$$\frac{dE_{l=0}}{d\,\mathrm{k}} = \frac{-2p}{p+1}\left(\frac{L}{2\,\mathrm{k}}\right)^{p+1}\alpha + \lambda_1\lambda_2.$$

If $\lambda_2 \geq \frac{2p}{p+1}\left(\frac{\lambda_1}{\alpha}\right)^{1/p}$ then $\frac{dE_{l=0}}{d\,\mathrm{k}} \geq 0$ for $\mathrm{k} > \bar{\mathrm{k}}$.

If $\lambda_2 < \frac{2p}{p+1}\left(\frac{\lambda_1}{\alpha}\right)^{1/p}$ then the point where the minimum is reached

$$\bar{\mathrm{k}}^*_{l=0} = \frac{L}{2}\left(\frac{(p+1)\lambda_1\lambda_2}{2p\alpha}\right)^{-\frac{1}{p+1}}$$

satisfies $\bar{\mathrm{k}}^*_{l=0} > \bar{\mathrm{k}}$ and thus belongs to the range considered. If $\bar{\mathrm{k}}^*_{l=0}$ is an integer then it is the minimizer of the energy, otherwise the minimizer is in the set $\{\lfloor\bar{\mathrm{k}}^*_{l=0}\rfloor, \lfloor\bar{\mathrm{k}}^*_{l=0}\rfloor + 1\}$. In all cases let us denote by $\mathrm{k}^*_{l=0}$ the minimizer of the energy: $\mathrm{k}^*_{l=0} = \arg\min_{k=\lfloor\bar{\mathrm{k}}^*_{l=0}\rfloor,\lceil\bar{\mathrm{k}}^*_{l=0}\rceil} E_{l=0}(k)$.

We note that there is a special case that $\mathrm{k}^*_{l=0} < \bar{\mathrm{k}}$. In that case the minimizer of the energy with exactly $\mathrm{k}^*_{l=0}+1$ components will be the one considered in the analysis of the $1 \leq \mathrm{k} \leq \bar{\mathrm{k}}$ case, and thus will have segments of positive length l^* given by formula (3.3.8).

To summarize, the optimal number of components will be

$$\begin{cases} 1 & \text{if} \quad \lambda_2 \geq \frac{2p}{p+1}\left(\frac{\lambda_1}{\alpha}\right)^{1/p} \\ \lfloor\bar{\mathrm{k}}\rfloor + 1 & \text{if} \quad \lambda_2 < \frac{2p}{p+1}\left(\frac{\lambda_1}{\alpha}\right)^{1/p}, \quad \text{and} \quad \mathrm{k}^*_{l=0} < \bar{\mathrm{k}} \\ \mathrm{k}^*_{l=0}+1 & \text{if} \quad \lambda_2 < \frac{2p}{p+1}\left(\frac{\lambda_1}{\alpha}\right)^{1/p}, \quad \text{and} \quad \mathrm{k}^*_{l=0} \geq \bar{\mathrm{k}}. \end{cases} \tag{3.3.10}$$

In the first case, there is just one single connected component. In the second case there are $\lfloor\bar{\mathrm{k}}\rfloor + 1$ components, each with equal positive length. We note that by Lemma 3.3.7 there exists a configuration with the same energy where one of these components has positive length, while the rest have zero length. The third case is that each of the components has length zero. We point out that if $\bar{\mathrm{k}}$ is integer-valued and $\lambda_2 < \frac{2p}{p+1}\left(\frac{\lambda_1}{\alpha}\right)^{1/p}$, then the minimizer will have $\mathrm{k}^*_{l=0}+1$ components.

We can now derive conclusions to the structure of minimizers if $L \gg 1$. From above we conclude that the minimizer will have one component (and be a continuous line) if $\lambda_2 \geq \frac{2p}{p+1}\left(\frac{\lambda_1}{\alpha}\right)^{1/p}$, and break up into at least $\lfloor\bar{\mathrm{k}}^*_{l=0}\rfloor + 1$ components otherwise. Rearranging, the condition also provides the critical density at which topological changes (gaps) in minimizers occur:

$$\alpha^* = \left(\frac{2p}{p+1}\right)^p \frac{\lambda_1}{\lambda_2^p}. \tag{3.3.11}$$

Finally we note that the typical gap length is $L/(\bar{\mathrm{k}}^*_{l=0})$ that is

$$L^* = 2\left(\frac{(p+1)\lambda_1\lambda_2}{2p\alpha}\right)^{\frac{1}{p+1}}. \tag{3.3.12}$$

3.4 Numerical algorithm for computing multiple penalized principal curves

For this section we assume the data measure μ is discrete, with points $x_1, x_2, ..., x_n \in \mathbb{R}^d$ and corresponding weights $w_1, w_2, ..., w_n \geq 0$. The weights are uniform $(1/n)$ for most applications, but we make note of our flexibility in this regard for cases when it is convenient to have otherwise.

For a piecewise linear curve $y = (y_1, ..., y_m)$, we consider projections of data to y_i's only. Hence, we approximate $d(x_i, y) \approx \min\{|x_i - y_j| : j = 1, ..., m\}$, unless otherwise stated. (Notation: when a is a vector, as are x_i, y_j in the previous line, $|a|$ denotes the Euclidean norm). Before addressing minimization of (MPPC), we first consider (PPC) where y represents a single curve. The discrete form is

$$\sum_{j=1}^{m}\sum_{i\in I_j} w_i|x_i - y_j|^2 + \lambda_1\sum_{j=1}^{m-1}|y_{j+1} - y_j| \tag{3.4.1}$$

where

$$I_j := \{i \ : \ (\forall k = 1, ..., m) \ |x_i - y_j| \leq |x_i - y_k|\} \tag{3.4.2}$$

represents the set indexes of data points for which y_j is the closest among $\{y_1, \ldots, y_m\}$. In case that the closest point is not unique an arbitrary assignment is made so that $I_1, ..., I_m$ partition $\{1, ..., n\}$ (for example set $\tilde{I}_j = I_j \backslash \bigcup_{i=1}^{j-1} I_i$).

3.4.1 Basic approach for minimizing (PPC)

Here we restrict our attention to performing energy decreasing steps for the (PPC) functional. We emphasize again that this minimization problem is non-convex. The projection assignments $I_1, ..., I_m$ depend on y itself. However, if the projection assignments are fixed, then the resulting minimization problem is convex. This suggests the following expectation-maximization algorithm outlined in Algorithm 1.

Note that if the minimization of (3.4.1) is solved exactly, then Algorithm 1 converges to a local minimum in finitely many steps (since there are finitely many projection states, which cannot be visited more than once).

Algorithm 1 Computing local minimizer of (PPC)

Input: data $x_1, ...x_n$, weights $w_1, ..., w_n$, initial curve $y_1, ..., y_m$, $\lambda_1 > 0$
repeat
1. compute $I_1, ..., I_m$ defined in (3.4.2)
2. minimize (3.4.1) for $I_1, ..., I_m$ fixed as described in Section 3.4.1

until convergence

Minimize functional with projections fixed

We now address the minimization of (3.4.1) with projections fixed (step 2 of Algorithm 1). One may observe that that his subproblem resembles that of a regression, and in particular the fused lasso [64].

To perform the minimization we apply the alternating direction method of multipliers (ADMM) [7], given in Section 2.5. We rewrite the total variation term as $|Dy|_{1,2} := \sum_{i=1}^{m-1} |(Dy)_i|$, where D is the difference operator, $(Dy)_i = y_{i+1} - y_i$ and $|\cdot|$ again denotes the Euclidean norm. An equivalent constrained minimization problem is then

$$\min_{y,z\,:\,z=Dy} \quad \sum_{j=1}^{m} \sum_{i \in I_j} w_i |x_i - y_j|^2 + \lambda |z|_{1,2}$$

Expanding the quadratic term and neglecting the constant, we obtain

$$\min_{y,z\,:\,z=Dy} \quad |y|_{\bar{w}}^2 - 2(y, \bar{x})_{\bar{w}} + \lambda |z|_{1,2} \tag{3.4.3}$$

where notation was introduced for total mass projecting to y_j by $\bar{w}_j = \sum_{i \in I_j} w_i$, center of mass $\bar{x}_j = \frac{1}{\bar{w}_j} \sum_{i \in I_j} w_i x_i$, and weighted inner product $(y, \bar{x})_{\bar{w}} = \sum_{j=1}^{m} \bar{w}_j (y_j, \bar{x}_j)$. One iteration of the ADMM algorithm then consists of the following updates:

1. $y^{k+1} = \text{argmin}_y \quad |y|_{\bar{w}}^2 - 2(y, \bar{x})_{\bar{w}} + \frac{\rho}{2} |Dy - z^k + b^k|^2$
2. $z^{k+1} = \text{argmin}_z \quad \lambda |z|_{1,2} + \frac{\rho}{2} |Dy^{k+1} - z + b^k|^2$
3. $b^{k+1} = b^k + Dy^{k+1} - z^{k+1}$

where $\rho > 0$ is a parameter that can be interpreted as penalizing violations of the constraint. As such, lower values of ρ tend to make the algorithm more adventurous, though the algorithm is known to converge to the optimum for any fixed value of $\rho > 0$.

The minimization in the first step is convex, and the first order conditions yield a tridiagonal system for y. The tridiagonal matrix to be inverted is the same for all subsequent iterations, so only one inversion is necessary, which can be done in $\mathcal{O}(md)$ time. In the second step, z decouples, and the resulting solution is given by block soft thresholding

$$z_i^{k+1} = \begin{cases} v_i^k - \frac{\lambda}{\rho} \frac{v_i^k}{||v_i^k||} & \text{if} \quad ||v_i^k|| > \frac{\lambda}{\rho} \\ 0 & \text{else} \end{cases}$$

where we have let $v_i^k = (Dy^{k+1})_i + b_i^k$. We therefore see that ADMM applied to (3.4.3) is very fast.

Note that one only needs for the energy to decrease in this step for Algorithm 1 to converge to a local minimum. This is typically achieved after one iteration of ADMM. In such cases few iterations may be appropriate, as finer precision typically gets lost once projections are updated. On the other hand, the projection step is more expensive, requiring $\mathcal{O}(nmd)$ operations to compute exactly. It may be worthwhile to investigate how to optimize alternating these steps, as well as more efficient methods for updating projections especially when changes in y are small. In our implementation we exactly recompute all projections, and if the resulting change in energy is small, we minimize (3.4.1) to a higher degree of precision (apply more iterations of ADMM before again recomputing projections).

3.4.2 Approach to minimizing (MPPC)

We now discuss how we perform steps that decrease the energy of the modified functional (MPPC). We allow $y = y_1, ..., y_m$ to consist of any number, k, of curves, and we denote them $y^1 = (y_1, ..., y_{m_1})$, $y^2 = (y_{m_1+1}, ..., y_{m_1+m_2})$, ..., $y^k = (y_{m-m_k+1}, ..., y_m)$, where $m_1 + m_2 + ... + m_k = m$. The indexes of the curve ends are $s_c = \sum_{j=1}^{c} m_j$ for $c = 1, ..., k$, and we set $s_0 = 0$. The discrete form of (MPPC) can then be written as

$$\sum_{j=1}^{m} \sum_{i \in I_j} w_i |x_i - y_j|^2 + \lambda_1 \sum_{c=0}^{k-1} \sum_{j=1}^{m_{c+1}} |y_{s_c+j+1} - y_{s_c+j}| + \lambda_1 \lambda_2 (k-1). \tag{3.4.4}$$

Our approach to (locally) minimizing the problem over y, k, $m_1, ..., m_k$ is to split the functional into parts that are decreased over different variables. Keeping k, $m_1, ..., m_k$ constant and minimizing over $y_1, ..., y_m$ we can decrease (3.4.4) by simply applying step 1 and step 2 of Algorithm 1 to each curve y^i, $i = 1, ..., k$ (note that step 2 can be run in parallel). To minimize over k, $m_1, ..., m_k$ we introduce topological routines below that disconnect, connect, add, and remove curves based on the resulting change in energy.

Disconnecting and connecting curves

Here we describe how to perform energy decreasing steps by connecting and disconnecting curves. We first examine the energy contribution of an edge $\{i, i'\} := [y_i, y_{i'}]$. To do so we compare the energies corresponding to whether or not the given edge exists. It is straightforward to check that the energy contribution of the edge $\{i, i'\}$ with respect to the continuum functional (MPPC) is

$$\Delta E_{i,i'} := \lambda_1 |y_{i'} - y_i| - \lambda_1 \lambda_2 - \sum_{j \in I_{i,i'}} w_j \min \left(|y_i - \Pi_{i,i'}(x_j)|, |y_{i'} - \Pi_{i,i'}(x_j)| \right)^2$$

where $I_{i,i'}$ is the set of data points projecting to the edge $\{i, i'\}$, and $\Pi_{i,i'}$ is the orthogonal projection onto edge $\{i, i'\}$. Our connecting and disconnecting routines will be based on

the sign of $\Delta E_{i,i'}$. We note that above criterion is based on the variation of the continuum functional rather than its discretization (3.4.4), in which projections to the vertices only (not edges) are considered. Our slight deviation here is motivated by providing a stable criterion that is invariant to further discretizations of the line segment $[y_i, y_{i'}]$. While we use the discrete functional to simplify computations in approximating the optimal fitting of curves, we will connect and disconnect curves based on the continuum energy (MPPC).

We first discuss disconnecting. We compute the energy contribution for each existing edge and if $\Delta E_{i,i'} < 0$, then we remove edge $\{i, i'\}$. Note this condition can only be true if the length of the edge is at least λ_2. It may happen that all edge lengths are less than λ_2, but that the energy may be decreased by removing a sequence of edges, whose total length is greater than λ_2. Thus, in addition to checking single edges, we implement an analogous check for sequences of edges. The energy contribution of a sequence of k edges $\{i, i+1\}, \{i+1, i+2\}, ..., \{i+k-1, i+k\}$ (including the corresponding interior vertices $y_{i+1}, ..., y_{i+k-1}$) is given by

$$\Delta E_{i:i+k} := \lambda_1 \left(\sum_{l=0}^{k-1} |y_{i+l+1} - y_{i+l}| - \lambda_2 \right) + \sum_{l=0}^{k-1} \sum_{j \in I_{i+l,i+l+1}} w_j \left((x_j - \Pi_{i+l,i+l+1}(x_j))^2 - (\min\{|x_j - y_i|, |x_j - y_{i+k}|\})^2 \right).$$

The routine for checking such edge sequences is outlined in Algorithm 2.

Algorithm 2 Removing appropriate edge sequences

Input: data $x_1, ...x_n$, weights $w_1, ..., w_n$, connected curve $y_1, ..., y_m$, projections I, $\lambda_1, \lambda_2 > 0$
set $i = 1$, $k = 1$, $len = |y_{i+1} - y_i|$
repeat
  repeat
    increment $k = k + 1$, $len = len + |y_{i+k} - y_{i+k-1}|$
  until $len > \lambda_2$ (or $i + k = m$, in which case break)
  compute $\Delta E_{i:i+k}$
  if $\Delta E_{i:i+k} < 0$ **then**
    remove edge sequence $\{i, i+1\}, \{i+1, i+2\}, ..., \{i+k-1, i+k\}$
    decrease $len = len - |y_{i+1} - y_i|$, advance $i = i + k - 1$, reset $k = 1$
  end if
  increment $i = i + 1$
until $i > m - 1$

Connecting is again based on the energy contribution of potential new edges. We use a greedy approach to adding the edges. That is, we compute $\Delta E_{i,i'}$ for each potential edge

$\{i, i'\}$, and add them in ascending order, connecting curves until no admissible energy-decreasing edges exist. We note that finding the globally optimal connections is essentially a traveling salesman problem, which is NP-hard. More sophisticated algorithms could be used here, but the greedy search is simple and has satisfactory performance.

Management of singletons:

Here we describe the procedures for topological changes via adding and removing components of the multiple curves. This is achieved by adding singletons (curves whose range is just a single point in $\mathbb{R}^d$), growing them into curves, and by removing singletons. Even if one is only interested in recovering one-dimensional structures, singletons may play a vital role. In particular, any low-density regions of the data (background noise or outliers) can often be represented by singletons in a minimizer of (MPPC), allowing the curves to be much less affected in approximating the underlying one-dimensional structure.

Below we provide effective routines for energy-decreasing transitions between configurations involving singletons. For checking whether (and where) singletons should be added, we examine each point y_i individually. If y_i is itself not a singleton, we compute the expected change in energy resulting from disconnecting y_i from its curve, placing it at the mean $\bar{x}_i$ of the data that project to it, and reconnecting the neighbors of y_i, so the number of components only increases by one. The change in the fidelity term will be exactly $-\bar{w}_i(\bar{x}_i - y_i)^2$, where $\bar{w}_i = \sum_{j \in I_j} w_j$ is the total mass projecting to y_i. Thus we add a singleton when

$$\lambda_1\lambda_2 < \bar{w}_i(\bar{x}_i - y_i)^2 + \lambda_1 \left(|y_i - y_{\max(1,i-1)}| + |y_i - y_{\min(m,i+1)}| - |y_{\max(1,i-1)} - y_{\min(m,i+1)}| \right).$$

If y_i is itself a singleton, then one cannot exactly compute the change in the energy due to adding another singleton in its neighborhood without knowing the optimal positions of both singletons. We restrict our attention to the data which project onto y_i, and note that if those points are the only ones that project to the new singleton, then adding the singleton may be advantageous only if the fidelity term associated to y_i is greater than $\lambda_1\lambda_2$. If that holds, we perturb y_i in the direction of one of its data points, place a new singleton opposite to y_i with respect to its original position, and apply a few iterations of Lloyd's k-means algorithm (with $k = 2$) to the data points that projected to y_i. We keep the two new points if and only if the energy decreases below that of the starting configuration with only y_i.

A singleton y_i gets removed if doing so decreases the energy. That is if

$$\lambda_1\lambda_2 > \sum_{j \in I_i} w_j \left(|x_j - y_i|^2 - d(x_j, y_{-i})^2 \right)$$

where $d(x_j, y_{-i}) := \min\{|x_j - y_{i'}| : i' \in [m],\, i' \neq i\}$.

Since singletons are represented by just a single point and cannot grow by themselves, we also check whether transitioning from singleton to short curve is advantageous. To do so we enforce that the average projection distance $\tilde{d}_i$ to a singleton y_i is less than $\sqrt{\lambda_1/\tilde{\alpha}}$,

which represents the expected spatial resolution, where $\tilde{\alpha} = \bar{w}_i/(4\tilde{d}_i)$ is an approximation to the potential linear density. Thus we add a neighboring point to y_i if

$$\tilde{d}_i = \sum_{j \in I_i} w_j |x_j - y_i| > \lambda_1 / \bar{w}_i.$$

Since this is based on an approximation, we also explicitly compute the posterior energy to make sure that it has indeed decreased, and only in this case keep the change.

Note that for each singleton y_j, minimizing the discrete energy (3.4.1) with projections fixed corresponds to placing y_j at its center of projected mass $\bar{w}_i$. Hence for singletons Algorithm 1 reduces to Lloyd's k-means algorithm.

In summary, we have fast and simple ways to perform energy decreasing steps involving the λ_2 term of the functional. Even when minimizers are expected to be connected, performing these steps may change the topological structure of the curve, keeping it in higher density regions of the data, and consequently evading several potential local minima of the original functional (PPC).

3.4.3 Re-parametrization of y

In applying the algorithm described thus far, it may, and often does, occur that some regions of y are represented with fewer points y_i than others, even if an equal amount of data are projected to those regions. That is, there is nothing that forces the nodes y_i to be well spaced along the discretized curve. To address this, we introduce criteria that $l_i \bar{w}_i$ be roughly constant for $i = 1, ..., m$, where $l_i = \frac{1}{2} \sum \{|y_i - y_j| : j \in \{i-1, i+1\} \cap [1, m]\}$ and $\bar{w}_i$ is the total weight of points projecting to y_i. This condition is motivated by finding for fixed m the optimal spacing of y_i's that minimizes the fidelity term of the discrete energy (3.4.4), under the assumption that the data are distributed with slowly changing density in a rectangular tube around straight line y.

3.4.4 Criteria for well-resolved curves

Here we discuss criteria for when a curve can be considered well-resolved with regard to the number of points m used to represent it. Namely, one would like to have an acceptable degree of resolution, without requiring m too large and needlessly increasing computational time. We suggest two such conditions.

One is related to the objective of obtaining an accurate topological representation of the minimizer, specifically the number of components. In order to have confidence in recovering components at a scale λ_2, the spacing between consecutive points on a discretized curve should be of the same scale. Thus we impose that the average of the edge lengths is at most $\frac{\lambda_2}{2}$.

Another approach for determining the degree of resolution of a curve is to consider its curvature. One may calculate the average turning angle and desire that it be less than some value (e.g. $\frac{\pi}{10}$). If λ_2 is not small enough, the first condition will not guarantee small

turning angles, and so we include this criterion as optional in our implementation. We note that in light of the possible lack of regularity of minimizers [59], it would not be reasonable to limit the maximal possible turning angle.

If either of the above criteria are not satisfied, we add more points to the curves where we expect they would decrease the discrete energy the most. Consistent with the criteria above for re-parametrization of the curves, we add points along the curve where $l_i \bar{w}_i$ is the largest.

3.4.5 Initialization

Finally, we discuss initialization. While the procedures described above enable the algorithm to evade many undesirable local minima, initialization can still impact the quality of the computed local minimizers. One of the simple ideas that we found to work very well is to initialize using singletons. We note that when the number of singletons is a fixed number k then minimizing (MPPC) reduces to minimizing the k-means functional. Thus to position the singletons for fixed k we use the standard Lloyd's algorithm to find the k-means cluster centers. We denote the (MPPC) energy of the k-means centers by $E(k)$. To determine a suitable value of k we perform a line search by starting with $k = 1$ and double it as long as $E(k)$ decreases, and then halve the intervals until a (local) minimizer k is found. We list the steps in Algorithm 3.

Algorithm 3 Initializing with singletons

Input: data $x_1, ...x_n$, weights $w_1, ..., w_n$, and $\lambda_1, \lambda_2 > 0$.
Set $k = 1/2$, $E(k) = +\infty$
repeat
  Let $k = 2k$
  Compute the k-means centers $C_k = \{c_1, \ldots, c_k\}$, and energy $E(k) := E_\mu^{\lambda_1,\lambda_2}(C_k)$
until $E(k) > E(k/2)$
Let $k' = \lfloor \frac{k}{2} \rfloor$, $k'' = k'$
repeat
  Let $k'' = \lfloor \frac{k''}{2} \rfloor$, $k = k' + k''$
  Compute the k-means centers $C_k = \{c_1, \ldots, c_k\}$, and energy $E(k) := E_\mu^{\lambda_1,\lambda_2}(C_k)$
  if $E(k) < E(k')$ **then**
    $k' = k$
  end if
until $k'' = 1$
Output: $y = C_{k'}$

Algorithm 4 Computing local minimizer of (MPPC) [Main Loop]

Input: data $x_1, ...x_n$, weights $w_1, ..., w_n$, initial curve $y_1, ..., y_m$, $\lambda_1, \lambda_2 > 0$.
set iter = 0
repeat
1. iter = iter + 1
2. compute $I_1, ..., I_m$, defined in (3.4.2)
3. run ADMM on non-singleton curves to decrease energy (3.4.4) as described in Section 3.4.1
4. replace singletons y_j by their center of mass: $y_j = \bar{x}_j$
if iter + 4 = 0 (mod top_period) **then**
remove appropriate edge sequences as described in Section 3.4.2 and Algorithm 2
else if iter + 3 = 0 (mod top_period) **then**
add or remove appropriate singletons as described in Section 3.4.2
else if iter + 2 = 0 (mod top_period) **then**
add appropriate connections as described in Section 3.4.2
else if iter + 1 = 0 (mod reparam_period) **then**
add points and re-parametrize the curves if needed as described in Sections 3.4.3, 3.4.4
end if
until convergence

3.4.6 Overview

Thus far we have described all of the main pieces of our algorithm to compute local minimizers of (MPPC). Here we describe how we put these pieces together. Algorithm 1, which includes ADMM for decreasing the discrete energy (3.4.1), computes approximate local minimizers of (PPC). To approximate local minimizers of (MPPC), we break up the minimization into separate parts. One consists of a "local" step that updates the placement of each curve, and is accomplished by running the ADMM step of Algorithm 1 on each curve. On the other hand, the inclusion of routines to disconnect, connect, add, and remove curves allows us to perform energy-decreasing steps of (MPPC) in a more global topological fashion.

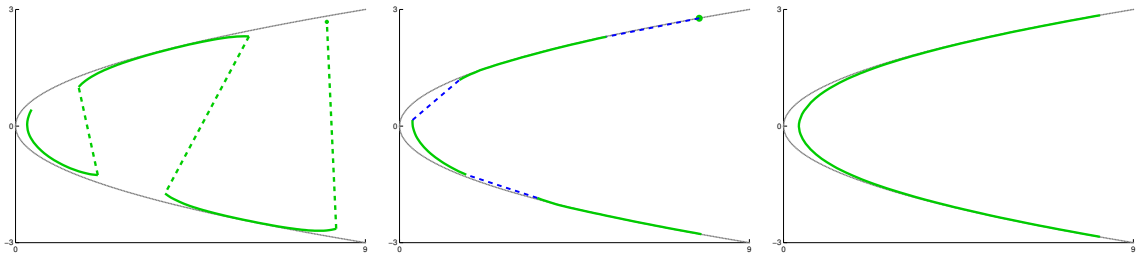
We provide an general outline for finding local minimizers of (MPPC) in Algorithm 4. The (potentially) topology-changing routines outlined in 3.4.2, 3.4.2 are run on a regular basis throughout the steps of Algorithm 1. In particular we run them every $top\_period = 10$ iterations and we run the reparameterization of curves every $reparam\_period = 5$ iterations. The performance for different values, as well as for different order of operations was similar.

Finally, we note that the the computational complexity of the algorithm is dominated by the step of computing the projections $I_1, ..., I_m$, which requires $\mathcal{O}(mnd)$ operations.

3.4.7 Further numerical examples

We present a couple of further computational examples which illustrate the behavior of the functionals and the algorithm. For some of the examples, we include comparisons with results from other approaches including the Subspace Constrained Mean Shift algorithm and Diffusion Maps.

Example 3.4.1. *Parabola.* We begin with an example that illustrates the cutting and reconnecting mechanism used in the Algorithm 4 for finding minimizers of (MPPC). We use data that are uniformly distributed on the graph of the parabola $x = y^2$ for $y \in [-3, 3]$ and set $\lambda_1 = 0.12$ and $\lambda_2 = 4/3$. For illustration, we first run Algorithm 4 for minimizing (PPC) (the same as main loop of Algorithm 4 without allowing any topological changes) starting from a small perturbation of the line segment $[0, 9] \times \{0\}$. The result is shown in Figure 3.1(a). We then turned on the cutting routine, described in Algorithm 2. The segments to be cut are indicated on Figure 3.1(a) as dashed lines. Figure 3.1(b) shows a subsequent configuration, after a few steps of ADMM relaxation, but prior to reconnecting. Edges that are about to be added in the reconnection step (described in Section 3.4.2) are shown as dashed blue lines.



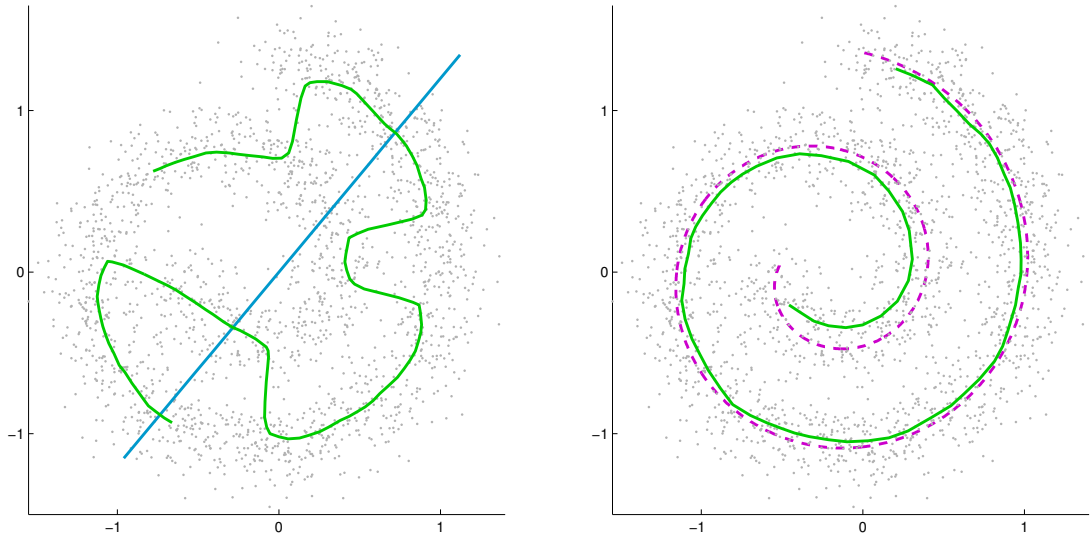
(a) Local minimizer of (PPC) with edges to be cut indicated.

(b) Subsequent configuration, with edges to be added indicated.

(c) Minimizer of (MPPC)

Figure 3.4.1:

Example 3.4.2. *Noisy spiral.* Here we consider data generated as noisy samples of the spiral $t \mapsto (t\cos(t), t\sin(t))$, $t \in [3, 14]$, shown as a dashed line in Figure 3.2(b). 2000 points are drawn uniformly with respect to arc length along the spiral. For each of these points, noise drawn independently from the normal distribution $1.5\mathcal{N}_2(0, 1)$ is added. In Figure 3.4.2, we show the results of algorithms for minimizing (PPC) and (MPPC). The initialization used for both experiments is a diagonal line corresponding to the fist principal component. The descent for (PPC) does not allow for topological changes of the curve and subsequently gets attracted to a local minimum. Meanwhile, Algorithm 4 for minimizing (MPPC) is able to recover the geometry of the data, via disconnecting and reconnecting the initial curve.

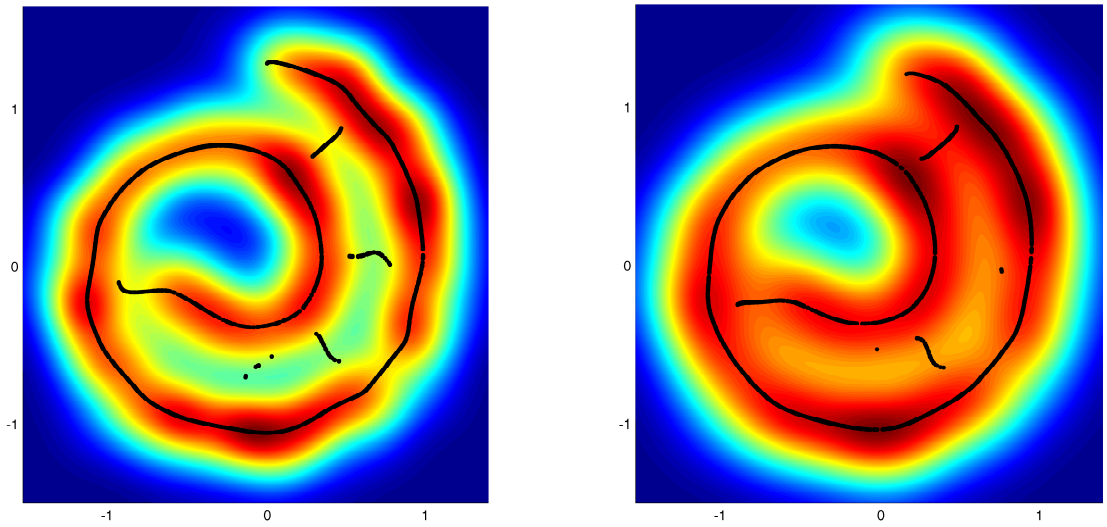


(a) Local minimizer of (PPC)

(b) Local minimizer of (MPPC)

Figure 3.4.2: Numerical results on data generated by a spiral (in purple) plus Gaussian noise. On the left, (PPC) is used to find the local minimizer given by the green curve, using the first principal component (blue) as initialization. On the right, (MPPC) (with the same initialization) is minimized to find the green curve, using critical linear density $\alpha^\star = .09$. In both cases $\lambda_1 = .01$.

For this dataset we also include results of the Subspace Constrained Mean Shift (SCMS) algorithm [53], also studied in [16, 17, 30] as means to find one-dimensional structure in data. SCMS seeks to find the ridges (of an estimate) of the underlying probability density of the data. The ridge set of a function $F : \mathbb{R}^d \to \mathbb{R}$ is defined as the set where ∇F is an eigenvector of the Hessian of F and the eigenvalues of all remaining eigenvectors are negative (the point is a local maximum along all orthogonal directions). In practice, given a random sample one uses a kernel density estimator (KDE) to approximate the probability distribution. SCMS algorithm takes a set of points as input, and successively updates each point until it converges to a ridge point of the KDE of a specified bandwidth. The output is then a list of unordered points that approximate the ridge set. We apply SCMS using a Gaussian kernel density estimate (KDE) for two different bandwidth, which both give good results. The algorithm is initialized with 2500 points on a 50x50 mesh in the range of the data. As shown in Figure 3.4.3, the algorithm does output points approximating the underlying one-dimensional structure. However, we note that mathematically there are a number of ridges of the KDE going between the layers of the spiral. The SCMC algorithm captures those with high enough density and large enough "basin of attraction". We note that as the kernel bandwidth increases, the number of undesirable ridges decreases, but their intensity increases (the density at the remaining ridges is higher). Removing points on the mesh that have density below a given threshold has been suggested for noisy data



(a) SCMS with bandwidth 0.15 (b) SCMS with bandwidth 0.2

Figure 3.4.3: There are more undesirable ridges for small bandwidths. Decreasing the bandwidth further can also result in gaps in the desirable spiral filament. There are fewer undesirable ridges at higher bandwidth, but they have higher density. Increasing the bandwidth further introduces a significant bias of the main ridge, compared to the generating curve.

[17, 30], and doing so can improve the results here by eliminating some of the undesirable ridges. However this introduces a parameter (density threshold) that needs to be chosen carefully (see appendix A of [17]).

Example 3.4.3. *Noisy grid with background clutter.* In the following example we illustrate the robustness of the proposed approach to background noise. We use data in $\mathbb{R}^3$ with an underlying grid-like structure, shown in Figure 3.4.4. The data consist of 2400 points generated by four intersecting lines with Gaussian noise, plus 2400 more points uniformly sampled from the background $[0, 3] \times [0, 3] \times [-.75, .75]$.

Since the linear density of data in the background noise is less than that of the intersecting lines, the computed minimizer approximates the data in the background by isolated points (in green). For the parameters we used, this is predicted by the discussion of the density threshold in Section 3.3. By choosing λ_1 and λ_2 so that the critical density threshold $\alpha^* = \left(\frac{4}{3}\right)^2 \frac{\lambda_1}{\lambda_2^2}$ is between the linear density of the background noise and the linear density of the lines, the background noise will be represented by isolated points, which allows the curves to appropriately approximate the intersecting lines.

We note that although the algorithm succeeds in approximating the one-dimensional structure of the data, it is not able to recover the intersections due to the simpler structure of configurations we consider in (MPPC). In such cases where our approach cannot identify the global topology, we presume it may be possible to use the obtained approximation as input for other approaches that aim to recover the topology of the data [58].

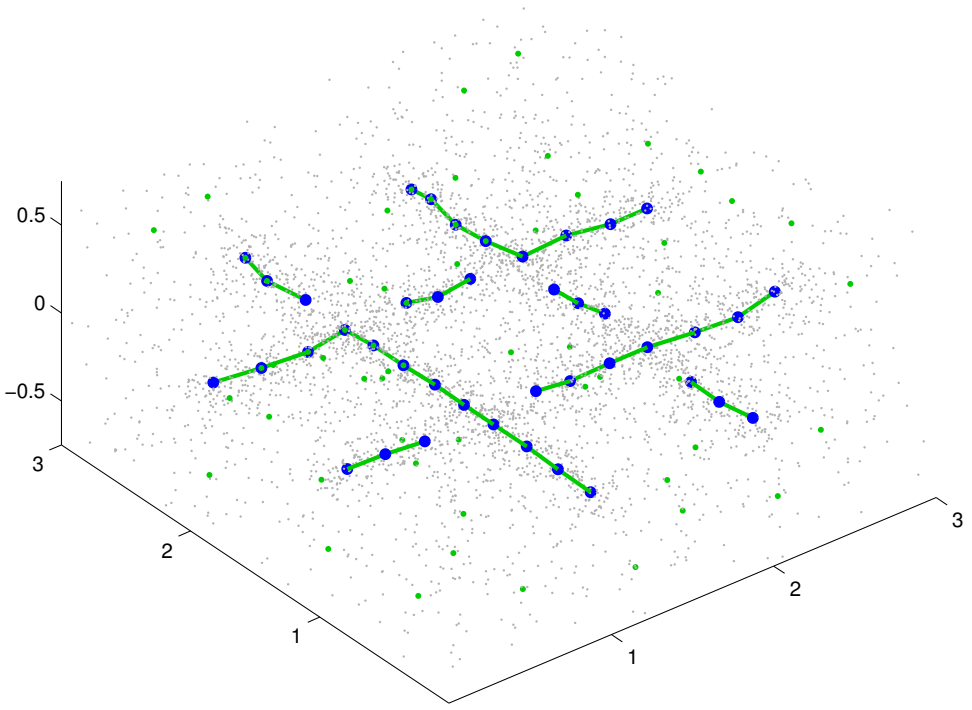


Figure 3.4.4: Computed minimizer on data generated from four intersecting lines forming a grid with Gaussian noise and background clutter in $\mathbb{R}^3$, with $\lambda_1 = 7 \times 10^{-4}$, $\lambda_2 = 0.2$. The minimizer consists of curves and isolated points (green). The larger blue dots represent the discretization (points y_i) of curves which are a part of the minimizer.

Example 3.4.4. *Zebrafish embryo images.* Here we demonstrate performance of the algorithm on a high-dimensional dataset that consists of grayscale images.

In [22] Dsilva et al. develop a technique for finding the temporal order of still images of a developmental process. They consider the problem where both the time ordering and the angular orientation of the images are unknown. To be able to handle both variables simultaneously they use vector diffusion maps [57]. One of the tests they performed to validate their approach was on images taken from a time-lapse movie that captures zebrafish embryogenesis [`https://zfin.org/zf_info/movies/Zebrafish.mov`] (Karlstrom and Kane [38]).

In this case the angle of rotation is fixed, and recovering the temporal order can be done using diffusion maps [18] alone, see Figure 3.6(b). Here we demonstrate that these images can also be ordered using our method.

As in [22], we apply our algorithm to 120 consecutive frames (roughly corresponding to seconds 6-17 in the movie) of 100x100 pixels in order to test how well it can recover the development trajectory. Thus each image is represented as a point in $\mathbb{R}^{10,000}$. We note that there is almost no noise in the dataset, but emphasize that the goal here is to recover a

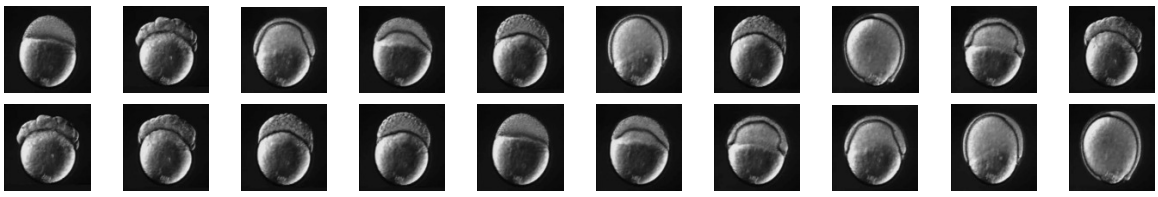
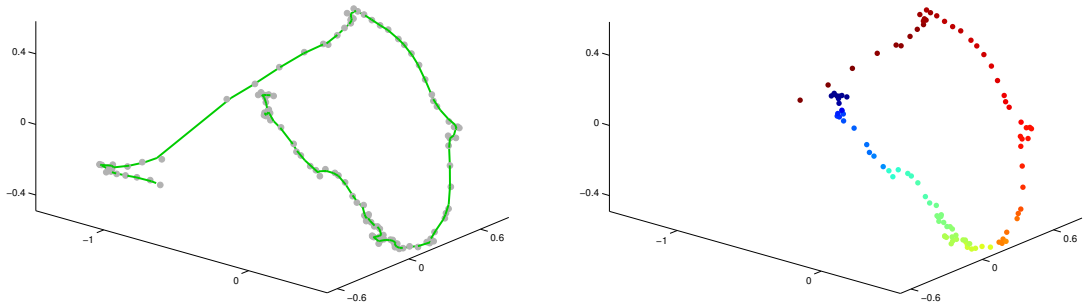


Figure 3.4.5: The top row shows 10 images of zebrafish embryos in random order. The bottom row shows 10 images ordered by the found curve that minimizes (MPPC).

single curve passing through data whose true order is not provided to the algorithm.

After normalizing the data, we run our algorithm with parameters $\lambda_1 = 10^{-3}$ and $\lambda_2 = 2$. The low value for λ_1 is appropriate given that there is virtually no noise. The high value of λ_2 ensures that a single curve is found, and so the functional (PPC) is also being minimized. Our algorithm outputs a curve that correctly ranks all of the original images. Figure 3.4.5 shows a random sample of the images used, along with their found true ordering. In Figure 3.6(a), we visualize the found curve in $\mathbb{R}^3$ using the first three principal components.

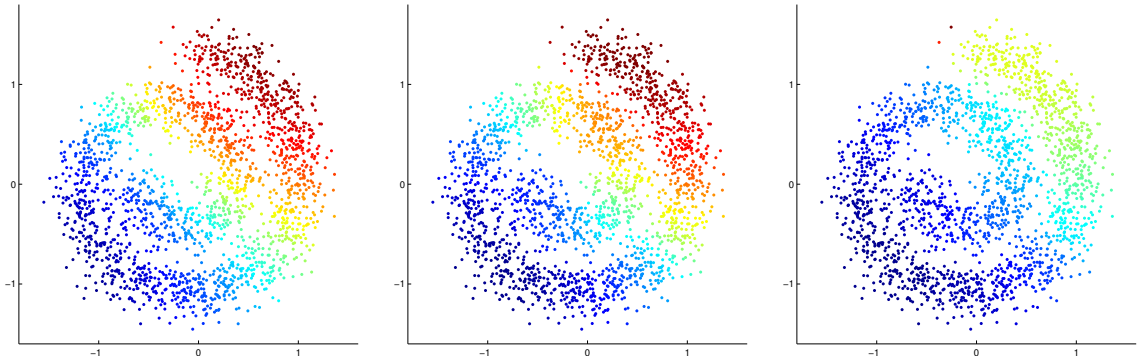


(a) The curve found minimizing (MPPC)

(b) Color-coded first embedding coordinate of the diffusion map

Figure 3.4.6: On both images the first three principal components are used for visualization. The (MPPC) algorithm was applied to all 120 images, while we applied diffusion maps to only the first 104 images due to a slight camera shift that resulted in relatively large euclidean distance between images 104 and 105. Both methods perfectly ranked their respective data, and some (simple) preprocessing done in [22] allows diffusion maps to work on the full 120 images.

Example 3.4.5. *Noisy spiral revisited.* In the previous example we discussed the feasibility of using nonlinear dimensionality reduction techniques such as diffusion maps to order



(a) $\epsilon = 4d^2$ (b) $\epsilon = \frac{d^2}{4}$ (c) $\epsilon = \frac{d^2}{16}$

Figure 3.4.7: Color-coded one-dimensional embeddings provided by the diffusion maps algorithm for three settings of the scaling parameter ϵ. No values for ϵ recover the intrinsic ordering of the data. Larger values (left and middle) cannot detect the finer structure, while a smaller value (right) separates two outliers at the top (colored red and brown) from the rest of the data.

the data. Since the data in Example 3.4.4 had almost no noise, one can obtain a good ordering using many different methods. Spectral dimensionality reduction techniques are often successful even when substantial noise is present. However, when there is significant overlap in the distribution of data whose generating points have large intrinsic distance, spectral methods can fail to recover the desired one-dimensional ordering. The example below illustrates this and indicates that in some situations minimizing (MPPC) gives better results in ordering the data than diffusion maps.

We revisit the noisy spiral data considered in Example 3.4.2, and run the diffusion maps algorithm using a range of scaling parameters $\epsilon = (\frac{d}{c})^2$, where d denotes the median of the pairwise distances of the data points. After testing a wide range of parameter values c, we found that for all values tested the spectral embedding fails to recover the desired one-dimensional ordering. We display the typical results (which correspond to $c = 0.5, 2,$ and 4) in Figure 3.4.7. Larger values of ϵ lead to an embedding that differentiates the data linearly from bottom left to top right, while smaller values lead to an embedding that separates outliers in the top from the rest of the data points. On the other hand minimizer of (MPPC) can correctly recover the one-dimensional structure, as shown in Figure 3.2(b).

3.5 Discussion and conclusions

In this chapter, we proposed a new objective functional (MPPC) for finding one-dimensional structures in data that allows for representation consisting of several components. The functional introduced is based on the average-distance functional and is a regularization of principal curves. It penalizes the approximation error, total length of the

curves, and the number of curves used. We have investigated the relationship between the data generated by one-dimensional signal with noise, the parameters of the functional, and the minimizer. Our findings provide guidance for the choice of parameters, and can further be used for multi-scale representation of the data. In addition, we have demonstrated that the zeroth-order term helps energy descent based algorithms converge to desirable configurations. In particular, energy descent approaches for (PPC) very often end up in undesirable local minima. The main reason for this is of topological nature – points on the approximate local minimizer represent the data points in an order which may be very different from the true ordering corresponding to the (unknown) generating curve. The added flexibility of being able to split and reconnect the curves provides a way for resolving such topological obstacles.

We have developed a fast numerical algorithm for estimating minimizers of (MPPC). It has computational complexity $\mathcal{O}(mnd)$, where n is the number of data points in $\mathbb{R}^d$, and m is the number of points used in the approximating curve(s). We demonstrated the effectiveness of the algorithm in recovering the underlying one-dimensional structure for real and synthetic data, in cases with significant noise and in very high dimensions. The robustness and computational efficiency of the algorithm compare favorably to existing methods, and further offer promise for scalability of approximating one-dimensional structures in very large and complex high-dimensional data.

Despite these advancements, we expect there are a few aspects of the functionals that can be further investigated. One of these concerns whether we can select the parameters of the functional in a more systematic way, without relying on any of the quantitative estimates of the data involved in the discussion in Section 3.3.3. In particular is the selection of λ_1, which has the role that minimizers are linearly stable as long as the mean squared projection distance is less than $\sqrt{\frac{\lambda_1}{2\alpha}}$. This may suggest that minimizers for which the mean squared projection distance is greater than $\sqrt{\frac{\lambda_1}{2\alpha}}$ are not overfitting, and hence are appropriately approximating the data. Thus, starting with a high value of λ_1, and decreasing it right until the minimizer becomes linear stable may be a strategy that is worth investigating.

Another direction for future work is in obtaining theoretical results on when the proposed algorithm can approximately recover the curve from which the data are generated. In the case that the data lie on a smooth curve with positive reach, the resolution estimates support the claim that the global minimizer will recover the generating curve (up to bias) for a small enough value of λ_1 and large enough value of λ_2. In this setting, we also believe that the generating curve can always be numerically recovered (up to bias), by computing a series of local minimizers for path of (MPPC) problems. In particular, the proposed algorithm can be seen as solving a sequence of problems with λ_2 increasing, if adding connections (done greedily) are limited to one at a time after the previous run converges. As is, the algorithm starts with singletons, and connects and grows them until it is advantageous to do so, based on the values of λ_1, λ_2. We expect that if one fixes λ_1 small enough, and starts with a low λ_2 value that is continuously increased up to a high enough value, the computed local minimizers should progress to a global minimizer that represents the

generating curve up to a controllable bias. Proving this in the simple case of no noise, and potentially extending it to cases with noise (under suitable assumptions on the reach) would further support our approach.

Recently there has been a significant effort to recover one-dimensional structures that are branching and intersecting, which are beyond the capability of multiple penalized principal curves (at the topological level). The ability to recover graph structures and the topology of the data is very valuable, and facilitates a number of data analysis tasks. Several notable works are based on Reeb graphs and related objects [15, 29, 58]. We note that these approaches are sensitive to noise, whose presence also significantly slows down the respective algorithms. We believe computing multiple penalized principal curves could be valuable as a pre-processing step for simplifying the data prior to applying graph-based approaches that find the connectivity network of the data set. Recalling the data from Example 3.4.3 and Figure 3.4.4, we see that although our approach does not recover the topological structure, it does identify and appropriately simplify the one-dimensional structure present in the data. The graph-based approaches should work much better on the simplified green curves or (or blue points) than on the original point cloud. We end by remarking that recovering more complicated one-dimensional structures of data will be further discussed in the following chapter.

Chapter 4

Optimal Networks for Selective-Transport

4.1 Introduction

In this chapter we investigate a new set of models for networks that represent well a given measure μ on $\mathbb{R}^d$. In contrast to the last chapter, for the most part here we will think of the measure as a population of agents, for which we want to find networks (one dimensional subsets of $\mathbb{R}^d$) that best meet the transportation needs of the population. We will investigate models where we assume that every member of the population is selective in their need to visit a distribution of locations of interest. This assumption will distinguish our approach from previous work on optimal transportation networks. The networks we seek will have a low cost that balances the total transportation cost of the population and the cost of building and maintaining the network. We will propose two models for optimal networks in this sense, and propose a third model for a setting in which the population measure μ is jointly sought with the network. We will establish existence of minimizers for the problems, and provide an algorithm for approximating optimal networks. Lastly, we will provide a couple of numerical examples that aim to illustrate how the complexity of the optimal networks increases as the population grows. One application to recovering one-dimensional of data will also be provided.

4.1.1 Background and related work

There are a number of works related to urban planning and the models we consider. However, most of the work we are aware of considers the problem of finding a network along which *two* given distributions μ and ν are optimally matched. In these models, one can view μ to represent a distribution of workers, and ν to represent a distribution of work locations. The approaches are based on the framework of optimal transportation, and inherently assume that any worker is suitable for any working location. Below we review some of the related approaches, starting with classical optimal transportation, in which there is no role

of networks.

Optimal transport

In 1781, Monge introduced the following problem [50]. Given two compactly supported (Borel) probability measures μ and ν on $\mathbb{R}^d$, let $\mathcal{T}$ denote the set of maps $T : \mathbb{R}^d \to \mathbb{R}^d$ such that $T_{\#}\mu = \nu$. That is, $T \in \mathcal{T}$ satisfies $\mu(T^{-1}(A)) = \nu(A)$ for all Borel sets $A \subseteq \mathbb{R}^d$. Given a cost $c(x, y) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ of transporting from x to y, Monge's problem is then

$$OT_M(\mu, \nu) := \inf_{T \in \mathcal{T}} \int_{\mathbb{R}^d} c(x, T(x)) d\mu(x). \tag{4.1.1}$$

A common choice for the cost function is $c(x, y) = |x - y|^p$ for $p \geq 1$.

Monge's problem faces some difficulties regarding existence of optimal transportation maps – maps that minimize (4.1.1). A major one is that the set of admissible maps $\mathcal{T}$ may be empty if the measure μ has atoms. Optimal transport maps are known to exist and be unique in the case that μ is absolutely continuous and $c(x, y)$ is a strictly convex function of $|x - y|$.

It was not until the 1940's that Kantorovich proposed another formulation [37] that alleviated some of these problems. Kantorovich's problem amounts to relaxing the class of transportation maps $\mathcal{T}$ to transportation plans, which allow for mass to be "split". Transportation plans are probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ with first marginal μ and second marginal ν. More precisely, the set of transportation plans is defined as

$$\Pi(\mu, \nu) = \left\{ \pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) : \pi(A \times \mathbb{R}^d) = \mu(A), \pi(\mathbb{R}^d \times B) = \nu(B) \ \forall A, B \subseteq \mathbb{R}^d \right\}$$

and Kantorovich's problem is then

$$OT_K(\mu, \nu) := \inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x, y) d\pi(x, y). \tag{4.1.2}$$

Note that for any given transportation map $T \in \mathcal{T}$, $(I \times T)_{\#}\mu$ defines a corresponding transportation plan, and so (4.1.2) is a relaxation of (4.1.1). In contrast to transportation maps, optimal transportation plans (minimizers of Kantorovich's problem) exist in the general case that c is a lower semi-continuous function that is bounded from below.

Using optimal transport one may define a distance between measures, known as the *Wasserstein distance*:

$$W_p(\mu, \nu) := \min \left\{ \int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y|^p d\pi(x, y) : \pi \in \Pi(\mu, \nu) \right\}^{\frac{1}{p}},$$

where $p \geq 1$. It can be shown the the Wasserstein distance is indeed a metric, and furthermore that it metrizes weak-$\star$ convergence of measures. That is, $\mu_n \overset{\star}{\rightharpoonup} \mu$ if and only if $W_p(\mu_n, \mu) \to 0$ for any $p > 1$. For reference on the facts stated here and further results on optimal transportation we point the reader to a recent book by Santambrogio [56], along with a couple of other good references [1], [66] .

Branched transport

We now turn our attention to finding optimal networks for transportation problems. The first model we describe is motivated by transport related phenomena that exhibit ramification or branching patterns. As in the Monge and Kantorovich problems, the branched transport problem deals with moving or transporting an initial distribution μ to a distribution ν, but differs in its core assumption that transporting mass together is less costly than transporting the mass separately.

When the measures are discrete, the branched transport problem can be formulated using a weighted and directed geometric graph. Suppose $\mu = \sum_{i=1}^{k} a_i \delta_{x_i}$, and $\nu = \sum_{j=1}^{l} b_i \delta_{y_j}$ are discrete measures with the same total mass. Let G be a graph with vertices $V(G)$ and straight edges $E(G)$ with a weight function $w : E(G) \to [0, \infty)$. We denote the in-degree at vertex v by

$$d_v := \sum_{u:(u,v)\in E(G)} w((u,v)) - \sum_{u:(v,u)\in E(G)} w((v,u)).$$

G is required to satisfy the mass preserving conditions that i) $d_{x_i} = -a_i$ for $i = 1, ..., k$, ii) $d_{y_j} = b_j$ for $j = 1, ..., l$, and iii) $d_v = 0$ for $v \in V(G) \setminus \{x_1, ..., x_k, y_1, ..., y_l\}$. One then seeks to minimize

$$BT^a(G) := \sum_{e\in E(G)} c(w(e))l(e)$$

where $l(e)$ denotes the length of edge e, and c is the concave cost $c(m) = m^a$. This formulation can be extended to the non-discrete measures, and is sometimes called the *flux-based* formulation, introduced by Xia in [68].

The problem in the general setting can be formulated in the following way, in terms of subsets Σ of $\mathbb{R}^d$.

$$BT^a(\Sigma) := \inf_{\substack{F:\Sigma\to\mathbb{R}^d\setminus\{0\} \\ \mathcal{F}=F\mathcal{H}^1|_\Sigma \\ \text{div } \mathcal{F}=\mu-\nu}} \int_\Sigma c(|F|)^a d\mathcal{H}^1 \qquad \text{with } c^a(m) = m^a \tag{4.1.3}$$

where $a \in (0, 1)$ is a parameter indicating the discount for transporting mass in bulk. The measure $\mathcal{F}$ describes the mass flux treating μ as a source and ν as a sink, and F is its density with respect to the one-dimensional Hausdorff measure restricted to Σ. This formulation (4.1.3) was provided in [10], where it was shown to be equivalent to original formulations [48, 68]. In [48], the authors take an approach based on patterns that describe the position of each particle at every time, in contrast to the approach of Xia in [68], where only the flux of the particles is described at every point (as above). Equivalence between the earlier formulations has also been studied in [46, 47].

We note that in the branched transport problem, any candidate network will contain the supports of the given measures, and therefore in the case of absolutely continuous measures, the minimizing set Σ will necessarily be of infinite length. This is a key difference from the urban planning problem that we consider next.

Urban planning

Here we consider a model for urban planning introduced in [9]. There, Brancolini and Buttazzo propose a problem for finding networks with low total transportation cost between given measures μ and ν. In contrast to the branched transport formulation (4.1.3), the networks considered in this approach do not cover all of the transportation. Instead, the cost of transportation is simply discounted when done over the network.

Let

$$\mathcal{G}_{x,y} := \{\Gamma \text{ closed and connected: } \{x, y\} \subseteq \Gamma \subseteq \mathbb{R}^d\}$$

denote the set of paths containing x and y. One may take $\mathcal{H}^1(\Gamma \cap \Sigma^C) + \alpha\mathcal{H}^1(\Gamma \cap \Sigma)$ as the cost associated to transport path Γ, where $\alpha \in [0, 1)$ represents the cost per distance when traveling along the network. The transportation cost from x to y can then be defined as

$$c_\Sigma(x, y) = \inf\{\mathcal{H}^1(\Gamma \cap \Sigma^C) + \alpha\mathcal{H}^1(\Gamma \cap \Sigma) : \Gamma \in \mathcal{G}_{x,y}\}. \tag{4.1.4}$$

Given a transportation plan $\pi \in \Pi$, the total transportation cost is then

$$TC_\pi(\Sigma) := \int_{\mathbb{R}^d \times \mathbb{R}^d} c_\Sigma(x, y) d\pi(x, y),$$

and the optimal network problem is then find Σ closed and connected that minimizes

$$ONT_{\mu,\nu}(\Sigma) := \inf_{\pi \in \Pi(\mu,\nu)} TC_\pi(\Sigma) + \lambda\mathcal{H}^1(\Sigma). \tag{4.1.5}$$

The second term can be seen as a penalty that represents the cost of building the network. Without it, or another regularization term, the problem would be ill-posed as the total transportation cost could be made arbitrarily small. In [9], Brancolin and Buttazzo prove existence of minimizers to the problem, allowing for costs that depend more generally on $\mathcal{H}^1(\Gamma \cap \Sigma)$, $\mathcal{H}^1(\Gamma \cap \Sigma^C)$, and $\mathcal{H}^1(\Sigma)$. We will focus our attention on the cost (4.1.4) unless otherwise indicated.

Though seemingly different from the branched transport problem, this urban planning problem can be cast in a branched transport setting through both pattern and flux-based frameworks in [10]. The difference from (4.1.3) is that the reformulated (4.1.5) would not be in terms of the cost $c^a(m) = m^a$, but the cost $c^{\lambda,\alpha}(m) = \min(m, \alpha m + \lambda)$ instead. The difference in cost functions here is crucial, and as the latter cost is not strictly subadditive, it does not promote branching to same effect.

The behavior of minimizers to the urban planning problem has been studied in some special cases, in particular when $\alpha = 0$ and when the length penalty is replaced by a length constraint. That special case closely resembles the constrained average-distance problem (1.1.1), the main difference being that not all mass is required to travel to the network. Like in the average-distance problem, it has been shown that minimizers in this special case will not have loops when the measures μ, ν are absolutely continuous. For this property and an Ahlfors regularity result we refer the reader to [13] (Theorem 3.8).

4.2 Optimal networks for selective-transport

We introduce a new set of models for selective-transportation. The main difference in the models we propose is that they assume fully specified, or otherwise constrained, transportation plans. In other words, the problems do not seek to merely match given distributions, but rather aim to minimize total transportation costs that take into account the agents' potentially different needs for their distribution of destinations.

4.2.1 Specified transport plan

Consider the setting of (4.1.5), but with a pre-specified transportation plan π whose first marginal μ corresponds to a distribution of agents. The problem of interest is then to minimize the functional

$$ONT_\pi(\Sigma) := TC_\pi(\Sigma) + \lambda \mathcal{H}^1(\Sigma) \tag{4.2.1}$$

over Σ closed and connected.

The second marginal of π can be seen as the distribution of destinations, aggregated across all agents. Taking the second marginal and the first marginal μ as given, the "only" difference from the urban planning problem (4.1.5) is that here there is no infimum over transportation plans, as one is instead specified. We note that this subtle change in the formulation does not only affect the problem interpretation and modeling aspects, but can also drastically affect the behavior of solutions and methods to compute them.

In the special case that $\pi = \mu \otimes \mu$, (ONT) represents a "universal" transport problem in which every agent seeks to visit every other agent.

4.2.2 Dynamic transport plan

Let us know consider an slightly different setting. Suppose that instead of the target distributions of all agents being pre-specified, they change depending on the available network, and that each agent may have a need of only visiting a certain fraction β of the total population, which would be selected in a cost-minimizing fashion. One may take the following approach to model this setting. Let

$$\overline{B}_\Sigma(x, r) := \{y : y \in \mathbb{R}^d, c_\Sigma(x, y) \leq r\}$$

denote the points within cost r of x, and let

$$r^\beta_{x,\Sigma} := \min\{r > 0 : \mu(\overline{B}_\Sigma(x, r)) \geq \beta\}$$

denote the minimum cost for which x is within reach of fraction β of the population. [The minimum above does exist since $\mu(\overline{B}_\Sigma(x, \cdot)$ is non-decreasing and upper semi-continuous,

and therefore the set over which the minimum is taken is compact.] The total transportation cost would then correspond to

$$\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} c_\Sigma(x,y)\mathbb{1}_{\overline{B}_\Sigma\left(x,r^\beta_{x,\Sigma}\right)}(y)d\mu(y)d\mu(x). \tag{4.2.2}$$

Note that if μ is absolutely continuous, then $r^\beta_{x,\Sigma}$ is such that $\mu(\overline{B}_\Sigma(x,r^\beta_{x,\Sigma})) = \beta$. However, if μ has atoms, then it may happen that $\mu(\overline{B}_c(x,r^\beta_{x,\Sigma})) > \beta$. To avoid difficulties with the lower semi-continuity of the transportation cost (4.2.2) and subsequent existence of minimizers, we alternatively set up the problem in the framework of transportation plans to allow for splitting of mass.

Given a distribution of agents μ, let $\Pi_\beta(\mu)$ denote the set of Borel measures on $\mathbb{R}^d\times\mathbb{R}^d$ such that $\pi \leq \mu\otimes\mu$ (on all Borel sets) and $\pi(A\times\mathbb{R}^d) = \beta\mu(A)$. We then set the transportation cost

$$TC^\beta_\mu(\Sigma) := \inf_{\pi\in\Pi_\beta(\mu)} TC_\pi(\Sigma), \tag{4.2.3}$$

and the optimal network for these dynamic transport plans then minimizes

$$ONDT^\beta_\mu(\Sigma) := TC^\beta_\mu(\Sigma) + \lambda\mathcal{H}^1(\Sigma). \tag{4.2.4}$$

Let us briefly show the equivalence of (4.2.2) and (4.2.3) when μ is absolutely continuous. We later prove (Proposition 4.4.5) that the infimum in (4.2.3) can be replaced with minimum, so let $\pi_\Sigma\in\Pi_\beta(\mu)$ denote the minimizing plan for a given network Σ. Since π_Σ is absolutely continuous with respect to $\mu\otimes\mu$ we have that

$$TC^\beta_\mu(\Sigma) = TC_{\pi_\Sigma}(\Sigma) = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} c_\Sigma(x,y)\frac{d\pi_\Sigma}{d(\mu\otimes\mu)}(x,y)d\mu(y)d\mu(x),$$

where $\frac{d\pi_\Sigma}{d(\mu\otimes\mu)}(x,y)\leq 1$. Moreover, for every $x\in\mathbb{R}^d$ and every $\epsilon>0$

$$\int_{B(x,\epsilon)}\int_{\mathbb{R}^d}\frac{d\pi_\Sigma}{d(\mu\otimes\mu)}(x,y)d\mu(y)d\mu(x) = \pi(B(x,\epsilon)\times\mathbb{R}^d) = \beta\mu(B(x,\epsilon)) = \int_{B(x,\epsilon)}\beta d\mu(x)$$

which implies that

$$\int_{\mathbb{R}^d}\frac{d\pi_\Sigma}{d(\mu\otimes\mu)}(x,y)d\mu(y) = \beta$$

for $\mu-$a.e. $x\in\mathbb{R}^d$. Combined with the fact that $\frac{d\pi_\Sigma}{d(\mu\otimes\mu)}(x,y)\leq 1$, this implies the minimizing plan π_Σ must satisfy

$$\frac{d\pi_\Sigma}{d(\mu\otimes\mu)}(x,y) = \mathbb{1}_{B_\Sigma\left(x,r^\beta_{x,\Sigma}\right)}(y)$$

$\mu\otimes\mu-$a.e., which is what we wanted to show. Lastly, we note that taking $\beta=1$ recovers the specified transportation functional $ONT_{\mu\otimes\mu}$.

4.3 Optimal settlement design

We consider an extended setting in which the population distribution given by μ is not specified. Namely, we seek both a measure and network that minimize a quantity measuring the overall burden or cost of the population. In allowing for flexibility in the settlement of the population, we introduce a term that measures aversion to congestion, or alternatively a desire for personal space. That is, we assume that as agents have a need of visiting others in the population, and prefer to pay minimal cost in doing so, they also prefer to reside in a low-density region. We note that a kind of balancing term is needed, otherwise the optimal settlement would collapse to a single point.

Let us consider an absolutely continuous measure μ with density ρ. We then seek (μ, Σ) that minimize

$$PC(\mu, \Sigma) := TC_{\mu\otimes\mu}(\Sigma) + \lambda_1 \mathcal{H}^1(\Sigma) + \lambda_2 \int_{\mathbb{R}^d} \rho^p dx \tag{4.3.1}$$

where $p > 1$ indicates a magnitude of aversion to higher-density regions. We note that this preference can also be interpreted to represent the amount of real estate that each agent has – they have less in higher density regions.

Note that when μ is discrete, one can define the problem using a smoothened out density via a convolution with a smoothly decaying function, also know as a kernel density estimate. Let $\mu = \sum_{i=1}^n \delta_{x_i} m_i$. Given a nonnegative smooth decreasing function $K : \mathbb{R} \to \mathbb{R}$ with $\int K(t)dt = 1$, a kernel density estimator for μ is

$$\hat{\rho}_h(x) = \int_{\mathbb{R}^d} K_h(|x - \tau|) d\mu(\tau) = \sum_{i=1}^n K_h(|x - x_i|) m_i$$

where $K_h(t) = \frac{1}{h} K(\frac{t}{h})$, and $h > 0$ is a smoothing parameter known as the bandwidth of the kernel K. In this setting we seek (μ, Σ) that minimize

$$PC_{n,h}(\mu, \Sigma) := TC_{\mu\otimes\mu}(\Sigma) + \lambda_1 \mathcal{H}^1(\Sigma) + \lambda_2 \int_{\mathbb{R}^d} \hat{\rho}_h^p dx \tag{4.3.2}$$

where μ is a discrete measure with at most n atoms.

4.4 Existence of minimizers

In order to prove the existence of minimizers for the problems above, we first show that the cost function we consider is lower semi-continuous with respect to Hausdorff convergence in Σ. We note that a proof for more general cost functions is given in [9], and here we provide an alternate, more direct proof for the cost function (4.1.4).

Our approach is to express the cost function in terms of geodesics on $\mathbb{R}^d$ with respect to a Riemannian metric g. We define g as

$$g_x^\Sigma(u, v) = \Big(1 - (1 - \alpha)\mathbb{1}_\Sigma(x)\Big)^2 (u \cdot v)$$

for $x, u, v \in \mathbb{R}^d$. Let

$$\mathcal{C} := \left\{ \gamma : [0, a] \to \mathbb{R}^d \,\middle|\, a \geq 0, \gamma \text{ is Lipschitz with } |\gamma'| \leq 1, \mathcal{L}^1 - a.e. \right\}$$

denote the space of curves that we will consider, and let

$$\mathcal{C}_{x,y} := \left\{ \gamma \in \mathcal{C} \,\middle|\, x, y \in \text{range}(\gamma) \right\}$$

denote the curves containing $x, y \in \mathbb{R}^d$.

Before proceeding further, let us also define the following metric on $\mathcal{C}$. Given $\gamma_1, \gamma_2 \in \mathcal{C}$ with respective domains $[0, a_1], [0, a_2]$ and $a_1 < a_2$, we define the extension of γ_1 to $[0, a_2]$ as

$$\tilde{\gamma}_1(t) = \begin{cases} \gamma_1(t) & \text{if } t \in [0, a_1] \\ \gamma_1(a_1) & \text{if } t \in (a_1, a_2]. \end{cases}$$

We then let

$$d_{\mathcal{C}}(\gamma_1, \gamma_2) = \max_{t \in [0, a_2]} |\tilde{\gamma}_1(t) - \gamma_2(t)|.$$

For $\gamma \in \mathcal{C}$ with domain $[0, a]$ we let

$$\delta_\Sigma(\gamma) := \int_0^a g^\Sigma_{\gamma(s)}(\gamma'(s), \gamma'(s))^{\frac{1}{2}} ds$$

denote the cost of curve γ. We can then express the cost (4.1.4) as

$$c_\Sigma(x, y) = \inf_{\gamma \in \mathcal{C}_{x,y}} \delta_\Sigma(\gamma).$$

In order to prove the lower semi-continuity of $c_{(\cdot)}(x, y)$, we first show the existence of geodesics. To do that, we need the following lower semi-continuity property of g.

Proposition 4.4.1. *Assume Σ and Σ_n are closed sets such that $\Sigma_n \xrightarrow{H} \Sigma$, and $\gamma, \gamma_n \in \mathcal{C}$ are arc-length parametrized with common domain $[0, a]$ and such that $\gamma_n \to \gamma$ with respect to $d_{\mathcal{C}}$. Then*

$$\liminf_{n \to \infty} g^{\Sigma_n}_{\gamma_n(s)}(\gamma_n'(s), \gamma_n'(s)) \geq g^\Sigma_{\gamma(s)}(\gamma'(s), \gamma'(s))$$

for all $s \in [0, a]$.

Proof. Fix $s \in [0, a]$. Note that $|\gamma'(s)| \leq |\gamma_n'(s)| = 1$ for all $n \in \mathbb{N}$. Thus it suffices to check that

$$\mathbb{1}_\Sigma(\gamma(s)) \geq \liminf_{n \to \infty} \mathbb{1}_{\Sigma_n}(\gamma_n(s)).$$

This is trivial if $\gamma(s) \in \Sigma$, so suppose that $\gamma(s) \notin \Sigma$. We want to show that there exists $n_0 \in \mathbb{N}$ such that for all $n > n_0$ $\gamma_n(s) \notin \Sigma_n$. Assume for contradiction that there exists a subsequence for which $\gamma_n(s) \in \Sigma_n$ for all n. Then

$$d(\gamma(s), \Sigma) \leq d(\gamma(s), \gamma_n(s)) + d_H(\Sigma_n, \Sigma) \to 0.$$

Since Σ is closed, this implies $\gamma(s) \in \Sigma$ which is a contradiction. □

Proposition 4.4.2. *Given closed $\Sigma \subseteq \mathbb{R}^d$ and $x, y \in \mathbb{R}^d$, there exists $\gamma \in \mathcal{C}_{x,y}$ such that $c_\Sigma(x, y) = \delta_\Sigma(\gamma)$.*

Proof. Let $\{\gamma_n\} \subseteq \mathcal{C}_{x,y}$ be a minimizing sequence for $c_\Sigma(x, y)$. We may assume that all γ_n are arc-length parametrized (re-parametrizing if necessary). We may further assume that the lengths of all γ_n are uniformly bounded by a constant L (depending on α). We extend the domain of all γ_n to $[0, L]$. Since the sequence $\{\gamma_n\}$ is uniformly bounded and equicontinuous, by the Arzelà-Ascoli theorem we may find a subsequence and a curve $\gamma : [0, L] \to \mathbb{R}^d$ such that $\gamma_n \to \gamma$ uniformly with respect to $d_{\mathcal{C}}$. Note that γ will be 1-Lipschitz, so $\gamma \in \mathcal{C}_{x,y}$. Moreover, by Proposition 4.4.1 and Fatou's lemma we have that

$$\delta_\Sigma(\gamma) \leq \liminf_{n\to\infty} \delta_\Sigma(\gamma_n),$$

which shows $c_\Sigma(x, y) = \delta_\Sigma(\gamma)$. □

Proposition 4.4.3 (Lower semi-continuity of cost)**.** *If $(x_n, y_n) \to (x, y) \in \mathbb{R}^d \times \mathbb{R}^d$, and $\{\Sigma_n\}_{n\in\mathbb{N}}$ is a sequence of closed subsets of $\mathbb{R}^d$ that converge to Σ in Hausdorff distance, then*

$$c_\Sigma(x, y) \leq \liminf_{n\to\infty} c_{\Sigma_n}(x_n, y_n).$$

Proof. Since the cost functions $\{c_{\Sigma_n}\}$ are 1-Lipschitz on $\mathbb{R}^d \times \mathbb{R}^d$, it suffices to prove that

$$c_\Sigma(x, y) \leq \liminf_{n\to\infty} c_{\Sigma_n}(x, y).$$

Let $\{\gamma_n\} \subseteq \mathcal{C}_{x,y}$ be such $\delta_{\Sigma_n}(\gamma_n) = c_{\Sigma_n}(x, y)$ for all n, and $\gamma \in \mathcal{C}_{x,y}$ such that $\delta_\Sigma(\gamma) = c_\Sigma(x, y)$. We may again assume the lengths of γ_n are uniformly bounded, and arc-length re-parametrize and extend the curves to a common domain $[0, L]$. By Arzelà-Ascoli we can find a subsequence and $\tilde{\gamma} \in \mathcal{C}_{x,y}$ with domain $[0, L]$ and such that $\gamma_n \to \tilde{\gamma}$ in $d_{\mathcal{C}}$. By Proposition 4.4.1 and Fatou's lemma we have

$$\begin{aligned}\liminf_{n\to\infty} \delta_{\Sigma_n}(\gamma_n) &= \liminf_{n\to\infty} \int_0^L g^{\Sigma_n}_{\gamma_n(s)}(\gamma_n'(s), \gamma_n'(s))^{\frac{1}{2}} ds \\ &\geq \int_0^L g^{\Sigma}_{\tilde{\gamma}(s)}(\tilde{\gamma}'(s), \tilde{\gamma}'(s))^{\frac{1}{2}} ds \geq \delta_\Sigma(x, y))\end{aligned}$$

□

4.4.1 Specified transport plan

We can now directly show existence of minimizers for the specified transport problem (ONT). We will let Π_C denote the set of compactly supported finite Borel measures on $\mathbb{R}^d$, and let $\mathcal{A}$ denote the set of compact and connected subsets of $\mathbb{R}^d$ with finite $\mathcal{H}^1$ measure.

Proposition 4.4.4. *Given $\lambda > 0$ and $\pi \in \Pi_C$, the functional* (ONT) *has a minimizer over the admissible set $\mathcal{A}$.*

Proof. Let $\{\Sigma_n\}$ be a minimizing sequence (of compact connected subsets of $\mathbb{R}^d$). Since π is compactly supported, we may assume all Σ_n are subsets of a compact set. By Blaschke's theorem, we find a subsequence (which we take to be the whole sequence) and a compact connected set Σ such that $\Sigma_n \xrightarrow{H} \Sigma$. By Gołąb's theorem we have that

$$\mathcal{H}^1(\Sigma) \leq \liminf_{n\to\infty} \mathcal{H}^1(\Sigma_n).$$

The lower semi-continuity of $c_{(\cdot)}(x, y)$ (Proposition 4.4.3) combined with Fatou's lemma gives

$$TC_\pi(\Sigma) \leq \liminf_{n\to\infty} TC_\pi(\Sigma_n),$$

and therefore Σ is a minimizer of ONT_π. □

4.4.2 Dynamic transport plan

We first show that for any given network, there is an optimal transportation plan satisfying the transport requirements.

Proposition 4.4.5. *For any probability measure μ with compact support, $\Sigma \subseteq \mathbb{R}^d$ compact, and $\beta > 0$, there exists $\pi \in \Pi_\beta(\mu)$ such that*

$$TC_\pi(\Sigma) = \inf_{\tilde{\pi}\in\Pi_\beta(\mu)} TC_{\tilde{\pi}}(\Sigma).$$

Proof. This follows from the fact that, in the sense of weak convergence of measures, $\Pi_\beta(\mu)$ is closed and $TC_{(\cdot)}(\Sigma)$ is lower semi-continuous. Indeed, let $\{\pi_n\}$ be a minimizing sequence for $\inf_{\tilde{\pi}\in\Pi_\beta(\mu)} TC_{\tilde{\pi}}(\Sigma)$. Since μ is compactly supported, the sequence $\{\pi_n\}$ is tight and by Prokhorov's theorem we may find a subsequence (which we take to be the whole sequence) that weakly converges to a Borel measure π.

To check that $\pi \in \Pi_\beta(\mu)$, note that for all continuous and bounded functions ϕ on $\mathbb{R}^d$

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} \phi(x)d\pi(x, y) = \lim_{n\to\infty} \int_{\mathbb{R}^d\times\mathbb{R}^d} \phi(x)d\pi_n(x, y) = \int_{\mathbb{R}^d} \phi(x)\beta\mu(x)$$

where we have used the fact that ϕ is continuous and bounded on $\mathbb{R}^d \times \mathbb{R}^d$. Since ϕ is arbitrary, this implies that $\pi(A \times \mathbb{R}^d) = \beta\mu(A)$ for any Borel set A. Similarly, we have that for all continuous and bounded functions ϕ on $\mathbb{R}^d \times \mathbb{R}^d$

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} \phi(x, y)d\pi(x, y) = \lim_{n\to\infty} \int_{\mathbb{R}^d\times\mathbb{R}^d} \phi(x, y)d\pi_n(x, y) \leq \int_{\mathbb{R}^d\times\mathbb{R}^d} \phi(x, y)d(\mu \otimes \mu)(x, y)$$

which shows that $\pi \leq \mu \otimes \mu$.

Finally, note that $c_\Sigma(x, y)$ is continuous and (therefore also) bounded on the support of $\mu \otimes \mu$. Thus,

$$TC_\pi(\Sigma) = \int_{\mathbb{R}^d\times\mathbb{R}^d} c_\Sigma(x, y)d\pi(x, y) = \lim_{n\to\infty} \int_{\mathbb{R}^d\times\mathbb{R}^d} c_\Sigma(x, y)d\pi_n(x, y) = \lim_{n\to\infty} TC_{\pi_n}(\Sigma)$$

□

The existence of minimizers for (4.2.4) now follows from the following generalized Fatou lemma (Proposition 2.3.6). (Alternatively, one can use the fact that the Wasserstein distance metrizes weak convergence of measures together with the Lipschtiz continuity of c_Σ.)

Proposition 4.4.6. *For any probability measure μ with compact support, $\beta \in (0,1)$, and $\lambda > 0$ the problem $ONDT_\mu^\beta$ has a minimizer over the admissible set $\mathcal{A}$.*

Proof. Let Σ_n be a minimizing sequence. Note that since μ is compactly supported, so is the sequence $\{\Sigma_n\}$. By Blaschke's theorem, there is a compact $\Sigma \subseteq \mathbb{R}^d$ such that $\Sigma_n \xrightarrow{H} \Sigma$ up to a subsequence. Let $\pi_n, \pi \in \Pi_\beta(\mu)$ denote the plans that minimize $TC_{(\cdot)}(\Sigma_n)$ and $TC_{(\cdot)}(\Sigma)$ respectively. Since the sequence $\{\pi_n\}$ is tight, by Prokhorov's theorem there exists $\tilde{\pi} \in \Pi_\beta(\mu)$ such that π_n weakly converges to $\tilde{\pi}$ up to a (further) subsequence. By the lower-semicontinuity of the cost functions (Proposition 4.4.3) and Fatou lemma (Proposition 2.3.6) we have that

$$\liminf_{n\to\infty} TC_{\pi_n}(\Sigma_n) \geq TC_{\tilde{\pi}}(\Sigma) \geq TC_\pi(\Sigma).$$

This together with Gołąb's theorem shows that Σ is a minimizer of $ONDT_\mu^\beta$.

□

4.4.3 Optimal settlement

In order to prove existence of minimizers of the optimal settlement model (4.3.1), we first prove the following compactness result for a minimizing sequence. In the following we let $\mathcal{M}_1$ denote the set of finite absolutely continuous probability measures $\mathbb{R}^d$.

Proposition 4.4.7. *Let $\{(\mu_n, \Sigma_n)\}_{n=1}^\infty$ be a minimizing sequence of* (4.3.1) *over the admissible set $\mathcal{M}_1 \times \mathcal{A}$. Then there exist $R > 0$ and $n_0 \in \mathbb{N}$ such that for $\forall n > n_0$, $\operatorname{supp}(\mu_n) \subseteq B(0,R)$ and $\Sigma_n \subseteq B(0,R)$.*

Proof. For any μ, Σ, it suffices to construct a lower energy configuration $(\tilde{\mu}, \tilde{\Sigma})$ contained in $B(0,R)$.

Let $R > r > 0$, and let $m_1 = \mu(B(0,R) \setminus B(0,r))$, $m_2 = \mu(B(0,R)^C)$, $\epsilon = \mu(B(0,r)^C) = m_1 + m_2$. For properly chosen r, R we obtain a competitor to (μ, Σ) with $\tilde{\mu} = \mu|_{B(0,R)} + \frac{m_2}{|B(0,r)|}\mathcal{L}^d|_{B(0,r)}$, and $\tilde{\Sigma}$ as the projection of Σ onto $B(0,R)$. Note that projecting onto a convex set decreases length, so $\mathcal{H}^1(\Sigma) \geq \mathcal{H}^1(\tilde{\Sigma})$.

We estimate the change in energy due to using $\tilde{\mu}$ in place of μ. The change consists of moving the mass m_2 from outside $B(0,R)$ to inside $B(0,r)$. The change in the transportation cost is bounded from above by

$$m_1 m_2 (R-r) - m_2(1-\epsilon)\alpha(R-r) = m_2(R-r)\left(m_1 - \alpha(1-\epsilon)\right).$$

On the other hand, the change in the density term is given by

$$\int_{B(0,R)} \left(\rho + \frac{m_2}{|B(0,r)|}\mathbb{1}_{B(0,r)}\right)^p - \rho^p d\mathcal{L}^d$$

$$\leq \frac{m_2}{|B(0,r)|}\int_{B(0,r)} p\left(\rho + \frac{m_2}{|B(0,r)|}\right)^{p-1} d\mathcal{L}^d = c_r m_2$$

where the inequality follows from Taylor's theorem, and c_r is a quantity that decreases as r increases. Therefore the total change in energy is bounded above by

$$m_2\left(c_r - (R-r)(\alpha(1-\epsilon) - m_1)\right).$$

Taking both r and R large enough, we can make the change in energy negative, which concludes the proof. □

The existence of minimizers of (4.3.1) now follows.

Proposition 4.4.8. *Given $\lambda_1, \lambda_2 > 0$ and $p > 1$, the functional* (4.3.1) *has a minimizer over the admissible set $\mathcal{M}_1 \times \mathcal{A}$.*

Let $\{(\mu_n, \Sigma_n)\}_{n=1}^{\infty}$ be a minimizing sequence, and let ρ_n denote the density of μ_n. By the previous proposition, we may assume that $\mathrm{supp}(\mu_n) \subseteq B(0,R)$ for all $n \in \mathbb{N}$ for some $R > 0$. By Prokhorov's theorem, there exists a subsequence of $\{\mu_n\}_{n=1}^{\infty}$, which we take to be the whole sequence, and μ (with density ρ) such that $\mu_n \overset{*}{\rightharpoonup} \mu$ (in the sense of weak convergence of measures). Furthermore, by Blaschke's theorem there exists a compact set Σ such that $\Sigma_n \xrightarrow{H} \Sigma$ up to a subsequence (which we again take to be the whole sequence). The lower semi-continuity of $\mathcal{H}^1(\cdot)$ follows by Gołąb's theorem. In addition, the lower semi-continuity of the density term

$$\liminf_{n\to\infty} \int_{\mathbb{R}^d} \rho_n^p dx \geq \int_{\mathbb{R}^d} \rho^p dx$$

holds from Theorem 2.3.7. It remains to check that the total transportation cost is lower semi-continuous with respect to Hausdorff convergence in Σ and weak convergence of μ. We have that

$$|TC_{\mu_n\otimes\mu_n}(\Sigma_n) - TC_{\mu\otimes\mu}(\Sigma_n)|$$

$$= \left|\int_{\mathbb{R}^d\times\mathbb{R}^d} c_{\Sigma_n}(x,y)d(\mu_n\otimes\mu_n)(x,y) - \int_{\mathbb{R}^d\times\mathbb{R}^d} c_{\Sigma_n}(x,y)d(\mu\otimes\mu)(x,y)\right|$$

$$\leq \sup_{f:Lip(f)\leq 1}\int_{\mathbb{R}^d\times\mathbb{R}^d} f(d(\mu_n\otimes\mu_n) - d(\mu\otimes\mu))(x,y)$$

since $\{c_{\Sigma_n}\}$ are Lipschitz continuous with Lipschitz constant 1. By weak convergence of measures we then have that the quantity vanishes in the limit. On the other hand, we have that

$$\liminf_{n\to\infty} TC_{\mu\otimes\mu}(\Sigma_n) \geq TC_{\mu\otimes\mu}(\Sigma)$$

from Proposition 4.4.3. Thus we can conclude that

$$\liminf_{n\to\infty} \left[TC_{\mu_n\otimes\mu_n}(\Sigma_n) - TC_{\mu\otimes\mu}(\Sigma_n) + TC_{\mu\otimes\mu}(\Sigma_n) - TC_{\mu\otimes\mu}(\Sigma)\right] \geq 0$$

and therefore (μ, Σ) is a minimizer of PC.

4.4.4 $\Gamma-$convergence

Here we briefly comment that the $\Gamma-$convergence of the specified and dynamic transport functionals with respect to weak convergence of the measures is immediate from the lower semicontinuity of the cost function.

Lemma 4.4.9. *Given $\beta \leq 1$, $\mu_n \overset{\star}{\rightharpoonup} \mu$ in $\mathcal{P}_R$, then $ONDT^\beta_{\mu_n} \overset{\Gamma}{\to} ONDT^\beta_\mu$ with respect to Hausdorff convergence of sets in $\mathcal{A}$. Likewise, if $\pi_n \overset{\star}{\rightharpoonup} \pi$ in $\mathcal{P}_R$, then $ONT_{\pi_n} \overset{\Gamma}{\to} ONT_\pi$ in the same sense.*

Proof. By definition of $\Gamma-$convergence we need to check the following two properties

- Lower semi-continuity: If $\mu_n \overset{\star}{\rightharpoonup} \mu$ and $\Sigma_n \to \Sigma$ in Hausdorff distance, then

$$\liminf_{n\to\infty} ONDT^\beta_{\mu_n}(\Sigma_n) \geq ONDT^\beta_\mu(\Sigma).$$

- Construction: If $\mu_n \overset{\star}{\rightharpoonup} \mu$ then for any $\Sigma \in \mathcal{A}$ there exists a sequence $\Sigma_n \in \mathcal{A}$ such that $\Sigma_n \to \Sigma$ in Hausdorff distance and

$$\lim_{n\to\infty} ONDT^\beta_{\mu_n}(\Sigma_n) = ONDT^\beta_\mu(\Sigma).$$

The lower semi-continuity property follows directly from the lower semi-continuity of the cost (Proposition 4.4.3) and the generalized Fatou lemma (Proposition 2.3.6). The construction property follows by taking $\Sigma_n = \Sigma$ and the definition of weak convergence of measure. The $\Gamma-$convergence of ONT is analagous. □

From $\Gamma-$convergence we have the following property regarding convergence of minimizers.

Corollary 4.4.10. *Given $\beta \leq 1$, $\mu_n \overset{\star}{\rightharpoonup} \mu$, and that Σ_n is a minimizer of $ONDT^\beta_{\mu_n}$, it follows that along a subsequence Σ_n converge in Hausdorff distance to a minimizer of $ONDT^\beta_\mu$. The analagous statement also holds for ONT.*

Proof. By Blaschke's Theorem, we can find a subsequence of Σ_n that converges to some $\Sigma \in \mathcal{A}$. The $\Gamma-$convergence will imply that Σ is a minimizer of $ONDT^\beta_\mu$. From the construction property, for any $\tilde{\Sigma} \in \mathcal{A}$ we can find a sequence $\tilde{\Sigma}_n$ converging to $\tilde{\Sigma}$ in Hausdorff distance, and such that

$$\lim_{n\to\infty} ONDT^\beta_{\mu_n}(\tilde{\Sigma}_n) = ONDT^\beta_\mu(\tilde{\Sigma}).$$

By the lower semi-continuity property and the fact that Σ_n is a minimizer of TNT_{μ_n} we have

$$ONDT_{\mu}^{\beta}(\Sigma) \leq \liminf_{n\to\infty} ONDT_{\mu_n}^{\beta}(\Sigma_n) \leq \liminf_{n\to\infty} ONDT_{\mu_n}^{\beta}(\tilde{\Sigma}_n) = ONDT_{\mu}^{\beta}(\tilde{\Sigma}),$$

which shows that Σ is a minimizer of $ONDT_{\mu}^{\beta}$. □

4.5 Numerical algorithm

In this section we outline a numerical algorithm for minimizing the optimal network transportation functional

$$TC_{\pi}(\Sigma) + \lambda \mathcal{H}^1(\Sigma). \tag{ONT}$$

We focus on the case $\pi = \mu \otimes \mu$, and remark that extension to the case of general π can be done without much difficulty. We start by introducing the discrete functional.

4.5.1 Discrete functional

Let $X = \{x_1, x_2, ..., x_n\}$ represent the locations of the agents, with masses $m_1, m_2, ..., m_n$ so that $\mu = \sum_{i=1}^{n} m_i \delta(x_i)$. We consider the transportation plan $\pi = \mu \otimes \mu$, and we restrict our attention to the case that Σ is piecewise linear. In particular, we assume Σ can be represented by an undirected geometric graph (Y, E_Y), with vertices $Y = \{y_1, y_2, ..., y_m\} \subseteq \mathbb{R}^d$, and E_Y a set of edges. Then the functional (ONT) in this discrete setting is

$$\sum_{i=1}^{n}\sum_{j=1}^{n} m_i m_j c_{\Sigma}(x_i, x_j) + \lambda \sum_{\{a,b\}\in E_Y} |a-b|, \tag{4.5.1}$$

which we minimize over $\Sigma = (Y, E_Y)$. We make a slight simplification of our definition of c_{Σ} in the continuous setting, and no longer consider the graph $\Sigma = (Y, E_Y)$ as a set, in which paths can enter and exit at any point. Instead, we will restrict paths to enter and exit Σ only from the vertices Y. Note that as the graph (Y, E_Y) becomes more refined, the difference between the two cost functions vanishes.

We now define a weighted graph from which we can compute the simplified cost function c_{Σ}. Let $G = (V, E, W)$ be an undirected weighted graph on $V = X \cup Y$, where we consider the complete edge set w, and define the weights $w_{a,b}$ as

$$w_{a,b} = \begin{cases} \alpha|a-b|, & \text{if } \{a,b\} \in E_Y \\ |a-b|, & \text{otherwise.} \end{cases}$$

Next for $x_i, x_j \in X$, we define $geod(x_i, x_j)$ as the set of edges contained in a (distinct) shortest path between x_i and x_j on G. We have that

$$c_\Sigma(x_i, x_j) = \sum_{\{a,b\} \in geod(x_i, x_j)} w_{a,b}.$$

We may denote the total mass through $\{a, b\} \in E$ as

$$m_{a,b} = \sum_{i=1}^{n} \sum_{j=1}^{n} m_i m_j \mathbb{1}\Big(\{a, b\} \in geod(x_i, x_j)\Big)$$

and its energy contribution with

$$M_{a,b} = m_{a,b} \left(\mathbb{1}\Big(\{a, b\} \notin E_Y\Big) + \left(\alpha + \frac{\lambda}{m_{a,b}}\right) \mathbb{1}\Big(\{a, b\} \in E_Y\Big) \right). \tag{4.5.2}$$

Then we may write (4.5.1) in the simple form

$$\frac{1}{2} \sum_{a \in V} \sum_{b \in V} |a - b| M_{a,b}. \tag{4.5.3}$$

(The $\frac{1}{2}$ is there since we have double counted each edge, and the weights $M_{a,b}$ already incorporate masses from both directions.) Our goal is to minimize (4.5.3) over (Y, E_Y). The first order conditions for minimality over Y are

$$\sum_{i=1}^{n} \frac{y_k - x_i}{|y_k - x_i|} M_{y_k, x_i} + \sum_{l=1}^{m} \frac{y_k - y_l}{|y_k - y_l|} M_{y_k, y_l} = 0 \qquad \forall y_k \in Y.$$

4.5.2 Minimization over Y

We consider a strategy for minimizing over Y, with E_Y fixed. In what follows enumerate $V = \{v_i\}_{i=1}^{m+n}$ where $v_i = y_i$ for $1 \le i \le m$, and $v_i = x_i$ for $m < i \le m + n$. For ease of notation we let $e_{i,j} = \{v_i, v_j\}$ and $M_{i,j} = M_{v_i, v_j}$.

Note that $M_{i,j}$ will be zero for most $i, j \in \{1, 2, ..., m + n\}$. Let $\tilde{E}$ denote the edges $e \not\subseteq X$ for which M_e is nonzero. In addition, let $\sigma : \tilde{E} \to \{1, 2, ., , n_{\tilde{E}}\}$ be an (bijective) enumeration of these edges, where $n_{\tilde{E}} := |\tilde{E}|$. We define the linear operator $D : \mathbb{R}^{m \times d} \to \mathbb{R}^{n_{\tilde{E}}}$ as

$$(Dy)_{\sigma(e_{i,j})} = \begin{cases} y_j - y_i & \text{if } m \ge i > j \ge 1 \\ y_j & \text{if } m + n \ge i > m \ge j \ge 1 \end{cases}$$

In other words, D is simply an incidence matrix for the vertices $Y = \{v_i\}_{i=1}^{m}$ in the graph $(V, \tilde{E})$. We also define the constant vector $c \in \mathbb{R}^{n_{\tilde{E}}}$

$$c_{\sigma(e_{i,j})} = \begin{cases} 0 & \text{if } m \ge i > j \ge 1 \\ -x_{i-m} & \text{if } m + n \ge i > m \ge j \ge 1. \end{cases}$$

Minimizing (4.5.3) over Y is then equivalent to

$$\min_{y,z\,:\,Dy+c=z} \quad \sum_{e=1}^{n_{\tilde{E}}} |z_e| M_{\sigma^{-1}(e)}.$$

We may apply ADMM to the above, and it consists of the following steps

$$y^k = \arg\min_y \frac{\rho}{2}|Dy + c - z^k + b^k|^2 \tag{4.5.4}$$

$$z^{k+1} = \arg\min_z |z|_{M,1,2} + \frac{\rho}{2}|Dy^k + c - z + b^k|^2 \tag{4.5.5}$$

$$b^{k+1} = b^k + Dy^{k+1} + c - z^{k+1} \tag{4.5.6}$$

where $|z|_{M,1,2} = \sum_{e=1}^{n_{\tilde{E}}} |z_e| M_{\sigma^{-1}(e)}$.

We note that the solution to the first step is given by

$$y^k = (D^T D)^{-1} D^T (z^k - b^k - c)$$

where $D^T D$ is the $m \times m$ Laplacian matrix of the vertices Y in $(V, \tilde{E})$ with entries

$$(D^T D)_{i,j} = \begin{cases} deg_{\tilde{E}}(y_i) & \text{if } i = j \\ -1 & \text{if } \{v_i, v_j\} \in \tilde{E} \\ 0 & \text{otherwise.} \end{cases}$$

where $deg_{\tilde{E}}(y_i)$ denotes the degree of y_i in $\tilde{E}$. In general, graph Laplacians are not invertible. However, as $D^T D$ is the Laplacian for only a subset of the vertices, it will be invertible as long as $(V, \tilde{E})$ is a connected graph and one of the rows is strictly diagonally dominant. The latter condition is satisfied as long as there there is an edge e in $\tilde{E}$ with $e \cap X \neq \emptyset$ and $e \cap Y \neq \emptyset$ (i.e. the network is being used by at least one data point), which holds in all non-degenerate cases.

The solution to the second step above is given by soft-thresholding. Letting $v_e^k = (Dy^k)_e + c_e + b_e^k$ for $e \in \{1, 2, ..., n_{\tilde{E}}\}$ we have

$$z_e^k = \begin{cases} v_e^k - \frac{M_{\sigma^{-1}(e)}}{\rho} \frac{v_e^k}{|v_e^k|} & \text{if } |v_e^k| > \frac{M_{\sigma^{-1}(e)}}{\rho} \\ 0 & \text{otherwise.} \end{cases}$$

4.5.3 Computing geodesics

The minimization over the vertices of the network Y requires computation of geodesics, which may change as soon the network's vertices or edges are modified. Recall that the geodesics needed are those between the data points X on the complete weighted graph $(X \cup Y, E, W)$. We therefore compute *all pairs shortest paths* on $(X \cup Y, E, W)$ after

Algorithm 5 All pairs shortest paths for $(X \cup Y, E, W)$ (Floyd-Warshall)

Input: $(m+n) \times (m+n)$ weight matrix W
initialize geodesic distance matrix $d_\Sigma = W$ and $(m+n) \times (m+n)$ matrix $next$ with $next(i,j) = j$ if $i \neq j$, 0 otherwise.
for $k = 1 : m$ **do**
  for $i = 1 : m+n$ **do**
    for $j = 1 : m+n$ **do**
      if $d_\Sigma(i,j) > d_\Sigma(i,k) + d_\Sigma(k,j)$ **then**
        $d_\Sigma(i,j) = d_\Sigma(i,k) + d_\Sigma(k,j)$
        $next(i,j) = next(i,k)$
      end if
    end for
  end for
end for
Output: d_Σ, $next$.

every change to the network graph (Y, E_Y). To do so, we use a slight simplification of the Floyd-Warshall algorithm (see for instance Section 25.2 of [20]) outlined in Algorithm 5.

Considering the initial geodesic pairwise distances W, we know that one only needs to loop through the vertices Y on the network in the outermost loop, since only by going through one of them may the current shortest distance d_Σ be improved. Using the output matrix $next$ one can easily reconstruct the geodesics, via the meaning that $next(i,j) = k$ implies k should be visited from i if on the way to j. The complexity of computing the shortest paths on the graph $(X \cup Y, E, W)$ is hence $\mathcal{O}(m(m+n)^2)$, or simply $\mathcal{O}(mn^2)$ since $m \leq n$. We also note that computing the weights W for the graph requires $\mathcal{O}((m+n)^2 d)$ time, and therefore the total complexity for computing geodesics is $\mathcal{O}((m+d)n^2)$.

4.5.4 Local minimization over E_Y

While the updates over the vertices Y update the geometry of the network, we also want to update the topology of the network through the edges E_Y. We observe the following simple strategy for decreasing the energy (4.5.3) over the edges E_Y if the vertices Y are fixed. In view of (4.5.2), we update E_Y corresponding to the (local) optimality condition

$$\{a,b\} \in E_Y \iff \alpha + \frac{\lambda}{m_{a,b}} < 1 \iff m_{a,b} > \frac{\lambda}{1-\alpha}. \tag{4.5.7}$$

We note that while this update does help achieve lower energy configurations for Σ, relying on it does get us far in finding low energy minima. The problem of course is that the masses $m_{a,b}$ depend on the geodesics, which depend on the network connectivity. In particular, it may happen only after adding (or removing) an edge $\{a,b\}$ and recomputing the geodesics that the edge mass $m_{a,b}$ favors the presence (respectively absence) of

the edge. This creates a significant combinatorial challenge in which several topological configurations might need to be tested in order to obtain a desirable local minimizer.

There are however very simple heuristics that may be of help. One that we consider consists of testing to add edges that have high ratio of geodesic distance to Euclidean distance. That is, for all non-edges $\{y_i, y_j\} \notin E_Y$ we compute $\frac{c_\Sigma(y_i,y_j)}{|y_i-y_j|}$, and temporarily add one such pair with probability proportional to the ratio, recompute geodesics, and check if the energy decreases. If the energy decreases we accept the change, otherwise we do not. Since the geodesics need to be recomputed for every such test, we limit their frequency to at most once per updating the geometry Y.

4.5.5 Algorithm overview

Here we provide an overview of the algorithm for computing local minimizers of (4.5.3). The algorithm consists of alternating updates of the vertices Y and the edges E_Y with re-computing the geodesics, and is outlined in Algorithm 6.

Algorithm 6 Computing local minimizer of (4.5.3)

Input: initial network graph $\Sigma^0 = (Y^0, E_Y^0)$

repeat

1. compute geodesics using Floyd-Warshall (Algorithm 5)
2. compute edge masses M given in (4.5.2)
3. update Y by applying ADMM steps until the energy (4.5.3) decreases
4. update the edges E_Y as given by the condition (4.5.7), and the heuristic described in 4.5.4

until convergence

Output: local minimizer $\Sigma = (Y, E_Y)$

The overall complexity of the algorithm is $\mathcal{O}((m + d)n^2)$, as it is dominated by the step of computing geodesics. Finally, note that all of the steps in Algorithm 6 decrease the energy, and so the algorithm does converge.

4.6 Numerical examples

In this section we demonstrate the Algorithm (6) on a few different discrete measures μ. One of the examples illustrates how the approach might be used to recover one-dimensional structure of data, while the second example aims to find the optimal transportation networks for uniformly distributed populations.

Example 4.6.1 (One-dimensional grid structure). We first revisit an earlier example in which data are concentrated around a grid-like structure. The example is similar to Example 3.4.3, except that we restrict the data to $\mathbb{R}^2$, and do not add background noise. The

data here consist of 240 points generated by four intersecting lines with Gaussian noise, and are shown on the left of Figure 4.6.1 along with the initialization for the algorithm.

Like the initialization strategy for computing multiple penalized principle curves, here we initialize with small components, but not singletons which do nothing for the (ONT) energy. Let us make the following observation to justify initializing with multiple components, given that we formally considered the (ONT) functional over connected (and compact) sets. Note that we may consider the functional over compact sets with a fixed bound on the number of components, which ensures the lower semicontinuity of the $\mathcal{H}^1$ measure, and therefore also the existence of minimizers. We note that in the numerical setting, the number of components will always be bounded by the number of data points n. Furthermore, if a minimizer to the problem over the larger configuration space is connected, then it is also a minimizer to our (ONT) problem.

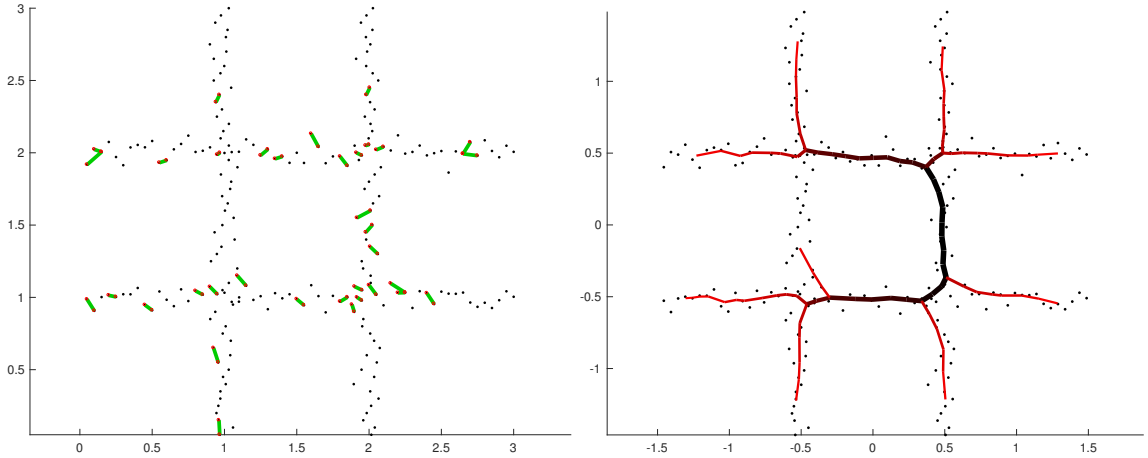


Figure 4.6.1: Data concentrated around a grid structure. The initialization of the algorithm is shown on the left. On the right is a computed minimizer for $\alpha = 0.1$, $\lambda = 0.02$, obtained by only using the criteria (4.5.7) for updating edges.

We first ran the algorithm only using the basic criteria (4.5.7) to update the edges of the network. The resulting configuration is shown on the right in Figure 4.6.1. Although the segments in the initialization are not distributed evenly along the data, the found network does approximate a large portion of the grid. Notably, there is a missing piece corresponding to the left vertical line, which the edge-update criteria (4.5.7) was not able to resolve.

Using the heuristic for adding edges described in Section (4.5.4) however, the algorithm is able to recover the full grid structure (albeit without the topology). The found minimizer, shown in Figure 4.6.2, does indeed have lower energy 0.4757, compared to 0.4975 for the minimizer without a loop (in Figure 4.6.1). In Figure 4.6.2, we also show the non-network edges with non-zero weight of the graph $(X \cup Y, E, W)$. With a close look one can observe that several of the data may enter the network through different points (or not

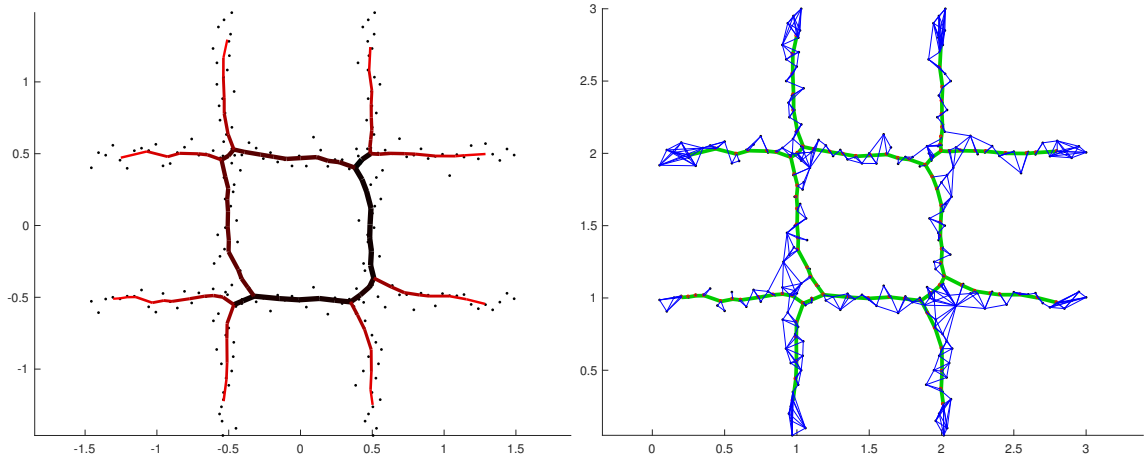


Figure 4.6.2: The computed local minimizer for the data and initialization shown in Figure 4.6.1, using the heuristic for adding edges described in (4.5.4). On the right the same minimizer is shown (green) together with the positive-weight edges on the full graph $(X \cup Y, E, W)$.

at all), depending on which which direction they are to travel to.

Example 4.6.2 (Optimal transportation network for uniform population distribution). In this example, our aim is to understand what the optimal transportation networks will look like for a uniformly distributed population. We will consider a uniform distribution inside a ball, and we are particularly interested in how optimal networks change as the population grows.

We note that have the following two scaling properties of the (ONT) functional. For the first, consider re-scaling the diameter of the measure μ by a factor of L, that is $\mu_L(A) := \mu(\frac{A}{L})$ for every Borel set A. Then we have that

$$E^{\lambda}_{\mu_L}(L\Sigma) = LE^{\lambda}_{\mu}(\Sigma),$$

since our cost function $c_\Sigma(x, y)$ also scales linearly with the length of the network and the diameter of the population. In other words, simply scaling the population only changes minimizers by the same scaling factor. On the other hand, if we re-scale the total mass of the measure μ by a, then we get the following property

$$E^{a^2\lambda}_{a\mu}(\Sigma) = a^2 E^{\lambda}_{\mu}(\Sigma).$$

In view of these scaling properties, we consider a sequence of population distributions with increasing total mass. We expect the optimal networks to become more complex as the mass increases, so we also refine the distribution (increase n) simultaneously. More precisely, we fix the mass of every point to $\frac{1}{200}$, and increase the number of data points n in successive runs. We also increase λ linearly with n ($\lambda = \frac{n}{3200}$), since experiments and heuristics otherwise show that the average distance to the network decreases as n grows (if

λ is fixed), causing the resolution of the computed minimizers to progressively worsen due to the discrete data.

That λ should scale like n can be supported by the following heuristic argument. Since the spacing between the points is kept fixed, the area scales like n, while the radius like $\sqrt{n}$. If we let ℓ denote the average projection distance to the network, then the total transportation cost behaves like $n^2(\ell + \alpha\sqrt{n})$, while the the total length of the network would be roughly $\frac{n}{\ell}$. First order conditions for the energy $n^2(\ell + \alpha\sqrt{n}) + \lambda\frac{n}{\ell}$ with respect to ℓ give $\ell^{*2} = \frac{\lambda}{n}$. Therefore, if the average distance ℓ to a minimizing network is to remain constant, then λ should behave like n. Note that in light of the second scaling property above, we have that

$$E^{n\lambda}_{n\mu}(\Sigma) = nE^{\lambda}_{\sqrt{n}\mu}(\Sigma)$$

and thus our experiments are equivalent to increasing the total mass at the rate $\sqrt{n}$ and keeping λ fixed.

We generate the data by evenly spacing a number of points on concentric circles (rings) of increasing size. The spacing between consecutive rings is fixed so that it is roughly the same as the distance between consecutive points on any ring.

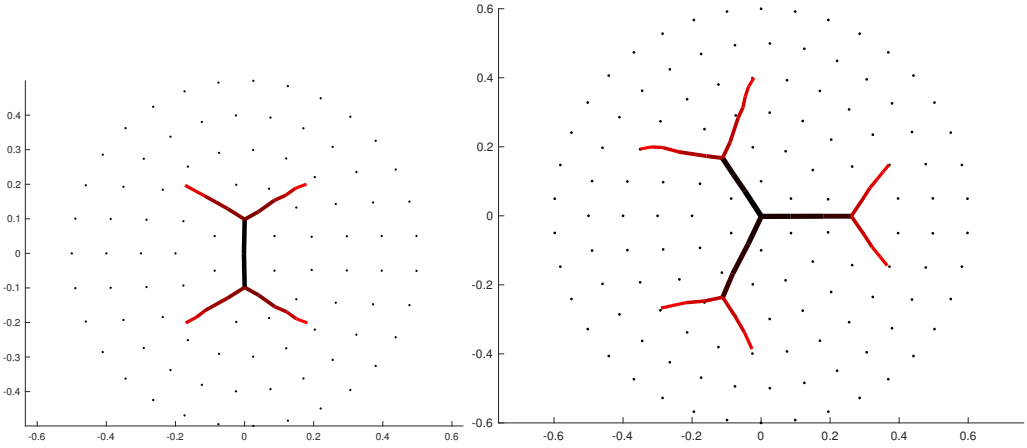


Figure 4.6.3: $n = 95$ (left) and $n = 133$ (right) data points shown with the lowest-energy local minimizers found.

We computed local minimizers for a range of n values (95, 133, 227, 347, 661, 856, 1321, 1887), and for each we ran the algorithm using a variety of initial configurations. Throughout we used $\alpha = 0.1$, and fixed the mass of each point to $\frac{1}{200}$, increasing $\lambda = \frac{n}{3200}$ along the way (as explained above). Shown in Figures 4.6.3, 4.6.2, 4.6.2, and 4.6.2 are the lowest energy local minimizers that we found. It is interesting for us to note that for increasingly larger cities (starting with $n = 347$), we found several configurations achieving energy comparable (within a few percent) to that of the minimizers shown. Indeed, for the larger cities, we make no claim that the topology of the shown configurations is similar to that of the global minimizers, but from our experiments we do believe that the found

networks are close in terms of the (ONT) energy. This perhaps has positive implications for applications in designing networks for transportation. Namely, there would be options for building networks with similar total costs, and careful design may not be so crucial.

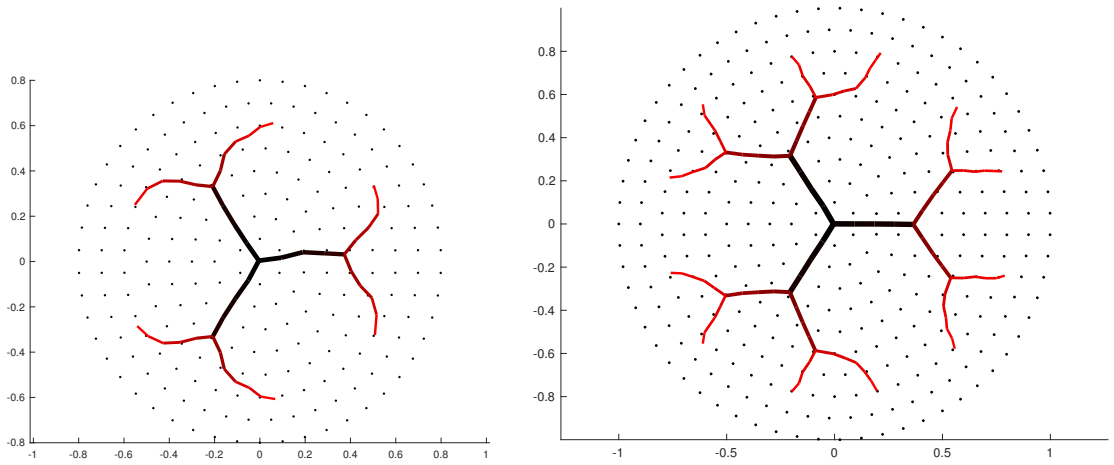


Figure 4.6.4: $n = 227$ (left) and $n = 347$ (right) data points shown with the lowest-energy local minimizers found.

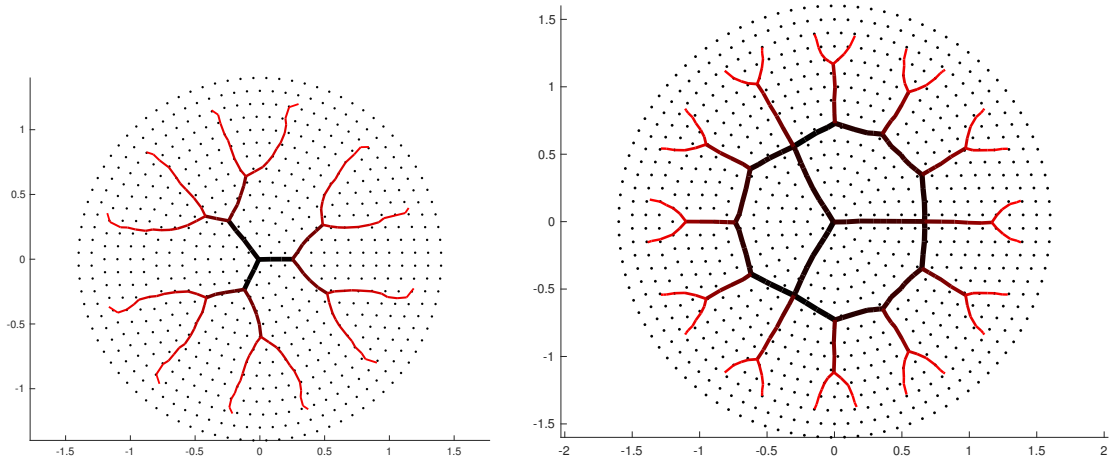


Figure 4.6.5: $n = 661$ (left) and $n = 856$ (right) data points shown with the lowest-energy local minimizers found.

In Figure 4.6.2 we illustrate how the energies and lengths of our computed minimizers relate to n via log-log plots. We note that the log-log plot for the energy seems to be remarkably linear. The slope of the least squares line is 2.17, while the slopes of the segments are 2.22, 2.19, 2.17, 2.15, 2.16, 2.14, 2.16, 2.16, 2.20. We note that the heuristics above

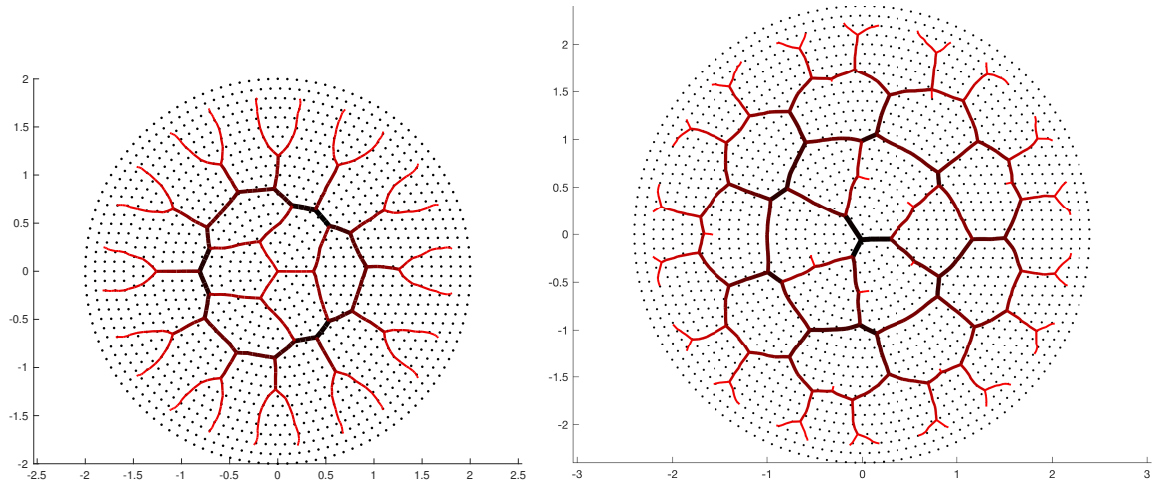


Figure 4.6.6: $n = 1321$ (left) and $n = 1887$ (right) data points shown with the lowest-energy local minimizer found.

predict the energy should scale like $n^{2.5}$. The consistent discrepancy from our experiments is intriguing, and is one that we do not have a good explanation for. On the other hand, the plot for the length is not as consistent, with the slope of the least squares line being 1.21 (individual slopes: 1.92, 1.44, 0.74, 1.39, 0.87, 1.02, 1.88, 1.03, 1.05). Perhaps this is not surprising, since as we noted earlier, there seem to be many network configurations with varying total length that achieve similar low energy. In both plots we included results for two more values of n (177, 491) for which we did not show found configurations, since they are similar to some shown above (for nearby values of n).

4.7 Further discussion

There are a number of directions that remain open for the selective-transport functionals that we have proposed.

Of significant interest is a better understanding of the relation between the parameters of the functional, the measure μ, and the properties of the minimizers, their topology in particular. This is not as direct to obtain as in the case of penalized curves, due to a more general configuration space and a transportation term whose influence is difficult to precisely understand. However, we believe some progress in this regard should be attainable.

On the algorithmic side, there a couple of areas where improvement would be very much valued. One of the biggest is in dealing with the non-convexity of the functionals and correctly identifying the topology of the global minimizers. In the case of data with a clear one-dimensional structure, one may expect that it can be appropriately recovered following

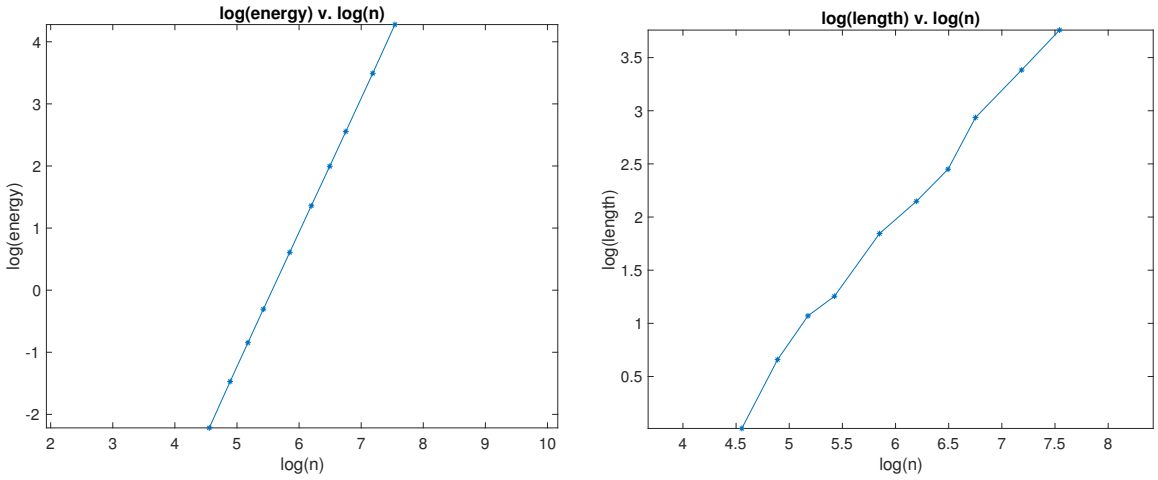


Figure 4.6.7: log-log plots of the energy (left) and length (right) of the computed minimizers against n.

an initialization strategy suggested by the algorithm for (MPPC), i.e. by initializing with short components. One might then hope that the components would grow and eventually connect to recover the underlying structure. However, there is a significant difference for the selective-transport problem we consider, which lies in our choice of the cost-function. Namely, for our geodesic cost-function the data are not forced to go through any part of the network, and therefore network components which are too small might be used for little or no transportation, preventing them to grow or connect. This seems to be the case in general, even though this initialization approach did a reasonable job in recovering the grid in Example 4.6.1. To address some of these concerns, one may consider more strict cost functions that force the data to visit a part of the network, and returning to functionals closer to the average-distance functional could be a starting point.

We conclude with a remark regarding the efficiency of the proposed algorithm for computing optimal networks for selective-transport. Although the algorithm has a feasible runtime (a few hours on a standard laptop) for computing local minimizers for problems on the order of $n \sim 1000$ data points, the complexity does scale quadratically with n. This could present difficulties for computing optimal transportation networks for much larger cities, and even more so for data analysis applications with a very large number of data points. We expect that the computational complexity may be improved, especially if one considers cost functions more closely related to the penalized principal curves and the average-distance problems.

Chapter 5

Conclusion

In this thesis we investigated different models and methods for approximating a measure μ with a one dimensional object. We focused on two approaches with different applications, but stemming from a common framework inspired by the average-distance problem.

We proposed a functional for computing multiple penalized principal curves, and demonstrated its value as a tool for recovering the one-dimensional structure of data, when it has such. We found quantitative relationships between the data and parameters of the functional that govern how the minimizing curves behave. In particular we understood the scale at which minimizers approximate the data and when they become linearly unstable, as well as when they consist of multiple components. On the computational side, we developed a fast numerical algorithm, and demonstrated its suitability through a number of examples that also illustrated its robustness to local minima in cases of significant noise.

One of the limitations of multiple penalized principal curves and minimizing networks of the average-distance problem is in the topologies that they can recover. We can say this was part of the motivation for introducing additional models in which the approximating objects are networks, although the problems are largely motivated by the independent application of finding networks that best serve the transportation needs of a given population. To this end, we proposed a few models for optimal networks for selective-transportation. We established existence of minimizers, and provided a numerical algorithm for approximating them. In contrast to multiple penalized principal curves, with the functionals we proposed for networks remain desires for better understanding the behavior of minimizers, and for dealing with non-convexity in computing them numerically. Extending analysis and efficient numerical algorithms from relatively simple curves to potentially complex networks has proven difficult, and is an interesting direction for future work.

Finally, we note that extending the framework to allow for higher dimensional approximations to measures remains very much open. In our setting of one-dimensional approximations, the functionals we consider have natural interpretations which make them appealing and perhaps easier to understand. A naive approach of using a $\mathcal{H}^2$ term in the ADP functional does not get far, as any set with finite $\mathcal{H}^2$ measure can be approximated arbitrary well by a set with zero $\mathcal{H}^2$ measure (or by a sequence of sets with positive $\mathcal{H}^2$ going to

zero). Careful investigation is needed to find an extension to the setting of two-dimensional (and higher) approximations that is both theoretically appealing and practically suitable for applications.

Bibliography

[1] L. AMBROSIO AND N. GIGLI, *A user's guide to optimal transport*, in Modelling and optimisation of flows on networks, Springer, 2013, pp. 1–155.

[2] L. AMBROSIO, N. GIGLI, AND G. SAVARÉ, *Gradient flows: in metric spaces and in the space of probability measures*, Springer Science & Business Media, 2008.

[3] L. AMBROSIO AND P. TILLI, *Topics on analysis in metric spaces*, no. 25 in Oxford Lecture Series in Mathematics and Its Applications, Oxford University Press, 2004.

[4] E. ARIAS-CASTRO, D. L. DONOHO, AND X. HUO, *Adaptive multiscale detection of filamentary structures in a background of uniform random points*, Ann. Statist., 34 (2006), pp. 326–349.

[5] M. BELKIN AND P. NIYOGI, *Laplacian eigenmaps for dimensionality reduction and data representation*, Neural Comput., 15 (2003), pp. 1373–1396.

[6] G. BIAU AND A. FISCHER, *Parameter selection for principal curves*, Information Theory, IEEE Transactions on, 58 (2012), pp. 1924–1939.

[7] S. BOYD, N. PARIKH, E. CHU, B. PELEATO, AND J. ECKSTEIN, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Found. Trends Mach. Learn., 3 (2011), pp. 1–122.

[8] A. BRAIDES, *Gamma-convergence for Beginners*, vol. 22, Clarendon Press, 2002.

[9] A. BRANCOLINI AND G. BUTTAZZO, *Optimal networks for mass transportation problems*, ESAIM: Control, Optimisation and Calculus of Variations, 11 (2005), pp. 88–101.

[10] A. BRANCOLINI AND B. WIRTH, *Equivalent formulations for the branched transport and urban planning problems*, Journal de Mathématiques Pures et Appliquées, 106 (2016), pp. 695–724.

[11] T. BRODERICK, B. KULIS, AND M. JORDAN, *Mad-bayes: Map-based asymptotic derivations from bayes*, in Proceedings of The 30th International Conference on Machine Learning, 2013, pp. 226–234.

[12] G. BUTTAZZO, E. OUDET, AND E. STEPANOV, *Optimal transportation problems with free dirichlet regions*, in Variational methods for discontinuous structures, Springer, 2002, pp. 41–65.

[13] G. BUTTAZZO, A. PRATELLI, S. SOLIMINI, AND E. STEPANOV, *Optimal urban networks via mass transportation*, Springer Science & Business Media, 2008.

[14] G. BUTTAZZO AND E. STEPANOV, *Optimal transportation networks as free dirichlet regions for the monge-kantorovich problem*, Annali della Scuola Normale Superiore di Pisa - Classe di Scienze, 2 (2003), pp. 631–678.

[15] F. CHAZAL AND J. SUN, *Gromov-hausdorff approximation of filament structure using reeb-type graph*, in Proceedings of the Thirtieth Annual Symposium on Computational Geometry, SOCG'14, New York, NY, USA, 2014, ACM, pp. 491:491–491:500.

[16] Y.-C. CHEN, C. R. GENOVESE, AND L. WASSERMAN, *Asymptotic theory for density ridges*, Ann. Statist., 43 (2015), pp. 1896–1928.

[17] Y.-C. CHEN, S. HO, P. E. FREEMAN, C. R. GENOVESE, AND L. WASSERMAN, *Cosmic web reconstruction through density ridges: method and algorithm*, Monthly Notices of the Royal Astronomical Society, 454 (2015), pp. 1140–1156.

[18] R. R. COIFMAN AND S. LAFON, *Diffusion maps*, Applied and Computational Harmonic Analysis, 21 (2006), pp. 5 – 30. Special Issue: Diffusion Maps and Wavelets.

[19] D. COMANICIU AND P. MEER, *Mean shift: A robust approach toward feature space analysis*, Pattern Analysis and Machine Intelligence, IEEE Transactions on, 24 (2002), pp. 603–619.

[20] T. H. CORMEN, C. E. LEISERSON, R. L. RIVEST, AND C. STEIN, *Introduction to Algorithms, Third Edition*, The MIT Press, 3rd ed., 2009.

[21] P. DELICADO, *Another look at principal curves and surfaces*, J. Multivariate Anal., 77 (2001), pp. 84–116.

[22] C. J. DSILVA, B. LIM, H. LU, A. SINGER, I. G. KEVREKIDIS, AND S. Y. SHVARTSMAN, *Temporal ordering and registration of images in studies of developmental dynamics*, Development, 142 (2015), pp. 1717–1724.

[23] T. DUCHAMP AND W. STUETZLE, *Geometric properties of principal curves in the plane*, in Robust statistics, data analysis, and computer intensive methods (Schloss Thurnau, 1994), vol. 109 of Lecture Notes in Statist., Springer, New York, 1996, pp. 135–152.

[24] D. EBERLY, *Ridges in image and data analysis*, vol. 7, Springer Science & Business Media, 1996.

[25] E. ESSER, *Applications of lagrangian-based alternating direction methods and connections to split bregman*, (2009).

[26] E. A. FEINBERG, P. O. KASYANOV, AND N. V. ZADOIANCHUK, *Fatou's lemma for weakly converging probabilities*, Theory of Probability & Its Applications, 58 (2014), pp. 683–689.

[27] C. FEUERSÄNGER AND M. GRIEBEL, *Principal manifold learning by sparse grids*, Computing, 85 (2009), pp. 267–299.

[28] I. FONSECA AND G. LEONI, *Modern Methods in the Calculus of Variations: L^p Spaces*, Springer Science & Business Media, 2007.

[29] X. GE, I. I. SAFA, M. BELKIN, AND Y. WANG, *Data skeletonization via reeb graphs*, in Advances in Neural Information Processing Systems 24, Curran Associates, Inc., 2011, pp. 837–845.

[30] C. R. GENOVESE, M. PERONE-PACIFICO, I. VERDINELLI, AND L. WASSERMAN, *Nonparametric ridge estimation*, Ann. Statist., 42 (2014), pp. 1511–1545.

[31] S. GERBER, T. TASDIZEN, AND R. WHITAKER, *Dimensionality reduction and principal surfaces via kernel map manifolds*, in Computer Vision, 2009 IEEE 12th International Conference on, IEEE, 2009, pp. 529–536.

[32] S. GERBER AND R. WHITAKER, *Regularization-free principal curve estimation*, The Journal of Machine Learning Research, 14 (2013), pp. 1285–1302.

[33] E. GILBERT AND H. POLLAK, *Steiner minimal trees*, SIAM Journal on Applied Mathematics, 16 (1968), pp. 1–29.

[34] T. GOLDSTEIN AND S. OSHER, *The split Bregman method for $L1$-regularized problems*, SIAM J. Imaging Sci., 2 (2009), pp. 323–343.

[35] T. HASTIE AND W. STUETZLE, *Principal curves*, J. Amer. Statist. Assoc., 84 (1989), pp. 502–516.

[36] M. HONG AND Z.-Q. LUO, *On the linear convergence of the alternating direction method of multipliers*, Mathematical Programming, 162 (2017), pp. 165–199.

[37] L. KANTOROVICH, *On the translocation of masses*, C.R. (Doklady) de Acad. Sci. URSS (N.S.), 37 (1942), pp. 199–201.

[38] R. O. KARLSTROM AND D. A. KANE, *A flipbook of zebrafish embryogenesis*, Development, 123 (1996), pp. 461–462.

[39] B. KEGL, A. KRZYZAK, T. LINDER, AND K. ZEGER, *Learning and design of principal curves*, Pattern Analysis and Machine Intelligence, IEEE Transactions on, 22 (2000), pp. 281–297.

[40] S. KIROV AND D. SLEPČEV, *Multiple penalized principal curves: Analysis and computation*, Journal of Mathematical Imaging and Vision, (2017).

[41] J. B. KRUSKAL, *Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis*, Psychometrika, 29 (1964), pp. 1–27.

[42] B. KULIS AND M. JORDAN, *Revisiting k-means: New algorithms via bayesian nonparametrics*, in Proceedings of the 29th International Conference on Machine Learning (ICML-12), J. Langford and J. Pineau, eds., ICML '12, New York, NY, USA, July 2012, Omnipress, pp. 513–520.

[43] A. LEMENANT, *A presentation of the average distance minimizing problem*, Journal of Mathematical Sciences, 181 (2012), pp. 820–836.

[44] X. Y. LU AND D. SLEPČEV, *Average-distance problem for parameterized curves*, ESAIM: COCV, (2015).

[45] X. Y. LU AND D. SLEPČEV, *Properties of minimizers of average-distance problem via discrete approximation of measures*, SIAM Journal on Mathematical Analysis, 45 (2013), pp. 3114–3131.

[46] F. MADDALENA AND S. SOLIMINI, *Transport distances and irrigation models*, J. Convex Anal, 16 (2009), pp. 121–152.

[47] ——, *Synchronic and asynchronic descriptions of irrigation problems*, Advanced Nonlinear Studies, 13 (2013), pp. 583–623.

[48] F. MADDALENA, G. TAGLIALATELA, AND J.-M. MOREL, *A variational model of irrigation patterns*, Interfaces and Free Boundaries, 5 (2003), pp. 391–416.

[49] C. MANTEGAZZA, A. C. MENNUCCI, ET AL., *Hamilton-Jacobi equations and distance functions on riemannian manifolds*, Applied Mathematics and Optimization, 47 (2003), pp. 1–26.

[50] G. MONGE, *Mémoire sur la théorie des déblais et des remblais*, Historie de l'Acad. des Sciences de Paris, (1781), pp. 666–704.

[51] R. NISHIHARA, L. LESSARD, B. RECHT, A. PACKARD, AND M. JORDAN, *A general analysis of the convergence of admm*, in International Conference on Machine Learning, 2015, pp. 343–352.

[52] S. OSHER, M. BURGER, D. GOLDFARB, J. XU, AND W. YIN, *An iterative regularization method for total variation-based image restoration*, Multiscale Modeling & Simulation, 4 (2005), pp. 460–489.

[53] U. OZERTEM AND D. ERDOGMUS, *Locally defined principal curves and surfaces*, The Journal of Machine Learning Research, 12 (2011), pp. 1249–1286.

[54] S. PULKKINEN, *Ridge-based method for finding curvilinear structures from noisy data*, Computational Statistics & Data Analysis, 82 (2015), pp. 89 – 109.

[55] S. T. ROWEIS AND L. K. SAUL, *Nonlinear dimensionality reduction by locally linear embedding*, Science, 290 (2000), pp. 2323–2326.

[56] F. SANTAMBROGIO, *Optimal transport for applied mathematicians*, Birkäuser, NY, (2015).

[57] A. SINGER AND H.-T. WU, *Vector diffusion maps and the connection Laplacian*, Comm. Pure Appl. Math., 65 (2012), pp. 1067–1144.

[58] G. SINGH, F. MEMOLI, AND G. CARLSSON, *Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition*, in Eurographics Symposium on Point-Based Graphics, 2007, pp. 91–100.

[59] D. SLEPČEV, *Counterexample to regularity in average-distance problem*, Annales de l'Institut Henri Poincare (C) Non Linear Analysis, 31 (2014), pp. 169 – 184.

[60] A. J. SMOLA, S. MIKA, B. SCHÖLKOPF, AND R. C. WILLIAMSON, *Regularized principal manifolds*, J. Mach. Learn. Res., 1 (2001), pp. 179–209.

[61] E. STEPANOV, *Partial geometric regularity of some optimal connected transportation networks*, Journal of Mathematical Sciences, 132 (2006), pp. 522–552.

[62] J. B. TENENBAUM, V. DE SILVA, AND J. C. LANGFORD, *A global geometric framework for nonlinear dimensionality reduction*, science, 290 (2000), pp. 2319–2323.

[63] R. TIBSHIRANI, *Principal curves revisited*, Stat. Comput., 2 (1992), pp. 182–190.

[64] R. TIBSHIRANI, M. SAUNDERS, S. ROSSET, J. ZHU, AND K. KNIGHT, *Sparsity and smoothness via the fused lasso*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67 (2005), pp. 91–108.

[65] J. VERBEEK, N. VLASSIS, AND B. KROSE, *A k-segments algorithm for finding principal curves*, Pattern Recognition Letters, 23 (2002), pp. 1009 – 1017.

[66] C. VILLANI, *Optimal transport: old and new*, vol. 338, Springer Science & Business Media, 2008.

[67] H. WANG AND T. C. LEE, *Automatic parameter selection for a k-segments algorithm for computing principal curves*, Pattern recognition letters, 27 (2006), pp. 1142–1150.

[68] Q. XIA, *Optimal paths related to transport problems*, Communications in Contemporary Mathematics, 5 (2003), pp. 251–279.