THÈSE

En vue de l'obtention du
DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par:
Université Toulouse III Paul Sabatier (UT3 Paul Sabatier)

Discipline ou spécialité:
Mathématiques appliquées

Présentée et soutenue par
Santiago Alejandro GALLÓN GÓMEZ
le: 28 juin 2013

Titre:
Template estimation for samples of curves and functional calibration estimation via the method of maximum entropy on the mean

---

École doctorale :
Mathématiques Informatique Télécommunications (MITT)

Unité de recherche:
UMR 5219

Directeurs de thèse:
Jean-Michel LOUBES (Professeur, Université Toulouse III Paul Sabatier)
Fabrice GAMBOA (Professeur, Université Toulouse III Paul Sabatier)

Rapporteurs:
Jean-François DUPUY (Professeur, INSA de Rennes)
Henryk GZYL (Professeur, Centro de Finanzas IESA)

Membres du jury:
Thibault ESPINASSE (Maitre de Conférences, Université Claude Bernard Lyon I)
David GINSBOURGER (Professeur, University of Bern)
Elie MAZA (Maitre de Conférences, Institut National de la Recherche Agronomique)
Anne RUIZ-GAZEN (Professeur, Université Toulouse I)

ED Mathématiques Informatique Télécommunication de Toulouse,
Université Toulouse III - Paul Sabatier,
118 route de Narbonne,
31062 Toulouse, France.

Institut de Mathématiques de Toulouse,
UMR CNRS 5219,
Université Toulouse III - Paul Sabatier,
118 route de Narbonne,
31062 Toulouse, France.

to my family
to Karoll

"Pain is temporary.
It may last a minute, or an hour, or a day, or a year,
but eventually it will subside and something else will take its place.
If I quit, however, it lasts forever"
L. Armstrong

"Success is achieved by converting each step in a goal and every goal in one step"
C.C. Cortéz

Acknowledgments

I am very grateful to my advisors Jean-Michel LOUBES and Fabrice GAMBOA for inviting me, when they were in Colombia, to work with them at the Institut de Mathématiques de Toulouse, Université Toulouse III - Paul Sabatier, and for giving me the opportunity to continue my doctoral training through the MELISSA (Mathematical E-Learning in Statistics for Latin and South America)[*] and PREFALC (Programme Régional France-Amérique Latine-Caraïbe)[†] projects. Thanks a lot for their hospitality, patience and orientation in my research, and for introducing me to fields of functional data analysis, statistical inverse problems, statistical learning, and many other topics in statistics and probability. This has been a great opportunity and experience that, for sure, I will share with my future colleagues and students.

I am also grateful to Elie MAZA, associate researcher of the Statistics and Probability Team at the Institut de Mathématiques de Toulouse, for sharing me his knowledge in curve warping models, manifold learning and programming.

Thank you very much to Jean-François DUPUY and Henryk GZYL for their willingness to evaluate this thesis, and for being members of the jury joint to Thibault ESPINASSE, David GINSBOURGER, Elie MAZA and Anne RUIZ-GAZEN, for them also my special thanks. I am grateful for their valuable comments and constructive suggestions also.

Also thanks to the administrative staff at the Institut de Mathématiques de Toulouse and the financial support from the University of Antioquia, Colombia.

Because the education is the best legacy of the parents to their children, I thank to my parents for your love and hard effort. Also thanks for my sister and her three daughters, for your words of encouragement. I have sacrificed too much time with my family staying far from them, but they know that all of this effort is for them also. A huge thank goes to my wife Karoll, with whom I began this long, hard but also exiting journey of the academic life, that always has believed in me.

A special thank goes to Jean-Michel family for their friendship and kindness. Also many thanks to my colleagues and friends at the University: Michael and his wife Laura, Thibaut, Chloé, Adil, Lilian, Salomón and Ricardo.

Finally, thanks to God because without him this could not be possible.

[*] http://www.math.univ-toulouse.fr/MELISSA_UPS/index_ang.html

[†] http://www.prefalc.msh-paris.fr/?lang=fr

Contents

List of figures

List of tables

Notation

$\mathbb{N}$	Set of natural numbers
$\mathbb{R}$	Set of real numbers
$\mathbb{R}_+$	Set of non-negative real numbers
$\mathbb{R}^m$	m-dimensional Euclidean space
$\mathbb{Z}$	Set of integer numbers
$\mathbb{Z}_{++}$	Set of strictly positive integer numbers
$\mathcal{F}$	Space of functions
$\mathcal{C}(\mathcal{I})$	Space of continuous functions on $\mathcal{I}$
$\mathcal{C}^\infty(\mathcal{I})$	Space of infinitely differentiable functions on $\mathcal{I}$ (space of smooth functions)
$(\Omega, \mathcal{A}, \mathbb{P})$	Probability space
$\mathcal{M}$	Manifold
$\mathcal{T}_p(\mathcal{M})$	Tangent space of $\mathcal{M}$ at p
$\mathcal{G}(V, E)$	A graph with sets of vertices V and edges E

$\langle \cdot, \cdot \rangle$	Inner product
$\lVert \cdot \rVert$	Euclidean norm
$\lVert \cdot \rVert_\infty$	Uniform norm, supremum norm
$\lVert X \rVert_{\mathrm{F}}$	Frobenius norm of a matrix X
$d(X_i, X_{i'})$	Distance between points X_i and $X_{i'}$
$W_p(\mu, \nu)$	Wasserstein distance between probability measures μ and ν
$d_{\mathcal{G}}(X_i, X_{i'})$	Graph distance between vertices X_i and $X_{i'}$
$\delta(X_i, X_{i'})$	Geodesic distance between points X_i and $X_{i'}$ on $\mathcal{M}$
$D(\nu \parallel \upsilon)$	Kullback-Leibler divergence between measures μ and ν
$S(\nu \parallel \upsilon)$	Entropy function
$B(X, \varepsilon)$	(Open) ball with center at X and radius ε

$F(x)$	Distribution function
$q(\alpha), F^{-1}(\alpha)$	Quantile function
$\mathbb{F}_m(x)$	Empirical distribution function
$\hat{q}(\alpha), \mathbb{F}_m^{-1}(\alpha)$	Empirical quantile function
$X_{(j)}, X_{j:m}$	jth order statistic in a sample of size m

$\mathbb{P}$	Probability
$\mathbb{E}(X)$	Expectation of X
$\mathrm{Med}(X)$	Median of X

$\mathrm{Var}(X)$	Variance of X
$\mathrm{Cov}(X,Y)$	Covariance between X and Y
$\mathcal{N}(\mu,\sigma^2)$	Gaussian distribution with mean μ and variance σ^2
$\mathcal{N}(\mu,\Sigma)$	Multivariate Gaussian distribution with mean μ and covariance matrix Σ
$\mathcal{U}[a,b]$	Uniform distribution on interval $[a,b]$
h, H	Warping functions
$f^{(k)}$	kth derivative of f
$\mathbb{1}_A$	Indicator function of A
δ_x	Dirac measure at x
$K(s,t)$	Kernel
exp	Exponential
min, max	Minimum, maximum
$\circ$	Composition of mappings
$\ll$	Absolutely continuous
$<<$	Much less than
$\otimes$	Product of measures
$\overset{d}{=}$	Equality in distribution
$\xrightarrow{\mathcal{D}}$	Converge in distribution
$\xrightarrow{\text{a.s.}}$	Almost sure converge
1_n	nth dimensional vector of ones
I_n	$n\times n$ identity matrix
$X^\top$	Transpose of a matrix X
$\mathrm{trace}(X)$	Trace of a matrix X
$\mathrm{diag}(\lambda_1,\ldots,\lambda_n)$	Diagonal matrix with diagonal elements $\lambda_1,\ldots,\lambda_n$
X^{-1}	Inverse of matrix X
$X^\dagger$	Moore-Penrose pseudoinverse of matrix X
$\mathrm{proj}_X Y$	Projection of vector Y onto matrix X

Introduction générale

L'analyse des données fonctionnelles est un champ de la statistique moderne ayant pour objet d'étude des fonctions aléatoires $X = \{X(t) \colon t \in \mathcal{T} \subset \mathbb{R}\}$ appartenant à un espace de fonctions $\mathcal{F}$ (i.e. des variables aléatoires à valeurs dans un espace fonctionnel de dimension infinie), dont les réalisations sont des courbes lisses. Les avancées technologiques dans la mesure, la collecte et le traitement de l'information ont permis un accès rapide et précis à ces données, présentes de nombreuses disciplines comme la bioinformatique, les sciences médicales, la physique, l'économie, la géologie, l'astronomie, la météorologie, pour n'en citer qu'une partie. Au cours des vingt dernières années, de nombreux nouveaux développements méthodologiques et théoriques ont eu lieu, en analyse des composantes principales et de corrélation canonique fonctionnel, en classification/régression fonctionnelle, sur les séries temporelles, les modèles additifs généralisés fonctionnels, les modèles non paramétriques fonctionnels, entre autres.

L'une des principales difficultés de l'analyse statistique des fonctions consiste à extraire un motif commun, quand il existe, qui synthétise l'information contenue par toutes les fonctions de l'échantillon lorsque les trajectoires individuelles varient en amplitude et en phase. C'est le sujet principal de cette thèse. Le motif est représenté par une courbe inconnue f qui préserve les caractéristiques structurelles de l'échantillon de courbes f_i, $i = 1, \ldots, n$.

Chapitre 2

Ce chapitre examine le problème d'identification d'une fonction qui représente le motif commun d'un échantillon de courbes en supposant que les données appartiennent à une variété ou en sont suffisamment proches, d'une variété $\mathcal{M}$ non linéaire de basse dimension intrinsèque munie d'une structure géométrique inconnue et incluse dans un espace de grande dimension. Sous cette hypothèse géométrique, le problème de l'estimation de la courbe commune est équivalent à considérer une mesure de la centralité de l'échantillon de fonctions par rapport à la distance géodésique intrinsèque δ associée à la variété. Pour répondre à cette question, nous proposons un nouvel algorithme basé sur une version modifiée de l'algorithme Isomap (isometric featuring mapping en anglais) de Tenenbaum et al. [96] pour l'approximation de la distance géodésique. Cette approximation est utilisée pour calculer la fonction médiane empirique de Fréchet correspondante. Cela fournit un estimateur intrinsèque robuste de la forme commune f. L'algorithme proposé est moins sensible à la présence de valeurs aberrantes et, à différence de l'algorithme Isomap original, ne requiert pas de choix des paramètres de voisinage pour la construction du graphe d'adjacence. Les comparaisons avec d'autres méthodes, sur des données simulées et réelles, montrent que notre algorithme fonctionne bien et surpasse d'autres méthodes.

Chapitre 3

Ce chapitre étudie les propriétés asymptotiques de la méthode de *normalisation quantile* développée par Bolstad et al. [12] qui est devenue l'une des méthodes les plus populaires pour aligner des courbes de densité en analyse de données de microarrays en bioinformatique. L'idée centrale de la normalisation quantile consiste à aligner les distributions empiriques des intensités. C'est-à-dire que le diagramme quantile-quantile entre deux vecteurs de données doit être au plus proche de la première bissectrice. Cette méthode est considérée comme un cas particulier de la procédure de la moyenne structurelle pour l'alignement des courbes proposée par Dupuy et al. [35]. Les propriétés sont démontrées à partir d'un modèle de déformation, dans lequel chaque fonction de distribution F_i, $i = 1, \ldots, n$ est obtenue par déformation d'une fonction de distribution commune F par un échantillon de fonctions de déformation H_i. Nous étudions la convergence forte et la distribution asymptotique de l'estimateur de normalisation quantile. De plus, nous montrons par des simulations que la méthode de normalisation quantile échoue dans certains cas. Ainsi, nous proposons une nouvelle méthode, pour faire face à ce problème. Cette méthode utilise l'algorithme développée dans le Chapitre 2.

Chapitre 4

Dans ce chapitre, nous étendons le problème d'estimation de calage pour la moyenne d'une population finie de la variable de sondage Y dans un cadre de données fonctionnelles. L'idée consiste à modifier les poids de base du plan de sondage de base $(d_i)_{i \in a}$ de l'estimateur sans biais d'Horvitz-Thompson fonctionnel par de nouveaux poids de calage fonctionnel $w_i(t)$ plus efficaces et suffisamment proches des d_i selon une certaine fonction de distance de dissimilarité $\mathcal{D}_a^*(w, d)$ satisfaisant la restriction de calage fonctionnel. Nous considérons le problème de l'estimation des poids de sondage fonctionnel à travers le principe du maximum d'entropie sur la moyenne (Gamboa [40], et Gamboa and Gassiat [41]). En particulier, l'estimation par calage est considérée comme un problème inverse linéaire de dimension infinie suivant la structure de l'approche du maximum d'entropie sur la moyenne. La méthode appliquée se concentre sur la reconstruction d'une mesure ν^* unique qui

maximise l'entropie relative $S(\nu \parallel \upsilon)$ par rapport à une mesure *a priori* υ sous une contrainte linéaire donnée par la restriction de calage fonctionnel. Nous donnons un résultat précis d'estimation des poids de calage fonctionnels pour deux types de mesures aléatoires *a priori*: la measure Gaussienne centrée et la measure de Poisson généralisée. Une étude de simulation simple, montre que notre estimateur de calage fonctionnel est plus précis que l'estimateur d'Horvitz-Thompson fonctionnel.

Chapter 1

General presentation

1.1 Motivation

In the domain of classical statistical practice, the sample of random elements is usually a set of finite-dimensional objects such as random variables, random vectors or random sequences. Nevertheless, in many real life scientific applications, these objects instead are properly assumed as real-valued random functions belonging to some infinite-dimensional (or functional) space $\mathcal{F}$. It is a sample of curves as realizations of a stochastic process $X \in \mathcal{F}$. So, as indicated by its name, in Functional Data Analysis −FDA− the variables are viewed as functions defined on some index set $\mathcal{I}$, i.e. $X = \{X(t) \colon t \in \mathcal{I} \subset \mathbb{R}\}$, where the dataset is obtained from observations of a smooth random process observed at discrete time points. The functional data assumption is appropriate in different cases including, for instance, irregularly spaced measurements, high-frequency data, analysis with derivatives of the functions, among others.

Recently, in the last twenty years, the statistical literature has witnessed numerous advances about statistical analysis of functions with its subsequent applications in a wide variety of scientific areas (e.g., in bioinformatics, medicine, physical sciences, economics, finance, marketing, education, ecology, astronomy, meteorology, geology, archeology, criminology, physiology, etc.), constituting itself as an important and dynamic area of modern statistics. Indeed, several methodological and theoretical developments have been directed toward, for instance, functional principal components analysis and canonical correlation, functional classification/regression, functional time series, functional additive models, functional nonparametric models, functional testing, among many others. The textbooks by Ramsay and Silverman [79], and Horváth and Kokoszka [51] offer detailed introductions to the branch of functional data analysis. The books by Bosq [13], and Bosq and Blanke [14] study its mathematical foundations and asymptotic analysis. Ferraty and Vieu [37] and Ramsay and Graves [76] cover the nonparametric statistical methods and computational aspects, respectively. Ramsay and Silverman [78] illustrate by case studies its application in several fields of science.

One of the main difficulties in statistical analysis of functions is the extraction of a meaningful common pattern that summarizes the information conveyed by all functions in the sample. This is the main subject of the present thesis. This task is often crucial to inference, prediction, or to apply any subsequent statistical analysis. The pattern is represented by an unknown curve f, called the *template* function, which preserves the structural features of the sample of curves f_i, $i = 1, \dots, n$. However, a serious difficulty arises when individual trajectories vary both in amplitude (variation on the y-axis) and phase (variation on the x-axis). In particular, when the phase variability is ignored the classical descriptive statistics such as the mean, variance, correlations, and standard multivariate statistical tools as principal component analysis are seriously affected (Ramsay and Silverman [79]). Thus the estimation of f through the classical pointwise mean is not a good strategy. Indeed it does not resemble any of the observed curves and fails to capture the structural characteristics in the sample (Ramsay and Li [77]).

One well-known method to obtain f in presence of this systematic phase variability between curves is though a synchronization or alignment process. In the statistical literature, this process is known as *curve registration* (also *curve alignment* in biology, *structural averaging* in computing, or *time warping* in engineering). Historically, the curve registration problem comes from the seminal papers by Kneip and Gasser [58] and Sakoe and Chiba [84] in statistics and engineering, respectively. Since these works, various curve registration methods have been proposed using different strategies. See, for instance, Kneip and Gasser [58], Wang and Gasser [106], Ramsay and Li [77], Kneip et al. [61], James [55], Tang and Müller [95], Kneip and Ramsay [59], and Dupuy et al. [35].

The problem setup may be formulated as follows. Let a collection of $n \geq 2$ units for which m_i observations on any variable X at $t_{ij} \in \mathcal{I} := [a, b] \subset \mathbb{R}$ are available, denoted by X_{ij}, $j = 1, \dots, m_i$, $i = 1, \dots, n$. The observations are considered as realizations generated by evaluating the set of unknown smooth functions f_i at points t_{ij}, i.e. $X_{ij} = f_i(t_{ij})$. Generally, an observational error term ϵ_{ij} is assumed to be present in the data collection process, so that X_{ij} satisfies the regression model

$$X_{ij} = f_i(t_{ij}) + \sigma_i \epsilon_{ij}, \qquad j = 1 \dots, m_i, \; i = 1, \dots, n. \tag{1.1}$$

Here ϵ_{ij} are centered independent and identically distributed (i.i.d.) random variables with $\mathbb{E}(\epsilon_{ij}^2) = 1$ and $\sigma_i > 0$ for all i. In presence of errors, the unknown smooth functions f_i can be estimated from the observed pairs (t_{ij}, X_{ij}) applying some nonparametric curve estimation method as local polynomial regression or projection estimation using some appropriate basis function system as Fourier, splines or wavelets basis (see for instance Ramsay and Silverman [79] and Tsybakov [99], and references therein).

The registration problem relies on the assumption that there exists an unknown common pattern f from which each individual function f_i may be deduced by warping f through a time-warping function $h_i(t)$, which monotonically shifts the (time) index set $\mathcal{I}$ into itself. If a nonparametric model is assumed, then each curve is given by a time-warping model $f_i(t) = f \circ h_i^{-1}(t)$. Thus, the model (1.1) can be rewritten as

$$X_{ij} = f \circ h_i^{-1}(t_{ij}) + \sigma_i \epsilon_{ij},$$

where h_i are i.i.d. invertible random functions from a general warping process $\mathcal{H}$. The template is found by taking the cross-sectional mean of the sample of warped curves, with respect to a given template f_0 of f, by using the estimated warping functions. Usually the first curve or the mean of the observed curves is used for f_0.

An alternative way to express the individual curves can be adopted appealing to a semi-parametric approach through a self-modeling regression framework $f_i(t) = f(t, \theta_i)$ (see Kneip and Gasser [57]), where all functions are deduced with respect to f by mean a finite-dimensional individual parameter vector $\theta_i \in \Theta \subset \mathbb{R}^p$ $(p \in \mathbb{Z}_{++})$,

$$X_{ij} = f(t_{ij}, \theta_i) + \sigma_i \epsilon_{ij}.$$

This model is appropriate when there exists certain homogeneity in structure of the sample of curves (Kneip and Gasser [57]). A common type of model that allows amplitude and time variations of f is given by a shape invariant model given by

$$X_{ij} = D_i + C_i f(B_i t_{ij} - A_i) + \sigma_i \epsilon_{ij}.$$

This particular specification is adopted, for example, by Silverman [88], Rønn [81], Gervini and Gasser [45], Gamboa et al. [42], Castillo and Loubes [18], Bigot and Gadat [9], Bigot et al. [10], and Trigano et al. [98]. Usually the estimation of f is made by a backfitting algorithm. The algorithm is based on two recursive steps assuming an initial estimate of f by a first guess. In the first step, the estimation of $\theta_i = (A_i, B_i, C_i, D_i)$, $i = 1, \ldots, n$ is performed. In the second step, the estimate of f is updated. In both steps, estimations are performed using a least squares criterion (Kneip and Gasser [57]).

In this thesis, all curves are assumed to be observed at the same time with the same occurrence, i.e. $t_{ij} = t_j$ and $m_i = m$, and the variance of the additive error term $\sigma_i \epsilon_{ij}$ is constant, i.e. $\sigma_i^2 = \sigma^2$.

Nevertheless, there exists a different approach for estimating the template f without assuming any deformation model for the individual curves as above. Instead, the template is obtained directly from the sample of observed curves without stressing any particular curve. The estimated function is assumed to be located at the *center* of the sample capturing its central amplitude behavior. This will be the approach assumed in Chapter 2.

For instance, López-Pintado and Romo [68], and Arribas-Gil and Romo [3] explore this idea proposing an estimator of f appealing to the concept of functional data depth as a measure of "centrality" of a function with respect to the sample of curves. In particular, López-Pintado and Romo [68] propose a new notion of band depth for functional data based on the graphic representation of the functions and making use of bands defined of their graphs on the plane. This notion is defined as follows.

Definition 1.1. For a set of functions $f_1, \dots, f_n$, the band depth for any of these curves f_i, $i = 1, \dots, n$ is defined by

$$\mathrm{BD}_{n,L}(f_i) = \sum_{l=2}^{L} \binom{n}{l}^{-1} \sum_{1 \le i_1 < \cdots < i_l \le n} \mathbb{1}\left\{G(f_i) \subseteq B(f_{i_1}, \dots, f_{i_l})\right\},$$

for some fixed value $L \in [2, n]$, where

$$\begin{aligned} B(f_{i_1}, \dots, f_{i_k}) &= \left\{(t, g) \colon t \in \mathcal{I},\ \min_{r=1,\dots,k} f_{i_r}(t) \le g \le \max_{r=1,\dots,k} f_{i_r}(t)\right\} \\ &= \left\{(t, g) \colon t \in \mathcal{I},\ g = \lambda \min_{r=1,\dots,k} f_{i_r}(t) + (1-\lambda) \max_{r=1,\dots,k} f_{i_r}(t),\ \lambda \in [0,1]\right\} \end{aligned}$$

is the band in $\mathbb{R}^2$ delimited by the curves $f_{i_1}, \dots, f_{i_k}$, $G(f_i) = \{(t, f_i(t) \colon t \in \mathcal{I})\}$ is the graph of the function f_i, and $\mathbb{1}(\cdot)$ denotes the indicator function.

Remark 1.1. This notion allows ordering the curves from the center-outward providing a generalization of L-statistics to the functional framework. Also the finite-dimensional version of the functional band depth provides a depth for multivariate data (López-Pintado and Romo [68]).

Remark 1.2. López-Pintado and Romo [68] recommend to use a $L = 3$ value in their band depth definition. Between the reasons for this choice are the stability and computational simplicity of the method. Additionally, the bands corresponding to large values of L do not resemble the shape of any of the curves in the sample.

Remark 1.3. A robust estimate of the template curve can be given by the function in the sample of curves with the highest depth corresponding to the median function, i.e. the deepest curve in the sample,

$$\hat{f}_n = \operatorname*{arg\,max}_{f \in \{f_1, \dots, f_n\}} \mathrm{BD}_{n,L}(f).$$

1.2 Template estimation based on manifold embedding

The Chapter 2 deals with the problem of finding a meaningful template function that represents the common pattern of a sample of curves from a manifold point of view. It is, assuming that functional data lie on, or close enough on, an intrinsically low-dimensional nonlinear submanifold $\mathcal{M}$ with an unknown underlying geometric structure embedding in a high-dimensional space. In other words, the observed functions are modeled as variables with values on a manifold that is nearly isomorphic to a low-dimensional Euclidean space. Although the manifold is unknown, the nice property is that its underlying geometric structure is contained in the sample of observed functions so that it can be reconstructed from the functional data. Below, we provided some preliminary definitions and results concerning the theory of manifolds. For more details, see, for instance, do Carmo [32], Small [90] and Jost [56].

Definition 1.2. Let $\mathcal{M}$ be a topological space with topology $\mathcal{U}$. A bijective mapping $\varphi\colon (U \subset \mathcal{M}) \rightarrow (V \subset \mathbb{R}^d)$ continuous in both directions is called a homeomorphism. It provides that both φ and φ^{-1} are continuous.

Consider a collection of open subsets $\{U_\alpha\}_{\alpha \in A}$ covering the topological space $\mathcal{M}$ (i.e. $\cup_{\alpha \in A} U_\alpha = \mathcal{M}$, where A is an arbitrary index set) with a corresponding collection of homeomorphisms $\varphi_\alpha\colon (U_\alpha \subset \mathcal{M}) \rightarrow (V \subset \mathbb{R}^d)$ called (coordinate) *charts* on $\mathcal{M}$. The family $\{U_\alpha, \varphi_\alpha\}_{\alpha \in A}$ is said to form an *atlas* on $\mathcal{M}$. Note that φ maps every point $p \in U$ to d coordinate points $\varphi(p) = (\varphi^1(p), \ldots, \varphi^d(p)) = (\varphi^1, \ldots, \varphi^d)$, which are considered as local coordinates on the manifold $\mathcal{M}$ when φ is a chart.

Definition 1.3. The set $\mathcal{M}$ together with its atlas $\{U_\alpha, \varphi_\alpha\}_{\alpha \in A}$ is called a topological manifold of dimension d. This is, a topological space that is locally homeomorphic to the Euclidean space (i.e. for each point $p \in \mathcal{M}$ there exists a neighborhood U that is homeomorphic to an open subset of $\mathbb{R}^d$).

Definition 1.4. An atlas $\{U_\alpha, \varphi_\alpha\}_{\alpha \in A}$ on a manifold is called differentiable if all chart transitions

$$\varphi_\beta \circ \varphi_\alpha^{-1}\colon \varphi_\alpha(U_\alpha \cap U_\beta) \rightarrow \varphi_\beta(U_\alpha \cap U_\beta) \quad \text{for all } \alpha, \beta \in A$$

are differentiable of class $\mathcal{C}^\infty$ (i.e. $\varphi_\beta \circ \varphi_\alpha^{-1}$ are diffeomorphisms). If the atlas is differentiable then the topological manifold of dimension d is called a smooth (or differentiable) manifold of dimension d. In other words, a smooth manifold of dimension d is a d-dimensional manifold that is locally diffeomorphic to $\mathbb{R}^d$.

Definition 1.5. Assume that $\mathcal{M}$ is a smooth manifold and let $\gamma(t)$ be a smooth path in $\mathcal{M}$ passing through a point $\gamma_0 = (\gamma^1(0), \ldots, \gamma^d(0))$ at $t = 0$. Suppose that a coordinate system is provided by a chart of $\mathcal{M}$ around γ_0 so that $\gamma(t) = (\gamma^1(t), \ldots, \gamma^d(t))$. The vector space of all tangent vectors to the manifold $\mathcal{M}$ at each given point $\gamma \in \mathcal{M}$ is called the *tangent space* of $\mathcal{M}$ at γ, denoted by $\mathcal{T}_\gamma(\mathcal{M})$. Here, a tangent vector $\dot{\gamma}(t)$ to the path $\gamma(t)$ at γ_0 is defined by the equivalence class of all coordinate paths $\zeta(t)$ in $\mathcal{M}$ such that $\zeta(0) = \gamma_0$ and $\zeta(t)$ is tangent to $\gamma(t)$ at $t = 0$.

Remark 1.4. The tangent vectors in $\mathcal{T}_\gamma(\mathcal{M})$ are spanned by the basis $\partial/\partial\gamma^1, \dots, \partial/\partial\gamma^d$. Thus any tangent vector v in $\mathcal{T}_\gamma(\mathcal{M})$ can be written as $v = \sum_{j=1}^d v^j \partial/\partial\gamma^j$. Hence, for example, the tangent vector $v = \dot\gamma(t)$ can be written as

$$\dot\gamma(t) = \sum_{j=1}^d \dot\gamma^j(t) \frac{\partial}{\partial\gamma^j},$$

where $\dot\gamma^j(t) = \mathrm{d}\gamma^j(t)/\mathrm{d}t$.

The geometric structure of a smooth manifold $\mathcal{M}$ is specified by a Riemannian metric, which is defined by the inner product of tangent vectors in $\mathcal{T}_\gamma(\mathcal{M})$ given by

$$\langle v, w\rangle = \left\langle \sum_{j=1}^d v^j \partial_j, \sum_{k=1}^d w^k \partial_k \right\rangle = \sum_{j=1}^d \sum_{k=1}^d g_{jk}(\gamma) v^j w^k = v^\top G(\gamma) w,$$

where ∂_j is a common shorter notation for $\partial/\partial\gamma^j$, and $G(\gamma) = \{g_{jk}(\gamma)\} = \{\langle\partial_j, \partial_k\rangle\}$ is positive definite symmetric matrix for all $\gamma \in \mathcal{M}$.

Definition 1.6. A smooth manifold $\mathcal{M}$ endowed with a Riemannian metric $\langle\cdot,\cdot\rangle$ is called a Riemannian manifold.

Definition 1.7. A mapping $g\colon \mathcal{M} \to \mathcal{S}$ of a d-dimensional manifold $\mathcal{M}$ into a m-dimensional manifold $\mathcal{S}$ is an embedding if it is a smooth homeomorphism to its image $g(\mathcal{M}) \subset \mathcal{S}$. If g is an embedding then $g(\mathcal{M})$ is an embedded submanifold of $\mathcal{S}$.

Now, with definitions given above, a distance measure between two points along the manifold may be established by mean the corresponding shortest path. This path is called a geodesic, and the length of the path is known as the geodesic distance. In local regions of the manifold the Euclidean distance converges to the geodesic distance as the radius of the region decreases, and for "far" points on a well-sampled Euclidean manifold, the geodesic may be approximated through graphical methods. The corresponding geodesic distance between two points on the manifold $\mathcal{M}$ is defined as the minimum length between all smooth curves (paths) connecting the two points. Formally, it is defined as:

Definition 1.8. Let $\mathcal{M}$ be a connected Riemannian manifold, i.e. $\mathcal{M}$ is pathwise connected in the sense that for any two points in $\mathcal{M}$ there exists a smooth path $\gamma(t)\colon [a,b] \to \mathcal{M}$ such that $\gamma(a) = p$ and $\gamma(b) = q$. The geodesic distance between points p and q is given by

$$\delta(p,q) = \delta\big(\gamma(a), \gamma(b)\big) = \inf_\gamma L\big(\gamma(t)\big),$$

where $L(\gamma)$ is the arc-length along the curve $\gamma(t)$ from a to b,

$$\begin{aligned} L\big(\gamma(t)\big) = \int_a^b \|\dot{\gamma}(t)\| \, \mathrm{d}t &= \int_a^b \left(\langle\dot{\gamma}(t), \dot{\gamma}(t)\rangle\right)^{1/2} \mathrm{d}t \\ &= \int_a^b \left(\sum_{j=1}^{d} \sum_{k=1}^{d} g_{jk}\left(\gamma(t)\right) \dot{\gamma}^j(t) \dot{\gamma}^k(t) \right)^{1/2} \mathrm{d}t. \end{aligned}$$

The Figure 1.1 by Tenenbaum et al. [96], corresponding to the well-known two-dimensional "Swiss roll" manifold, illustrates how the geodesic distance captures the geometry of manifolds. Indeed for two arbitrary distant points on the nonlinear manifold, the geodesic distance (solid curved line) reflects appropriately the intrinsic nonlinear geometric structure of the manifold taking into account its curvature, unlike the Euclidean distance (dashed line) which obscures the intrinsic manifold structure. Thus the next result is stated.

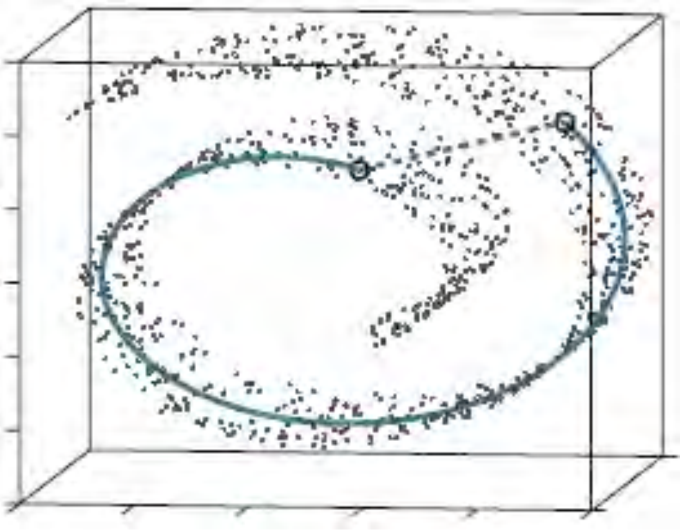

Figure 1.1: "Swiss roll" manifold

Theorem 1.1. *The topology on $\mathcal{M}$ induced by the distance function δ coincides with the original manifold topology of $\mathcal{M}$.*

Proof. See Jost [56]. ■

Under the assumed geometric framework, any statistical analysis of random objects lying in a smooth manifold should be carried out carefully. Indeed, new definitions, for example, for probability density and distribution functions, measures of location and dispersion, etc., must be revised due to these involve integrals with respect to the manifold. In particular, the problem of template curve estimation is then equivalent to consider a central location of data with respect to the intrinsic geodesic distance associated to the manifold. Following Pennec [75], a general notion of central location value for a probability measure Q on a metric space $\mathcal{M}$ with metric δ is through the mean deviation at order p or Fréchet function.

Definition 1.9. Let $(\mathcal{M}, \delta)$ be a metric space and let $p > 0$ be a given real number. For a given probability measure Q defined on the Borel σ-field of $\mathcal{M}$, the Fréchet function of Q is defined as

$$F_p(\mu) = \int_{\mathcal{M}} \delta^p(X, \mu) Q(\mathrm{d}x), \qquad \mu \in \mathcal{M}.$$

If the function $F_p(\mu)$ has a unique minimizer, then it is called the Fréchet central point at order p, and when $\mathcal{M}$ is a geodesically connected and complete Riemannian manifold, it is refereed as the intrinsic central point at order p, denoted by $\mu_{\mathrm{I}}^{p}(Q)$. For instance, when p takes the 1 and 2 values the intrinsic median and mean are obtained, respectively. Also for $p \to \infty$, the "barycenter" of the distribution support is obtained. Statistical analysis of a probability measure Q on a differentiable manifold has diverse applications in morphometrics, medical diagnostics and image analysis (see, e.g., Small [90], Bhattacharya and Patrangenaru [7], and Pennec [75]).

Bhattacharya and Patrangenaru [7] prove that every probability measure Q on $\mathcal{M}$ has a unique intrinsic mean, provided $F_2(\mu)$ is finite for some μ. Furthermore, they show the strong consistency of the Fréchet sample mean toward the Fréchet mean. A central limit theorem for Fréchet sample means is derived by Bhattacharya and Patrangenaru [8], leading to an asymptotic distribution theory of intrinsic sample means on Riemannian manifolds. The existence and uniqueness of intrinsic medians $\mu_{\mathrm{I}}^{1}(Q)$, and the strong consistency of Fréchet sample medians in compact Riemannian manifolds have been recently proved by Yang [111].

Chapter 2 deals particularly with the intrinsic median $\mu_{\mathrm{I}}^{1}(Q)$ in order to obtain a robust estimate for the template $f \in \mathcal{M}$, following Koenker [62]. The corresponding empirical intrinsic median is

$$\widehat{\mu}_{\mathrm{I}}^{1} = \arg\min_{\mu \in \mathcal{M}} \sum_{i=1}^{n} \delta\left(X_i, \mu\right). \tag{1.2}$$

However, this estimator depends on the unobserved manifold $\mathcal{M}$ and on its underlying geodesic distance δ that need to be approximated. One method for this goal, is by mean the Isometric featuring mapping –Isomap– algorithm developed by Tenenbaum et al. [96]. It is one of the most known and applied procedures for manifold learning in high-dimensional data analysis or nonlinear dimensionality reduction. Manifold learning consists in finding a low-dimensional representation of the data (i.e. to "learn" a manifold from the data points). Formally, it assumes that observed data lie on a d-dimensional manifold $\mathcal{M}$ embedded into a high-dimensional space $\mathbb{R}^m$ with $d << m$. Therefore, the main problem consists in mapping a given m-dimensional data set $X_i \in \mathcal{M} \subset \mathbb{R}^m$ into a d-dimensional data set $Y_i \in \mathbb{R}^d$, $i = 1, \dots, n$ preserving the intrinsic (local or global) geometry of the original manifold as much as possible.

There are other algorithms developed over the last decade for nonlinear dimensionality reduction and data representation. Some of these are, for instance, the Local Linear Embedding –LLE– (Roweis and Saul [82]), Laplacian Eigenmap (Belkin and Niyogi [5]), Hessian Eigenmap (Donoho and Grimes [33]), Diffusion maps (Coifman and Lafon [21]) and Local Tangent Space Alignment –LTSA– (Zhang and Zha [116]). For surveys on manifold algorithms see Cayton [19], Lee and Verleysen [66], and Izenman [54]. All of these procedures, except the diffusion map, are based on graph-based algorithms summarized in three steps: *1*) find k-nearest neighbors (or ϵ-neighborhood) for each point X_i, *2*) estimate local properties of the manifold by looking at found neighborhoods, and *3*) find a global embedding that preserves the properties found in the previous step.

Isometric featuring mapping –Isomap–

The Isomap algorithm relies on the assumptions that the manifold is compact and convex (holes are not allowed), and that there exists an isometric coordinate chart $\varphi \colon \mathcal{M} \to \mathbb{R}^d$, i.e. a chart that preserves the distances,

$$\|\varphi(X_i) - \varphi(X_{i'})\| = \delta(X_i, X_{i'}), \quad \text{for all } \; i \neq i' \in \{1, \dots, n\}\,.$$

The algorithm approximates the unknown geodesic distance δ between any pairs of points in $\mathcal{M}$ in terms of its shortest path distance between the points on a proximity graph $\mathcal{G}$ constructed from the data points $X_i \in \mathbb{R}^m$, $i = 1, \dots, n$. Few definitions related to some concepts of graph theory are given in the Appendix

Definition 1.10. Given n points $X_1, \dots, X_n$ with $X_i \in \mathbb{R}^m$. A proximity graph is a (undirected) weighted graph $\mathcal{G}$ with vertices $\{X_1, \dots, X_n\}$. If there exists an edge between X_i and $X_{i'}$, then the weight of this edge is given by a distance $d(X_i, X_{i'})$ function (e.g. the Euclidean distance). There are two popular proximity graphs:

- k-nearest neighbors graph (k-rule). Let $1 \leq k \leq n$ be an integer. In a k-nearest neighbors graph $\mathcal{G}$, there is an edge between X_i and $X_{i'}$ if and only if $X_{i'}$ is one of the k closest neighbors of X_i.

- ϵ-ball graph (ϵ-rule). Let $\epsilon > 0$ be a real number. In a ϵ-ball graph $\mathcal{G}$, there is an edge between X_i and $X_{i'}$ if and only if $X_{i'}$ lies in a ball of radius ϵ with center X_i, $B(X_i, \epsilon)$.

The Isomap method consists of three steps summarized in the Algorithm 1.1. As we will see, Isomap can be considered as an extension of the Multidimensional Scaling method.

Based on the Euclidean distances $d(X_i, X_{i'}) = \|X_i - X_{i'}\|$ between points X_i and $X_{i'}$, a weighted neighborhood graph $\mathcal{G} = \mathcal{G}(\mathcal{X}, E)$ with set of nodes $\mathcal{X} = \{X_1, \dots, X_n\}$ and set of edges $E = \{d(X_i, X_{i'})\}$ is constructed according to a k-rule or ϵ-rule. The choice of the parameter k or ϵ controls the neighborhood size and therefore the success of algorithm.

In the next step, the unknown geodesic distances $\delta(X_i, X_{i'})$ between all pairs of points in the manifold $\mathcal{M}$ are estimated by computing the matrix of graph distances $D_{\mathcal{G}} = \{d_{\mathcal{G}}(X_i, X_{i'})\}$, which are the shortest path distances between all pairs of points in the graph $\mathcal{G}$. Algorithms for computing the graph distances between every pair of vertices in a graph are the Floyd's and Dijkstra's algorithms (see Lee and Verleysen [66]). A description of the Dijkstra's algorithms, which is used in Chapter 2, is also provided in the Appendix.

Finally, the embedding of the data in a d-dimensional Euclidean space preserving, as much as possible, the intrinsic geometry of the estimated manifold is obtained. This is done by applying the classical Multidimensional Scaling $-$cMDS$-$ method. It is based on the connection between the space of Euclidean distance matrices and the space of inner product (Gram) matrices, which permits convert an Euclidean distance matrix into a Gram matrix. For Isomap, the input to cMDS is the matrix of pairwise squared geodesic distances $S = \left\{D^2_{\mathcal{G}}(i, i')\right\}$. Thus, cMDS converts the matrix of geodesic distances $D_{\mathcal{G}}$ into a $n \times n$ Gram matrix

$$\tau(D_{\mathcal{G}}) = -\frac{1}{2} HSH,$$

where $H = I_n - n^{-1} 1_n 1_n^\top$ is the centering (positive semidefinite) matrix, where 1_n is the n-dimensional vector of ones. More details on MDS can be found in Cox and Cox [22] or Izenman [54].

The embedding coordinates in the d-dimensional Euclidean space are obtained by computing the spectral decomposition of $\tau(D_{\mathcal{G}})$, $\tau(D_{\mathcal{G}}) = V \Lambda V^\top$, and choosing the d eigenvectors of V, $V_d = (v_1, \dots, v_d)$ corresponding to the d largest eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d$, i.e. the embedding is given by

$$\widehat{Y} = (\widehat{Y}_1, \dots, \widehat{Y}_n) = \Lambda_d^{1/2} V_d^\top = \left(\lambda_1^{1/2} v_1, \dots, \lambda_d^{1/2} v_d\right).$$

One of the most important stages of the Isomap algorithm involves the estimation of the geodesic distance δ between pair of data points in $\mathcal{M}$ based upon the graph distance $d_{\mathcal{G}}$ with respect to the graph $\mathcal{G}$. Bernstein et al. [6] show that these two distance metrics approximate asymptotically each other arbitrarily closely under some sampling condition and some conditions on the graph $\mathcal{G}$, and therefore the

Algorithm 1.1 Isomap algorithm

```
Require: $X_i \in \mathbb{R}^m$, $i = 1, \dots, n$ and $k$
Ensure: $\widehat{Y}_i \in \mathbb{R}^d$, $i = 1, \dots, n$
1: Find the $k$-nearest neighbors (or $\epsilon$-neighborhoods) for each $X_i$ based on the Euclidean distances $d(X_i, X_{i'}) = \|X_i - X_{i'}\|$ between $X_i$, and $X_{i'}$, and construct a graph $\mathcal{G} = \mathcal{G}(\mathcal{X}, E)$ with sets of vertices and edges $\mathcal{X} = \{X_1 \dots, X_n\}$ and $E = \{d(X_i, X_{i'})\}$, respectively.
2: Obtain the matrix of shortest path distances $D_{\mathcal{G}} = d_{\mathcal{G}}(X_i, X_{i'})$ between all pairs of points in $\mathcal{G}$ applying the Floyd's or Dijkstra's algorithm.
3: Obtain $\widehat{Y} = (\widehat{Y}_1, \dots, \widehat{Y}_n)$ by applying the cMDS method to $S = \left\{D^2_{\mathcal{G}}(i, i')\right\}$.
```

Isomap recovers the true structure of manifolds that are isometric to a convex subset of the Euclidean space. The next theorem due to Bernstein et al. [6] gives conditions for this convergence.

Theorem 1.2. *Let $\mathcal{M}$ be a compact submanifold of $\mathbb{R}^m$ and let $X_i \in \mathcal{M}$, $i = 1, \dots, n$ be a set of data points. Assume that a graph $\mathcal{G}$ is constructed from X_i, and there are positive real numbers $\lambda_1, \lambda_2 < 1$, ϵ_{min}, ϵ_{max} and τ. Suppose that:*

1. *$\mathcal{G}$ contains all edges $X_iX_{i'}$ of length $\|X_i - X_{i'}\| \leq \epsilon_{min}$,*
2. *All edges of $\mathcal{G}$ have length $\|X_i - X_{i'}\| \leq \epsilon_{max}$,*
3. *For every point $p \in \mathcal{M}$ there is a point X_i for which $\delta(p, X_i) \leq \tau$,* and*
4. *$\mathcal{M}$ is geodesically convex, i.e. any two points $X_i, X_{i'} \in \mathcal{M}$ are connected by $\delta(X_i, X_{i'})$.*

Then provided that:

5. *$\epsilon_{max} < s_0$, where s_0 is the minimum branch separation of $\mathcal{M}$,†*
6. *$\epsilon_{max} \leq (2/\pi) r_0 \sqrt{24\lambda_1}$, where r_0 is the minimum radius of curvature of $\mathcal{M}$,‡*
7. *$\tau \leq \lambda_2 \epsilon_{min}/4$,*

it follows that the inequalities

$$(1 - \lambda_1)\delta(X_i, X_{i'}) \leq d_{\mathcal{G}}(X_i, X_{i'}) \leq (1 + \lambda_2)\delta(X_i, X_{i'})$$

are valid for all $X_i, X_{i'} \in \mathcal{M}$.

Proof. Bernstein et al. [6]. ∎

* This condition is referred as the "τ-sampling condition" in $\mathcal{M}$ for the data set $X_i \in \mathbb{R}^m$, $i = 1, \dots, n$.

† The minimum branch separation s_0 is defined as the largest positive number for which $\|X_i - X_{i'}\| < s_0$ implies $\delta(X_i, X_{i'}) \leq \pi r_0$ for all $X_i, X_{i'} \in \mathcal{M}$.

‡ The minimum radius of curvature r_0 is defined by $r_0^{-1} = \max_{\gamma(t)} \{\|\ddot{\gamma}(t)\|\}$, where the path $\gamma \colon \mathcal{I} \to \mathbb{R}^m$ varies over all unit-speed geodesics in $\mathcal{M}$.

Local linear embedding –LLE–

The local linear embedding algorithm (Roweis and Saul [82]) is similar to the Isomap algorithm, but it attempts to preserve local neighborhood information on the Riemannian manifold without estimating the geodesic distances. Thus LLE is viewed as a local method rather than as a global one as is the Isomap. The philosophy of the method comes from visualizing a manifold as a collection of overlapping coordinate patches. If the neighborhood sizes are small and the manifold is sufficiently smooth, then these patches will be approximately linear.

The method starts representing each point X_i as a linear combination of its k nearest neighbors $N_i = N_{X_i}$, $X_i = \sum_{i' \in N_i} w_{ii'} X_{i'}$, where $w_{ii'}$ is a scalar weight with constraints $\sum_{i'} w_{ii'} = 1$, and $w_{ii'} = 0$ if $i' \notin N_i$, reflecting that the method is local. The weight (sparse) $n \times n$ matrix W is estimated by solving the constrained least squares fit

$$\widehat{W} = \arg\min_{W} \sum_{i=1}^{n} \left\| X_i - \sum_{i' \in N_i} w_{ii'} X_{i'} \right\|^2$$

subject to $\sum_{i'} w_{ii'} = 1$, and $w_{ii'} = 0$ if $i' \notin N_i$.

By mean the Lagrange multipliers method (see Izenman [54] for the details), the estimated local weights for each point are given by

$$\hat{w}_i = \frac{C^{-1} 1_n}{1_n^\top C^{-1} 1_n} = \frac{\sum_{i*} C^{-1}_{i'i^*}}{\sum_{i'} \sum_{i^*} C^{-1}_{i'i^*}}, \quad i = 1, \ldots, n,$$

where $C = \{C_{i'i^*}\}$ is a symmetric and positive semi-definite $n \times n$ local matrix with $C_{i'i^*} = (X_i - X_{i'})^\top (X_i - X_{i^*})$ for $i', i^* \in N_i$.

Finally the global embedding $d \times n$ matrix Y is found by solving the problem

$$\min_{Y} \sum_{i=1}^{n} \left\| Y_i - \sum_{i'=1}^{n} \hat{w}_{ii'} Y_{i'} \right\|^2 = \min_{Y} \operatorname{trace} \left\{ Y \left(I_n - \widehat{W} \right)^\top \left(I_n - \widehat{W} \right) Y^\top \right\},$$

subject to the constraints $\sum_i Y_i = Y 1_n = 0$, which centers the embedding on the origin, and $\sum_i Y_i Y_i^\top = Y Y^\top = I_d$, which forces the solution to be of range d.

The unique global minimum of the objective function is given by the eigenvectors $V = (v_0, v_1, \ldots, v_d)$ corresponding to the smallest $d + 1$ eigenvalues $\Lambda = (\lambda_0, \lambda_1, \ldots, \lambda_d)$ of the sparse, symmetric, and positive semi-definite $n \times n$ matrix $\left(I_n - \widehat{W}\right)^\top \left(I_n - \widehat{W}\right)$. The smallest eigenvalue is zero with corresponding eigenvector 1. Thus the final coordinate of each point is identical. To avoid this degenerate dimension, the d-dimensional embedding is given by the smallest non-constant d eigenvectors, obtaining that

$$\widehat{Y} = (\widehat{Y}_1, \ldots, \widehat{Y}_n) = (v_1, \ldots, v_d)^\top.$$

The LLE method is summarized in the Algorithm 1.2.

Algorithm 1.2 Local linear embedding algorithm

```
Require: $X_i \in \mathbb{R}^m$, $i = 1, \dots, n$, $k$ and $d$
Ensure: $\widehat{Y}_i \in \mathbb{R}^d$, $i = 1, \dots, n$
1: Find the $k$-nearest neighbors $N_i$ for each $X_i$.
2: Compute the weight matrix $\widehat{W}$, where each column is
   $$\hat{w}_i = \frac{C^{-1} 1_n}{1_n^\top C^{-1} 1_n} = \frac{\sum_{i^*} C^{-1}_{i' i^*}}{\sum_{i'} \sum_{i^*} C^{-1}_{i' i^*}}, \quad i = 1, \dots, n,$$
   where $C = \{C_{i' i^*}\}$ with $C_{i' i^*} = (X_i - X_{i'})^\top (X_i - X_{i^*})$ for $i', i^* \in N_i$.
3: Obtain the smallest non-constant $d$ eigenvectors $v_1, \dots, v_d$ associated to the nonzero eigenvalues $\lambda_1, \dots, \lambda_d$ of the matrix $\big(I_n - \widehat{W}\big)^\top \big(I_n - \widehat{W}\big)$, and set $\widehat{Y} = (v_1, \dots, v_d)^\top$.
```

Laplacian eigenmap

The Laplacian eigenmaps algorithm is closely related to local linear embedding although, it tackles the problem in a different way. The method relies on the spectral graph theory, and particularly on the concept of Laplacian operator of a graph. To begin with, given a graph $\mathcal{G}$ and an edge weight $n \times n$ matrix $W = \{w_{ii'}\}$, the (unnormalized) graph Laplacian matrix is defined as

$$L = D - W,$$

where $D = \operatorname{diag}(\sum_{i'} w_{ii'})$. L is a symmetric, positive semidefinite matrix that provides a natural measure on the graph vertices (Belkin and Niyogi [5]).

The eigenvalues and eigenvectors of L provide information about whether the graph is complete or connected. Therefore, it can be used to capture the local information on the manifold. Belkin and Niyogi [5] propose to use a Gaussian heat kernel with scale parameter σ to define the entries of the local adjacency matrix W, defined as

$$W = \{w_{ii'}\} = \begin{cases} \exp\left\{-\|X_i - X_{i'}\|\right\}^2 / \sigma & \text{if } X_{i'} \in N_i \\ 0 & \text{if } X_{i'} \notin N_i. \end{cases}$$

Note that σ is an additional parameter that has to be taking into account in the algorithm, join to the parameters k and d.

The low-embedding $d \times n$ matrix Y is found in a similar way to the LLE method by solving the problem

$$\min_Y \sum_i \sum_{i'} w_{ii'} \|Y_i - Y_{i'}\|^2 = \min_Y \operatorname{trace}\left\{Y L Y^\top\right\}$$

subject to the constrain $YDY^\top = I_d$, which forces Y to be of full dimensionality, i.e. prevents to have a solution onto a subspace of fewer than $d-1$ dimensions.

This minimization problem tries to ensure that if X_i and $X_{i'}$ are close then Y_i and $Y_{i'}$ are close as well. This problem reduces to solving the generalized eigenvalue problem

$$Lv = \lambda Dv.$$

As in the LLE method, the d-dimensional embedding is given by the d smallest non-constant eigenvectors, $\widehat{Y} = (v_1, \dots, v_d)^\top$. The method is summarized in the Algorithm 1.3

Algorithm 1.3 Laplacian eigenmap algorithm

```
Require: $X_i \in \mathbb{R}^m$, $i = 1, \dots, n$, $k$, $\sigma$ and $d$
Ensure: $\widehat{Y}_i \in \mathbb{R}^d$, $i = 1, \dots, n$
1: Find the $k$-nearest neighbors $N_i$ for each $X_i$.
2: Compute the weight matrix $W = \{w_{ii'}\}$, where each entry is
     $$w_{ii'} = \begin{cases} \exp\{-\|X_i - X_{i'}\|\}^2/\sigma & \text{if } X_{i'} \in N_i \\ 0 & \text{if } X_{i'} \notin N_i. \end{cases}$$
3: Obtain non-constant eigenvectors $v_1, \dots, v_d$ associated to non-zero $d$ eigenvalues $\lambda_1, \dots, \lambda_d$ of the generalized eigenequation $Lv = \lambda Dv$, where $L = D - W$ with $D = \mathrm{diag}(\sum_{i'} w_{ii'})$, and set $\widehat{Y} = (v_1, \dots, v_d)^\top$.
```

Hessian eigenmap

This method is closely related to LLE and Laplacian eigenmap algorithms, substituting a quadratic form based on the Hessian instead of one based on the Laplacian. The method has been proposed for recovering manifolds of high-dimensional libraries of articulated images where the underlying parameter space is often not convex (Donoho and Grimes [33]).

The method assumes that there is a smooth mapping $\varphi\colon \Theta \to \mathbb{R}^m$ such that the smooth manifold is given by $\mathcal{M} = \varphi(\Theta) \subset \mathbb{R}^m$ (i.e. the manifold of articulated images). Thus, given $X_i \in \mathcal{M}$, $i = 1, \dots, n$, the aim consists in recovering the mapping φ and the parameter points θ_i. However, this problem is ill-posed, because if ψ is one solution, and $\zeta : \mathbb{R}^d \to \mathbb{R}^d$ is a morphing of $\mathbb{R}^d$, then $\psi \circ \zeta$ is also a solution. Thus additional hypotheses are required. Donoho and Grimes [33] replace the convexity and isometry assumptions of Isomap by weaker conditions: *i*) the parameter space $\Theta \subset \mathbb{R}^d$ is open and connected, and *ii*) in a small neighborhood of each point in $\mathcal{M}$, the geodesic distances to nearby points $X_{i'} \in \mathcal{M}$ are equal to the Euclidean distances between the corresponding parameter points θ_i and $\theta_{i'}$,

$$\delta(X_i, X_{i'}) = \|\theta_i - \theta_{i'}\|, \quad \text{for all } X_i = \varphi(\theta_i) \text{ and } X_{i'} = \varphi(\theta_{i'}).$$

The theoretical solution under this framework relies on the properties of a quadratic form $\mathcal{H}(f)$ defined on functions $f\colon \mathcal{M} \to \mathbb{R}$, expressed as

$$\mathcal{H}(f) = \int_{\mathcal{M}} \|H_f^{\tan}(X)\|_F^2 \mathrm{d}X,$$

where $\|H\|_{\mathrm{F}}^2 = \sum_i \sum_{i'} H_{ii'}^2$ is the Frobenius norm of a square matrix.

This form is based on the tangent space $\mathcal{T}_X(\mathcal{M})$ at $X \in \mathcal{M}$. Viewing it as a subspace of $\mathbb{R}^m$, an orthonormal coordinate system can be associated to each tangent space $\mathcal{T}_X(\mathcal{M}) \subset \mathbb{R}^m$ through the inner product on $\mathbb{R}^m$. Let N_X be a neighborhood of X such that each point $X_{i'} \in N_X$ has a unique closest point $\xi_{i'} \in \mathcal{T}_X(\mathcal{M})$ (it is a point in N_X with local tangent coordinates $\xi = \xi(X) = (\xi^1(X), \dots, \xi^d(X))$. Thus if the point $X_{i'} \in N_X$ has local coordinates $\xi \in \mathbb{T}^d$, then the rule $g(\xi) = f(X)$ defines a twice continuously differentiable function $g\colon U \to \mathbb{R}$, where U is a neighborhood of $0 \in \mathbb{R}^m$. Hence the $d \times d$ tangent Hessian matrix $H_f^{\tan}(X)$, that measures the "curviness" of f at $X \in \mathcal{M}$, is given by

$$H_f^{\tan}(X) = \left\{ \left. \frac{\partial^2 g(\xi)}{\partial \xi^i \partial \xi^{i'}} \right|_{\xi=0} \right\}.$$

Finally, Donoho and Grimes [33] show that $\mathcal{H}(f)$ has a $(d+1)$-dimensional null space consisting of a constant function and a d-dimensional space of functions spanned by the original isometric coordinates. Hence, the isometric coordinates can be recovered, up to a rigid motion, from the null space of $\mathcal{H}(f)$. The steps to get a discrete approximation of $\mathcal{H}$ and obtain the embedding d-dimensional Euclidean space are given in the Algorithm 1.4.

Diffusion map

This method relies on a diffusion processes framework for finding meaningful geometric descriptions of the data. Coifman and Lafon [21] show that eigenfunctions of Markov transition probability matrices can be used to construct coordinates that generate efficient representations of complex geometric structures. These coordinates are obtained by defining a family of mappings, known as diffusion maps, that embed the data points into a Euclidean space.

Based on a Gaussian heat kernel (called also isotropic Gaussian kernel) $W = \{w_{ii'}\}$, $w_{ii'} = \exp\left\{-\|X_i - X_{i'}\|\right\}^2/\sigma$, a diffusion process is constructed by renormalizing the symmetric and positive semi-definite matrix W as $A = D^{-1}W$,

Algorithm 1.4 Hessian eigenmap algorithm

```
Require: $X_i \in \mathbb{R}^m$, $i = 1, \dots, n$, $k$ and $d$, with $k > d$.
Ensure: $\widehat{Y}_i \in \mathbb{R}^d$, $i = 1, \dots, n$
1: Find the $k$-nearest neighbors $N_i$ for each $X_i$.
2: Form an $n \times k$ matrix $M_i = (X_{i_1} - \bar{X}_i, \dots, X_{i_k} - \bar{X}_i)$ with $\bar{X}_i = \sum_{i' \in N_i} X_{i'}$, and obtain the tangent coordinates through a singular value decomposition of $M_i$, $M_i = U_i \Sigma V_i^\top$, where the $k \times k$ matrix $V_i$ approximates the tangent space at $X_i$. Next, construct the $k \times (1 + d + d(d+1)/2)$ matrix
   $$Z_i = \left(1, V_{i,1} \dots, V_{i,d}, V_{i,1}^2, \dots, V_{i,d}^2, V_{i,1}V_{i,2}, \dots, V_{i,d-1}V_{i,d}\right),$$
   where $V_{i,j}V_{i,j'}$ is the pointwise (Hadamard) product between $V_{i,j}$ and $V_{i,j'}$. Perform a Gram-Schmidt orthogonalization on $Z_i$,
   $$\widetilde{Z}_i = \left(1, \widetilde{v}_{i,1} \dots, \widetilde{v}_{i,d}, \widetilde{w}_{i,1}, \dots, \widetilde{w}_{i,d(d+1)/2}\right),$$
   and obtain the tangent Hessian matrix given by the transpose of last $d(d+1)/2$ orthonormal columns of $\widetilde{Z}_i$, $\widehat{H}_i = \left(\widetilde{w}_{i,1}, \dots, \widetilde{w}_{i,d(d+1)/2}\right)^\top$.
3: Obtain non-constant eigenvectors $v_1, \dots, v_d$ associated to non-zero $d$ eigenvalues $\lambda_1, \dots, \lambda_d$ of the $m \times m$ matrix,
   $$\widehat{\mathcal{H}}_{jj'} = \sum\nolimits_i \sum\nolimits_{i'} \left(\widehat{H}_i\right)_{i'j} \left(\widehat{H}_i\right)_{i'j'}.$$
   Finally set $\widehat{Y} = (v_1, \dots, v_d)^\top$.
```

where $D = \operatorname{diag}(\sum_{i'} w_{ii'})$, and satisfying $\sum_{i'=1}^n A_{ii'} = 1$ for all $i \in \{1, \dots, n\}$. Hence the matrix A constitutes a row Markov matrix with transition probabilities $A_{ii'} = \mathbb{P}(\mathcal{X}_{t+1} = X_i | \mathcal{X}_t = X_{i'}) = P(X_i, X_{i'}) \geq 0$.

By using the spectral decomposition of the matrix A, $A = \Phi \Lambda \Psi^\top$, Coifman and Lafon [21] prove that it induces a diffusion distance at time $t \geq 0$, defined as

$$d_t(X_i, X_{i'}) = \|\Phi_t(X_i) - \Phi_t(X_{i'})\| = \left(\sum_{l \geq 1} \lambda_l^{2t} \Big(\phi_l(X_i) - \phi_l(X_{i'}) \Big)^2 \right)^{1/2},$$

where $\Phi_t = (\lambda_1^t \phi_1(X_i), \lambda_2^t \phi_2(X_i), \dots)$ is called the diffusion map.

The diffusion distance can be approximated by retaining a finite number of terms by considering a truncate diffusion map with first d eigenfunctions, which embeds the data points into $\mathbb{R}^d$ in an approximately isometric fashion, with respect to the diffusion distance. Therefore this provides an embedding $\widehat{Y}_i = \Phi_{t,d}(X_i) = (\lambda_1^t \phi_1(X_i), \dots, \lambda_d^t \phi_d(X_i))$ for each $i = 1, \dots, n$.

Coifman and Lafon [21] consider a generalization of the method proposing a general family of normalizations and their corresponding diffusions based on a specific anisotropic kernel. The steps of the method for this generalization are gathered in the Algorithm 1.5, where the definitions given above are recovered by setting the α value, stated in the algorithm below, to 0.

Algorithm 1.5 Diffusion map algorithm

```
Require: $X_i \in \mathbb{R}^m$, $i = 1, \dots, n$, $\sigma$, $\alpha \geq 0$.
Ensure: $\widehat{Y}_i \in \mathbb{R}^d$, $i = 1, \dots, n$
1: Compute a Gaussian kernel $W = \{w_{ii'}\}$, with $w_{ii'} = \exp\left\{-\|X_i - X_{i'}\|\right\}^2 / \sigma$.
2: Construct the family of kernels:
       $w^{(\alpha)}_{ii'} = \frac{w_{ii'}}{p_i^\alpha p_{i'}^\alpha}, \quad \text{where } p_i = \sum_{i'} w_{ii'}.$
   Form the Markov matrix (anisotropic transition kernel)
       $A^{(\alpha)} = \left(D^{(\alpha)}\right)^{-1} W^{(\alpha)}, \quad \text{where } D^{(\alpha)} = \operatorname{diag}\left(\sum_{i'} w^{(\alpha)}_{ii'}\right).$
3: Compute the singular value decomposition of the matrix $A^{(\alpha)}$, $A^{(\alpha)} = \Phi \Lambda \Psi^\top$,
   and determine the embedding using the first $d$ eigenvectors and eigenvalues as,
   $\widehat{Y}_i = \Phi_{t,d}(X_i) = (\lambda_1^t \phi_1(X_i), \dots, \lambda_d^t \phi_d(X_i))$, for each $i = 1, \dots, n$.
```

Local tangent space alignment –LTSA–

The goal of this method is to preserve the local coordinates of the data points in the neighborhood with respect to the tangent space at each data point, which provides a low-dimensional linear approximation of the local geometric structure of $\mathcal{M}$. Those local tangent coordinates are aligned in the low-dimensional space by different local affine transformations to obtain a global coordinate system. Zhang and Zha [116] consider the problem of nonlinear dimensionality reduction in a parameterized d-dimensional manifold $\mathcal{M} = \varphi(\Theta)$ defined by a unknown mapping $\varphi\colon \Theta \subset \mathbb{R}^d \to \mathbb{R}^m$. They assume that points are sampled with noise from $\mathcal{M}$, $X_i = \varphi(\theta_i) + \epsilon_i$, $i = 1, \dots, n$, where ϵ_i is the noise.

To recover the θ_i's from X_i's, they approximate the set of k-nearest neighbors N_i of each X_i using a d-dimensional (affine) linear subspace

$$X_{i_{i'}} \approx \bar{X}_i + Q_i \varrho^{(i)}_{i'}, \qquad i' \in N_i = \{1, \dots, k_i\},$$

where $\bar{X}_i = \sum_{i' \in N_i} X_{i'}$ and $\varrho^{(i)}_{i'}$ are the local coordinates of $X_{i_{i'}}$'s associated with the orthonormal $m \times d$ matrix Q_i. The optimal fit is determined by solving the problem:

$$\min_{c, Q, \varrho_{i'}} \sum_{i' \in N_i} \|X_{i_{i'}} - (c + Q\varrho_{i'})\|_2^2 \quad \text{subject to} \quad Q^\top Q = I_d.$$

The optimal solution is given by the singular value decomposition of the centered matrix $X^{(i)} = (X_{i_1} - \bar{X}_i, \dots, X_{i_k} - \bar{X}_i)$ with $X^{(i)} = U_i \Sigma_i V_i^\top$. The optimal $\widehat{Q}_i$ is then given by the matrix of the d left eigenvectors corresponding to the d largest eigenvalues, $\widehat{Q}_i = (U_{i1}, \dots, U_{id})$, and the corresponding optimal $\widehat{\varGamma}_i = \left(\hat{\varrho}^{(i)}_{i_1}, \dots, \hat{\varrho}^{(i)}_{i_{k_i}}\right)$ with $\hat{\varrho}^{(i)}_{i'} = \widehat{Q}_i^\top (X_{i'} - \bar{X}_i)$.

Finally, a local affine transformation (linear alignment) L_i is postulated such that, in each neighborhood, the corresponding global parameter vectors $\Theta_i = (\theta_{i_1}, \dots, \theta_{i_{k_i}})$ are represented in terms of the local ones $\widehat{\varGamma}_i = \left(\hat{\varrho}^{(i)}_{i_1}, \dots, \hat{\varrho}^{(i)}_{i_{k_i}}\right)$. To preserve the local geometry in the low-dimensional space the optimal parameters have to solve

$$\min_{\Theta_i, L_i} \sum_i \| \Theta_i (I_k - k^{-1} 1 1^\top) - L_i \widehat{\varGamma}_i \|_F^2 = \min_{\Theta} \operatorname{trace} \left\{ \Theta \widetilde{\varGamma} \Theta^\top \right\},$$

subject to $\Theta \Theta^\top = I_d$, where $\widetilde{\varGamma} = \sum_i S_i W_i W_i^\top S_i^\top$ is the $n \times n$ alignment matrix with S_i the 0-1 selection matrix such that $\Theta_i = \Theta S_i$, and $W_i = (I_k - k^{-1} 1 1^\top)(I_k - \widehat{\varGamma}_i^{+} \widehat{\varGamma}_i)$, were $\widehat{\varGamma}_i^\dagger$ is the Moore-Penrose pseudoinverse of $\widehat{\varGamma}_i$.

The optimal $\widehat{\Theta} = (\hat{\theta}_1, \dots, \hat{\theta}_n)$ is given by the d eigenvectors of $\widetilde{\varGamma}$ corresponding to the eigenvalues $\lambda_2, \dots, \lambda_{d+1}$ of $\widetilde{\varGamma}$ (see Zhang and Zha [116] for the details).

Algorithm 1.6 Local tangent space alignment algorithm

```
Require: $X_i \in \mathbb{R}^m$, $i = 1, \dots, n$ and $k$.
Ensure: $\widehat{Y}_i \in \mathbb{R}^d$, $i = 1, \dots, n$
  Find the $k$-nearest neighbors $N_i$ for each $X_i$.
  Form an $n \times k$ matrix $X^{(i)} = (X_{i_1} - \bar{X}_i, \dots, X_{i_k} - \bar{X}_i)$ with $\bar{X}_i = \sum_{i' \in N_i} X_{i'}$, and obtain its singular value decomposition $X^{(i)} = U_i \Sigma_i V_i^\top$. Define the matrix of the $d$ left eigenvectors corresponding to the $d$ largest eigenvalues, $\widehat{Q}_i = (U_{i1}, \dots, U_{id})$, and the local coordinates $\widehat{\varGamma}_i = \left(\hat{\varrho}^{(i)}_{i_1}, \dots, \hat{\varrho}^{(i)}_{i_{k_i}}\right)$ with $\hat{\varrho}^{(i)}_{i'} = \widehat{Q}_i^\top (X_{i'} - \bar{X}_i)$.
  Form the alignment matrix $\widetilde{\varGamma} = \sum_i S_i W_i W_i^\top S_i^\top$, where $S_i$ is the 0-1 selection matrix such that $\Theta_i = \Theta S_i$, and $W_i = (I_k - k^{-1} 1 1^\top)(I_k - \widehat{\varGamma}_i^{+} \widehat{\varGamma}_i)$, where $\widehat{\varGamma}_i^\dagger$ is the Moore-Penrose pseudoinverse of $\widehat{\varGamma}_i$. Compute the singular value decomposition of $\widetilde{\varGamma}$, and determine the embedding by using the $d$ eigenvectors $v_2, \dots, v_{d+1}$ corresponding to the first $\lambda_2, \dots, \lambda_{d+1}$ eigenvalues. Finally set $\widehat{Y} = (v_2, \dots, v_{d+1})^\top = \widehat{\Theta} = (\hat{\theta}_1, \dots, \hat{\theta}_n)$.
```

A graphical comparison of the nonlinear dimensionality reduction methods described above is provided in the Figure 1.2 for the “Swiss roll” data set using the MANIfold Learning Matlab Demo provided by Todd Wittman free available at `http://www.math.ucla.edu/~wittman/mani/index.html`.

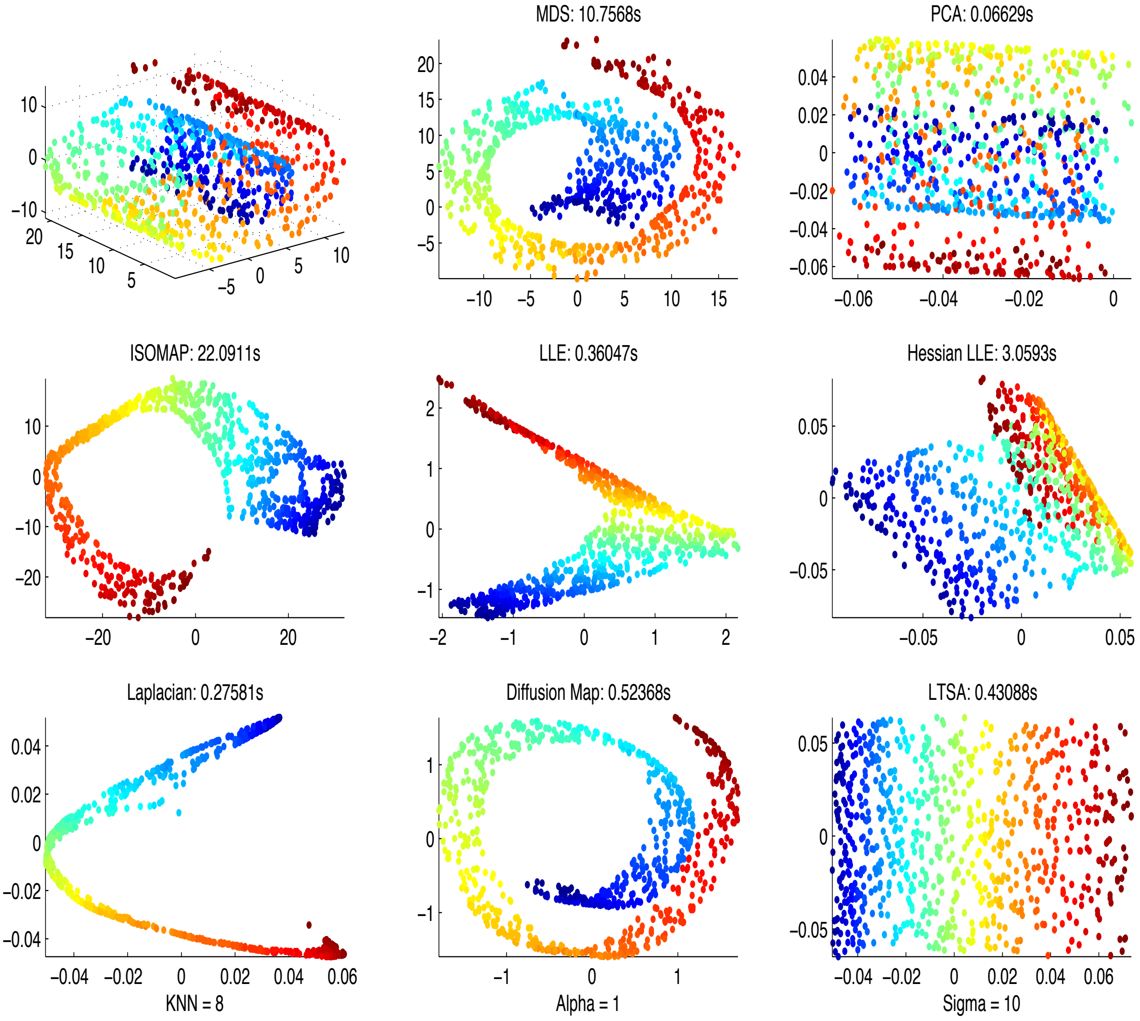

Figure 1.2: Comparison of manifold learning methods on the “Swiss roll” data set

Although we have described several methods for nonlinear dimensionality reduction, our original main goal is to approximate the geodesic distances to calculate the empirical intrinsic median given in (1.2) as a robust estimator of the template curve f. Hence, we use the Isomap algorithm to carry out this issue. In Chapter 2, a modified version (of the first two steps) of the Isomap algorithm applied to functional data is proposed. The proposed algorithm has the advantage of being less sensitive to outliers and does not require any additional tuning parameter, as in the original Isomap procedure, making it much easier to handle. The algorithm has been applied successfully to obtain the common pattern function on several samples of curves in different scientific areas. For example, the Figure 1.3 illustrates the template functions estimated through the robust estimator based upon the manifold framework corresponding to some samples of real curves which are used in this thesis.

As we can see, for each of the examples of samples of curves, the corresponding estimated template functions capture appropriately the common structural behavior of the samples.

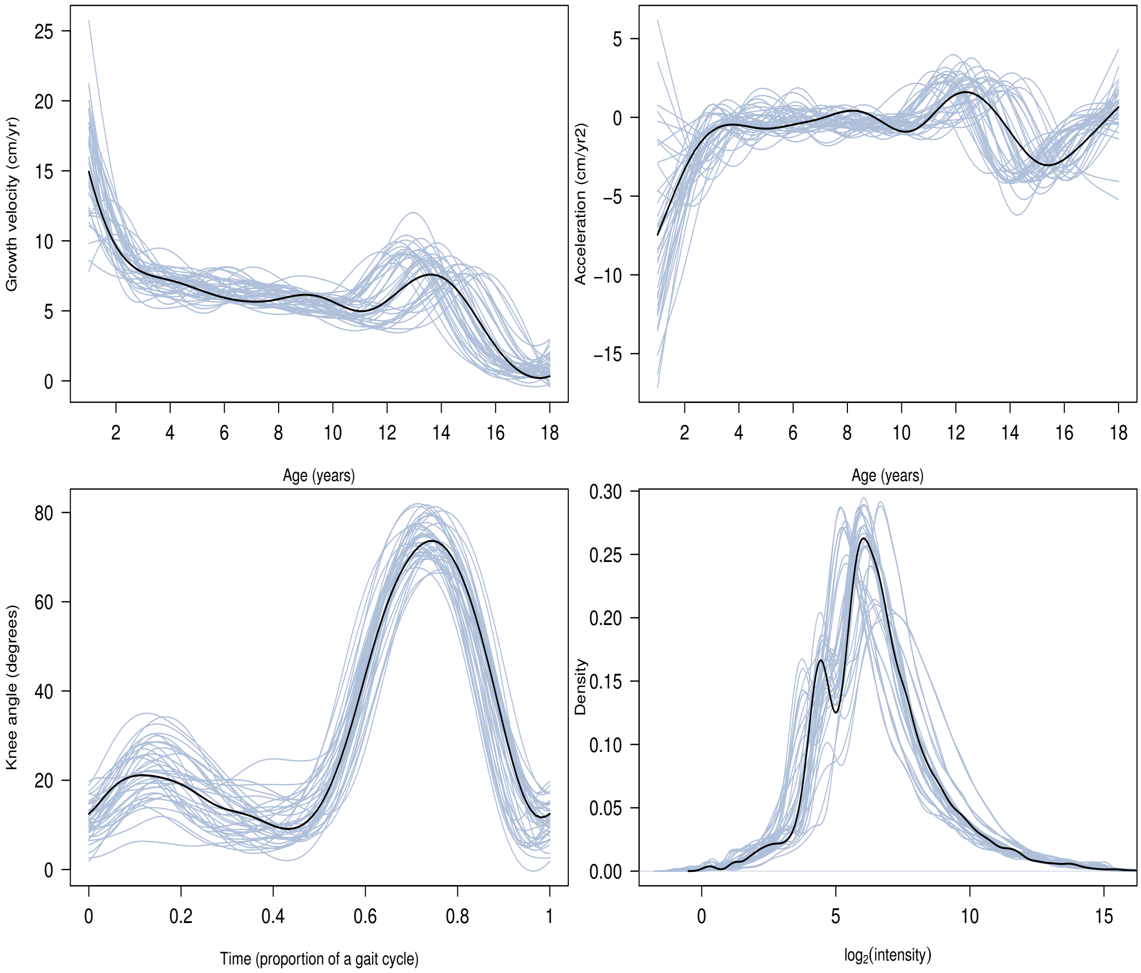

Figure 1.3: Examples of samples of curves and their respective template curve (bold solid line) estimated through the empirical intrinsic median.

1.3 Density curve alignment

In Chapter 3, the curve alignment problem of a sample of probability density functions is considered. This particular situation arises for example in bioinformatics and economics, where usually the goal is to study experiments that involve multiple (often thousands) high density oligonucleotide arrays (Bolstad et al. [12]) or analyze the densities of a relative household income of the household head over several years (Kneip and Utikal [60]).

1.3.1 Quantile normalization

In bioinformatics, the observed variability among density curves of expression (also intensity) measures of oligonucleotide arrays usually can not be directly correlated to the biological phenomenon under study. This non-biological variations are associated, principally, to different efficiencies of reverse transcription, labeling, hybridization reactions, physical problems with arrays, reagent batch effects, and laboratory conditions. Therefore, the intensities need to be adjusted to give accurate measurements of specific hybridization, i.e. remove (or at least reduce) from microarray data the effects which arise from systematic technical variations in the technology, ensuring that differences in intensity indeed reflect the differential gene expressions. This process is known as *normalization*, which can be considered as an alignment method for density functions. Unless arrays are appropriately normalized, comparisons from different arrays can lead to misleading results. Smyth and Speed [93], and Irizarry et al. [53] offer details about the normalization process.

One of most popular and wide used normalization method is the *quantile normalization* developed by Bolstad et al. [12]. Its popularity is due to its simplicity and nice estimation results. A comparison study of a number of normalization methods to high-density oligonucleotide data is provided by Bolstad et al. [12]. The central idea of quantile normalization consists in to enforce the same empirical distribution of intensities to each array. This imply that the quantile-quantile plot between two data vectors should be close to the 45-degree diagonal line. The Figure 1.4 illustrates an example of how the quantile normalization method works in the simple case of $n = 2$ vectors.

Thereby, in n dimensions, and based on sample quantiles, the "mean" distribution is obtained through the projection of the empirical quantile vector of the j-th sample quantiles, $\hat{q}_j = (\hat{q}_{1,j}, \ldots, \hat{q}_{n,j})^\top$, onto the vector $d = (1/\sqrt{n}, \ldots, 1/\sqrt{n})^\top$, given by $\text{proj}_d \hat{q}_j = (n^{-1}\sum_{i=1}^n \hat{q}_{i,j}, \ldots, n^{-1}\sum_{i=1}^n \hat{q}_{i,j})^\top$. The quantile normalization method is given in the Algorithm 1.7. This method can be understood as a quantile-quantile plot extended to n dimensions such that if all n data vectors share the same distribution, then the plot gives a straight line along the line d.

Algorithm 1.7 Quantile normalization algorithm

```
Require: A matrix $X$ of $n$ vectors of $m$ observations.
Ensure: A normalized matrix $X_{\text{norm}}$.
1: Form the $n$-vector $d = (1/\sqrt{n}, \ldots, 1/\sqrt{n})^\top$
2: Obtain the matrix of sample quantiles $\widehat{Q}$ by sorting the columns of $X$
3: Obtain a matrix $\widehat{Q}_{\text{proj}}$, where each row is the projection of each row of $\widehat{Q}$,
   $\hat{q}_j = (\hat{q}_{1,j}, \ldots, \hat{q}_{n,j})^\top$, $j = 1, \ldots, m$, onto $d$. It is, each row of $\widehat{Q}_{\text{proj}}$ is given
   by $\text{proj}_d \hat{q}_j = (n^{-1}\sum_{i=1}^n \hat{q}_{i,j}, \ldots, n^{-1}\sum_{i=1}^n \hat{q}_{i,j})^\top$.
4: Get $X_{\text{norm}}$ by rearranging each column of $\widehat{Q}_{\text{proj}}$ to have the same ordering as $X$.
```

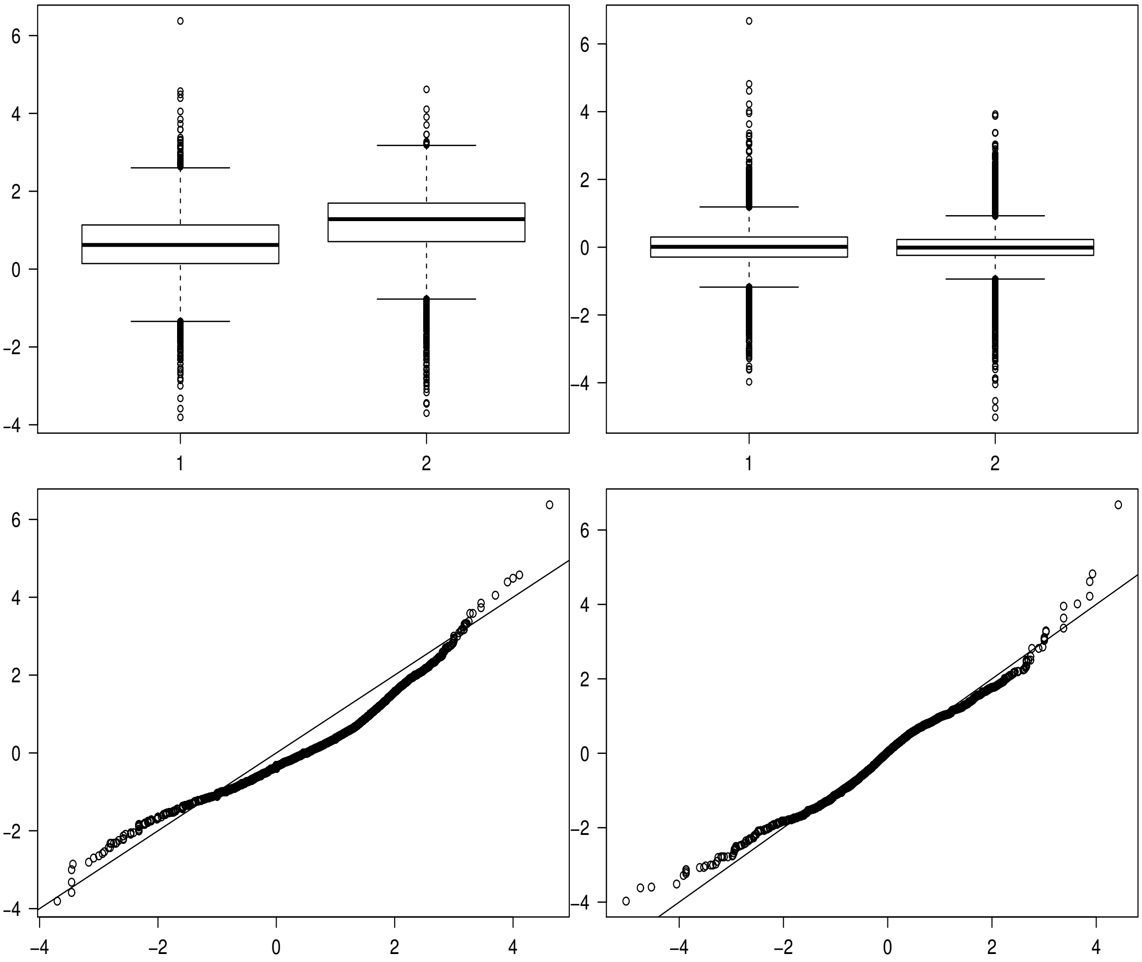

Figure 1.4: Boxplots and quantile-quantile plots for unnormalized (on the left) and normalized (on the right) data with the quantile normalization method.

However, despite its popularity, its large sample properties have not yet been studied. This is one of the goal of this thesis. To obtain the asymptotic properties of the quantile normalization method which is a particular case of the structural mean curve alignment procedure proposed by Dupuy et al. [35]. The properties are proved starting from a warping model in which each distribution function F_i is obtained by warping a common distribution function F by a sample of invertible and differentiable warping mappings H_i

$$F_i(t) = F \circ H_i^{-1}(t), \qquad i = 1, \dots, n, \; j = 1, \dots, m,$$

where $H_1, \dots, H_n$ are i.i.d functions from a stochastic process $\mathcal{H}$ with mean function ϕ and variance function ϑ.

The estimation problem of F and H from F_i, $i = 1, \dots, n$ based on the warping model is not identifiable. More precisely, the unknown distribution function F

and the unknown warping process $\mathcal{H}$ cannot be uniquely estimated. Indeed, let $\widetilde{H} : \mathcal{I} \to \mathcal{I}$ be an increasing continuous function, then we have, for all $i \in \{1, \ldots, n\}$,

$$F_i(t) = F \circ H_i^{-1}(t) = F \circ \widetilde{H}^{-1} \circ \widetilde{H} \circ H_i^{-1}(t) = \widehat{F} \circ \widehat{H}_i^{-1}(t).$$

Hence the function $\widehat{F} = F \circ \widetilde{H}^{-1}$ associated with the warping process $\mathcal{H} \circ \widetilde{H}_i^{-1}$ is also a solution. In fact there are infinitely many different representations of the same observed process for any $\widetilde{H}$.

To overcome the identifiability problem the method by Dupuy et al. [35] is followed, where an archetype function, observed in the center of the sample, representing the common behavior of the sample is defined, without stressing any particular curve but taking into account the information conveyed by the warping process itself. It is the archetype is directly obtained from the data. Hence, we consider the definition of *structural expectation* (*SE*) for the quantile function, given by

$$q_{SE}(\alpha) := F_{SE}^{-1}(\alpha) = \mathbb{E}(H_i) \circ F^{-1}(\alpha) = \phi \circ F^{-1}(\alpha), \qquad 0 \leq \alpha \leq 1,$$

where its estimator is

$$\overline{q_n(\alpha)} = \frac{1}{n} \sum_{i=1}^{n} q_i(\alpha) = \frac{1}{n} \sum_{i=1}^{n} H_i \circ F^{-1}(\alpha).$$

As the distribution function is not observed, the corresponding order statistics $X_{i,1:m} \leq \cdots \leq X_{i,m:m}$ of random samples $X_{i,1}, \ldots, X_{i,m}$ from F_i are used to estimate the quantile functions $q_i(\alpha)$ by $\hat{q}_{i,m}(\alpha) = X_{i,j:m}$ for $(j-1)/m < \alpha \leq j/m$, $j = 1, \ldots, m$, such that the estimator for $q_{SE}(\alpha)$ is given by

$$\bar{\hat{q}}_j = \frac{1}{n} \sum_{i=1}^{n} \hat{q}_{i,j} = \frac{1}{n} \sum_{i=1}^{n} X_{i,j:m}, \qquad j = 1, \ldots, m.$$

In Chapter 3, strong consistency and asymptotic distribution of the quantile normalization estimator $\bar{\hat{q}}_j$ are proved: $\bar{\hat{q}}_j \xrightarrow{a.s} q_{SE}(\alpha_j)$, as $m, n \to \infty$ for $j = 1, \ldots, n$; and

$$\sqrt{m} \begin{bmatrix} \bar{\hat{q}}_{j_1} - q_{SE}(\alpha_1) \\ \vdots \\ \bar{\hat{q}}_{j_K} - q_{SE}(\alpha_K) \end{bmatrix} \xrightarrow[m,n\to\infty]{\mathcal{D}} \mathcal{N}_K(\mathbf{0}, \mathbf{\Sigma})$$

for any $K \in \mathbb{N}$ and fixed $(\alpha_1, \ldots, \alpha_K) \in [0,1]^K$, where the (k,k')-element of matrix $\mathbf{\Sigma}$ is $\Sigma_{k,k'} = \vartheta\big(q(\alpha_k), q(\alpha_{k'})\big)$ for all $(\alpha_k, \alpha_{k'}) \in [0,1]^2$ with $\alpha_k < \alpha_{k'}$.

1.3.2 Manifold normalization

A natural distance to measure the proximity between two cumulative distribution functions $F(x) = \mu\,[X \leq x]$ and $G(y) = \nu\,[Y \leq y]$ is the well-known Wasserstein metric (see, for instance Villani [104]) given by

Definition 1.11. The Wasserstein distance of order $p \geq 1$ between two Borel probability measures μ and ν on a given a metric space $(\mathcal{X}, d)$ is defined by

$$\begin{aligned} W_p(\mu, \nu) &= \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X} \times \mathcal{X}} d(x,y)^p \mathrm{d}\pi(x,y) \right)^{1/p} \\ &= \inf \left(\mathbb{E}\left[d(X,Y)^p \right] \right)^{1/p}, \end{aligned}$$

where $\Pi(\mu, \nu)$ denotes the collection of all probability measures on $\mathcal{X} \times \mathcal{X}$ with marginals $\mu = \mathrm{law}(X)$ and $\nu = \mathrm{law}(Y)$ having finite pth moments.

Upon this framework, the problem of finding a *mean* probability measure can be posed, for example, in the space of probability measures with finite second order moments (i.e. the Wasserstein space)

$$P_2(\mathcal{X}) := \left\{ \mu \in P(\mathcal{X}) \colon \int_{\mathcal{X}} d(x_0, x)^2 \mu(\mathrm{d}x) < +\infty \text{ for some } x_0 \in \mathcal{X} \right\},$$

by solving the minimization problem

$$\min_{\mu \in P_2(\mathcal{X})} \frac{1}{n} \sum_{i=1}^{n} W_2^2(\mu_i, \mu), \tag{1.3}$$

with the Mallows-Wasserstein metric

$$W_2^2(\mu_i, \mu) = \int_0^1 \left| F_i^{-1}(\alpha) - F^{-1}(\alpha) \right|^2 \mathrm{d}\alpha, \qquad 0 \leq \alpha \leq 1$$

where F_i and F are the distributions functions of μ_i, $i = 1, \ldots, n$ and μ in $P_2(\mathcal{X})$, respectively.

The solution of the problem is refereed as the barycenter of the measures μ_i. The existence and uniqueness of such a minimizer is a difficult task in a general framework. Indeed these have been recently proved by Agueh and Carlier [1]. However, for one-dimensional distributions, an explicit solution can be given, which corresponds to the *structural expectation* defined in Dupuy et al. [35].

An alternative criterion to the Mallows-Wasserstein metric in (1.3) can be explored proposing any other distance function on the inverse of distribution functions. In Chapter 3, we appeal to the assumption made in Chapter 2, considering that functional data belong to a manifold $\mathcal{M}$. Thus, the geodesic distance δ provides also a natural way to compare two objects upon this framework. Therefore, the manifold embedding approach for a sample of density curves, that we rename *manifold normalization*, is applied as an interesting alternative to the problem. This gives rise to the problem

$$\min_{F \in \mathcal{M}} \frac{1}{n} \sum_{i=1}^{n} \delta^2\left(F^{-1}, F_i^{-1}\right).$$

For this, based on the observed random variables $X_{i,j}$, $i = 1, \dots, n$, $j = 1, \dots, m$, the approximation of the geodesic distance between the inverse of the distribution functions, $F_i^{-1}(\alpha)$, $(j-1)/m < \alpha \leq j/m$, is conducted by using the corresponding order statistics $X_{i,j:m}$. Denoting the sorted vector $X_{i,1:m}, \dots, X_{i,m:m}$ for $i = 1, \dots, n$ by $X_{(i).}$, the intrinsic mean is defined as

$$\hat{\mu}_{\hat{\delta}} = \underset{x \in \{X_{(i).}, \, i=1,\dots,n\}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \hat{\delta}^2 \big(X_{(i).}, x \big),$$

where $\hat{\delta}$ is the approximation of the geodesic distance between vectors $X_{(i).}$, using the modified version of the algorithm developed in Chapter 2.

1.4 Maxentropic functional calibration estimation

Usually the way in which data are collected is rarely considered by the statistician, assuming that data are independent replications from a common distribution. That is, the sample consists in several simultaneous measurements of the same random experiment. However, sometimes the sample a of n elements comes from a survey sampling strategy over a finite survey population $U_N = \{1, \dots, N\}$, which must be taken into account in the estimation of unknown parameters. Assume that associated with the ith population element there is a random variable denoted by Y_i, $i \in U_N$. A finite population parameter of particular interest is the finite population mean given by

$$\mu_Y = \frac{1}{N} \sum_{i \in U_N} Y_i.$$

The problem of obtaining an estimate for the unknown finite population mean parameter of the survey variable Y is conducted by using the information contained in a subset of U_N rather that U_N itself (i.e. based on a sample a of n elements selected from the set U_N). The selection of the sample is carry out according to a probabilistic selection scheme. A wide variety of random sampling selection methods exists in the literature, see for instance Särndal et al. [86]. Based on a given sample selection scheme is possible to define the sampling design $p(a) = \mathbb{P}(A = a)$, where $a \in \mathcal{A}$ with $\mathcal{A}$ denoting the collection of all subsets A of U_N that contains all possible samples of n different elements randomly drawn from U_N, and $\mathbb{P}$ is a probability measure on $\mathcal{A}$. The function $p(a)$ gives the probability of selecting a sample a under the scheme used.The sampling design specifies the probability structure of selecting an element i in the sample, obtaining the inclusion probabilities given by

$$\pi_i = \mathbb{P}(i \in a) = \sum_{a \in A(i)} p(a),$$

where $A(i)$ is the set of samples that contain the ith element.

A popular design-unbiased and consistent estimator of μ_Y, assuming that all inclusion probabilities are strictly positive for all $i \in U_N$, is given by the linear estimator (commonly known as the Horvitz-Thompson or π estimator, see Horvitz and Thompson [52]), defined as

$$\hat{\mu}_Y^{\mathrm{HT}} = \frac{1}{N} \sum_{i \in a} \pi_i^{-1} Y_i = \frac{1}{N} \sum_{i \in a} d_i Y_i.$$

The Horvitz-Thompson estimator although design-unbiased has low precision, especially if the sample is small, implying that the sample does not describe well the behavior of the variable of interest in the total population. So, to prevent biased estimation due to bad sample selection, inference on the sample can be achieved by considering a modification of the weights of the individuals chosen in the sample. One modification method to obtain an improvement of the estimator can be conducted by incorporating auxiliary information observed for each element in the population, $i \in U_N$. The method, proposed by Deville and Särndal [30], is called calibration estimation and consists in modifying the standard sampling design weights $d_i = \pi_i^{-1}$ of the Horvitz-Thompson estimator by new weights w_i close enough to d_i's according to some distance function $\mathcal{D}(w, d)$, satisfying a linear calibration equation, in which the auxiliary information is taken into account,

$$\frac{1}{N} \sum_{i \in a} w_i X_i = \mu_X,$$

where the q-dimensional vector X_i represents the auxiliary information.

The calibration weights w_i will generally result in estimates that are design consistent, and that have a smaller variance than the Horvitz-Thompson estimator (see, for instance, Deville and Särndal [30] and Särndal [85]). The idea of calibration has been extended to estimate other finite population parameters.

The calibration estimator for μ_Y based on the calibration weights is expressed by the linear weighted estimator $\hat{\mu}_Y = N^{-1} \sum_{i \in a} \hat{w}_i Y_i$. Note that there are two basic components in the construction of calibration estimators: the dissimilarity function and the set of calibration equations. Different calibration estimators can be obtained depending on the chosen distance function (Deville and Särndal [30]). However, it is well known that all of calibration estimators are asymptotically equivalent to the one obtained through the use of the chi-square distance function $\mathcal{D}_a(w, d) = \sum_{i \in a} (w_i - d_i)^2 / 2 d_i q_i$, where q_i is an individual given positive weight uncorrelated with d_i.

In Chapter 4, the estimation of the finite population mean of a survey variable is covered under a infinite-dimensional setting. Particularly, the problem of functional calibration estimation for the finite population mean of the functional survey variable $Y(t)$, with values in the space of continuous functions on $[0,1]$, $\mathcal{C}([0,1])$, is considered using the information provided by functional auxiliary information $X(t) \in \mathcal{C}([0,1]^q)$, $q \geq 1$, in order to correct the bad sample effect and thus improve its accuracy in terms of efficiency. As was described above, the goal is how to obtain a design consistent estimator for the unknown functional finite population mean $\mu_Y(t) = N^{-1} \sum_{i \in U_N} Y_i(t)$ based on the calibration method. The idea consists in modify the basic sampling design weights d_i of the unbiased functional Horvitz-Thompson estimator $\hat{\mu}_Y^{\mathrm{HT}}(t) = N^{-1} \sum_{i \in a} d_i Y_i(t)$ for new more efficient functional calibration weights $w_i(t)$. These must to be sufficiently close to d_i's according to some dissimilarity distance function $\mathcal{D}_a^*(w,d)$, satisfying the functional calibration restriction

$$\frac{1}{N} \sum_{i \in a} w_i(t) X_i(t) = \mu_X(t). \tag{1.4}$$

The functional calibration estimator for $\mu_Y(t)$ is then expressed as

$$\hat{\mu}_Y(t) = \frac{1}{N} \sum_{i \in a} w_i(t) Y_i(t).$$

In this thesis the functional calibration sampling weights are obtained by matching the calibration estimation problem with the maximum entropy on the mean principle. In particular, the calibration estimation is viewed as an infinite-dimensional linear inverse problem following the structure of the maximum entropy on the mean approach as follows.

1.4.1 Principle of the Maximum entropy on the mean

Consider the following linear inverse problem

$$y = \mathcal{K}x, \tag{1.5}$$

where $\mathcal{K}\colon \mathcal{X} \to \mathcal{Y}$ is a known bounded linear operator between separable Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$. Here y and x are respectively the observed and the unknown data.

The goal is to build a solution for x denoted by $\hat{x}$. However, its well-known that, in general, this kind of inverse problems suffers of ill-posedness in the sense of Hadamard (see for instance Engl et al. [36] and Gzyl and Velásquez [49]). Therefore, a regularization method must be applied in order to get a unique and stable solution. In addition, to the popular Tikhonov's regularization method (see Engl et al. [36]), there also exist alternative methods such as probabilistic-based approaches which instead of finding an explicit solution to the equation (1.5), searching for a probability distribution ν such that the possible solution is the mean of a random variable X. Among these, the Maximum Entropy on the Mean approach arises as a powerful alternative method to solve constrained linear inverse problems (Gamboa [40], Dacunha-Castelle and Gamboa [26], and Gzyl and Velásquez [49]).

The Maximum Entropy on the Mean approach considers x as the mean value of a random element X, such that the equation (1.5) becomes

$$y = \mathcal{K}\mathbb{E}_{\nu}(X),$$

where the expectation is over an unknown probability measure ν.

The solution of the inverse problem $\hat{x}$ is then defined as the mathematical expectation of X under a probability measure ν^* that must be determined. More precisely, we study the case where this measure is constructed by the maximization, over the (convex) set of all probability measures, of the entropy functional

$$S(\nu \parallel \upsilon) = -D(\nu \parallel \upsilon),$$

subject to the constraint

$$y = \mathcal{K}\mathbb{E}_{\nu}(X).$$

Here, $D(\nu \parallel \upsilon)$ is the Kullback-Leibler divergence between a feasible finite measure ν with respect to a given prior probability measure υ, whose definition is as follows:

Let $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ be a measurable space, where $\mathcal{B}(\mathcal{X})$ is the Borel σ-field of $\mathcal{X}$, and $\mathcal{P}(\mathcal{X})$ set of probability measures on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$.

Definition 1.12. For $\nu, \upsilon \in \mathcal{P}(\mathcal{X})$, the Kullback-Leibler information divergence (also I-divergence and relative entropy) of ν with respect to υ is defined by,

$$D(\nu \parallel \upsilon) = \begin{cases} \int_{\mathcal{X}} \log\left(\frac{\mathrm{d}\nu}{\mathrm{d}\upsilon}\right) \mathrm{d}\nu & \text{if } \nu \ll \upsilon \\ +\infty & \text{otherwise,} \end{cases}$$

where $\nu \ll \upsilon$ means that ν is absolutely continuous with respect to υ.

Remark 1.5. It is important to remark that although the Kullback-Leibler divergence is always nonnegative and is zero if and only if $\nu = \upsilon$, it is not a true distance since it is not symmetric and does not satisfy the triangle inequality. Nevertheless, it is often useful to think of Kullback-Leibler divergence as a "distance" between measures.

The uniqueness of the probability measure ν^* is guaranteed by the theorem below. This theorem relies on the concept of I-projection (Csiszár [24, 25]).

Definition 1.13. Let $\upsilon \in \mathcal{P}(\mathcal{X})$ be a probability measure and $\mathcal{C}$ a convex subset of $\mathcal{P}(\mathcal{X})$ such that $D(\mathcal{C} \parallel \upsilon) < +\infty$. A probability measure ν^* is called the I-projection (or entropic projection) of υ on $\mathcal{C}$ if $\nu^* \in \mathcal{C}$ and

$$D(\nu^* \parallel \upsilon) = \inf_{\nu \in \mathcal{C}} D(\nu \parallel \upsilon).$$

This condition, in terms of the entropy function S, is given by

$$S(\nu^* \parallel \upsilon) = -D(\nu^* \parallel \upsilon) = \inf_{\nu \in \mathcal{C}} -D(\nu \parallel \upsilon) = \sup_{\nu \in \mathcal{C}} S(\nu \parallel \upsilon).$$

Now we let $\mathcal{X}$ be a locally convex topological vector space. The following Theorem due to Csiszár [25] characterizes the entropic projection of a given probability measure on a convex set. For further details see also Gozlan [47, Chapter 2].

Theorem 1.3. *Let υ be a probability measure on $\mathcal{X}$ and $\mathcal{C} \subset \mathcal{P}(\mathcal{X})$ a convex set of $\mathcal{X}$ whose interior has a non-empty intersection with the convex hull of the support of υ. Let $\Pi(\mathcal{C}) = \left\{\nu \in \mathcal{P}(\mathcal{X})\colon \int_{\mathcal{X}} x \mathrm{d}\nu(x) \in \mathcal{C}\right\}$. Then the I-projection ν^* of υ on $\Pi(\mathcal{C})$ is given by the (exponential family) relation defined by*

$$\frac{\mathrm{d}\nu^*}{\mathrm{d}\upsilon} = \frac{\exp\left\{\langle \lambda^*, \mathcal{K}x\rangle\right\}}{Z_\upsilon(\lambda^*)},$$

where $\lambda^* \in \mathcal{Y}'$ (dual space of $\mathcal{Y}$) minimizes the function

$$H_\upsilon(\lambda) = \log Z_\upsilon(\lambda) - \inf_{y \in \mathcal{Y}} \langle \lambda, y\rangle,$$

with $Z_\upsilon(\lambda) = \mathbb{E}_\upsilon\left[\exp\left\{\langle \lambda, \mathcal{K}x\rangle\right\}\right] = \int_{\mathcal{X}} \exp\left\{\langle \lambda, \mathcal{K}x\rangle\right\} \mathrm{d}\upsilon(x) < +\infty$.

In Chapter 4, the problem of estimation of functional survey weights $w_i(t)$ is conducted by mean the maximum entropy on the mean principle, expressing the calibration constraint $\sum_{i\in a} w_i\,(t)\, X_i(t) = N\mu_X(t)$ as an infinite-dimensional linear inverse problem, writing $w_i(t)$ as

$$w_i(t) = \int_0^1 K(s,t)\varpi_i\,(s)\,\mathrm{d}s + d_i, \quad \text{for each } i \in a,$$

where $K(s,t)$ is a known continuous, real-valued and bounded kernel function and $\varpi_i(s)$ the mean value of a stochastic process $\mathcal{W}_i(s)$.

Thus, we have the following inverse problem

$$\begin{aligned} N\mu_X(t) &= \mathcal{K}\mathbb{E}_\nu\left[\mathcal{W}\right] \\ &= \sum_{i\in a} \mathbb{E}_\nu\left[\int_0^1 K(s,t)\mathrm{d}\mathcal{W}_i\,(s) + d_i\right] X_i(t) \\ &= \int_0^1 \sum_{i\in a} K(s,t) X_i(t) \mathbb{E}_\nu\left[\mathrm{d}\mathcal{W}_i\,(s)\right] + \sum_{i\in a} d_i X_i(t), \quad t \in [0,1], \end{aligned}$$

which takes the form of a Fredholm integral equation of the first kind. In general, integral equations of the first kind with continuous kernels provides typical examples for ill-posed problems (see, e.g. Kress [64]).

The probability measure ν^* is given by applying the Theorem above, where the corresponding inner product is given by

$$\langle \lambda, \mathcal{K}\mathcal{W} \rangle = \int_0^1 \lambda^\top(\mathrm{d}t) \left(\int_0^1 \sum_{i \in a} K(s,t) X_i(t) \mathrm{d}\mathcal{W}_i(s) + \sum_{i \in a} d_i X_i(t) \right).$$

The estimated functional calibration weights are

$$\hat{w}_i(t) = \int_0^1 K(s,t) \varpi_i^*(s) \mathrm{d}s + d_i, \quad i \in a,$$

where $\varpi_i^*(s)\mathrm{d}s = \mathrm{d}\mathbb{E}_{\nu^*}\left[\mathcal{W}_i(s)\right]$ depends of the stochastic processes $\mathcal{W}$ assumed as prior.

In the present thesis, two prior measures are assumed: the centered Gaussian and compound Poisson random measures.

Appendix

Some definitions related to some concepts of graph theory are given below. For details see, for instance, West [108].

Definition 1.14. A graph $\mathcal{G} = \mathcal{G}(V, E)$ consists of a vertex set $V(\mathcal{G})$ and edge set $E(\mathcal{G})$. Each edge $e \in E$ is said to join two vertices called its endpoints.

Definition 1.15. A loop is an edge whose endpoints are equal. Multiple edges are edges having the same pair of endpoints. A simple graph is a graph having no loops or multiple edges. A simple graph is specified by its vertex set $V(\mathcal{G})$ and edge set $E(\mathcal{G})$, treating $E(\mathcal{G})$ as a set of unordered pairs of vertices. When x and x' are the endpoints of an edge $e = xx'$, they are adjacent and are neighbors.

Definition 1.16. A graph $\mathcal{H}$ is a subgraph of $\mathcal{G}$ if $V(\mathcal{H}) \subseteq V(\mathcal{G})$ and $E(\mathcal{H}) \subseteq E(\mathcal{G})$ such that for all $e = xx' \in E(\mathcal{H})$, we have that $x, x' \in V(\mathcal{G})$, denoted as $\mathcal{H} \subseteq \mathcal{G}$.

Definition 1.17. A (x_0, x_n)-walk in $\mathcal{G}$ is an alternating sequence of vertices and edges $[x_0, e_1, x_1, e_2, \ldots, x_{n-1}, e_n, x_n]$, with $e_i = x_{i-1}x_i$. In a closed walk, $x_0 = x_n$. A trail is a walk where all edges are distinct; a path is a trail where also all vertices are distinct. A cycle is a closed trail where all vertices except x_0 and x_n are distinct.

Definition 1.18. Two vertices x and x' in a graph $\mathcal{G}$ are connected if there exists a (x, x')-path in $\mathcal{G}$. $\mathcal{G}$ is connected if all pairs of distinct vertices are connected.

Definition 1.19. A complete graph is a simple graph whose vertices are pairwise adjacent.

Definition 1.20. The neighbor set $N(x)$ of a vertex $x \in V(\mathcal{G})$ is the set of all its adjacent vertices, i.e. $N(x) := \{x' \in V(\mathcal{G}) | x \neq x', \exists e \in V(\mathcal{G}) \colon e = xx'\}$.

Definition 1.21. A simple, connected graph having no cycles is called a tree, i.e. an acyclic graph in which any two vertices are connected by exactly one simple path. A spanning subgraph of $\mathcal{G}$ is a subgraph with vertex set $V(\mathcal{G})$. A spanning tree is a spanning subgraph that is a tree.

Definition 1.22. A weighted graph $\mathcal{G}$ is a graph for which each edge $e = xx' \in E(\mathcal{G})$ has an associated real number $w(e) = w(x, x')$ called its weight. For any subgraph $\mathcal{H} \subseteq \mathcal{G}$, the weight of $\mathcal{H}$ is the sum of weights of its edges, $w(\mathcal{H}) = \sum_{e \in E(\mathcal{H})} w(e)$.

Definition 1.23. Let P be a (x, x')-path with minimal weight among all (x, x')-paths in a graph $\mathcal{G}$. The weight of P is known as the (geodesic) distance $d_{\mathcal{G}}(x, x')$, and P is the shortest path, or geodesic between x and x'.

Definition 1.24. A minimum spanning tree is a spanning tree with minimal weight among all spanning trees in $\mathcal{G}$.

Dijkstra's algorithm

This algorithm is an efficient method for computing the shortest paths from a given vertex x to all other vertices in a given graph. It is probably the most popular algorithm for this issue. The algorithm is more efficient when graphs are large and have many edges. The algorithm solves the problem using the observation that an $(x, x^\star)$-portion of a shortest (x, x')-path must be a shortest $(x, x^\star)$-path. It finds optimal paths from x to other vertices x' in increasing order of $d_{\mathcal{G}}(x, x')$.

Algorithm 1.8 Dijkstra's algorithm

```
Require: $\mathcal{G}$ with $w(v, v') > 0$, $\forall vv' \in E(\mathcal{G})$, and an initial vertex $x$. Let
    $w(v, v') = \infty$ if $vv' \notin E(\mathcal{G})$.
Ensure: Shortest paths from one vertex to all other vertices in a weighted graph
  Idea: Keep the set $S$ of vertices to which a shortest path from $x$ is known,
    enlarging $S$ to include all vertices, maintaining a tentative distance $t(x')$ from $x$
    to each $x' \notin S$, being the length of the shortest $(x, x')$-path yet found.
  Initialization: Set $S = \{x\}$; $t(x) = 0$; $t(x') = w(x, x')$ for $x' \neq x$.
  Iteration: Select a vertex $x^*$ outside $S$ such that $t(x^*) = \min_{x' \notin S} t(x')$. Add
    $x^*$ to $S$. Explore edges from $x^*$ to update tentative distances: for each edge
    $x^*x'$ with $x' \notin S$, update $t(x')$ to $\min\{t(x'), t(x^*) + w(x^*, x')\}$. The iteration
    continues until $S = V(\mathcal{G})$ or until $t(x') = \infty$ for every $x' \notin S$. At the end, set
    $d_{\mathcal{G}}(x, x^*) = t(x^*)$ for all $x^*$.
```

Chapter 2

A robust algorithm for template curve estimation based on manifold embedding

joint work with C. Dimeglio*, J-M. Loubes† and E. Maza‡

Abstract: This chapter considers the problem of finding a meaningful template function that represents the common pattern of a sample of curves. To address this issue, a novel algorithm based on a robust version of the isometric featuring mapping (Isomap) algorithm is developed. Assuming that the functional data lie on an intrinsically low-dimensional smooth manifold with unknown underlying structure, we propose an approximation of the geodesic distance. This approximation is used to compute the corresponding empirical Fréchet median function, which provides an intrinsic estimator of the template function. Unlike the Isomap method, the algorithm has the advantage of being parameter free and easier to use. Comparisons with other methods, with both simulated and real datasets, show that the algorithm works well and outperforms these methods.

Key Words: Fréchet median; functional data analysis; Isomap; manifold learning.

[*] Institut de Mathématiques de Toulouse, Université Paul Sabatier - Toulouse III, Toulouse, France. E-mail: cd@geosys.com

[†] Institut de Mathématiques de Toulouse, Université Paul Sabatier - Toulouse III, Toulouse, France. E-mail: jean-michel.loubes@math.univ-toulouse.fr

[‡] École Nationale Supérieure Agronomique de Toulouse, and Genomic & Biotechnology of the Fruit Laboratory. UMR 990 INRA/INP-ENSAT. E-mail: elie.maza@ensat.fr

2.1 Introduction

Nowadays, experiments where the outcome constitutes a sample of functions $\{f_i(t)\colon t \in \mathcal{T} \subset \mathbb{R},\ i = 1, \ldots, n\}$ are more and more frequent. Such kind of functional data are now commonly encountered in speech signal recognition in engineering, growth curves analysis in biology and medicine, microarray experiments in molecular biology and genetics, expenditure and income studies in economics, just to name a few.

However, extracting the information conveyed by all the curves is a difficult task. Indeed when finding a meaningful representative function that characterizes the common behavior of the sample, capturing its inner characteristics (as trends, local extrema and inflection points), a major difficulty comes from the fact that usually there are both amplitude (variation on the y-axis) and phase (variation on the x-axis) variations with respect to the common pattern, as pointed out in Ramsay and Li [77], Ramsay and Silverman [79], or Vantini [101] for instance. Hence, in the two last decades, there has been a growing interest for statistical methodologies and algorithms to remove the phase variability and recover a single template conveying all the information in the data since the classical cross-sectional mean is not a good representative of the data (see for instance Kneip and Gasser [58]).

Two different kinds of methods have been developed for template function estimation. The first group relies on the assumption that there exists a mean pattern from which all the observations differ, i.e an unknown function f such that each observed curve is given by $f_i(t) = f \circ h_i(t)$, where h_i are deformation functions. Hence finding this patten is achieved by aligning all curves f_i. This method is known as *curve registration*. In this direction, various curve registration methods have been proposed using different strategies. When the warping operator is not specified, we refer for instance to Kneip and Gasser [58], Wang and Gasser [106] Kneip et al. [61], James [55], Tang and Müller [95], and Kneip and Ramsay [59] or Dupuy et al. [35]. When a parametric model for the deformation is chosen, the statistical problem requires a semi-parametric approach through a self-modeling regression framework $f_i(t) = f(t, \theta_i)$ (see Kneip and Gasser [57]), where all functions are deduced with respect to the template f by mean a finite-dimensional individual parameter vector θ_i. This point of view is also followed in Silverman [88], Rønn [81], Gamboa et al. [42], Castillo and Loubes [18], Bigot et al. [10] and Trigano et al. [98].

The second category of methods do not assume any deformation model for the individual functions. The purpose is to select a curve which is assumed to be located at the *center* of the functions and estimate it directly from the data without stressing any particular curve. This is achieved for instance by López-Pintado and Romo [68] and Arribas-Gil and Romo [3] estimating the template based on the concept of depth for functional data as measure of centrality of the sample.

In this chapter, we propose an alternative way based on the ideas of manifold learning theory. We assume that the observed functions can be modeled as variables with values in a manifold $\mathcal{M}$ with an unknown geometry. Although the manifold is unknown, the key property is that its underlying geometric structure is contained in the sample of observed curves so that the geodesic distance can be reconstructed directly from the data. The template curve estimation is then equivalent to consider a location measure of the data with respect to this geodesic distance, hence approximating the Fréchet mean or median of the data. Recently, Chen and Müller [20] have also adopted a similar methodology appealing to the nonlinear manifold representation for functional data. Several algorithms have been developed over the last decade in order to reconstruct the natural embedding of data onto a manifold. Some of these are, for instance, the Isometric featuring mapping −Isomap− (Tenenbaum et al. [96]), Local Linear Embedding −LLE− (Roweis and Saul [82]), Laplacian Eigenmap (Belkin and Niyogi [5]), Hessian Eigenmap (Donoho and Grimes [33]), Diffusion maps (Coifman and Lafon [21]), Local Tangent Space Alignment −LTSA− (Zhang and Zha [116]), among others. In the following, we propose a robust version of the Isomap algorithm devoted to functional data, less sensitive to outliers and easier to handle. The performance of the algorithm is evaluated both on simulations and real data sets.

The chapter is organized as follows. The frame of our study is described in Section 2.2. Section 2.3 is devoted to the robust modification of the Isomap algorithm proposed to the metric construction of the approximated geodesic distance based on the observed curves. In Section 2.4 we analyze the template estimation problem in a shape invariant model, showing that this issue can be solved using the manifold geodesic approximation procedure. In Section 2.5, the performance of our algorithm is studied using simulated data. In Section 2.6, several applications on real functional data sets are performed. Some concluding remarks are given in Section 2.7.

2.2 Template estimation with a manifold embedding framework

Consider discrete realizations of functions f_i observed at time $t_{ij} \in \mathcal{T}$, with $\mathcal{T}$ a bounded interval of $\mathbb{R}$. For simplicity, we assume that all curves are observed at the same time with the same occurrence, i.e. $t_{ij} = t_j$ and $j = 1, \ldots, m$. Set $X_i = \{f_i(t_{ij}), j = 1, \ldots, m\} \in \mathbb{R}^m$ for $i = 1, \ldots, n$. We assume that the data have a common structure which can be modeled as a manifold embedding. Hence the sample $\mathcal{E} = \{X_1, \ldots, X_n\}$ consists of i.i.d random variables sampled from a law $Q \in \mathcal{M}$,where $\mathcal{M}$ is an unknown connected smooth submanifold of $\mathbb{R}^m$, endowed with the geodesic distance δ induced by the Riemannian metric g on $\mathcal{M} \subset \mathbb{R}^m$ (see for instance do Carmo [32]).

Under this geometrical framework, the statistical analysis of the curves should be carried out carefully, using the intrinsic geodesic distance and not the Euclidean distance, see for instance Pennec [75] . In particular, an extension of the usual notion of central value from Euclidean spaces to arbitrary manifolds is based on the Fréchet function, defined by

Definition 2.1 (Fréchet function)**.** Let $(\mathcal{M}, \delta)$ be a metric space and let $\alpha > 0$ be a given real number. For a given probability measure Q defined on the Borel σ-field of $\mathcal{M}$, the Fréchet function of Q is given by

$$F_\alpha(\mu) = \int_{\mathcal{M}} \delta^\alpha(X, \mu) Q(\mathrm{d}x), \qquad \mu \in \mathcal{M}.$$

For $\alpha = 1$ and $\alpha = 2$, the minimizers of $F_\alpha(\mu)$, if there exist, are called the Fréchet (or intrinsic) median and mean respectively. Following Koenker [62], in this chapter we will particularly deal with the intrinsic median, denoted by $\mu_I^1(Q)$ to obtain a robust estimate for the template function $f \in \mathcal{M}$. Hence define the corresponding empirical intrinsic median as

$$\widehat{\mu}_I^1 = \arg\min_{\mu \in \mathcal{M}} \sum_{i=1}^{n} \delta\left(X_i, \mu\right). \tag{2.1}$$

However, the previous estimator relies on the unobserved manifold $\mathcal{M}$ and its underlying geodesic distance δ. A popular estimator is given by the Isomap algorithm for δ. The idea is to build a simple metric graph from the data, which will be close enough from the manifold. Hence the approximation of the geodesic distance between two points depends on the length of the edges of the graph which connect these points. The algorithm approximates the unknown geodesic distance δ between all pairs of points in $\mathcal{M}$ in terms of shortest path distance between all pairs of points in a nearest neighbor graph $\mathcal{G}$ constructed from the data points $\mathcal{E}$. If the discretization of the manifold contains enough points with regards to the curvature of the manifold, hence the graph distance will be a good approximation of the geodesic distance. For details about the Isomap algorithm, see Tenenbaum et al. [96], Bernstein et al. [6], and de Silva and Tenenbaum [29].

The construction of the weighted neighborhood graph in the first step of the Isomap algorithm requires the choice of a parameter which controls the neighborhood size and therefore its success. This is made according to a $K-$rule (connecting each point with its K nearest neighbors) or $\epsilon-$rule (connecting each point with all points lying within a ball of radius ϵ) which are closely related to the local curvature of the manifold. Points which are too distant to be connected to the biggest graph are not used, making the algorithm unstable (see Balasubramanian and Schwartz [4]). In this chapter we propose a robust version of this algorithm which leads to an approximation of the geodesic distance, $\hat{\delta}$. Our version does not exclude any point and does not require any additional tuning parameter. This algorithm has been applied with success to align density curves in microarray data analysis (task known as normalization in bioinformatics) in Gallón et al. [39]. The construction of the approximated geodesic distance is detailed in Section 2.3.

Once an estimator of the geodesic distance is built, we propose to estimate the empirical Fréchet median by its approximated version

$$\widehat{\mu}^{1}_{I,n} = \arg\min_{\mu \in \mathcal{G}} \sum_{i=1}^{n} \hat{\delta}\left(X_i, \mu\right). \tag{2.2}$$

This estimator is restricted to stay within the graph $\mathcal{G}$ since the approximated geodesic distance is only defined on the graph. Hence we choose as a pattern of the observation the point which is at the *center* of the dataset, where center has to be understood with respect to the inner geometry of the observations.

2.3 The robust manifold learning algorithm

Let X be a random variable with values in an unknown connected and geodesically complete Riemannian manifold $\mathcal{M} \subset \mathbb{R}^m$, and a sample $\mathcal{E} = \{X_i \in \mathcal{M}, i = 1, \dots, n\}$ with distribution Q. Set d the Euclidean distance on $\mathbb{R}^m$ and δ the induced geodesic distance on $\mathcal{M}$. Our aim is to estimate the geodesic distance between two points on the manifold $\delta\left(X_i, X_{i'}\right)$ for all $i \neq i' \in \{1, \dots, n\}$.

The Isomap algorithm proposes to learn the manifold topology from a neighborhood graph. In the same way, our purpose is to approximate the geodesic distance δ between a pair of data points by the graph distance on the shortest path between the pair on the neighborhood graph. The main difference between our algorithm and the Isomap algorithm lies in the treatment of points which are far from the others. Indeed, the first step of the original Isomap algorithm consists in constructing the K-nearest neighbor graph or the ϵ-nearest neighbor graph for a given positive integer K or a real $\epsilon > 0$, respectively and then to exclude points which are not connected to the graph. Such a step is not present in our algorithm since we consider that a distant point is not always considered an outlier. Hence, we do not exclude any points. Moreover, a sensitive issue of the Isomap algorithm is that it requires the choice of the neighbor parameter (K or ϵ) which is closely related to the local curvature of the manifold, determining the quality of the construction (see, for instance, Balasubramanian and Schwartz [4]). In our algorithm, we give a tuning parameter free process to simplify the analysis.

The algorithm has three steps. The first step constructs a complete weighted graph associated to $\mathcal{E}$ based on Euclidean distances $\mathrm{d}(X_i, X_{i'})$ between all pairwise points $X_i, X_{i'} \in \mathbb{R}^m$. It is a complete Euclidean graph $\mathcal{G}_{\mathrm{E}} = (\mathcal{E}, E)$ with set of nodes

$\mathcal{E}$ and set of edges $E = \{\{X_i, X_{i'}\}, \ i = 1, \ldots, n-1, \ i' = i+1, \ldots, n\}$ weighted with the corresponding Euclidean distances.

In the second step, the algorithm obtains the Euclidean Minimum Spanning Tree $\mathcal{G}_{\mathrm{MST}} = (\mathcal{E}, E_{\mathrm{T}})$ associated to $\mathcal{G}_{\mathrm{E}}$, i.e. the spanning tree that minimizes the sum of the weights of the edges in the spanning tree of $\mathcal{G}_{\mathrm{E}}$, $\sum_{\{X_i, X_{i'}\} \in E_{\mathrm{T}}} \mathrm{d}(X_i, X_{i'})$. The underlying idea in this construction is that, if two points X_i and $X_{i'}$ are relatively close, then we have that $\delta(X_i, X_{i'}) \approx \mathrm{d}(X_i, X_{i'})$. This may not be true if the manifold is very twisted and/or if too few points are observed, and may induce bad approximations. So the algorithm will produce a good approximation for relatively regular manifolds. This drawback is well known when dealing with graph-based approximations of the geodesic distance (Tenenbaum et al. [96], and de Silva and Tenenbaum [29]).

An approximation of $\delta(X_i, X_{i'})$ is provided by the sum of all the Euclidean distances of the edges of the shortest path on $\mathcal{G}_{\mathrm{MST}}$ connecting X_i to $X_{i'}$, i.e. $\hat{\delta}(X_i, X_{i'}) = \min_{g_{ii'} \in \mathcal{G}_{\mathrm{MST}}} L(g_{ii'})$, where $L(g_{ii'})$ denotes the length of a path $g_{ii'}$ connecting X_i to $X_{i'}$ on $\mathcal{G}_{\mathrm{MST}}$. However, this construction is highly unstable since the addition of new points may change completely the structure of the graph.

To cope with this problem, we propose in the third stage to add more robustness in the construction of the approximation graph. Actually, in our algorithm we add more edges between the data points to add extra paths and thus to cover better the manifold. The underlying idea is that paths which are close to the ones selected in the construction of the $\mathcal{G}_{\mathrm{MST}}$ could also provide good alternate ways of connecting the edges. Closeness here is understood as lying in open balls $B(X_i, \epsilon_i) \subset \mathbb{R}^m$ around the point X_i with radius $\epsilon_i = \max_{\{X_i, X_i\} \in E_{\mathrm{T}}} \mathrm{d}(X_i, X_{i'})$. Hence, these new paths between the data are admissible and should be added to the edges of the graph. Finally, we obtain a new robustified graph $\mathcal{G}' = (\mathcal{E}, E')$ defined by

$$\{X_i, X_{i'}\} \in E' \Longleftrightarrow \overline{X_i X_{i'}} \subset \bigcup_{i=1}^{n} B(X_i, \epsilon_i),$$

where

$$\overline{X_i X_{i'}} = \{X \in \mathbb{R}^m, \ \exists \lambda \in [0,1], \ X = \lambda X_i + (1-\lambda) X_{i'}\}.$$

Finally, $\mathcal{G}'$ is the graph which gives rise to our estimator of δ, given by

$$\hat{\delta}(X_i, X_{i'}) = \min_{g_{ii'} \in \mathcal{G}'} L(g_{ii'}). \tag{2.3}$$

Hence, $\hat{\delta}$ is the distance associated with $\mathcal{G}'$, that is, for each pair of points X_i and $X_{i'}$, we have $\hat{\delta}(X_i, X_{i'}) = \mathrm{L}(\hat{\gamma}_{ii'})$ where $\hat{\gamma}_i$ is the minimum length path between X_i and $X_{i'}$ associated to $\mathcal{G}'$. We point out that all points of the data sets are connected in the new graph $\mathcal{G}'$.

Algorithm 2.1 Robust approximation of δ

```
Require: $\mathcal{E} = \{X_i \in \mathbb{R}^m,\ i = 1, \dots, n\}$
Ensure: $\hat{\delta}$
  1: Calculate $\mathrm{d}(X_i, X_{i'}) = \|X_i - X_{i'}\|_2$ between all pairwise data points $X_i$ and $X_{i'}$,
     $i = 1, \dots, n-1,\ i' = i+1, \dots, n$, and construct the complete Euclidean graph
     $\mathcal{G}_{\mathrm{E}} = (\mathcal{E}, E)$ with set of edges $E = \{\{X_i, X_{i'}\}\}$.
  2: Obtain the Euclidean Minimum Spanning Tree $\mathcal{G}_{\mathrm{MST}} = (\mathcal{E}, E_{\mathrm{T}})$ associated to
     $\mathcal{G}_{\mathrm{E}}$.
  3: For each $i = 1, \dots, n$ calculate $\epsilon_i = \max_{\{X_i, X_{i'}\} \in E_{\mathrm{T}}} \mathrm{d}(X_i, X_{i'})$, and open balls
     $B(X_i, \epsilon_i) \subset \mathbb{R}^m$ of center $X_i$ and radius $\epsilon_i$. Construct a graph $\mathcal{G}' = (\mathcal{E}, E')$
     adding more edges between points according to the rule
         $$\{X_i, X_{i'}\} \in E' \Longleftrightarrow \overline{X_i X_{i'}} \subset \bigcup_{i=1}^{n} B(X_i, \epsilon_i),$$
     where $\overline{X_i X_{i'}} = \{X \in \mathbb{R}^m,\ \exists \lambda \in [0,1],\ X = \lambda X_i + (1-\lambda) X_{i'}\}$.
  4: Estimate the geodesic distance between two points by the length of the shortest
     path in the graph $\mathcal{G}'$ between these points using the Floyd's or Dijkstra's
     algorithm (see, e.g. Lee and Verleysen [66]).
```

A summary of the procedure is gathered in the Algorithm 2.1 and its corresponding R code is given in the Appendix B.

Note that, the 3-step algorithm described above contains widespread graph-based methods to achieve our purpose. In this article, all graph-based calculations, such as Minimum Spanning Tree estimation or shortest path calculus, were carried out with the igraph package for network analysis by Csárdi and Nepusz [23].

An illustration of the algorithm and its behavior when the number of observations increases are displayed respectively in Figures 2.1 and 2.2. In Figure 2.1, points $(X_i^1, X_i^2)_i$ are simulated as follows:

$$X_i^1 = \frac{2i-n-1}{n-1} + \epsilon_i^1, \quad \text{and} \quad X_i^2 = 2\left(\frac{2i-n-1}{n-1}\right)^2 + \epsilon_i^2, \tag{2.4}$$

where ϵ_i^1 and ϵ_i^2 are independent and normally distributed with mean 0 and variance 0.01 for $i = 1, \dots, n$ and $n = 30$. In Figure 2.2, some results of graph $\mathcal{G}'$ for $n = 10, 30, 100$ are given. We can see that graph $\mathcal{G}'$ tends to be close to the true manifold $\{(t, t^2) \in \mathbb{R}^2,\ t \in \mathbb{R}\}$ when n increases.

Obviously, this estimation shows that the recovered structures in Figures 2.1 and 2.2 are pretty sensitive to noise. Nevertheless, to estimate a representative of a sample of curves, a prior smoothing step is almost always carried out as in Ramsay and Silverman [79]. This is done in Section 2.6 for our real data sets.

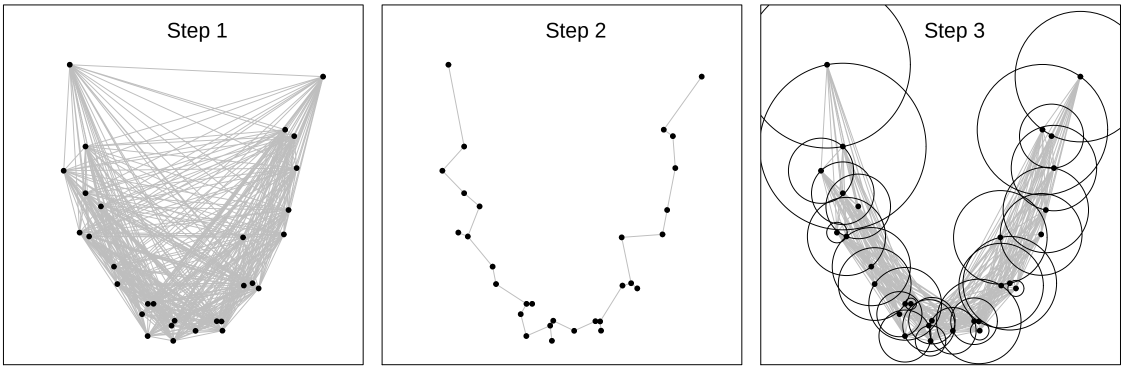

Figure 2.1: The 3-step construction of a subgraph $\mathcal{G}'$ from Simulation (2.4). On the left, the simulated data set (black dots) and the associated complete Euclidean graph $\mathcal{G}_{\mathrm{E}}$ (Step 1). On the middle, the $\mathcal{G}_{\mathrm{MST}}$ associated with the complete graph $\mathcal{G}_{\mathrm{E}}$ (Step 2). On the right, the associated open balls and the corresponding subgraph $\mathcal{G}'$ (Step 3).

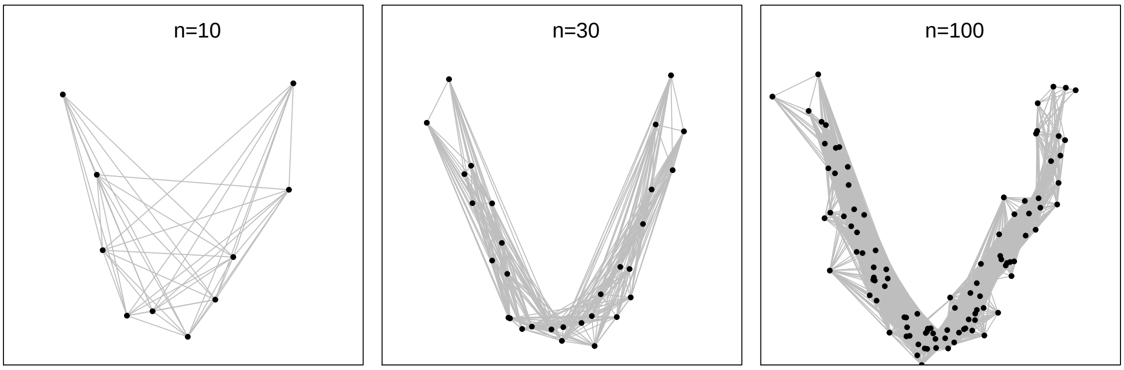

Figure 2.2: Evolution of graph $\mathcal{G}'$ from Simulation (2.4) for $n = 10, 30, 100$

2.4 Application: template estimation in a shape invariant model

In this section, we consider the case where the observations are curves warped from an unknown template $f\colon \mathcal{T} \to \mathbb{R}$. We want to study whether the *central* curve defined previously as the median of the data with respect to the geodesic distance provides a good pattern of the curves. Good means, in that particular case, that the intrinsic median should be close to the pattern f.

We consider a translation model indexed by a real valued random variable A with unknown distribution on an interval $(b, c) \subset \mathbb{R}$

$$X_{ij} = f_i(t_j) = f\left(t_j - A_i\right),\ i = 1, \dots, n,\ j = 1, \dots, m, \tag{2.5}$$

where $(A_i)_i$ are i.i.d random variables drawn with distribution A which models the unknown shift parameters. This specification is an special case of the self-modeling regression mentioned in the introduction.

Under a nonparametric registration model, Maza [70] and Dupuy et al. [35] define the *structural expectation* function of a sample of curves and build a registration procedure in order to estimate it. Following the same philosophy, but for the case of the translation effect model (2.5), we propose to use as a good pattern of the dataset the *Structural Median* function f_{SM} defined as

$$f_{\mathrm{SM}} = f\left(\cdot - \operatorname{med}(A)\right), \tag{2.6}$$

where $\operatorname{med}(A)$ denotes the median of A.

We will see that the manifold embedding point of view enables to recover this pattern. Actually, define a one-dimensional function in $\mathcal{M} \subset \mathbb{R}^m$ parameterized by a parameter $a \in (b, c) \subset \mathbb{R}$ as

$$\begin{aligned} X \colon (b,c) &\to \mathbb{R}^m \\ a \quad &\mapsto X(a) = \left(f\left(t_1 - a\right), \ldots, f\left(t_m - a\right)\right), \end{aligned}$$

and set $\mathcal{C} = \{X(a) \in \mathbb{R}^m,\ a \in (b,c)\}$.

As soon as X is a regular curve, that is, if its first derivative never vanishes,

$$X' \neq 0 \Longleftrightarrow \forall a \in (b,c),\ \exists j \in \{1, \ldots, m\},\ f'\left(t_j - a\right) \neq 0, \tag{2.7}$$

then, the smooth mapping $X \colon a \mapsto X(a)$ provides a natural parametrization of $\mathcal{C}$ which can thus be seen as a submanifold of $\mathbb{R}^m$ of dimension 1 (do Carmo [32]). The corresponding geodesic distance is given by

$$\delta\left(X(a_1), X(a_2)\right) = \left| \int_{a_1}^{a_2} \|X'(a)\| \,\mathrm{d}a \right|, \tag{2.8}$$

with $X'(a) = \mathrm{d}X(a)/\mathrm{d}a = \left(\mathrm{d}X_1(a)/\mathrm{d}a, \ldots, \mathrm{d}X_m(a)/\mathrm{d}a\right)^\top$.

The observation model (2.5) can then be seen as a discretization of the manifold $\mathcal{C}$ for different values $(A_i)_i$. Hence, finding the intrinsic median of all shifted curves can be done by understanding the geometry of space $\mathcal{C}$, and thus, by approximating the geodesic distance between observed curves. Define the intrinsic median with respect to the geodesic distance (2.8) on $\mathcal{C}$, that is

$$\widehat{\mu}_I^1 = \underset{\mu \in \mathcal{C}}{\arg\min} \sum_{i=1}^{n} \delta\left(X_i, \mu\right). \tag{2.9}$$

The following theorem gives a minimizer, whose proof is provided in the Appendix A.

Theorem 2.1. *Under the assumption (2.7) that X is a regular curve, we get that*

$$\widehat{\mu}_I^1 = \left(f\left(t_1 - \widehat{\operatorname{med}}(A)\right), \ldots, f\left(t_m - \widehat{\operatorname{med}}(A)\right)\right),$$

where $\widehat{\operatorname{med}}(A)$ is the empirical median.

Hence as soon as we observe a sufficient number of curves to ensure that the median and the empirical median are close, the intrinsic median is a natural approximation of (2.6). Therefore, the manifold framework provides a geometrical interpretation of the structural median of a sample of curves. The estimator is thus given by

$$\widehat{\mu}_{I,n}^1 = \arg\min_{\mu \in \mathcal{E}} \sum_{i=1}^{n} \widehat{\delta}\left(X_i, \mu\right), \tag{2.10}$$

where $\widehat{\delta}$ is an approximation of the unknown underlying geodesic distance, that is estimated by the algorithm described in Section 2.3.

We point out that in many situations, giving a particular model for the deformations corresponds actually to consider a particular manifold embedding for the data. Once the manifold is known, its corresponding geodesic distance may be properly computed, as done in the translation case. So in some particular cases, the minimization in (2.9) can give an explicit formulation and then it is possible to identify the resulting Fréchet median. Hence previous theorem may be generalized to such cases as done in Gallón et al. [39].

Note first that this case only holds for the Fréchet median ($\alpha = 1$) but not the mean for which the so-called structural expectation and the Fréchet mean are different. Moreover, the choice of the median is also driven by the need for a robust method, whose good behavior will be highlighted in the simulations and applications in the following sections.

As shown in the simulation study below, when the observations can be modeled by a set of curves warped from an unknown template by a general deformation process, estimate (2.10) enables to recover the main pattern in a better way than classical methods. Obviously, the method relies on the assumption that all the observed data belong to an embedded manifold whose geodesic distance can be well approximated by the proposed algorithm.

2.5 Simulation study

IIn this section, the numerical properties of our estimator, called Robust Manifold Embedding (RME), defined by the equation (2.10) in Section 2.4 are studied using simulated data. The estimator is compared to those obtained with the Isomap algorithm and the Modified Band Median (MBM) estimator proposed by Arribas-Gil and Romo [3], which is based on the concept of depth for functional data (see López-Pintado and Romo [68]). The behavior of the estimator when the number of curves increases is also analyzed.

Four different types of simulations of increasing warping complexity for the single shape invariant model were carried out, observing $n = 15, 30, 45, 60$ curves on $m = 100$ equally spaced discrete points $(t_j)_j$ in the interval $[-10, 10]$. The experiment was conducted with $R = 100$ repetitions. The template function f and shape invariant model, for each simulation, are given as follows:

***Simulation 1*:** One-dimensional manifold defined by $f(t) = 5\sin(t)/t$ and

$$X_{ij} = f\left(t_j + A_i\right),$$

where $(A_i)_i$ are i.i.d uniform random variables on interval $[-5, 5]$.

***Simulation 2*:** Two-dimensional manifold given by $f(t) = 5\sin(t)$ and

$$X_{ij} = f\left(A_i t_j + B_i\right),$$

where $(A_i)_i$ and $(B_i)_i$ are independent and (respectively) i.i.d uniform random variables on intervals $[0.7, 1.3]$ and $[-1, 1]$.

***Simulation 3*:** Four-dimensional manifold given by $f(t) = t\sin(t)$ and

$$X_{ij} = A_i f\left(B_i t_j + C_i\right) + D_i,$$

where $(A_i)_i$, $(B_i)_i$, $(C_i)_i$ and $(D_i)_i$ are independent and (respectively) i.i.d uniform random variables on intervals $[0.7, 1.3]$, $[0.7, 1.3]$, $[-1, 1]$ and $[-1, 1]$.

***Simulation 4*:** Four-dimensional manifold given by $f(t) = \phi t + t\sin(t)\cos(t)$ with $\phi = 0.9$, and

$$X_{ij} = A_i f\left(B_i t_j + C_i\right) + D_i,$$

where $(A_i)_i$, $(B_i)_i$, $(C_i)_i$ and $(D_i)_i$ as in the Simulation 3.

Figure 2.3 illustrates the simulated data sets from Simulations 1-4 with $n = 30$ curves for one simulation run (one of 100 repetitions). For Simulation 1, where there is only phase variability, all methods follows the structural characteristics of the sample of curves, where the template estimated by the robust manifold approach is the closest curve to the theoretical function. The same conclusion can be inferred from Simulation 2. Indeed, for this simulation type, and for this particular simulation run, the RME estimator coincides with the theoretical template function. For the four-dimensional manifolds in Simulations 3 and 4, where there is an additional amplitude variability, the robust manifold estimator captures better the structural pattern in the sample of curves followed by the Isomap estimator. Note that in the Simulation 4, both approaches coincide. Although the MBM estimator follows the shape of the theoretical template, the estimator deviates from it in the cases 2-4.

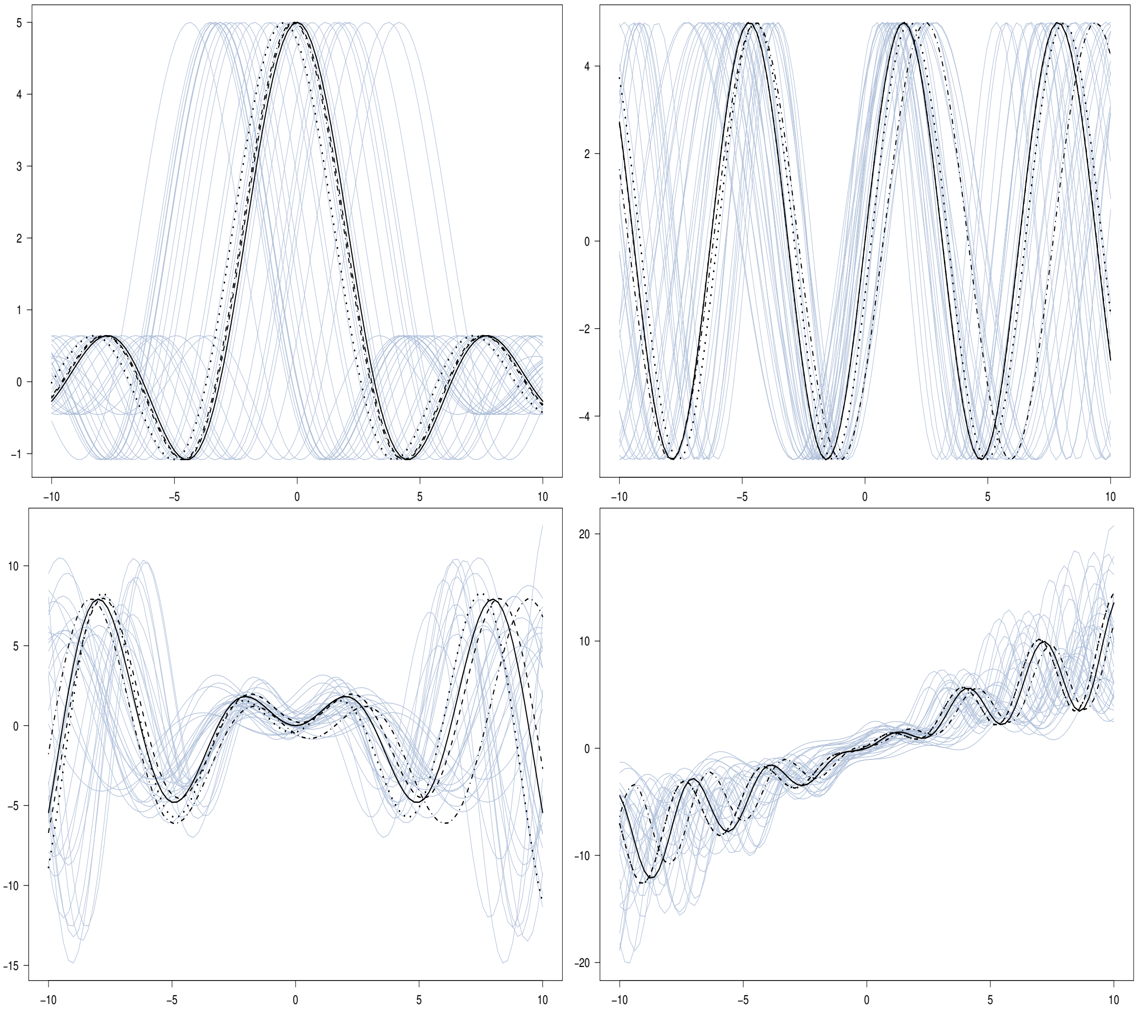

Figure 2.3: Simulated curves (gray) from Simulation 1 (top left), 2 (top right), 3 (bottom left) and 4 (bottom right) for one simulation run, and the respective target template function f (solid line), MBM (dash-dotted line), Isomap (dotted line), and RME (dashed line) template estimators.

In order to compare more accurately the estimators described above, we calculate, for each one, the empirical mean squared error obtained on the $R = 100$ repetitions of each type of simulation. We recall the definition, for estimator $\hat{f}$ of a given type of simulation, of the mean squared error:

$$\text{Mean Squared Error}\left(\hat{f}\right) = \frac{1}{R}\sum_{r=1}^{R} \|\hat{f}_r - f\|_2^2,$$

where, $\hat{f}_r$ is the estimation from the r-th repetition of the given simulation type, f is the true template function and $\|\cdot\|_2$ is the classical Euclidean norm. We also highlight, for our comparisons, the fact that

$$\text{Mean Squared Error}\left(\hat{f}\right) = \underbrace{\frac{1}{R}\sum_{r=1}^{R}\|\hat{f}_r - \bar{f}\|_2^2}_{\text{Variance}} + \underbrace{\|\bar{f} - f\|_2^2}_{\text{Squared bias}},$$

where $\bar{f}$ is the mean of all R obtained estimations.

Table 2.1 shows the mean squared errors, variances and squared biases of each estimator for simulations 1, 2, 3 and 4, and for different number on curves $n = 15, 30, 45, 60$ in the sample. Values have been rounded to zero decimal places to facilitate the comparisons, and the minimum values are signed in bold.

Table 2.1: Comparison of estimators for simulations 1-4 with different sample sizes.

n	**Statistic**	**Simulation1**			**Simulation 2**		
		MBM	Isomap	RME	MBM	Isomap	RME
15	MSE	**136**	389	335	790	**400**	435
	Bias2	**23**	152	118	141	**35**	46
	Variance	**113**	236	217	649	**366**	389
30	MSE	**30**	108	92	666	338	**268**
	Bias2	**8**	13	10	98	34	**18**
	Variance	**22**	95	82	568	304	**249**
45	MSE	**24**	139	66	669	244	**155**
	Bias2	10	23	**5**	120	20	**9**
	Variance	**14**	116	60	549	224	**147**
60	MSE	**14**	85	55	634	168	**136**
	Bias2	5	9	**4**	161	5	**4**
	Variance	**10**	76	51	472	163	**132**
n	**Statistic**	**Simulation 3**			**Simulation 4**		
		MBM	Isomap	RME	MBM	Isomap	RME
15	MSE	1350	**1152**	1171	**876**	890	893
	Bias2	**251**	375	441	**394**	512	522
	Variance	1098	777	**730**	483	378	**371**
30	MSE	1025	**673**	721	911	**861**	876
	Bias2	212	**160**	248	**470**	536	554
	Variance	813	513	**473**	441	325	**323**
45	MSE	1034	553	**498**	**820**	827	868
	Bias2	223	**105**	141	**397**	524	569
	Variance	811	449	**356**	423	303	**299**
60	MSE	965	572	**402**	879	**776**	842
	Bias2	168	**97**	122	**458**	474	563
	Variance	797	475	**280**	421	302	**279**

From the table, we observe that when the number of curves in the sample is small ($n = 15$) the MBM estimator has better results in terms of the MSE, Bias2 and variance for the Simulation 1. The same is true when $n = 30$. With $n = 45, 60$

curves the MBM estimator has minimum mean squared error and variance, and our estimator has smaller bias. Comparing the Isomap and RME methods only, the latter overcomes the former. For Simulation 2, the RME estimator overcomes the MBM and Isomap estimators for $n = 30, 45, 60$ curves, except when $n = 15$, where the Isomap estimator is better. However, in this case there are not big differences between Isomap and RME estimators. As we expected, when the geometry of the curves is more complex, i.e. when we have a four-dimensional manifolds, the results are more variated. For Simulation 3, the RME estimator has a good performance with $n = 45, 60$. With $n = 15, 30$ the better results are shared by the MBM and Isomap methods. In Simulation 4, the MBM estimator has, in general, better results. Finally, note that although the theorem in Section 2.4 is valid for one-dimensional manifolds generated by time shifts (Simulation 1), we can see that the intrinsic sample median estimator by approximating the corresponding geodesic distance with the robust algorithm performs well for manifolds of higher dimension (Simulations 2-4).

2.5.1 Robustness analysis

In order to assess the robustness of the RME estimator, we carried out an additional simulation study generating atypical curves in the functional data sets. In particular, from $n = 15, 30, 45, 60$ curves we generated 10% of them as atypical according to the single shape invariant model in the four type of simulations considered above, modifying the corresponding individual shift parameters but preserving the geometric structure of the curves. So, for each simulation, the non-atypical curves X_{ij} with $i = 1, \ldots, (n - \lceil 0.10n \rceil)$ are generated as above, and the atypical curves $\tilde{X}_{ij}$ with $i = (n - \lceil 0.10n \rceil) + 1, \ldots, n$ were generated as:

***Simulation 1*:** One-dimensional manifold defined by $f(t) = 5\sin(t)/t$ and

$$\tilde{X}_{ij} = f\left(t_j + \tilde{A}_i\right),$$

where $\left(\tilde{A}_i\right)_i$ are i.i.d uniform random variables on interval $[4.5, 6]$.

***Simulation 2*:** Two-dimensional manifold given by $f(t) = 5\sin(t)$ and

$$\tilde{X}_{ij} = f\left(\tilde{A}_i t_j + \tilde{B}_i\right),$$

where $\left(\tilde{A}_i\right)_i$ and $\left(\tilde{B}_i\right)_i$ are independent and (respectively) i.i.d uniform random variables on intervals $[0.35, 0.65]$ and $[-0.5, 0.5]$.

***Simulation 3*:** Four-dimensional manifold given by $f(t) = t\sin(t)$ and

$$\tilde{X}_{ij} = \tilde{A}_i f\left(\tilde{B}_i t_j + \tilde{C}_i\right) + \tilde{D}_i,$$

where $\left(\tilde{A}_i\right)_i$, $\left(\tilde{B}_i\right)_i$, $\left(\tilde{C}_i\right)_i$ and $\left(\tilde{D}_i\right)_i$ are independent and (respectively) i.i.d uniform random variables on intervals $[1.3, 1.4]$, $[0.7, 1.3]$, $[-1.5, -1]$ and $[1, 1.5]$.

***Simulation 4*:** Four-dimensional manifold given by $f(t) = \phi t + t \sin(t) \cos(t)$ with $\phi = 0.9$, and

$$\tilde{X}_{ij} = \tilde{A}_i f\left(\tilde{B}_i t_j + \tilde{C}_i\right) + \tilde{D}_i,$$

where $\left(\tilde{A}_i\right)_i$, $\left(\tilde{B}_i\right)_i$, $\left(\tilde{C}_i\right)_i$ and $\left(\tilde{D}_i\right)_i$ are independent and (respectively) i.i.d uniform random variables on intervals $[1.05, 1.95]$, $[1.05, 1.95]$, $[-1, 1]$ and $[-1, 1]$.

Figure 2.4 illustrates the simulated data sets from Simulations 1-4 with $n = 30$ curves for one replication. The curves signed in red color correspond to the atypical curves. For this particular simulation run, we see that the atypical curves has influence over the Isomap estimator for all types of simulation. For the one and two-dimensional manifolds in Simulations 1 and 2 respectively we observe that the RME estimator has a good performance. For example, note, as in the simulation study without atypical curves developed above, the RME estimator coincides with the theoretical template function for the simulation type 1, and for this particular simulation run. For complex shape functions as in Simulations 3 and 4, our estimator captures adequately the common pattern of the sample in presence on atypical curves. As expected, the depth-based estimator is robust against atypical observations.

The mean squared errors, variances and squared biases of each estimator for Simulations 1-4 and different number on curves $n = 15, 30, 45, 60$ including atypical curves are showed in the Table 2.2. For Simulation 1, the MBM estimator has the best results for all number of curves. In this case, the RME method overcomes its not robust version estimator (Isomap). Additionally, when the warping complexity increases, the RME estimator has minimum mean squared errors in most cases for Simulations 2-4. As expected, only the when the number of curves is small ($n = 15$) the estimator performs less well.

2.6 Applications

In this section we apply the proposed robust manifold learning algorithm to extract the template function of a sample of curves on three real datasets of functional data: the well-known Berkeley Growth and Gait data in functional data applications (Ramsay and Silverman [79]), and a reflectance data of two landscape types. Our algorithm is compared with the Isomap and Modified Band Median methods.

2.6.1 Berkeley growth study

The data of the Berkeley's study consist in 31 height measurements for 54 girls and 38 boys recorded between the ages of 1 and 18 years. Intervals between measurements range from 3 months (age 1-2 years), to yearly (age 3-8), to half-yearly (age 8-18).

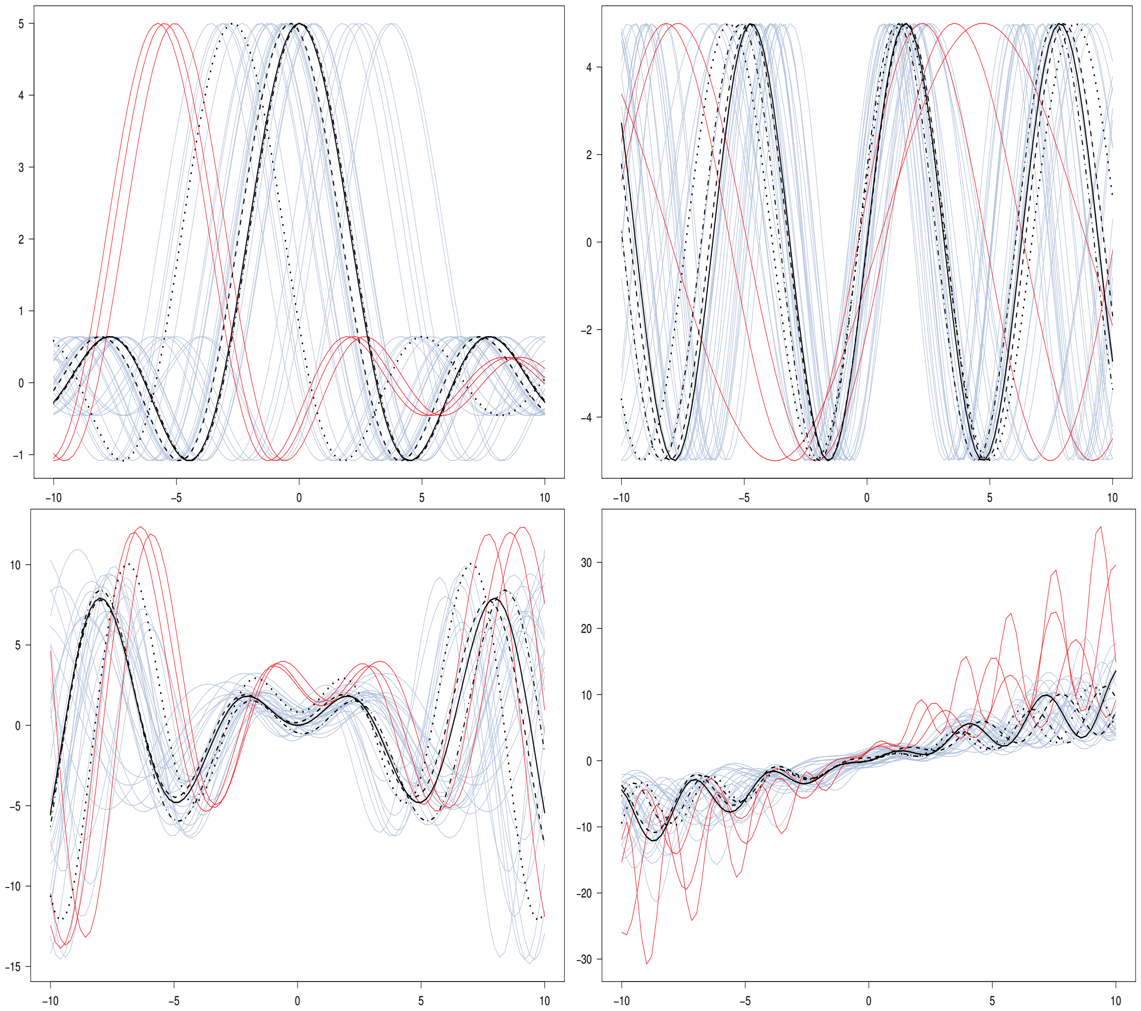

Figure 2.4: Simulated curves (gray) from simulation type 1 (top left), 2 (top right), 3 (bottom left) and 4 (bottom right) including atypical curves (red) for one simulation run, and the respective target template function f (solid line), MBM (dash-dotted line), Isomap (dotted line), and RME (dashed line) template estimators.

One of the goals with this kind of data is the pattern analysis of growth velocity and acceleration curves, represented by the first and second derivatives of the height functions, in order to characterize its spurts and trends during years. The velocity and acceleration curves for girls and boys were obtained by taking the first and second order differences, respectively, of the height curves, whose functional representations were made using a B-spline smoothing (see Ramsay and Silverman [79] for details).

Figure 2.5 provides the smoothed velocity curves (on the top) measured in centimeters per year (cm/year) and the smoothed acceleration curves (on the bottom) measured in centimeters per squared year (cm/year^2) of height for girls (on the left) and for boys (on the right). It is evident that all individuals exhibit a

Table 2.2: Comparison of estimators for simulations 1-4 including atypical curves for different sample sizes.

n	Statistic	Simulation1			Simulation 2		
		MBM	Isomap	RME	MBM	Isomap	RME
15	MSE	**107**	452	400	830	**570**	680
	Bias2	**21**	215	169	186	**76**	116
	Variance	**85**	237	232	645	**495**	564
30	MSE	**36**	177	166	649	409	**300**
	Bias2	**8**	49	46	115	39	**20**
	Variance	**28**	128	120	535	370	**280**
45	MSE	**21**	121	81	523	296	**212**
	Bias2	**9**	32	26	89	23	**11**
	Variance	**13**	89	56	433	273	**200**
60	MSE	**19**	151	90	551	276	**212**
	Bias2	**9**	45	40	98	63	**46**
	Variance	**10**	106	51	453	213	**166**
n	Statistic	Simulation 3			Simulation 4		
		MBM	Isomap	RME	MBM	Isomap	RME
15	MSE	1387	**1093**	1098	990	983	**963**
	Bias2	**257**	327	396	**485**	561	538
	Variance	1129	766	**702**	505	**422**	425
30	MSE	1370	**856**	857	901	861	**856**
	Bias2	260	**204**	312	**434**	505	511
	Variance	1110	652	**545**	467	355	**345**
45	MSE	1206	640	**547**	874	863	**860**
	Bias2	234	**155**	197	556	541	**406**
	Variance	972	484	**350**	**317**	322	454
60	MSE	963	585	**462**	924	864	**861**
	Bias2	154	**118**	165	**500**	537	561
	Variance	809	468	**297**	424	327	**301**

common velocity and acceleration pattern throughout years, but features as peaks, troughs and inflection points occur at different times for each child.

From all of the graphs in the Figure 2.5, we see, in general, that all the template estimators obtain a curve situated in the middle of the samples of curves capturing its common shape pattern appropriately. For the case of sample velocity curves of girls (top-left graph) the RME and MBM estimators coincide. The Isomap estimator deviates slightly from the center. In the case of samples of velocity and acceleration curves of boys, both the RME and Isomap estimators choose the same template function. Only in the case of acceleration curves of girls (bottom-left graph), the three methods choose different functions. In summary, we infer that the RME estimator seems to perform a good work extracting a meaningful shape curve.

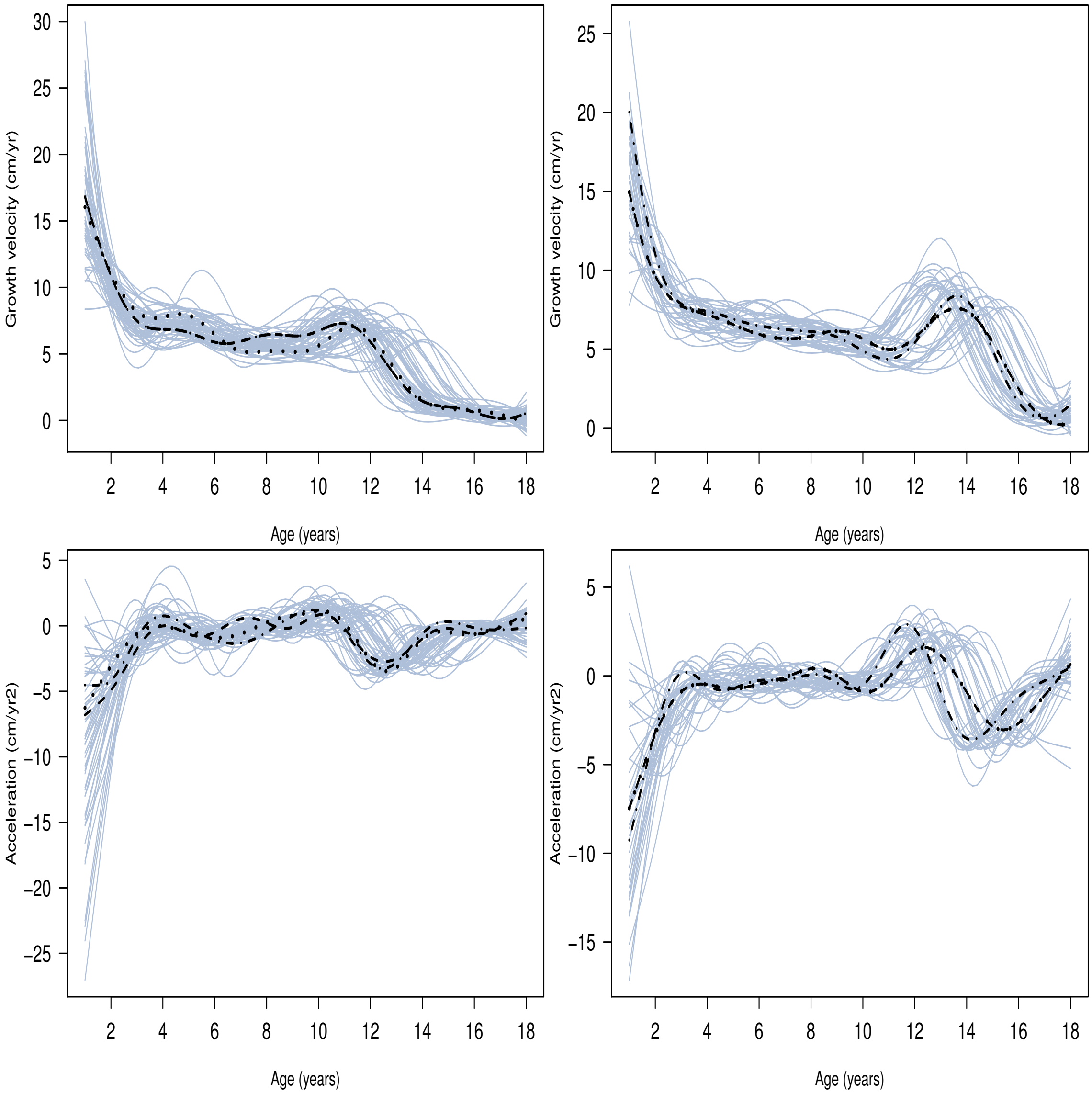

Figure 2.5: Velocity (on the top) and acceleration (on the bottom) curves of 54 girls (on the left) and 31 boys (on the right) in the Berkeley growth study (gray lines). The estimated template functions with the MBM (dashed line), Isomap (dashed-dotted line), and RME (solid line) methods.

2.6.2 Gait cycle data

For this application, we consider the data of angle measurements (in degrees) in the sagittal plane formed by the hip and knee of 39 children through a gait cycle, where time is measured in terms of the child's gait cycle such that every curve is given for values ranging between 0 and 1. The smoothed curves were obtained by fitting a Fourier basis system following the analysis of Ramsay and Silverman [79] for this data, where both sets of curves are periodic. Figure 2.6 displays the curves of hip (on the left) and knee (on the right) angles observed during the gait. As we can see, a two-phase process can be identified for the knee motion, while for the hip motion there is a single-phase. Also, both sets of curves share a common pattern around which there are both phase and amplitude variability.

For this application, the template functions obtained by the Robust Manifold Estimator based on our algorithm seem to capture the salient features of the sample of hip and knee angle curves. Note also that the same template, located in the center of the samples, was chosen by all the estimators.

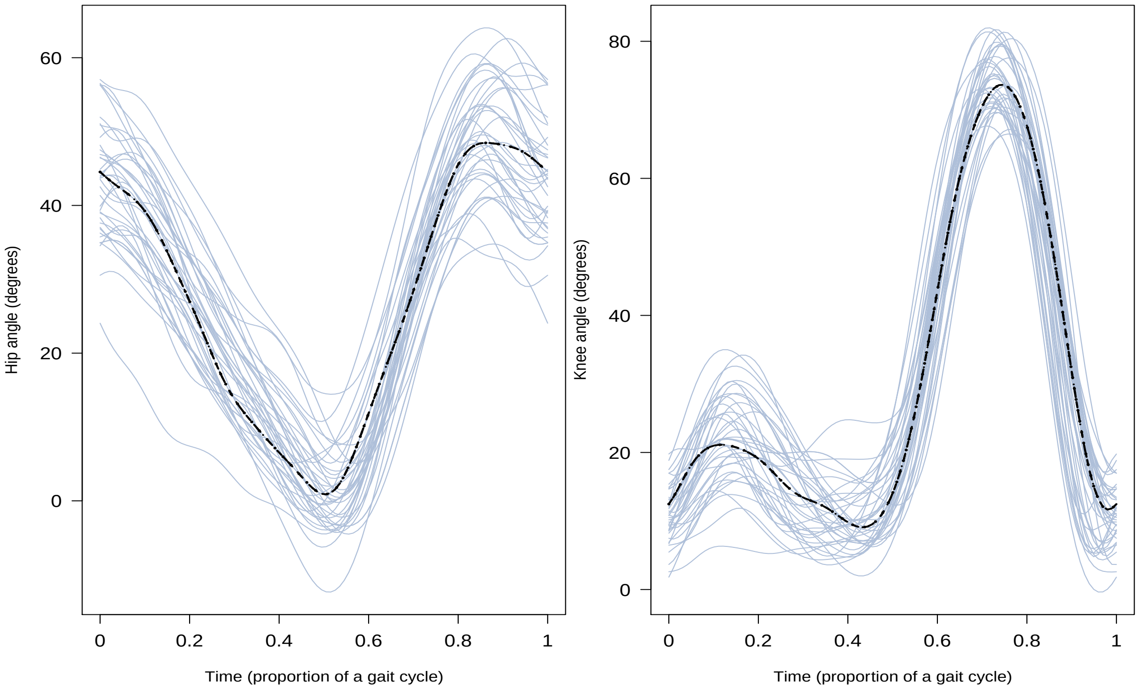

Figure 2.6: Angle curves formed by the hip (on the left) and knee (on the right) of 39 children through a gait cycle, and the MBM (dash-dotted line), Isomap (dotted line), and RME (dashed line) template estimators.

2.6.3 Landscape reflectances data

Finally, we consider two data sets where the corresponding observed curves represent the weekly reflectance profiles of two particular landscapes (corn and wheat). The reflectance is a measure of the incident electromagnetic radiation that is reflected by a given interface. For these data, there are 23 and 124 curves for corn and wheat landscapes respectively. The aim consists in extracting a representative curve of a type of landscape while observing the reflectance profiles of different landscapes of the same type. In Figure 2.7, the smoothed curves corresponding to reflectance patterns of two landscape types in the same region in the same period are showed. The smoothing was obtained from discrete data with B-spline basis system. The reflectance depends on the vegetation whose growth depends on the weather condition and the soil behavior. It is therefore relevant to consider that these profiles are deformations in translation, scale and amplitude of a single representative function of the reflectance behavior of each landscape type in this region at this time.

In Figure 2.7, we observe that for the corn landscape case, where there are relatively a few number of curves, the robust manifold estimator chooses a meaningful template curve which seems to appear at the center of curve sample, which coincides with the curve obtained by the modified band median estimator. The Isomap estimator chooses a different curve as representative function which is slightly away from the center of the sample. For the wheat landscape, all of three estimators choose a different template curve. Although all the estimated template curves follow the structural features of the sample, the RME estimator select a curve that is located more in the middle. In this application domain, extracting a curve by RME is best able to report data as reflecting their structure and thus to obtain a better representative and improve further future functional analysis.

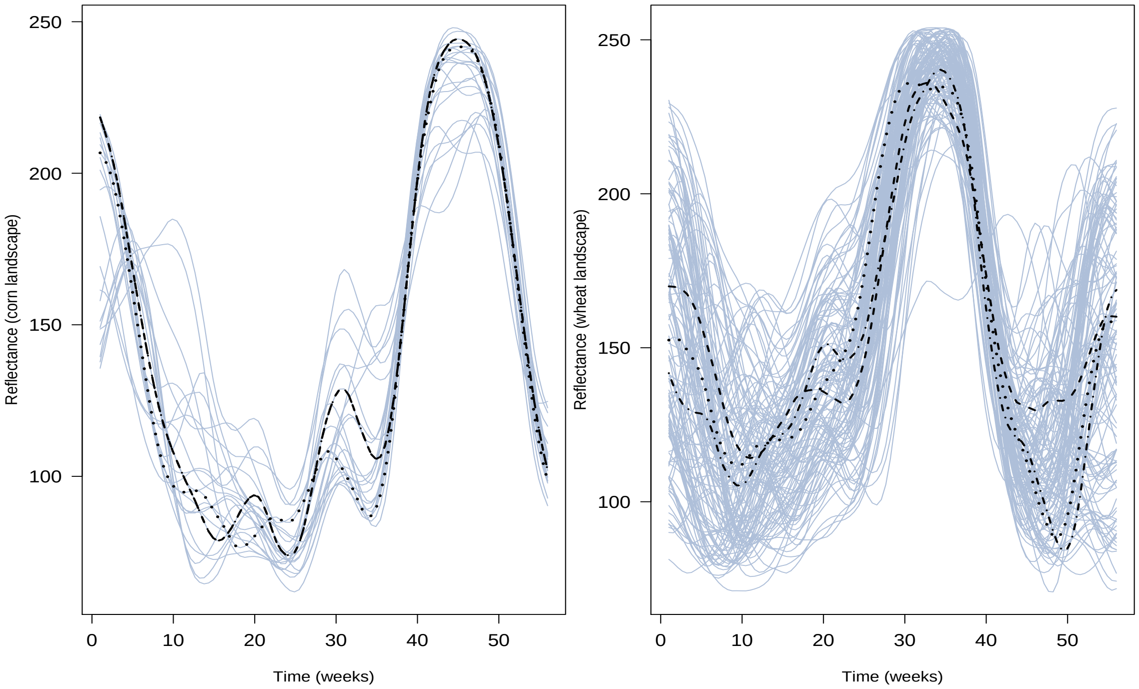

Figure 2.7: Reflectance curves of corn (left) and wheat (right) landscapes, and MBM (dash-dotted line), Isomap (dotted line), and RME (dashed line) template estimators.

2.7 Concluding remarks

In this chapter, we have proposed a robust algorithm to approximate the geodesic distance of the underlying manifold. This approximated distance is used to build an empirical Fréchet median of the functions. This function is a meaningful template curve for a sample of functions, which have both amplitude and time deformations.

Our approach relies on the fundamental paradigm of functional data analysis which involves treating the entire observed curve as a unit of observation rather than individual measurements from the curve. Indeed, we show that, when the structure of the deformations entails that the curve can be embedded into a manifold, finding a representative of a sample of curves corresponds to calculate an intrinsic statistic of observed curves on their unknown underlying manifold. Moreover in a translation model, i.e where the curves are actually warped from an unknown pattern, both methodologies coincide since the structural median of a sample of curves corresponds to the intrinsic median on a one-dimensional manifold. Moreover, we show that our method improves the performance of other pattern extraction methods, for simulated and real data sets.

From a computational point of view, our method is inspired by the ideas of the Isomap algorithm. We note that we have also used the Isomap algorithm in the simulation study and the applications with some similar results with to respect to our algorithm. Hence, our algorithm has the advantage of being parameter free and then it is of easiest use. One of the major drawbacks of these methodologies are that a relatively high number of data are required in order to guarantee a good approximation of the geodesic distance at the core of this work (see Tenenbaum et al. [96]). This drawback is clearly related with the high variance of our estimator discussed previously and should be outperformed with further work. But, anyway, we show that our method improves the performance of other classical ones.

Appendix A

Proof of Theorem 2.1. Let X be defined by

$$\begin{aligned} X \colon (b,c) &\to \mathbb{R}^m \\ a &\mapsto X(a) = \left(f\left(t_1 - a\right), \dots, f\left(t_m - a\right)\right) \end{aligned}$$

and set $\mathcal{C} = \{X(a) \in \mathbb{R}^m,\ a \in (b,c)\}$.

By assumption (2.7), $\mathcal{C}$ can be seen as a submanifold of $\mathbb{R}^m$ of dimension 1 with corresponding geodesic distance defined by (2.8).

Take $\mu = X(\alpha)$ with $\alpha \in (b, c)$, thus we can write

$$\begin{aligned}\widehat{\mu}_{\mathrm{I}}^{1} &= \operatorname*{arg\,min}_{X(\alpha)\in\mathcal{C}} \sum_{i=1}^{n} \delta\left(X\left(A_i\right), X(\alpha)\right)\\ &= \operatorname*{arg\,min}_{\mu\in\mathcal{C}} \sum_{i=1}^{n} D\left(A_i, \alpha\right) = \operatorname*{arg\,min}_{\mu\in\mathcal{C}} C(\alpha),\end{aligned}$$

where D is the distance on (b, c), given by

$$D\left(A_i, \alpha\right) = \left| \int_{A_i}^{\alpha} \|X'(a)\| \,\mathrm{d}a \right| .$$

In the following, let $\left(A_{(i)}\right)_i$ be the ordered parameters. That is $A_{(1)} < \cdots < A_{(n)}$. Then, for a given $\alpha \in (b, c)$ such that $A_{(j)} < \alpha < A_{(j+1)}$, we get that

$$\begin{aligned}C(\alpha) = jD\left(\alpha, A_{(j)}\right) &+ \sum_{i=1}^{j-1} iD\left(A_{(i)}, A_{(i+1)}\right)\\ &+ (n-j)D\left(\alpha, A_{(j+1)}\right) + \sum_{i=j+1}^{n-1} (n-i)D\left(A_{(i)}, A_{(i+1)}\right) .\end{aligned}$$

For the sake of simplicity, let $n = 2q + 1$. It follows that $\widehat{\mathrm{Med}}(A) = A_{(q+1)}$. Moreover, let $\alpha = A_{(j)}$ with $j < q+1$. By symmetry, the case $j > q+1$ holds. Then, we rewrite $C\left(\alpha\right)$ as

$$C\left(\alpha\right) = \sum_{i=1}^{j-1} iD\left(A_{(i)}, A_{(i+1)}\right) + \sum_{i=j}^{n-1} (n-i)D\left(A_{(i)}, A_{(i+1)}\right)$$

and, by introducing $A_{(q+1)}$, we get that

$$\begin{aligned}C(\alpha) = &\sum_{i=1}^{j-1} iD\left(A_{(i)}, A_{(i+1)}\right) + \sum_{i=j}^{q} iD\left(A_{(i)}, A_{(i+1)}\right)\\ &+ \sum_{i=j}^{q} (n-2i)D\left(A_{(i)}, A_{(i+1)}\right) + \sum_{i=q+1}^{n-1} (n-i)D\left(A_{(i)}, A_{(i+1)}\right) .\end{aligned}$$

Finally, we notice that

$$C(\alpha) = C\left(A_{(q+1)}\right) + \sum_{i=j}^{q} (n-2i)D\left(A_{(i)}, A_{(i+1)}\right) > C\left(A_{(q+1)}\right) .$$

And the result follows since

$$\widehat{\mu}_{\mathrm{I}}^{1} = \operatorname*{arg\,min}_{\mu\in\mathcal{C}} C(\alpha) = X\left(A_{(q+1)}\right) = X\left(\widehat{\mathrm{Med}}(A)\right) = \widehat{f}_{\mathrm{SM}}.$$

■

Appendix B

```
library(igraph)

# Function to calculate the Minimum Spanning Tree of a data set X
mst <- function(X) {
    D <- as.matrix(dist(X))
    G <- graph.adjacency(D, "undirected", weighted = TRUE)
    T <- minimum.spanning.tree(G)
    A <- get.adjacency(T)
    return(A)
}

# Function to obtain an extended graph
manif <- function(X) {
    D <- as.matrix(dist(X))
    A <- mst(X)
    DA <- D*A
    R <- apply(DA, 1, max)
    n <- dim(X)[1]
    d <- dim(X)[2]
    A <- matrix(0, nrow=n, ncol=n)
    L <- 100
    f <- function(l,k) {
        (1-l)*X[i,k]+l*X[j,k]
    }
    for (i in 1:(n-1)) {
        for (j in (i+1):n) {
            P <- outer(1:L/(L+1), 1:d, f)
            E <- as.matrix(dist(rbind(P, X)))
            F <- E[(L+1):(L+n), 1:L]
            if (all(apply(F<R, 2, any))) A[i,j]=1
        }
    }
    A <- A + t(A)
    return(A)
}

# Function to calculate the estimated intrinsic mean
imean <- function(X, moment=2) {
    n <- dim(X)[1]
    D <- as.matrix(dist(X))
    A <- manif(X)
    DA <- D*A
    G <- graph.adjacency(DA, "undirected", weighted = TRUE)
    GD <- shortest.paths(G)
    GD <- GD^moment
    d <- apply(GD, 2, sum)
    im <- which.min(d)
    return(im)
}
```

Chapter 3

Statistical properties of the quantile normalization method for density curve alignment[*]

joint work with J-M. Loubes[†] and E. Maza[‡]

Abstract: The chapter investigates the large sample properties of the quantile normalization method by Bolstad et al. [12] which has become one of the most popular methods to align density curves in microarray data analysis. We prove consistency of this method which is viewed as a particular case of the structural expectation procedure for curve alignment, which corresponds to a notion of barycenter of measures in the Wasserstein space. Moreover, we show that, this method fails in some case of mixtures, and we propose a new methodology to cope with this issue.

Key Words: Curve registration; Manifold registration; Microarray data analysis; Normalization; Order statistics; Structural expectation; Wasserstein distance.

3.1 Introduction

We consider a density estimation problem in the particular situation where the data are samples of density curves, observed with some variations which are not directly correlated to the studied studied phenomenon. This situation occurs often in

[*] The results of this chapter have been recently published in *Mathematical Biosciences*, 242(2), 129–142, 2013.

[†] Institut de Mathématiques de Toulouse, Université Paul Sabatier - Toulouse III, Toulouse, France. E-mail: jean-michel.loubes@math.univ-toulouse.fr

[‡] École Nationale Supérieure Agronomique de Toulouse, and Genomic & Biotechnology of the Fruit Laboratory. UMR 990 INRA/INP-ENSAT. E-mail: elie.maza@ensat.fr

biology, for example when considering gene expression data obtained from microarray technologies, which is used to measure genome wide expression levels of genes in a given organism. A microarray may contain thousands of spots, each one containing a few million copies of identical DNA molecules that uniquely correspond to a gene. From each spot, a measure is obtained and then one of the most popular applications is to compare gene expression levels on different conditions, which leads to millions of measures of gene expression levels on technical and biological samples. However, before performing any statistical analysis on such data, it is necessary to process raw data in order to remove any systematic bias inhering to the microarray technology: differential efficiency of the two fluorescent dyes, different amounts of starting mRNA material, background noise, hybridization reactions and conditions. A natural way to handle this phenomena is to try remove these variations in order to align the measured densities, which proves difficult since the densities are unknown. In bioinformatics and computational biology, a method to reduce this kind of variability is known as normalization.

Among the many normalization methods, the quantile normalization method proposed by Bolstad et al. [12] has received a large interest. The procedure consists in assuming that there is an underlying common distribution followed by the measures. Then, for $i = 1, \ldots, n$ samples of $j = 1, \ldots, m$ of i.i.d random variables X_{ij}, the *mean* distribution is achieved by projecting the j-th empirical vector of sample quantiles, $\hat{\boldsymbol{q}}_j = (\hat{q}_{1,j}, \ldots, \hat{q}_{n,j})^\top$, onto the vector $\boldsymbol{d} = (1/\sqrt{n}, \ldots, 1/\sqrt{n})^\top$. This gives $\mathrm{proj}_{\boldsymbol{d}} \hat{\boldsymbol{q}}_j = (n^{-1} \sum_{i=1}^n \hat{q}_{i,j}, \ldots, n^{-1} \sum_{i=1}^n \hat{q}_{i,j})^\top$, which is such that if all n data vectors X_i, $i = 1, \ldots, n$, share the same distribution, then the plot of the quantiles gives a straight line along the line $\boldsymbol{d}$. We refer to Bolstad et al. [12] and Irizarry et al. [53] for some applications of this method. An example of this method is given in Figure 3.1, where the densities of a sample of 18 two-color microarrays are plotted after normalization of the expression log-ratios within two-color arrays. The dot-dashed and solid lines through densities corresponds to cross-sectional mean and quantile normalization of the log intensities across the arrays, respectively. The quantile normalization method has the advantages to be simple and quick with respect to others normalization procedures and yet providing very good estimation results. However its statistical properties have not been derived yet up to our knowledge.

Actually, normalization of density samples may be seen as the empirical version of a warping problem between distribution functions. This issue has received a growing attention in the last decade where many authors tackle the problem of recovering an unknown curve observed with both amplitude (variation in the y-axis) or phase (variation in the $x-$axis) variations, which prevent any direct extraction of classical statistics such as the mean or the median. Indeed the classical cross-sectional mean does not provide a consistent estimate of the function of interest when the phase variations are ignored since it fails to capture the characteristics of the sample of curves as quoted in Ramsay and Li [77]. Therefore curve registration

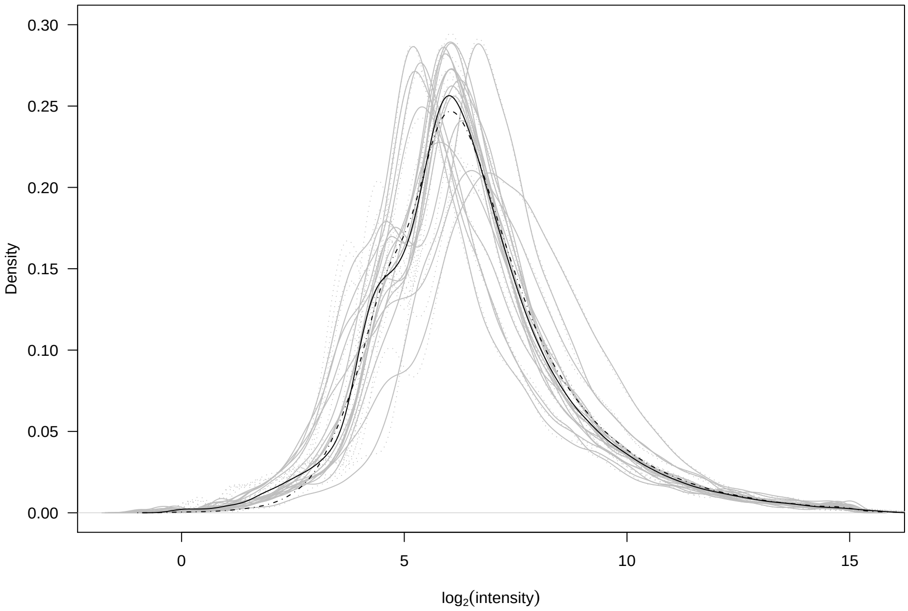

Figure 3.1: Densities for individual-channel intensities for two-color microarray data after normalization within arrays. Dotted and solid gray lines correspond to the “green” and “red” color arrays, respectively.

(also called curve alignment, structural averaging, and time warping) methods have been proposed in the statistical literature, among them we refer, for example, to Kneip and Gasser [58], Silverman [88], Gasser and Kneip [44], Wang and Gasser [106, 107], Ramsay and Li [77], Liu and Müller [67], Gamboa et al. [42], James [55], Kneip and Ramsay [59], and Dupuy et al. [35] and references therein.

Hence, in this chapter we point out that the quantile normalization can be seen as a particular case of the structural mean procedure, described in Dupuy et al. [35], which corresponds to a notion of barycenter of measures in the Wasserstein space as described in Boissard et al. [11]. We study the large sample properties of the quantile normalization method. In addition, when this procedure fails, using the analogy with warping issues, we propose a variation of this method to still recover a mean density and thus improving one pointed drawback of the quantile normalization method.

The outline of the chapter is as follows. In Section 3.2 we describe a nonparametric warping functional model, which is related with the quantile normalization method. In Section 3.3 we present the quantile estimation method and derive the asymptotic properties of the quantile normalization method. Section 3.4 is devoted to present the connection between normalization and distribution function alignment, which enables to improve quantile normalization method. Simulations are shown in Section 3.5. Finally, in Section 3.6 we apply the methods to normalize two-channel spotted microarray densities and evaluate its utility to identify differentially expressed genes. All proofs are gathered in Appendix A.

3.2 Statistical model for density warping

Let $X_{i,j},\ i = 1,\dots,n,\ j = 1,\dots,m_i$ be a sample of n independent real valued random variables of size m_i with density function $f_i\colon \mathbb{R} \to [0,+\infty)$ and distribution function $F_i\colon \mathbb{R} \to [0,1]$. We assume without loss of generality that $m_i = m$ for all units $i = 1,\dots,n$. The random variables are assumed to model the same phenomena with a variation effect modeled as follows.

Each distribution function F_i is obtained by warping a common distribution function $F\colon \mathbb{R} \to [0,1]$ by an invertible and differentiable warping function H_i, of the following manner:

$$F_i(t) = \mathbb{P}(X_{i,j} \le t) = F \circ H_i^{-1}(t), \qquad i = 1,\dots,n,\ j = 1,\dots,m, \tag{3.1}$$

where H_i is random, in the sense that $(H_1,\dots,H_n)$ is an i.i.d random sample from a (non parametric) warping stochastic process $\mathcal{H}\colon \Omega \to \mathcal{C}(\mathbb{R})$ defined on an unknown probability space $(\Omega,\mathcal{A},\mathbb{P})$, while $\mathcal{C}(\mathbb{R})$ denotes the space of all continuous functions defined on $\mathbb{R}$. Define ϕ its mean and let ϑ be its variance which is assumed to be finite. This model is also considered in Gamboa et al. [42] and in Dupuy et al. [35].

Since the model (3.1) to estimate the function f is not identifiable (see Dupuy et al. [35]), we consider the *structural expectation (SE)* of the quantile function to overcome this problem as

$$q_{SE}(\alpha) := F_{SE}^{-1}(\alpha) = \phi \circ F^{-1}(\alpha), \qquad 0 \le \alpha \le 1. \tag{3.2}$$

Inverting equation (3.1) leads to

$$q_i(\alpha) = F_i^{-1}(\alpha) = H_i \circ F^{-1}(\alpha), \qquad 0 \le \alpha \le 1, \tag{3.3}$$

where $q_i(\alpha)$ is the population quantile function (the left continuous generalized inverse of F_i), $F_i^{-1}\colon [0,1] \to \mathbb{R}$, given by

$$q_i(\alpha) = F_i^{-1}(\alpha) = \inf\left\{x_{ij} \in \mathbb{R}\colon F_i(x_{ij}) \ge \alpha\right\}, \qquad 0 \le \alpha \le 1.$$

Hence the natural estimator of the structural expectation (3.2) is given by

$$\overline{q_n(\alpha)} = \frac{1}{n}\sum_{i=1}^{n} q_i(\alpha), \qquad 0 \le \alpha \le 1. \tag{3.4}$$

In order to get the asymptotic behavior of the estimator, the following assumptions on the warping process $\mathcal{H}$ and on the distribution function F are considered:

Assumption 1. There exists a constant $C_1 > 0$ such that for all $(\alpha, \beta) \in [0,1]^2$, we have

$$\mathbb{E}\left[\left|H(\alpha) - \mathbb{E}H(\alpha) - \big(H(\beta) - \mathbb{E}H(\beta)\big)\right|^2\right] \leq C_1 \left|\alpha - \beta\right|^2.$$

Assumption 2. There exists a constant $C_2 > 0$ such that, for all $(\alpha, \beta) \in [0,1]^2$, we have

$$\mathbb{E}\left[\left|F^{-1}(\alpha) - F^{-1}(\beta)\right|^2\right] \leq C_2 \left|\alpha - \beta\right|^2.$$

The following theorem deals with the asymptotic behavior of the estimator (3.4).

Theorem 3.1. *The estimator $\overline{q_n(\alpha)}$ is consistent is the sense that*

$$\left\|\overline{q_n(\alpha)} - \mathbb{E}\left(\overline{q_n(\alpha)}\right)\right\|_\infty = \left\|\overline{q_n(\alpha)} - q_{SE}(\alpha)\right\|_\infty \xrightarrow[n\to\infty]{\text{a.s.}} 0.$$

Moreover, under assumptions 1 and 2, the estimator is asymptotically Gaussian, for any $K \in \mathbb{N}$ and fixed $(\alpha_1, \dots, \alpha_K) \in [0,1]^K$,

$$\sqrt{n}\begin{bmatrix} \overline{q_n(\alpha_1)} - q_{SE}(\alpha_1) \\ \vdots \\ \overline{q_n(\alpha_K)} - q_{SE}(\alpha_K) \end{bmatrix} \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}_K\left(\mathbf{0}, \mathbf{\Sigma}\right),$$

where the (k, k')-element of the asymptotic variance-covariance matrix $\mathbf{\Sigma}$ is given by $\Sigma_{k,k'} = \vartheta\big(q(\alpha_k), q(\alpha_{k'})\big)$ for all $(\alpha_k, \alpha_{k'}) \in [0,1]^2$ with $\alpha_k < \alpha_{k'}$.

3.3 Quantile estimation and the quantile normalization method

The distribution function is not observed and only random samples $X_{i,1}, \dots, X_{i,m}$ from $F_i(x)$ for $i = 1, \dots, n$ are observed. The i-th empirical quantile function is a natural estimator of F_i^{-1} when there is not any information on the underlying distribution function F_i. Consider the order statistics $X_{i,1:m} \leq X_{i,2:m} \leq \cdots \leq X_{i,m:m}$, hence the estimation of the quantile functions, $q_i(\alpha)$, is obtained by

$$\begin{aligned} \hat{q}_{i,m}(\alpha) &= \mathbb{F}_{i,m}^{-1}(\alpha) = \inf\left\{x_{ij} \in \mathbb{R} \colon \mathbb{F}_{i,m}(x_{ij}) \geq \alpha\right\} \\ &= X_{i,j:m} \quad \text{for} \quad \frac{j-1}{m} < \alpha \leq \frac{j}{m}, \qquad j = 1, \dots, m, \end{aligned}$$

where $\mathbb{F}_{i,m}^{-1}$ is the ith empirical quantile function.

Finally, the estimator of the structural quantile is given by

$$\bar{\hat{q}}_j = \frac{1}{n}\sum_{i=1}^{n} \hat{q}_{i,j} = \frac{1}{n}\sum_{i=1}^{n} X_{i,j:m}, \qquad j = 1, \dots, m. \tag{3.5}$$

Note that, this procedure corresponds to the so-called quantile normalization method proposed by Bolstad et al. [12].

Based on sample quantiles we can obtain a "mean" distribution through the projection of the empirical quantile vector of the j-th sample quantiles, $\hat{\boldsymbol{q}}_j = (\hat{q}_{1,j}, \dots, \hat{q}_{n,j})^\top$, onto the vector $\boldsymbol{d} = (1/\sqrt{n}, \dots, 1/\sqrt{n})^\top$, given by $\mathrm{proj}_{\boldsymbol{d}}\hat{\boldsymbol{q}}_j = (n^{-1}\sum_{i=1}^n \hat{q}_{i,j}, \dots, n^{-1}\sum_{i=1}^n \hat{q}_{i,j})^\top$. The quantile normalization method can be understood as a quantile-quantile plot extended to n dimensions such that if all n data vectors share the same distribution, then the plot of the quantiles gives a straight line along the line $\boldsymbol{d}$.

The asymptotic behavior of the quantile normalization estimator (3.5) is established by the next theorem.

Theorem 3.2. *The quantile normalization estimator $\bar{\hat{q}}_j$ is strongly consistent, $\bar{\hat{q}}_j \xrightarrow{\text{a.s}} q_{SE}(\alpha_j)$ as soon as $n, m \to \infty$, $j = 1, \dots, m$. Also under the assumptions of compactly central data, $|X_{i,j:m} - \mathbb{E}(X_{i,j:m})| \le L < \infty$ for all i and j, and $\sqrt{n}/m \to 0$, it is asymptotically Gaussian. Actually, for any $K \in \mathbb{N}$ and fixed $(\alpha_1, \dots, \alpha_K) \in [0,1]^K$,*

$$\sqrt{n}\begin{bmatrix} \bar{\hat{q}}_{j_1} - q_{SE}(\alpha_1) \\ \vdots \\ \bar{\hat{q}}_{j_K} - q_{SE}(\alpha_K) \end{bmatrix} \xrightarrow[n,m\to\infty]{\mathcal{D}} \mathcal{N}_K(\mathbf{0}, \boldsymbol{\Sigma}),$$

where the (k, k')-element of the asymptotic variance-covariance matrix $\boldsymbol{\Sigma}$ is given by $\Sigma_{k,k'} = \vartheta\big(q(\alpha_k), q(\alpha_{k'})\big)$ for all $(\alpha_k, \alpha_{k'}) \in [0,1]^2$ with $\alpha_k < \alpha_{k'}$.

This theorem relies on the asymptotic behavior of the quantile estimator, $\hat{q}_{i,m}(\alpha)$, given by the following proposition.

Proposition 3.1. *Assume F_i is continuously differentiable at the αth population quantile $q_i(\alpha)$ which is the unique solution of $F_i(q_i(\alpha)-) \le \alpha \le F_i(q_i(\alpha))$, and $f_i\big(q_i(\alpha)\big) > 0$ for a fixed $0 < \alpha < 1$. Also assume $m^{-1/2}(j/m - \alpha) = o(1)$. Then, for $i = 1, \dots, n$, the estimator $\hat{q}_{i,m}(\alpha)$ is strongly consistent, $\hat{q}_{i,m}(\alpha) \xrightarrow{\text{a.s.}} q_i(\alpha)$ as $m \to \infty$; and asymptotically Gaussian*

$$\sqrt{m}\big(X_{i,j:m} - H_i \circ q(\alpha)\big) \xrightarrow[m\to\infty]{\mathcal{D}} \mathcal{N}\left(0, \frac{\alpha(1-\alpha)}{\Big(f \circ H_i^{-1}\big(H_i \circ q(\alpha)\big) \cdot \big(H_i^{-1}\big)'\big(H_i \circ q(\alpha)\big)\Big)^2}\right),$$

where $\big(H_i^{-1}\big)'(z) = \mathrm{d}H_i^{-1}(z)/\mathrm{d}z = \big\{H_i' \circ H_i^{-1}(z)\big\}^{-1}$.

3.4 Density alignment as a registration problem

As we have seen in the previous sections, quantile normalization amounts to finding a mean distribution that fits the data density. Indeed, if the distribution function were known, hence, given respectively F_i's the distribution functions and μ_i's the distributions of the i.i.d sample $X_{i,j}$, $i = 1, \ldots, n, j = 1, \ldots, m$, the problem consists in finding a distribution function F and a probability μ which plays the role of a *mean* function but close enough to the data. This corresponds to the usual registration problem of the F_i's function restricted to the set of distribution functions.

One of the major issue in registration problem is to find the fitting criterion which enables to give a sense to the notion of mean of a sample of points. A natural criterion is in this framework given by the Wasserstein distance and this problem can be rewritten as finding a measure μ which minimizes

$$\mu \mapsto \frac{1}{n} \sum_{i=1}^{n} W_2^2(\mu, \mu_i), \tag{3.6}$$

where W_2^2 stands for the 2-Wasserstein distance

$$W_2^2(\mu, \mu_i) = \int \left| F_i^{-1}(\alpha) - F^{-1}(\alpha) \right|^2 d\alpha.$$

The existence and the uniqueness of such a minimizer is a difficult task in a general framework, which has been proved very recently under some technical conditions on the μ_i's in Agueh and Carlier [1]. However, for one-dimensional distributions, an explicit solution can be given, which corresponds to the *structural expectation* defined in Dupuy et al. [35]. Here, the F_i's and the μ_i's are not observed and only their empirical version are available. The estimation counterpart is considered in Section 3.3.

As pointed here, Wasserstein distance appears as a natural way to model distance between distribution functions which are warped one from another. Nevertheless, other criterion than (3.6) can be investigated. Indeed, for any distance d on the inverse of distribution functions, we can define a criterion to be minimized

$$F \mapsto \frac{1}{n} \sum_{i=1}^{n} d\big(F^{-1}, F_i^{-1}\big).$$

Each choice of d implies different properties for the minimizers. Recall that the choice of the L^2 loss corresponds to the Wasserstein distance between the distributions. Another choice, when dealing with warping problems, is to consider that the functional data belong to a non euclidean set, and to look for the most suitable corresponding distance. Hence, a natural framework is given by considering that the functions belong to a manifold using a manifold embedding and, in this context, the geodesic distance provides a natural way to compare two objects. This point of view has been developed in Dimeglio et al. [31] where $\hat{\delta}$, an approximation of the geodesic distance δ, is provided using an Isomap-type graph approximation, following Tenenbaum et al. [96]. This gives rise to the criterion

$$F \mapsto \frac{1}{n}\sum_{i=1}^{n} \hat{\delta}\big(F^{-1}, F_i^{-1}\big).$$

Only the approximation of the distribution function remains.

A theoretical study of this framework is difficult, mainly due to the problems of both choosing a good manifold embedding and then approximating the geodesic distance.

Many authors have considered this issue but results on the consistency of minimizers of such criterion are very scarce. Hence, we provide here a feasible algorithm to compute it and compare the performances of the corresponding estimator. For this, recall that we observe $X_{i,j}$, $i = 1, \dots, n$, $j = 1, \dots, m$ random variables. In order to mimic the geodesic distance between the inverse of the distribution functions, we will directly estimate $F_i^{-1}(\alpha)$, for $(j-1)/m < \alpha \leq j/m$ by the corresponding order statistics $X_{i,j:m}$. Hence, we sort the observations for each sample i, and denote by $X_{(i).}$ the sorted vector $X_{i,1:m}, \dots, X_{i,m:m}$ and thus we obtain an array of sorted observations $(X_{(1).}, \dots, X_{(n).})$. We then consider $\hat{\delta}$ an approximation of the geodesic distance between the vectors $X_{(i).}$ and define the corresponding geodesic mean as the minimizer over all the observation vectors $x \in \{X_{(i).},\ i = 1, \dots, n\}$ of the criterion $x \mapsto n^{-1}\sum_{i=1}^{n} \hat{\delta}\big(x, X_{(i).}\big)$.

Even if the theoretical properties of this estimate are hard to understand due to the difficulties inherent to the graph-type geodesic approximation, its practical properties for density normalization will be studied in the next section.

3.5 Simulation study

In this section, we illustrate by mean of simulated data the cases in which the quantile normalization method by Bolstad et al. [12] works and the situation in which it has problems to represent properly the behavior of the sample of density curves.

We simulated a sample of n mixture density functions as linear combinations of three Gaussian probability density functions $\phi_{il}(x; \mu_{il}, \sigma_{il})$, $l = 1, 2, 3$,

$$f_i(x) = \sum_{l=1}^{3} \omega_{il}\phi_{il}(x; \mu_{il}, \sigma_{il}), \qquad i = 1, \dots, n,$$

where $\omega_{il} \in [0, 1]$ are the probability weights with $\sum_{l=1}^{3} \omega_{il} = 1$, $i = 1, \ldots, n$.

The simulated sample of mixture density functions were generated following the next procedure:

1. For each $i = 1, \ldots, n$ three samples of size m of random observations are drawn from a Gaussian distribution.
2. A sampling (with replacement) of size m is carried out on the three samples based on the probability weights for obtaining the elements for each i.
3. Finally, for each i a kernel density estimate is obtained.

The values assumed to the location parameters were $\mu_{i1} = 1$, $\mu_{i2} = 4$ and $\mu_{i3} = 7$; to the scale parameters $\sigma_{i1} = 0.7$, $\sigma_{i2} = 0.8$, and $\sigma_{i3} = 0.9$; and to the probability weights $\omega_{i1} = 0.4$, $\omega_{i2} = 0.3$, and $\omega_{i3} = 0.3$. The number of simulated curves and observations assumed were $n = 50$ and $m = 1000$ respectively. The variability for the sample of curves was generated according to the next cases:

Case 1 (location variations): $U(\mu_{il} - 0.15, \mu_{il} + 0.15)$, $l = 1, 2, 3$.

Case 2 (scale variations): $U(\sigma_{il} - 0.35, \sigma_{il} + 0.35)$, $l = 1, 2$, and $U(\sigma_{i3} - 0.5, \sigma_{i3} + 0.5)$.

Case 3 (location and scale variations): Cases 1 and 2 together.

Case 4 (probability weight variations): $U(\omega_{il} - 0.1, \omega_{il} + 0.1)$, $l = 1, 2$,

where U is a uniformly distributed random variable.

Figure 3.2 shows the simulated density and distribution functions for each case. The estimated "mean" density and distribution functions using the quantile and manifold normalization methods corresponds to the solid and dash lines, respectively. The R code for the manifold normalization is given in the Appendix B. From the graphs, we can see that the quantile normalization estimate represents the variability among the density curves for the cases 1, 2 and 3, i.e when the probability weights do not vary among the densities, $\omega_{il} = \omega_{i'l}$, $l = 1, 2, 3$ for $i, i' = 1, \ldots, n$. In the case 4, on the contrary, there are large differences between quantile and manifold normalization methods, where the based quantile method does not capture the structural characteristics across the set of densities.

To overcome the drawback corresponding to case 4, we propose to apply the manifold embedding approach to estimate the structural mean pattern f based on an approximation of the induced geodesic distance on an unknown connected and geodesically complete Riemannian manifold $\mathcal{M} \subset \mathbb{R}^m$ by Dimeglio et al. [31]. As we can see in the Figure 3.2, the estimation of the "mean" density f through the manifold normalization improves the normalization of the sample of densities for the case of variations in probability weights (case 4) capturing properly the structural mean behavior of sample of curves.

3.6 Application to identification of differentially expressed genes

In this section, we apply the quantile and manifold methods to normalize two-channel (two-color) spotted microarrays in order to remove, from the expression measures, the systematic variations which arise from the microarray technology rather than from the differences between the probes, retaining the biological signals. For a description on two-channel spotted microarrays see Yang and Thorne [114] and Yang and Paquet [112]. We also evaluate the new manifold normalization method with respect to its ability to identify differentially expressed genes. For this we use two data sets of Tomato and Swirl experiments.

3.6.1 Tomato data set

The two-channel spotted microarray expression data comes from an experiment carried out by Wang et al. [105] in the Génomique et Biotechnologie des Fruits (GBF) laboratory at the Institut National Polytechnique-Ecole Nationale Superieure Agronomique de Toulouse (INP-ENSAT), which studies the underlying molecular mechanisms of the process of fruit set (i.e. the transition from flower-to-fruit) of tomato plants (*Solanum lycopersicum*). The data are provided by the experiment E-MEXP-1617 downloaded from the ArrayExpress database of functional genomic experiments at the European Bioinformatics Institute (EBI).

The data set contains 11860 spots (probes) and 18 arrays. The Bioconductor `limma` package (`http://www.bioconductor.org/`) based on the `R` programming language was used to read and carry out the quality assessment of the intensity data (Smyth and Speed [93] and Smyth [92]). Figure 3.3 shows the density plots for individual-channel intensities of two-color microarrays. Dotted and solid lines correspond to densities of “green” and “red” color intensities for each array, respectively.

We normalize the two-channel microarray data applying the single-channel normalization method by Yang and Thorne [114], which removes the systematic intensity bias from the red and green channels separately, both within and between arrays. The method proceeds in two stages: a within-array normalization followed by a between-array (between all channels from multiple arrays) normalization. The first stage normalizes the expression log-ratios (M-values, $M = \log_2(R/G)$, where

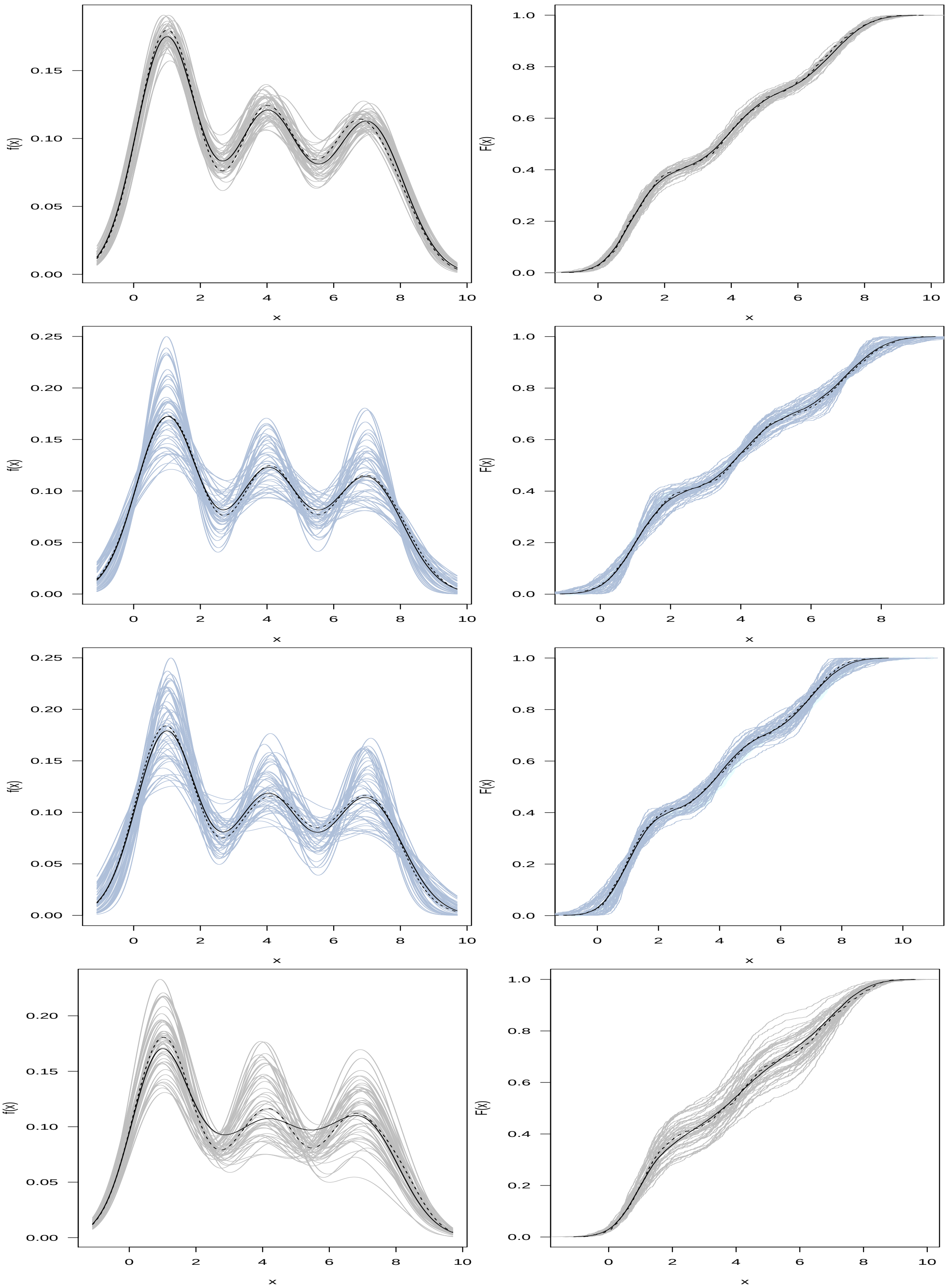

Figure 3.2: Simulated density (left side) and distribution (right side) functions. Quantile (bold solid) and manifold (dash) normalizations. Cases 1, 2, 3 and 4 from the top to bottom

R and G are the red and green intensities, respectively) from two-color arrays such that these average to zero within each array separately. The advantage of using the log-ratios for measuring relative gene expression within two samples on the same slide rather than log-intensity values is due to these are considered to be more stable than the absolute intensities across slides (Yang and Thorne [114]). The second stage normalizes the log intensities across arrays ensuring that these have the same empirical distribution across arrays and across channels. Procedures for within-array and between-array normalizations are implemented in the `normalizeWithinArrays` and `normalizeBetweenArrays` functions from the `limma` package, respectively.

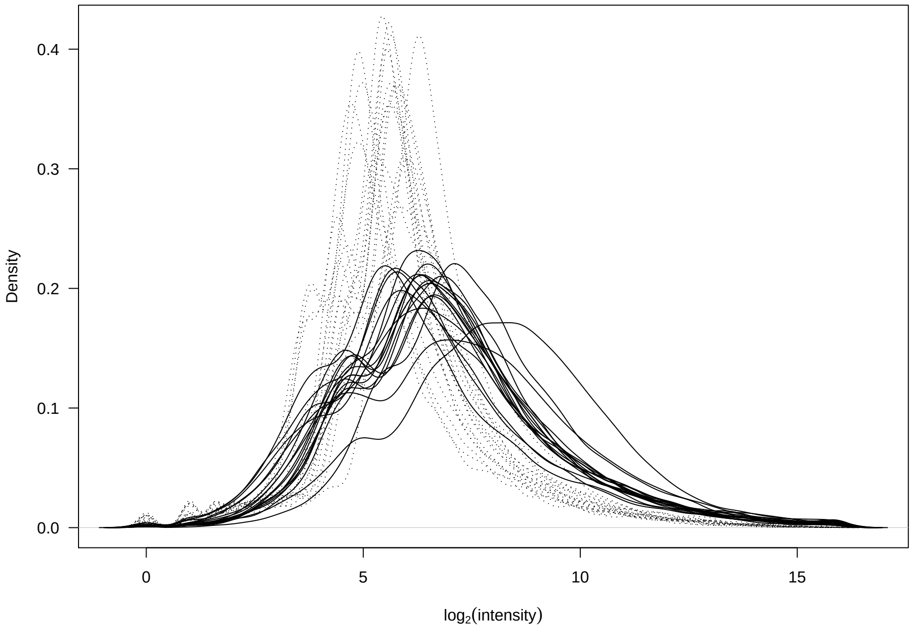

Figure 3.3: Densities for individual-channel intensities for two-color microarray data. Dotted and solid lines correspond to the "green" and "red" color arrays, respectively.

The log-ratios within arrays were normalized using the loess method (see Smyth and Speed [93] and Yang and Paquet [112]). Figure 3.4 plots the densities for each array after loess normalization. The normalization between arrays applying the quantile and manifold normalization are plotted in the same figure in solid and dashed lines. As we can see, the manifold normalization captures better the structural characteristics of the densities, in particular those that corresponding to the inflection points present in the individual arrays.

Now we evaluate the usefulness of the manifold normalization to identify differentially expressed genes. One of the aims of the tomato experiment in Wang et al. [105] is to identify gene expression in the (MicroTom) tomato lines downregulated in the expression of the Indole Acetic Acid 9 gene (AS-IAA9) and the wild type at three developmental stages during fruit set: flower bud, anthesis (i.e. the period during which the flower is fully open and functional), and post-anthesis.

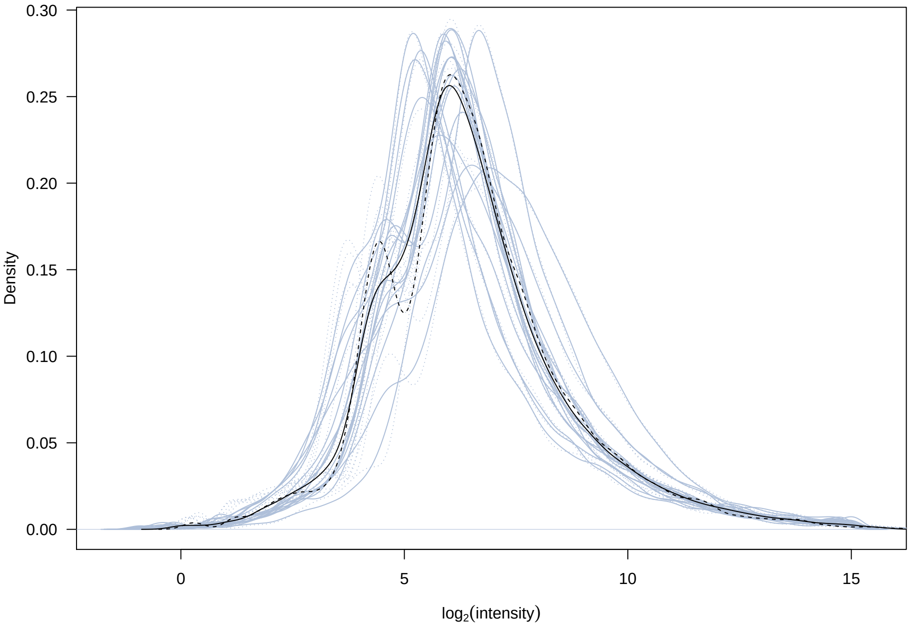

Figure 3.4: Densities for individual-channel intensities for two-color microarray data after loess normalization within arrays. Solid and dashed lines correspond the normalization between arrays applying the quantile and manifold normalization, respectively.

Thus, there are three experiments of identification of genes taking each tomato fruit stage separately. Hence, the experimental designs were based on six arrays for each corresponding stage, in two dye-swap pairs.

The statistical tool used for the identification of differentially expression genes in designed microarray experiments was the procedure based on the fit of gene-wise linear models and the application of empirical Bayes methods developed by Smyth [91]. The method relies on the "moderated" t-statistic across genes, a classical t-statistic improved by moderation of the standard errors, i.e., posterior estimators that shrunk the standard errors towards a common prior value using a Bayesian model (see Smyth [91] and Smyth et al. [94] for details). The tables for each stage show the top 30 differentially expression genes identified using the expression intensities normalized with the quantile and manifold normalizations methods, respectively. In the subsequent tables (Table 3.2-3.6) are included, for each identified gene, the M-value, the moderate t-statistic, the adjusted p-value and the B-statistic (log-odds that the gene is differentially expressed). The ranking of genes with significant differential expression are reported in order of increasing B-values. To adjust the p-values for multiple testing the Benjamini-Hochberg's method was used to control the expected false discovery rate (FDR) (see Smyth et al. [94]). The number of differentially expressed genes detected, for each stage of the tomato fruit, by the use of normalized log-ratios through the two normalization methods are reported in the Table 3.1, according to an assumed threshold value of 0.05 for adjusted p-values. In the same table are also reported the number of common genes shared by both methods, and the number of genes identified with the quantile (manifold) normalization but not with the manifold (quantile) method.

Table 3.1: Number of differentially expressed genes identified for each stage of tomato fruit assuming an adjusted p-value less than 0.05

	Tomato stage		
	Bud	Anthesis	Post-anthesis
Quantile	93	1291	262
Manifold	68	1274	254
Quantile ∩ Manifold	68	1250	252
Quantile − Manifold	25	41	10
Manifold − Quantile	0	24	2

$A - B$ denotes the difference set between sets A and B.
$A \cap B$ denotes the intersection between sets A and B.

Bud stage

For the bud stage of tomato fruit, the number of differentially expressed genes identified employing the normalized expression log-ratios through quantile and manifold normalization methods were 93 and 68, respectively, with a common number of genes of 68. The top 30 of differentially expressed genes detected are shown in the Table 3.2. The ordering of genes of first 30 genes is more or less parallel between both normalization methods. Some common genes have a quite different position, e.g. genes 5812, 6848, 9173, 3786, 7180, 4646 and 7859. Mostly of these top genes are common, except the genes 11454 and 11019 in the quantile normalization, and genes 8254 and 12181 in the manifold method.

Important features on genes can be found by means of the scatterplot between the average of $\log_2$ fold changes against the average of log-intensity $A = \log_2 \sqrt{R \times G}$ for each probe over all arrays in the experiment (MA-plot). There are other plots over which the identification can be contrasted, e.g., scatterplots between the moderated t-statistics and the average of log-intensity A, between the B-statistics against average of $\log_2$ fold changes (volcano plot), and quantile-quantile plots of moderated t-statistics (Yang and Speed [113]). Although we choose the MA-plots to save space, the results were practically the same for these graphs. The MA-plot for the respective normalization method are in the Figure 3.5. The black symbols correspond to differentially expressed genes with adjusted p-value less than 0.05. From the MA-plots is clear that these symbols are well separated from the clouds such that the corresponding genes are likely to be differentially expressed (Yang and Speed [113]). The genes detected with the normalized expressions by the quantile normalization that are not identified with the manifold method are signed in red points for the manifold MA-plot (on the bottom) for comparison. The number of theses genes are reported in the Table 3.1.

Table 3.2: Top 30 differentially expressed genes identified in the bud stage

Quantile normalization					Manifold normalization				
Gene	M-val.	mod. t	adj. p-val.	B	Gene	M-val.	mod. t	adj. p-val.	B
3733	2.86	29.87	0.001	7.958	3733	2.854	18.77	0.002	6.488
9737	2.484	13.67	0.011	5.246	9737	2.371	13.39	0.013	5.056
12795	-0.897	-11.89	0.017	4.533	12795	-0.866	-11.72	0.023	4.382
10960	0.896	11.44	0.017	4.324	10905	0.815	10.68	0.027	3.883
10905	0.876	11.27	0.017	4.244	8334	0.989	10.57	0.027	3.823
5812	0.91	10.27	0.023	3.725	10960	0.86	10.34	0.027	3.701
8334	1.046	10.21	0.023	3.692	3786	0.769	9.184	0.034	3.028
6768	0.881	10.12	0.023	3.640	6768	0.828	9.166	0.034	3.017
8712	0.906	9.849	0.024	3.485	7338	0.662	8.999	0.034	2.909
6848	0.903	9.387	0.031	3.205	8712	0.878	8.933	0.034	2.867
9173	0.765	9.01	0.031	2.964	7859	-0.847	-8.846	0.034	2.809
7338	0.665	8.783	0.031	2.811	7180	1.057	8.748	0.034	2.744
2266	0.706	8.7	0.031	2.755	2266	0.678	8.742	0.034	2.740
12214	-0.814	-8.611	0.031	2.693	4646	0.622	8.547	0.034	2.607
2489	0.775	8.6	0.031	2.686	12214	-0.79	-8.542	0.034	2.604
3786	0.786	8.584	0.031	2.674	4603	0.649	8.517	0.034	2.587
4603	0.669	8.525	0.031	2.633	5812	0.886	8.401	0.034	2.505
3426	-0.752	-8.509	0.031	2.622	12787	1.317	8.387	0.034	2.495
7180	1.04	8.498	0.031	2.614	7948	0.757	8.102	0.034	2.289
4646	0.627	8.302	0.031	2.473	2489	0.75	8.079	0.034	2.273
12787	1.358	8.265	0.031	2.447	6826	0.616	8.077	0.034	2.270
4192	0.863	8.25	0.031	2.435	3426	-0.755	-8.068	0.034	2.264
7859	-0.842	-8.077	0.031	2.308	4192	0.817	8.051	0.034	2.252
7948	0.77	8.044	0.031	2.283	2432	0.572	8.048	0.034	2.249
6826	0.632	7.995	0.031	2.245	8254	0.882	7.972	0.034	2.192
2432	0.586	7.992	0.031	2.243	12181	-0.613	-7.91	0.034	2.146
11454	0.639	7.929	0.031	2.195	6848	0.842	7.847	0.034	2.098
6474	0.591	7.822	0.031	2.113	87	0.676	7.844	0.034	2.096
87	0.691	7.822	0.031	2.112	9173	0.731	7.831	0.034	2.085
11019	-0.63	-7.796	0.031	2.092	6474	0.562	7.743	0.034	2.018

Anthesis stage

The number of differentially expressed genes detected in the anthesis stage with the expression intensities normalized with the quantile and manifold methods were 1291 and 1274, respectively, with 1250 genes in common. As is illustrated in the Table 3.3, the first 30 genes identified with both normalization methods are almost the same, except the gene 9582 for the quantile normalization and the gene 4209 for the manifold normalization. As in the bud stage, the position of genes in this ranking is fairly parallel for both methods, where only the position of genes 7825, 1127, 132, 4312 and 7164 is slightly different.

Although there are not big differences in the MA-plots between both normalization methods shown in Figure 3.5, the identification with the normalized intensities with the manifold method is a little bit sparser with respect to the quantile method, identifying 17 genes less. In the plots, the genes detected with the quantile normalization that are not identified with the manifold method are signed in red points for the manifold MA-plot (on the bottom), and the genes identified with the manifold normalization but not detected by the quantile method are represented by red pluses for the quantile MA-plot (on the top).

Table 3.3: Top 30 differentially expressed genes identified in the anthesis stage

Quantile normalization					Manifold normalization				
Gene	M-val.	mod. t	adj. p-val.	B	Gene	M-val.	mod. t	adj. p-val.	B
9306	-2.1	-35.99	0	11.17	9306	-2.137	-34.99	0	10.92
10855	1.39	20.62	0.0004	8.521	10855	1.334	21.04	0.0004	8.541
4051	1.6	19.43	0.0004	8.169	7075	1.383	19.03	0.0004	7.962
7075	1.435	19.29	0.0004	8.127	4051	1.537	18.54	0.0004	7.806
7	-1.488	-18.52	0.0004	7.880	8180	-1.423	-17.53	0.0004	7.467
8180	-1.474	-17.63	0.0004	7.578	7	-1.424	-17.45	0.0004	7.440
7825	-2.703	-17.58	0.0004	7.561	6884	1.098	17.43	0.0004	7.430
6884	1.124	17.53	0.0004	7.542	7904	0.986	17.38	0.0004	7.415
12861	1.178	17.37	0.0004	7.483	12861	1.149	17.35	0.0004	7.403
8334	1.172	17.30	0.0004	7.461	8334	1.15	17.31	0.0004	7.390
7904	1.004	17.19	0.0004	7.418	10617	1.17	16.99	0.0004	7.275
10617	1.193	16.70	0.0004	7.238	7825	-2.604	-16.84	0.0004	7.217
9963	1.537	16.47	0.0004	7.146	9963	1.487	16.10	0.0006	6.936
1127	-1.801	-16.45	0.0004	7.139	4040	-2.253	-15.82	0.0006	6.822
4040	-2.265	-16.43	0.0004	7.132	12795	-2.852	-15.75	0.0006	6.794
12795	-2.985	-15.83	0.0005	6.891	7686	-0.952	-15.57	0.0006	6.718
7686	-0.972	-15.78	0.0005	6.871	6218	-0.964	-15.24	0.0006	6.582
6218	-0.992	-15.60	0.0005	6.795	4209	1.406	15.14	0.0006	6.541
9582	1.177	15.18	0.0005	6.617	7164	0.908	14.72	0.0006	6.357
12911	-1.804	-15.12	0.0005	6.592	11406	0.94	14.71	0.0006	6.353
9457	0.97	15.05	0.0005	6.563	1127	-1.7	-14.67	0.0006	6.335
132	-1.235	-15.05	0.0005	6.560	9457	0.936	14.65	0.0006	6.325
4312	1.156	14.95	0.0005	6.516	7118	0.826	14.63	0.0006	6.316
11406	0.952	14.74	0.0005	6.425	3117	-1.126	-14.63	0.0006	6.314
3117	-1.153	-14.71	0.0005	6.412	12911	-1.752	-14.60	0.0006	6.302
7164	0.948	14.66	0.0005	6.387	8241	-1.395	-14.43	0.0006	6.226
7118	0.841	14.60	0.0005	6.363	2339	-1.158	-14.37	0.0006	6.198
8241	-1.421	-14.43	0.0005	6.281	132	-1.193	-14.35	0.0006	6.186
9876	-1.735	-14.26	0.0006	6.204	9876	-1.694	-14.25	0.0006	6.142
2339	-1.219	-14.23	0.0006	6.192	4312	1.118	14.24	0.0006	6.138

Post-anthesis stage

For the post-anthesis stage, 262 and 254 differentially expressed genes were identified with the quantile and manifold normalization, respectively, where the number of genes shared by both methods was 252. The top 30 genes are reported in the Table 3.4. As in the two previous stages, the position of these genes in table is parallel, especially for the top 10 genes. After, the position changes a little, specially for genes 7038, 6785, 2282, 6234 and 6806. There exist only four no common genes in the first 30 identified genes (gene 11474 and 8456 for the quantile method and 10240 and 6497 for the manifold normalization). The MA-plots for both methods are shown in Figure 3.5. As in the bud and anthesis stages, mostly of detected genes are relatively far from the to zero line on the M-axis.

Table 3.4: Top 30 differentially expressed genes identified in the post-anthesis stage

Quantile normalization					Manifold normalization				
Gene	M-val.	mod. t	adj. p-val.	B	Gene	M-val.	mod. t	adj. p-val.	B
3161	-1.669	-21.43	0.0013	7.380	3161	-1.637	-21.26	0.0012	7.305
7953	-2.826	-19.88	0.0013	7.073	7953	-2.744	-20.20	0.0012	7.099
5626	-1.119	-15.30	0.0051	5.851	5626	-1.08	-15.75	0.0045	5.961
8339	-1.329	-14.75	0.0051	5.662	8339	-1.295	-14.25	0.0054	5.447
980	-1.439	-14.14	0.0054	5.441	980	-1.399	-14.22	0.0054	5.435
4789	-0.85	-13.44	0.0064	5.163	4789	-0.822	-13.02	0.0062	4.957
11217	-1.455	-12.74	0.0064	4.866	914	-0.806	-12.95	0.0062	4.927
914	-0.843	-12.53	0.0064	4.776	11217	-1.393	-12.57	0.0062	4.761
8590	-0.814	-12.44	0.0064	4.732	8590	-0.796	-12.44	0.0062	4.704
6927	-0.872	-12.34	0.0064	4.688	7038	-1.151	-12.42	0.0062	4.696
9637	-0.752	-12.27	0.0064	4.653	6927	-0.841	-12.08	0.0062	4.537
8912	-0.918	-11.93	0.0064	4.496	9637	-0.733	-12.06	0.0062	4.528
3211	-1.005	-11.83	0.0064	4.447	3211	-0.986	-11.96	0.0062	4.477
6253	-0.962	-11.81	0.0064	4.435	6253	-0.934	-11.73	0.0063	4.365
7038	-1.1	-11.63	0.0064	4.346	8912	-0.88	-11.60	0.0063	4.304
4040	-1.187	-11.55	0.0064	4.306	985	-0.705	-11.54	0.0063	4.271
5405	-0.712	-11.48	0.0064	4.269	4040	-1.149	-11.34	0.0063	4.199
985	-0.723	-11.41	0.0064	4.234	5405	-0.684	-11.34	0.0063	4.169
2649	-0.753	-11.25	0.0067	4.150	6785	1.731	14.57	0.0062	4.164
5672	0.813	11.01	0.0073	4.022	2649	-0.742	-11.27	0.0063	4.134
12787	1.85	10.71	0.0082	3.860	12787	1.796	11.23	0.0063	4.114
7342	-0.808	-10.45	0.0082	3.713	6234	-0.921	-10.72	0.0081	3.836
6785	1.816	12.96	0.0079	3.711	5672	0.764	10.59	0.0081	3.764
2282	-0.8	-10.41	0.0082	3.688	10240	0.572	10.41	0.0081	3.666
6992	0.607	10.36	0.0082	3.662	7342	-0.766	-10.41	0.0081	3.662
11474	-0.799	-10.29	0.0082	3.621	6806	-0.781	-10.35	0.0081	3.629
10032	-1.138	-10.27	0.0082	3.606	6992	0.58	10.31	0.0081	3.607
8456	-0.838	-10.23	0.0082	3.584	10032	-1.091	-10.31	0.0081	3.607
6234	-0.965	-10.12	0.0082	3.566	2282	-0.783	-10.31	0.0081	3.601
6806	-0.805	-10.18	0.0082	3.553	6497	-0.745	-10.19	0.0084	3.534

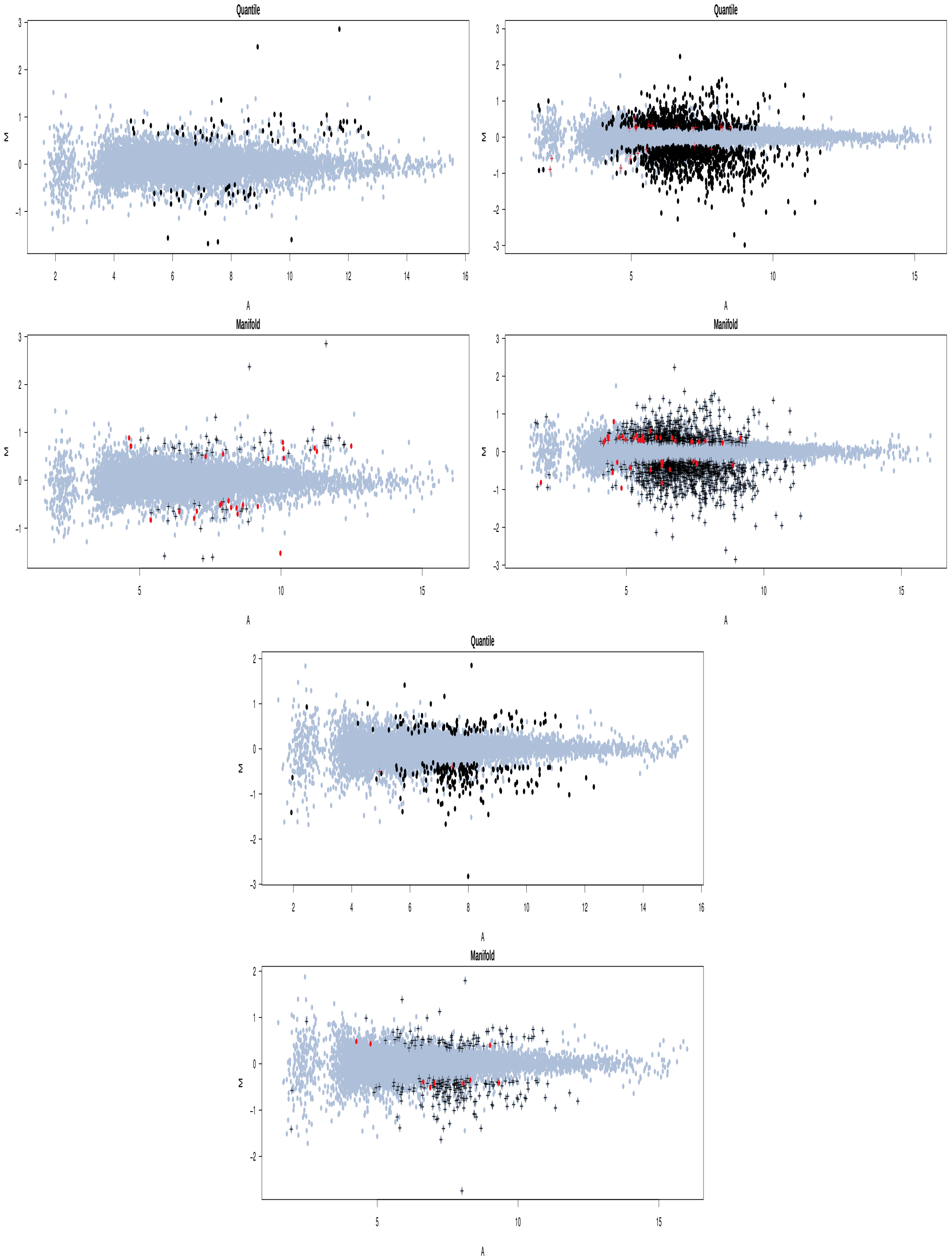

Figure 3.5: Graphical illustration of the differentially expressed genes identified for bud (top-left), anthesis (top-right) and post-anthesis (bottom) stages using the normalized expressions with the quantile and manifold methods. Black points and pluses correspond to the detected genes with a assuming an adjusted p-value less than 0.05. The red points (pluses) symbols correspond to the genes identified with the quantile (manifold) normalization but not with the manifold (quantile) method (see Table 3.1).

Validation by qRT-PCR data

Finally, in order to validate the microarray analysis of tomato data set in terms of the accuracy to detect differentially expressed genes during the fruit set using the normalized log ratios with the quantile and manifold normalization methods, the results of a quantitative Real-Time Polymerase Chain Reaction (qRT-PCR) analysis over 28 genes carried out by Wang et al. [105] were employed. In the Table 3.5, the qRT-PCR column, for each of developmental stage of fruit set, indicates whether the corresponding analyzed gene was validated (categorized as "yes") or not in the analysis by Wang et al. [105]. The columns of the quantile and manifold methods indicate whether the respective gene was detected as differentially expressed in each stage. Numbers within parenthesis indicate the position of the identified gene in the ranking of genes.

The comparison between the identification results with the normalization methods and the validation results by the qRT-PCR shows that the detection of genes using normalized expressions with both of normalization methods have a good accuracy. In general, for all stages of fruit set, there is a proportion of 75.5% of favorable cases (i.e. identified gene matching with validated gene or not identified gene matching with not validated gene). Additionally, the manifold method seems to have a higher significance, identifying first the validated gene with respect to the quantile method; 33 cases of validated genes are identified fist by the manifold normalization (82.5%).

3.6.2 Swirl zebrafish data set

The same exercise of identification of differentially expressed genes was carried out with the popular Swirl data set, which can be downloaded from http://bioinf.wehi.edu.au/limmaGUI/DataSets.html. This experiment was conducted using zebrafish (*Brachydanio rerio*) as a model organism to study early development in vertebrates. Swirl is a point mutation in the BMP2 gene that affects the dorsal-ventral body axis. Ventral fates such as blood are reduced, whereas dorsal structures such as somites and the notochord are expanded. One of the goals of the experiment is to identify genes with altered expression in the swirl mutant compared to wild-type zebrafish. See Dudoit and Yang [34], Yang and Speed [113] or Smyth [91] for detailed information about this experiment. A total of four arrays were performed in two dye-swap pairs with 8448 probes. Smyth [91] normalized the expression of log-ratios within-arrays using the print-tip loess normalization with a window span of 0.3 and three robustifying iterations. We follow his method, but instead of between arrays scale normalization of log intensities, here, of course, the quantile and manifold normalization are applied to compare both methods. The Figures 3.6 and 3.7 show the densities of unnormalized individual-channel intensities for two-color microarrays and its corresponding print-tip loess normalization within arrays, respectively. Solid

Table 3.5: Validation by qRT-PCR of Tomato experiment

Gene	Bud			Anthesis			Post-anthesis		
	Bolstad	Manifold	qRT-PCR	Bolstad	Manifold	qRT-PCR	Bolstad	Manifold	qRT-PCR
10318	no	no	yes	yes(88)	yes(96)	yes			
10617				yes(12)	yes(11)	yes			
10960	yes(4)	yes(6)	no						
12580	no	no	yes	yes(51)	yes(51)	yes			
12722				no	no	yes			
2266	yes(13)	yes(13)	yes	yes(103)	yes(77)	yes			
6848									
3733	yes(1)	yes(1)	yes	yes(145)	yes(137)	yes	yes(48)	yes(45)	yes
9737	yes(2)	yes(2)	yes						
9876				yes(29)	yes(29)	yes			
11243				yes(275)	yes(268)	yes			
12861				yes(9)	yes(9)	yes			
7468				yes(85)	yes(75)	yes			
7392				yes(303)	yes(299)	yes			
6299	no	no	yes	no	no	yes			
3161							yes(1)	yes(1)	yes
2	no	no	no	yes(147)	yes(171)	yes			
2020							yes(81)	yes(79)	yes
5626							yes(3)	yes(3)	no
2238							yes(158)	yes(155)	yes
12413	no	no	yes	yes(161)	yes(172)	yes			
10258				yes(44)	yes(40)	yes	yes(83)	yes(74)	yes
3043	yes(33)	yes(31)	yes	yes(83)	yes(80)	yes	yes(36)	yes(36)	yes
11189				yes(138)	yes(134)	yes			
2402				yes(109)	yes(103)	yes			
8241				yes(28)	yes(26)	no			
12911				yes(20)	yes(25)	yes	yes(35)	yes(39)	yes
10032				yes(115)	yes(110)	yes	yes(27)	yes(28)	yes

and dashed lines in Figure 3.7 correspond to densities after normalization between arrays applying the quantile and manifold normalization, respectively.

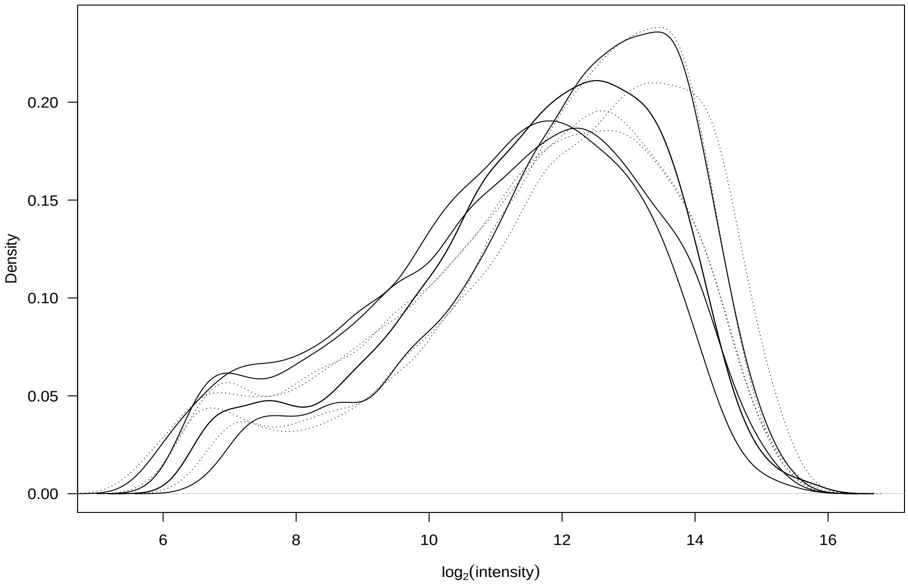

Figure 3.6: Densities for individual-channel intensities for two-color microarray data. Dotted and solid lines correspond to the “green” and “red” color arrays, respectively.

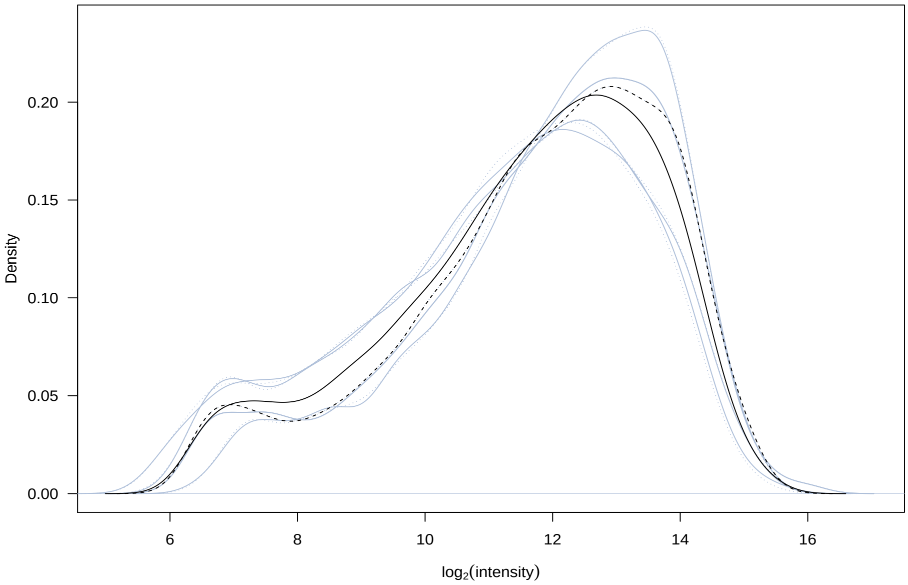

Figure 3.7: Densities for individual-channel intensities of of swirl data after print-tip loess normalization within arrays. Solid and dashed lines denotes the normalization between arrays with the quantile and manifold normalization, respectively.

The top 30 differentially expressed genes based on the quantile and manifold normalizations are reported in Table 3.6, respectively. With a threshold value of 0.05 for adjusted p-values, the number of differentially expressed genes identified with the quantile and manifold methods were 168 and 150, respectively. The 150 genes detected using the manifold method are also identified with the quantile normalization. The additional 18 genes detected with the quantile method are signed in the MA-plot for the manifold case in red points. As in the Tomato data, comparing with the quantile normalization, the detection of differential expressed genes based on the intensities normalized with the manifold method restricts a bit more the number of genes, being a more conservative (sparse) method. The MA-plots are shown in Figure 3.8.

Unfortunately, for the swirl zebrafish experiment there is not exist qRT-PCR data to validate the results of identification of differentially expressed genes found when the expression intensities normalized with the quantile and manifold normalizations methods are used.

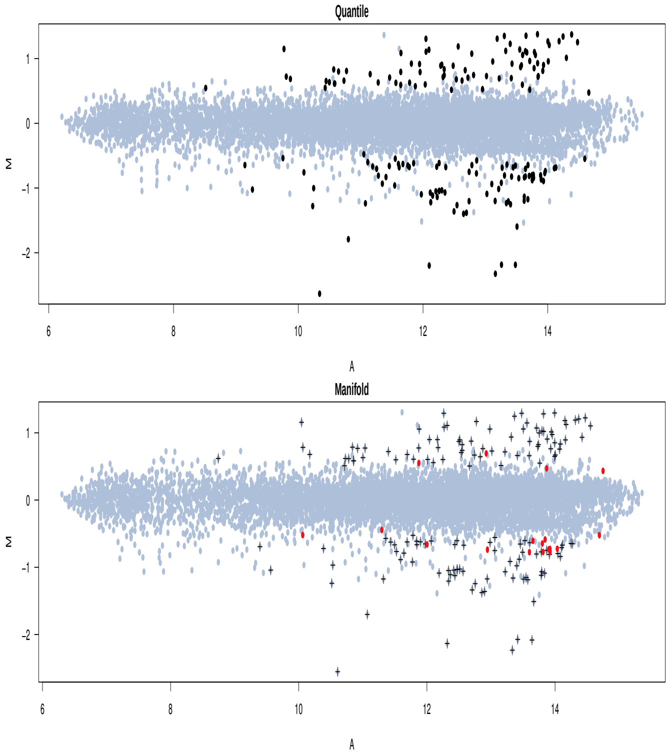

Figure 3.8: Graphical illustration of differentially expressed genes identified using the normalized expressions with the quantile and manifold methods. Red symbols correspond to genes identified with the quantile normalization but not with the manifold method.

Table 3.6: Top 30 differentially expressed genes identified from swirl zebrafish data

Quantile normalization					Manifold normalization				
Gene	M-val.	mod. t	adj. p-val.	B	Gene	M-val.	mod. t	adj. p-val.	B
2961	-2.633	-17.198	0.002	6.966	2961	-2.553	-16.684	0.0033	6.415
3723	-2.185	-16.415	0.002	6.713	3723	-2.072	-15.647	0.0033	6.079
1611	-2.186	-15.736	0.002	6.479	1611	-2.082	-15.471	0.0033	6.018
3721	-2.198	-14.329	0.0024	5.939	3721	-2.133	-14.282	0.0038	5.580
1609	-2.325	-13.71	0.0024	5.677	1609	-2.233	-14.008	0.0038	5.472
7602	1.21	13.121	0.0024	5.412	8295	1.291	13.442	0.0038	5.237
8295	1.306	13.07	0.0024	5.388	319	-1.243	-13.275	0.0038	5.165
319	-1.265	-13.05	0.0024	5.378	515	1.246	13.08	0.0038	5.080
515	1.308	12.835	0.0024	5.277	7602	1.124	12.738	0.0039	4.925
5075	1.373	12.795	0.0024	5.258	3790	1.168	12.598	0.0039	4.860
3790	1.187	12.356	0.0024	5.042	157	-1.701	-12.216	0.0039	4.678
157	-1.792	-12.301	0.0024	5.014	5931	-1.066	-12.105	0.0039	4.624
7307	1.228	12.253	0.0024	4.989	7307	1.147	11.987	0.0039	4.565
7036	1.376	12.018	0.0024	4.869	7491	1.282	11.74	0.0039	4.440
2276	1.253	11.978	0.0024	4.848	1697	1.057	11.726	0.0039	4.433
7491	1.353	11.907	0.0024	4.810	3726	-1.238	-11.698	0.0039	4.419
3726	-1.28	-11.873	0.0024	4.792	683	1.29	11.608	0.0039	4.372
5931	-1.091	-11.857	0.0024	4.784	7036	1.295	11.549	0.0039	4.341
683	1.35	11.657	0.0026	4.676	5084	-1.049	-11.118	0.0044	4.110
1697	1.119	11.534	0.0026	4.609	5075	1.223	11.11	0.0044	4.106
4380	1.265	11.415	0.0027	4.543	4188	-1.206	-11.044	0.0044	4.069
7542	1.141	11.287	0.0028	4.471	4380	1.181	10.965	0.0044	4.025
4032	1.341	10.884	0.0034	4.239	7542	1.072	10.89	0.0044	3.983
4188	-1.22	-10.827	0.0034	4.205	2276	1.104	10.861	0.0044	3.967
5084	-1.072	-10.731	0.0035	4.147	4032	1.206	10.808	0.0044	3.937
6903	-1.251	-10.585	0.0036	4.059	6903	-1.185	-10.436	0.0053	3.720
6023	1.012	10.238	0.0044	3.843	4017	-1.042	-10.202	0.0059	3.579
3695	1.057	10.167	0.0045	3.798	6023	0.933	10.1	0.0061	3.516
4546	1.269	10.012	0.0048	3.697	3695	0.998	10.031	0.0062	3.474
2679	-1.233	-9.75	0.0052	3.524	4546	1.189	9.744	0.0071	3.292

Appendix A

Proof of Theorem 3.1. In order to prove the almost sure convergence of the estimator $\overline{q_n}(\alpha)$ given in the equation (3.4) the following corollary by Ledoux and Talagrand [65] is required.

Corollary (Corollary 7.10, Ledoux and Talagrand [65])**.** Let W_i be a sequence of independent and identically distributed Borel random variables distributed like W with values in a separable Banach space $\mathcal{B}$, and the nth partial sum $S_n = \sum_{i=1}^{n} W_i$. Then the sequence $S_n/n \to 0$ strongly as $n \to \infty$ if and only if $\mathbb{E}\|W\| < \infty$ and $\mathbb{E}(W) = 0$.

Now first note, from the equation (3.3), that

$$\begin{aligned}\mathbb{E}\big(q_i(\alpha)\big) &= \mathbb{E}\big(F_i^{-1}(\alpha)\big)\\ &= \mathbb{E}\big(H_i \circ F^{-1}(\alpha)\big)\\ &= \mathbb{E}(H_i) \circ F^{-1}(\alpha)\\ &= \phi \circ F^{-1}(\alpha) = F_{SE}^{-1}(\alpha)\\ &= \phi \circ q(\alpha) = q_{SE}(\alpha),\end{aligned}$$

where $q(\alpha) = F^{-1}(\alpha) = \inf\{x \in \mathbb{R} \colon F(x) \geq \alpha\}\,,\ 0 \leq \alpha \leq 1.$

Thus we have

$$\begin{aligned}\overline{q_n(\alpha)} - \mathbb{E}\left(\overline{q_n(\alpha)}\right) &= \frac{1}{n}\sum_{i=1}^{n} H_i \circ F^{-1}(\alpha) - \phi \circ F^{-1}(\alpha)\\ &= \frac{1}{n}\sum_{i=1}^{n}(H_i - \phi) \circ F^{-1}(\alpha)\\ &= \frac{1}{n}\sum_{i=1}^{n}(H_i - \phi) \circ q(\alpha).\end{aligned}$$

Setting $S_n = \sum_{i=1}^{n} W_i$, where $W_i = (H_i - \phi) \circ q(\alpha)$ is a sequence of independent and identically distributed random variables in a separable Banach space $\mathcal{B} = \mathcal{C}([0,1])$, and applying the above Corollary, the almost sure convergence of $\overline{q_n(\alpha)}$ is guaranteed.

The asymptotic normality of $\overline{q_n(\alpha)}$ is now obtained applying the multivariate central limit theorem. For any $K \in \mathbb{N}$, and fixed $(\alpha_1, \ldots, \alpha_K) \in [0,1]^K$,

$$\sqrt{n}\begin{bmatrix}\overline{q_n(\alpha_1)} - \mathbb{E}\overline{q_n(\alpha_1)}\\ \vdots\\ \overline{q_n(\alpha_K)} - \mathbb{E}\overline{q_n(\alpha_K)}\end{bmatrix} = \sqrt{n}\begin{bmatrix}\frac{1}{n}\sum\limits_{i=1}^{n}(H_i - \phi) \circ q(\alpha_1)\\ \vdots\\ \frac{1}{n}\sum\limits_{i=1}^{n}(H_i - \phi) \circ q(\alpha_K)\end{bmatrix} \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}_K\left(\mathbf{0}, \boldsymbol{\Sigma}\right),$$

where the (k,k')-element of the asymptotic variance-covariance matrix $\boldsymbol{\Sigma}$ is given by $\Sigma_{k,k'} = \vartheta\big(q(\alpha_k), q(\alpha_{k'})\big)$ for all $(\alpha_k, \alpha_{k'}) \in [0,1]^2$ with $\alpha_k < \alpha_{k'}$, which is obtained as

$$\begin{aligned}\operatorname{Cov}\left(\overline{q_n(\alpha_k)}, \overline{q_n(\alpha_{k'})}\right) &= \operatorname{Cov}\left(\frac{1}{n}\sum_{i=1}^{n} q_i(\alpha_k), \frac{1}{n}\sum_{i=1}^{n} q_i(\alpha_{k'})\right)\\ &= \frac{1}{n^2}\sum_{i=1}^{n}\operatorname{Cov}\big(H_i \circ q(\alpha_k), H_i \circ q(\alpha_{k'})\big)\\ &= \frac{1}{n}\vartheta\big(q(\alpha_k), q(\alpha_{k'})\big),\end{aligned}$$

where $\vartheta\big(q(\alpha_k), q(\alpha_{k'})\big)$ is the autocovariance function of H_i, $i = 1, \ldots, n$.

Finally, following van der Vaart and Wellner [100], the tightness moment condition to weak convergence is given by

$$\begin{aligned}
&\mathbb{E}\left[\left|\sqrt{n}\left(\overline{q_n(\alpha)} - \mathbb{E}\overline{q_n(\alpha)}\right) - \sqrt{n}\left(\overline{q_n(\beta)} - \mathbb{E}\overline{q_n(\beta)}\right)\right|^2\right] \\
&= \mathbb{E}\left[\left|\sqrt{n}\left(\left(\overline{q_n(\alpha)} - \mathbb{E}\overline{q_n(\alpha)}\right) - \left(\overline{q_n(\beta)} - \mathbb{E}\overline{q_n(\beta)}\right)\right)\right|^2\right] \\
&= \mathbb{E}\left[n\left|\left(\frac{1}{n}\sum_{i=1}^{n} H_i \circ q(\alpha) - \phi \circ q(\alpha)\right) - \left(\frac{1}{n}\sum_{i=1}^{n} H_i \circ q(\beta) - \phi \circ q(\beta)\right)\right|^2\right] \\
&= \mathbb{E}\left[n\left|\frac{1}{n}\sum_{i=1}^{n}(H_i - \phi) \circ \big(q(\alpha) - q(\beta)\big)\right|^2\right] \\
&\leq C_1 C_2 \left|\alpha - \beta\right|^2,
\end{aligned}$$

if assumptions 1 and 2 are satisfied. ■

Proof of Proposition 3.1. The proof is a direct application of the following theorems of strong consistency and asymptotic normality for quantile estimators. See Serfling [87] or David and Nagaraja [28] for its proofs.

Theorem (Strong consistency of quantile estimator). If the αth population quantile, $q(\alpha)$, is the unique solution of $F(x-) \leq \alpha \leq F(x)$, then $\hat{q}_m(\alpha) \xrightarrow{\text{a.s.}} q(\alpha)$ as soon as $m \to \infty$.

Therefore $\hat{q}_{i,m}(\alpha) \xrightarrow{\text{a.s.}} q_i(\alpha)$ as $m \to \infty$ for $i = 1, \ldots, n$.

Theorem (Asymptotic normality of order statistics). For a fixed $0 < \alpha < 1$, assume F is continuously differentiable at the αth population quantile, $q(\alpha)$, $f\big(q(\alpha)\big) > 0$, and $m^{-1/2}(j/m - \alpha) = o(1)$. Then $\sqrt{m}\big(X_{j:m} - q(\alpha)\big) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, \alpha(1-\alpha)/f^2(q(\alpha))\big)$ as $m \to \infty$, where $X_{j:m} = X_{[\alpha m]+1}$ is the jth sample quantile, and $[\alpha m]$ denotes the greatest integer less or equal than αm.

In consequence we have for $i = 1, \ldots, n$

$$\sqrt{m}\big(X_{i,j:m} - q_i(\alpha)\big) \xrightarrow[m\to\infty]{\mathcal{D}} \mathcal{N}\left(0, \frac{\alpha(1-\alpha)}{f_i^2\big(q_i(\alpha)\big)}\right),$$

that conditioned to a fixed H_i implies

$$\sqrt{m}\big(X_{i,j:m} - H_i \circ q(\alpha)\big) \xrightarrow[m\to\infty]{\mathcal{D}} \mathcal{N}\left(0, \frac{\alpha(1-\alpha)}{\Big(f \circ H_i^{-1}\big(H_i \circ q(\alpha)\big) \cdot \big(H_i^{-1}\big)'\big(H_i \circ q(\alpha)\big)\Big)^2}\right),$$

where $\left(H_i^{-1}\right)'(z) = \mathrm{d}H_i^{-1}(z)/\mathrm{d}z = \left\{H_i' \circ H_i^{-1}(z)\right\}^{-1}$. ■

The moments of order statistics are hard to compute for many distributions so these can be approximated reasonably using a linear Taylor series expansion of the relation $X_{i,j:m} \overset{d}{=} F_i^{-1}(U_{i,j:m})$ around the point $\mathbb{E}(U_{i,j:m}) = \alpha_j = j/(m+1)$, where $U_{i,j:m}$ denotes the jth order statistic in a sample of size m from the uniform $(0,1)$ distribution. The approximated means, variances and covariances of order statistics for $i = 1, \dots, n$ are given by (see, for instance, David and Nagaraja [28] or Arnold et al. [2])

$$\begin{aligned}\mathbb{E}\Big(X_{i,j:m}\Big|H_i\Big) &= q_{i,j} + \frac{\alpha_j(1-\alpha_j)}{2(m+2)} q_{i,j}^{''} \\ &+ \frac{\alpha_j(1-\alpha_j)}{(m+2)^2}\left[\frac{1}{3}\big((1-\alpha_j)-\alpha_j\big)q_{i,j}^{'''} + \frac{1}{8}\alpha_j(1-\alpha_j)q_{i,j}^{(4)}\right] + O\left(\frac{1}{m^2}\right),\end{aligned} \tag{3.7}$$

$$\begin{aligned}\mathrm{Var}\Big(X_{i,j:m}\Big|H_i\Big) &= \frac{\alpha_j(1-\alpha_j)}{m+2} q_{i,j}^{'2} + \frac{\alpha_j(1-\alpha_j)}{(m+2)^2} \\ &\times \left[2\big((1-\alpha_j)-\alpha_j\big)q_{i,j}^{'}q_{i,j}^{''} + \alpha_j(1-\alpha_j)\left(q_{i,j}^{'}q_{i,j}^{'''} + \frac{1}{2}q_{i,j}^{''2}\right)\right] + O\left(\frac{1}{m^2}\right)\end{aligned} \tag{3.8}$$

and

$$\begin{aligned}\mathrm{Cov}\Big(X_{i,j:m}, X_{i,s:m}\Big|H_i\Big) &= \frac{\alpha_j(1-\alpha_s)}{m+2} q_{i,j}^{'}q_{i,s}^{'} + \frac{\alpha_j(1-\alpha_s)}{(m+2)^2} \\ &\times \Bigg[\big((1-\alpha_j)-\alpha_j\big)q_{i,j}^{''}q_{i,s}^{'} + \big((1-\alpha_s)-\alpha_s\big)q_{i,j}^{'}q_{i,s}^{''} + \frac{1}{2}\alpha_j(1-\alpha_j)q_{i,j}^{'''}q_{i,s}^{'} \\ &+ \frac{1}{2}\alpha_s(1-\alpha_s)q_{i,j}^{'}q_{i,s}^{'''} + \frac{1}{2}\alpha_j(1-\alpha_s)q_{i,j}^{''}q_{i,s}^{''}\Bigg] + O\left(\frac{1}{m^2}\right),\end{aligned} \tag{3.9}$$

where, since $\alpha_j = F_i(q_{i,j})$, we have

$$\begin{aligned} q_{i,j}^{'} &= \frac{\mathrm{d}q_{i,j}}{\mathrm{d}\alpha_j} = \frac{1}{f_i(q_{i,j})} < \infty, \\ q_{i,j}^{''} &= -\frac{f_i^{'}(q_{i,j})}{f_i^2(q_{i,j})} = -\frac{\mathrm{d}f_i(q_{i,j})}{\mathrm{d}q_{i,j}}\frac{1}{f_i^3(q_{i,j})} < \infty, \quad \text{and so on,}\end{aligned}$$

where $f_i(q_{i,j}) > C$ with $C > 0$ is the density-quantile function of X evaluated at $q_{i,j} = q_i(\alpha_j) = H_i \circ F^{-1}(\alpha_j)$ with $\alpha_j = j/(m+1)$, $j = 1, \dots, m$. $\left|f_i^{'}\right| < M$, $\left|f_i^{''}\right| < M$, and $\left|f_i^{'''}\right| < M$.

This approximation method is due to David and Johnson [27], where $(m+2)^{-3}$ order approximations are derived. The asymptotic means, variances, and covariances correspond to the first terms of (3.7), (3.8) and (3.9), respectively (David and Nagaraja [28]).

Using the approximation in equation (3.7), the mean of $\bar{\hat{q}}_j$ is calculated as

$$\begin{aligned}
\mathbb{E}\Big(\bar{\hat{q}}_j\Big) &= \mathbb{E}\Big[\mathbb{E}\Big(\bar{\hat{q}}_j\big|H_i\Big)\Big] \\
&= \mathbb{E}\left[\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_{i,j:n}\big|H_i\right)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\mathbb{E}\Big(X_{i,j:m}\big|H_i\Big)\Big] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[q_{i,j} + \frac{\alpha_j(1-\alpha_j)}{2(m+2)}q_{i,j}^{''} + O\left(\frac{1}{m^2}\right)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\mathbb{E}\left(q_{i,j}\right) + \frac{\alpha_j(1-\alpha_j)}{2(n+2)}\mathbb{E}\left(q_{i,j}^{''}\right) + O\left(\frac{1}{m^2}\right)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[q_{SE}(\alpha_j) + \frac{\alpha_j(1-\alpha_j)}{2(m+2)}\mathbb{E}\left(\frac{-\mathrm{d}f_i(q_{i,j})}{\mathrm{d}q_{i,j}}\frac{1}{f_i^3(q_{i,j})}\right) + O\left(\frac{1}{m^2}\right)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[q_{SE}(\alpha_j) + \frac{1}{8(m+2)}\left(\frac{-M}{C^3}\right) + O\left(\frac{1}{m^2}\right)\right] \\
&= q_{SE}(\alpha_j) + \frac{1}{8(m+2)}\left(\frac{-M}{C^3}\right) + O\left(\frac{1}{m^2}\right),
\end{aligned}$$

where $\left|\mathrm{d}f_i(q_{i,j})/\mathrm{d}q_{i,j}\right| < M$ and $f_i^3(q_{i,j}) > C$.

While through equation (3.9), the covariance between of $\bar{\hat{q}}_{j_k}$ and $\bar{\hat{q}}_{j_{k'}}$ for $k \neq k'$ $k = 1, \dots, K$ is given by

$$\begin{aligned}
&\mathrm{Cov}\Big(\bar{\hat{q}}_{j_k}, \bar{\hat{q}}_{j_{k'}}\Big) \\
&= \mathrm{Cov}\left(\frac{1}{n}\sum_{i=1}^{n} X_{i,j_k:m}, \frac{1}{n}\sum_{i=1}^{n} X_{i,j_{k'}:m}\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Cov}\Big(X_{i,j_k:m}, X_{i,j_{k'}:m}\Big) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\Big\{\mathbb{E}\left[\mathrm{Cov}\Big(X_{i,j_k:m}, X_{i,j_{k'}:m}\big|H_i\Big)\right] + \mathrm{Cov}\left[\mathbb{E}\Big(X_{i,j_k:m}\big|H_i\Big), \mathbb{E}\Big(X_{i,j_{k'}:m}\big|H_i\Big)\right]\Big\} \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\Big\{\mathbb{E}\left[\frac{\alpha_{j_k}(1-\alpha_{j_{k'}})}{m+2}q_{i,j_k}^{'}q_{i,j_{k'}}^{'} + O\left(\frac{1}{m^2}\right)\right] \\
&+ \mathrm{Cov}\left[q_{i,j_k} + \frac{\alpha_{j_k}(1-\alpha_{j_k})}{2(m+2)}q_{i,j_k}^{''} + O\left(\frac{1}{m^2}\right), q_{i,j_{k'}} + \frac{\alpha_{j_{k'}}(1-\alpha_{j_{k'}})}{2(m+2)}q_{i,j_{k'}}^{''} + O\left(\frac{1}{m^2}\right)\right]\Big\} \\
&\vdots
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n^2}\sum_{i=1}^{n}\Bigg\{\mathbb{E}\left[\frac{\alpha_{j_k}(1-\alpha_{j_{k'}})}{m+2}\frac{1}{f_i^2(q(\alpha_{j_k}))}\frac{1}{f_i^2(q(\alpha_{j_{k'}}))}+O\left(\frac{1}{m^2}\right)\right] \\
&\quad + \mathrm{Cov}\Bigg[H_i(q(\alpha_{j_k}))+\frac{\alpha_{j_k}(1-\alpha_{j_k})}{2(m+2)}\left(-\frac{\mathrm{d}f_i(q_{i,j_k})}{\mathrm{d}q_{i,j_k}}\frac{1}{f_i^3(q_{i,j_k})}\right)+O\left(\frac{1}{m^2}\right), \\
&\qquad\qquad H_i(q(\alpha_{j_{k'}}))+\frac{\alpha_{j_{k'}}(1-\alpha_{j_{k'}})}{2(m+2)}\left(-\frac{\mathrm{d}f_i(q_{i,j_{k'}})}{\mathrm{d}q_{i,j_{k'}}}\frac{1}{f_i^3(q_{i,j_{k'}})}\right)+O\left(\frac{1}{m^2}\right)\Bigg]\Bigg\} \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\Bigg\{\mathbb{E}\left[\frac{1}{4(m+2)}\frac{1}{C^2}\frac{1}{C^2}+O\left(\frac{1}{m^2}\right)\right] \\
&\quad + \mathrm{Cov}\Bigg[H_i(q(\alpha_{j_k}))+\frac{1}{8(m+2)}\left(\frac{-M}{C^3}\right)+O\left(\frac{1}{m^2}\right), \\
&\qquad\qquad H_i(q(\alpha_{j_{k'}}))+\frac{1}{8(m+2)}\left(\frac{-M}{C^3}\right)+O\left(\frac{1}{m^2}\right)\Bigg]\Bigg\} \\
&= \frac{1}{n}\left[\frac{1}{4(m+2)}\frac{1}{C^4}+O\left(\frac{1}{m^2}\right)\right]+\frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Cov}\left[H_i(q(\alpha_{j_k})),H_i(q(\alpha_{j_{k'}}))\right] \\
&= \frac{1}{n}\left[\frac{1}{4(m+2)}\frac{1}{C^4}+O\left(\frac{1}{m^2}\right)\right]+\frac{1}{n}\vartheta\big(q(\alpha_{j_k}),q(\alpha_{j_{k'}})\big),
\end{aligned}
$$

for all $(\alpha_k,\alpha_{k'})\in[0,1]^2$ with $\alpha_k<\alpha_{k'}$.

From above equations we have that

$$\mathbb{E}\big(\bar{\hat{q}}_j\big)\xrightarrow[m\to\infty]{} q_{SE}(\alpha_j)$$

and

$$\mathrm{Cov}\Big(\bar{\hat{q}}_{j_k},\bar{\hat{q}}_{j_{k'}}\Big)\xrightarrow[m\to\infty]{}\frac{1}{n}\vartheta\big(q(\alpha_{j_k}),q(\alpha_{j_{k'}})\big).$$

Proof of Theorem 3.2. The almost sure convergence of $\bar{\hat{q}}_j$ is established applying the results of strong consistency of $\overline{q_n(\alpha)}$ and $\hat{q}_{i,m}(\alpha)$ from Theorem 3.1 and Proposition 3.1, respectively.

The asymptotic normality of $\bar{\hat{q}}_j$ is obtained as follows

$$
\begin{aligned}
\sqrt{n}\frac{\Big(\bar{\hat{q}}_j-q_{SE}(\alpha_j)\Big)}{\sqrt{\vartheta\big(q(\alpha_j)\big)}} &= \sqrt{n}\frac{\left(\frac{1}{n}\sum\limits_{i=1}^{n}X_{i,j:m}-q_{SE}(\alpha_j)\right)}{\sqrt{\vartheta\big(q(\alpha_j)\big)}} \\
&= \frac{\sqrt{n}\left(\frac{1}{n}\sum\limits_{i=1}^{n}\Big(X_{i,j:m}-\mathbb{E}\left(X_{i,j:m}\right)\Big)\right)}{\sqrt{\vartheta\big(q(\alpha_j)\big)}}+\frac{\sqrt{n}\left(\frac{1}{n}\sum\limits_{i=1}^{n}\mathbb{E}\left(X_{i,j:m}\right)-q_{SE}(\alpha_j)\right)}{\sqrt{\vartheta\big(q(\alpha_j)\big)}} \\
&\ \vdots
\end{aligned}
$$

$$= \frac{\left(\sum_{i=1}^{n}\Big(X_{i,j:m} - \mathbb{E}\left(X_{i,j:m}\right)\Big)\right)\sqrt{\frac{1}{n}\sum_{i=1}^{n}\operatorname{Var}\left(X_{i,j:m}\right)}}{\sqrt{\sum_{i=1}^{n}\operatorname{Var}\left(X_{i,j:m}\right)}\sqrt{\vartheta\big(q(\alpha_j)\big)}} + \frac{\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(X_{i,j:m}\right) - q_{SE}(\alpha_j)\right)}{\sqrt{\vartheta\big(q(\alpha_j)\big)}}$$

$$= \frac{\left(\sum_{i=1}^{n} X_{i,j:m} - \sum_{i=1}^{n}\mathbb{E}\left(X_{i,j:m}\right)\right)}{\sqrt{\sum_{i=1}^{n}\operatorname{Var}\left(X_{i,j:m}\right)}}\frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\operatorname{Var}\left(X_{i,j:m}\right)}}{\sqrt{\vartheta\big(q(\alpha_j)\big)}} + \frac{\sqrt{n}\left(\frac{1}{8(m+2)}\left(\frac{-M}{C^3}\right) + O\left(\frac{1}{m^2}\right)\right)}{\sqrt{\vartheta\big(q(\alpha_j)\big)}}.$$

Given that $\operatorname{Var}\left(X_{i,j:m}\right) \to \vartheta\big(q(\alpha_j)\big)$ as $m \to \infty$, and under the assumption $\sqrt{n}/m \to 0$ we obtain, by the Lindeberg-Feller's central limit theorem for independent but not identically distributed random variables to independent random variables $X_{1,j:m}, \ldots, X_{n,j:m}$, that

$$\sqrt{n}\frac{\left(\bar{\hat{q}}_j - q_{SE}(\alpha_j)\right)}{\sqrt{\vartheta\big(q(\alpha_j)\big)}} \xrightarrow[n,m\to\infty]{\mathcal{D}} \mathcal{N}\left(0,1\right).$$

In multivariate terms it is expressed as

$$\sqrt{n}\begin{bmatrix}\bar{\hat{q}}_{j_1} - q_{SE}(\alpha_1)\\ \vdots \\ \bar{\hat{q}}_{j_K} - q_{SE}(\alpha_K)\end{bmatrix} \xrightarrow[n,m\to\infty]{\mathcal{D}} \mathcal{N}_K\left(\mathbf{0}, \mathbf{\Sigma}\right),$$

where $(\alpha_1, \ldots, \alpha_K) \in [0,1]^K$ and the (k,k')-element of $\mathbf{\Sigma}$ is given by $\Sigma_{k,k'} = \vartheta\big(q(\alpha_{j_k}), q(\alpha_{j_{k'}})\big)$.

The Lindeberg-Feller's central limit theorem holds if the Lyapunov's condition

$$\frac{1}{\left(\sqrt{\sum_{i=1}^{n}\operatorname{Var}\left(X_{i,j:m}\right)}\right)^{2+\delta}}\sum_{i=1}^{n}\mathbb{E}\left|X_{i,j:m} - \mathbb{E}\left(X_{i,j:m}\right)\right|^{2+\delta} \xrightarrow[n,m\to\infty]{} 0$$

is satisfied for some $\delta > 0$. Indeed for $\delta = 1$ and under the compactly central data hypothesis, $\left|X_{i,j:m} - \mathbb{E}\left(X_{i,j:m}\right)\right| \leq L < \infty$ for all i and j, we have

$$\begin{aligned}&\frac{1}{\left(\sqrt{\sum_{i=1}^{n}\operatorname{Var}\left(X_{i,j:m}\right)}\right)^{2+1}}\sum_{i=1}^{n}\mathbb{E}\left|X_{i,j:m} - \mathbb{E}\left(X_{i,j:m}\right)\right|^{2+1}\\ &\leq \frac{L}{\left(\sqrt{\sum_{i=1}^{n}\operatorname{Var}\left(X_{i,j:m}\right)}\right)^{2+1}}\sum_{i=1}^{n}\mathbb{E}\left|X_{i,j:m} - \mathbb{E}\left(X_{i,j:m}\right)\right|^{2}\\ &\vdots\end{aligned}$$

$$= \frac{L}{\left(\sqrt{\sum_{i=1}^{n} \mathrm{Var}\left(X_{i,j:m}\right)}\right)^{2+1}} \sum_{i=1}^{n} \mathrm{Var}\left(X_{i,j:m}\right)$$

$$= \frac{L}{\left(\sqrt{\sum_{i=1}^{n} \mathrm{Var}\left(X_{i,j:m}\right)}\right)} \xrightarrow[n,m\to\infty]{} 0,$$

given that $\mathrm{Var}\left(X_{i,j:m}\right) \to \vartheta\big(q(\alpha_j)\big)$ as $m \to \infty$.

Therefore the Lyapunov's condition is satisfied. ■

Appendix B

This `R` code is based on a slight modification of the `normalizeQuantiles` function in the package `limma` intended to normalize single channel microarray intensities between arrays, allowing for missing values and treating ties carefully. The code also depends on the functions given in the Appendix B to Chapter 2.

```
# Manifold normalization
normanif <- function(X, ties = TRUE) {
   n <- dim(X)
   if (is.null(n))
      return(X)
   if (n[2] == 1)
      return(X)
   O <- S <- array(, n)
   nobs <- rep(n[1], n[2])
   i <- (0:(n[1] - 1))/(n[1] - 1)
   for (j in 1:n[2]) {
       Si <- sort(X[, j], method = "quick", index.return = TRUE)
       nobsj <- length(Si$x)
       if (nobsj < n[1]) {
          nobs[j] <- nobsj
          isna <- is.na(X[, j])
          S[, j] <- approx((0:(nobsj - 1))/(nobsj - 1), Si$x, i,
                       ties = "ordered")$y
          O[!isna, j] <- ((1:n[1])[!isna])[Si$ix]
       }
       else {
            S[, j] <- Si$x
            O[, j] <- Si$ix
       }
   }
   m <- S[, imean(t(S), 2)]
   for (j in 1:n[2]) {
       if (ties)
          r <- rank(X[, j])
       if (nobs[j] < n[1]) {
          isna <- is.na(X[, j])
          if (ties)
             X[!isna, j] <- approx(i, m, (r[!isna] - 1)/(nobs[j] -
                                 1), ties="ordered")$y
          else
             X[O[!isna,j], j] <- approx(i, m, (0:(nobs[j]-1))/
                                    (nobs[j]-1), ties="ordered")$y
       }
       else {
            if (ties)
               X[, j] <- approx(i, m, (r - 1)/(n[1] - 1),
                            ties = "ordered")$y
            else
               X[O[, j], j] <- m
       }
   }
   X
}
```

Chapter 4

Functional calibration estimation via the maximum entropy on the mean principle

joint work with F. Gamboa* and J-M. Loubes†

Abstract: We extend the problem of obtaining an estimator for the finite population mean parameter incorporating complete auxiliary information through calibration estimation in survey sampling but considering a functional data framework. The functional calibration sampling weights of the estimator are obtained by matching the calibration estimation problem with the maximum entropy on the mean principle. In particular, the calibration estimation is viewed as an infinite-dimensional linear inverse problem following the structure of the maximum entropy on the mean approach. We give a precise theoretical setting and estimate the functional calibration weights assuming, as prior measures, the centered Gaussian and compound Poisson random measures. Additionally, through a simple simulation study, we show that our functional calibration estimator improves its accuracy compared with the Horvitz-Thompson estimator.

Key Words: Auxiliary information; Functional calibration weights; Functional data; Infinite-dimensional linear inverse problems; Maximum entropy; Survey sampling.

* Institut de Mathématiques de Toulouse, Université Paul Sabatier - Toulouse III, Toulouse, France. E-mail: gamboa@math.univ-toulouse.fr

† Institut de Mathématiques de Toulouse, Université Paul Sabatier - Toulouse III, Toulouse, France. E-mail: jean-michel.loubes@math.univ-toulouse.fr

4.1 Introduction

In survey sampling, the well-known calibration estimation method proposed by Deville and Särndal [30] allows to construct an estimate for the finite population total or mean of a survey variable by incorporating complete auxiliary information on the study population in order to improve its efficiency. The main idea of the calibration method consists in modifying the standard sampling design weights d_i of the unbiased Horvitz-Thompson estimator (Horvitz and Thompson [52]) by new weights w_i close enough to d_i's according to some distance function $\mathcal{D}(w, d)$, while satisfying a linear calibration equation in which the auxiliary information is taken into account. The sources of this information may come, for example, from census data, administrative registers, and previous surveys (e.g., Deville and Särndal [30], Särndal et al. [86], Montanari and Ranalli [72]). The estimator based on these new calibration weights is asymptotically design unbiased and consistent with a variance smaller than the Horvitz-Thompson one.

The idea of calibration has been extended to estimate other finite population parameters, such as finite population variances, distribution functions and quantiles (see, e.g., Rao et al. [80], Kovaĉević [63], Théberge [97], Singh [89], Wu and Sitter [110], Wu [109], Harms and Duchesne [50], Rueda et al. [83], Särndal [85], and references therein). Recent developments have also been conducted toward, for example, the approach of (parametric and non-parametric) non-linear relationships between the survey variable and the set of auxiliary variables for the underlying assisting model, and a broad classes of conceivable calibration constraints functions (Breidt and Opsomer [15], Wu and Sitter [110], Wu [109], Montanari and Ranalli [72]).

One interesting extension emerges when both the survey and auxiliary variables are considered as infinite-dimensional objects such as random functions. This generalization relies on the fact that, due to improvements in data collection technologies, large and complex databases are being registered frequently at very fine time scales, regarded these as functional datasets. This kind of data are collected in many scientific fields as molecular biology, astronomy, marketing, finance, economics, among many other. A depth overview on functional data analysis can be found in Ramsay and Silverman [78], Ramsay and Silverman [79] and Horváth and Kokoszka [51]. Functional versions of the Horvitz-Thompson estimator have been proposed recently by Cardot and Josserand [16] and Cardot et al. [17] for the cases of error free and noisy functional data, respectively.

The purpose of the present chapter is to extend the problem of obtaining calibration sampling weights using functional data. This is conducted through the generalization of the work by Gamboa et al. [43], where the calibration estimation problem, which is considered as a linear inverse problem following Théberge [97], is matched with the maximum entropy on the mean approach under a finite dimensional setting. The maximum entropy on the mean principle applied to our goal focuses on reconstructing a unique posterior measure ν^* that maximizes the entropy $S(\nu \parallel \upsilon)$ between a feasible finite measure ν relative to a given prior measure υ subject to a linear constraint. Finally, the functional calibration sampling weights are defined as the mathematical expectation with respect to ν^* of a random variable with mean equal to the standard sampling design weights d_i. In this chapter, we reconstruct ν^* adopting the random measure approach by Gzyl and Velásquez [49] under an infinite-dimensional context.

The maximum entropy method on the mean was introduced by Navaza [73, 74] to solve an inverse problem in crystallography, and has been further investigated, from a mathematical point of view, by Gamboa [40], Dacunha-Castelle and Gamboa [26] and Gamboa and Gassiat [41]. Complementary references on the approach are Mohammad-Djafari [71], Maréchal [69], Gzyl [48], Gzyl and Velásquez [49] and Golan and Gzyl [46]. Maximum entropy solutions, as an alternative to the Tikhonov's regularization of ill-conditioned inverse problems, provide a very simple and natural way to incorporate constraints on the support and the range of the solution (Gamboa and Gassiat [41]), and its usefulness has been proven, e.g., in crystallography, seismic tomography and image reconstruction.

The chapter is organized as follows. Section 4.2, presents the calibration estimation framework for the functional finite population mean. In Section 4.3, the connection between calibration and maximum entropy on the mean approaches is established, and the functional calibration sampling weights are obtained assuming two prior measures. In Section 4.4, the respective approximations of the functional maximum entropy on the mean estimators are derived. The performance of the estimator is studied through a simple simulation study in Section 4.5. Some concluding remarks are given in Section 4.6. Finally, the proofs of the technical results are gathered in the Appendix.

4.2 Calibration estimation for the functional finite population mean

Let $U_N = \{1, \dots, N\}$ be a finite survey population from which a realized sample a is drawn with fixed-size sampling design $p_N(a) = \mathbb{P}(A = a)$. Here $a \in \mathcal{A}$, where $\mathcal{A}$ is the collection of all subsets A of U_N that contains all possible samples of n_N different elements randomly drawn from U_N according to a given sampling selection scheme, and $\mathbb{P}$ a probability measure on $\mathcal{A}$. The first order inclusion probabilities, $\pi_{iN} = \mathbb{P}(i \in a) = \sum_{a \in A(i)} p_N(a)$, where $A(i)$ represents the set of samples that contain the ith element, are assumed to be strictly positive for all $i \in U_N$. See Särndal et al. [86] and Fuller [38] for details about survey sampling.

Associated with the ith element in U_N there exists a unique functional random variable $Y_i(t)$ with values in the space of all continuous real-valued functions defined on $[0,T]$ with $T < +\infty$, $\mathcal{C}([0,T])$. However, only the sample functional data, $Y_i(t)$, $i \in a$ are observed. Additionally, an auxiliary q-dimensional functional vector is available for each $i \in U_N$, $\boldsymbol{X}_i(t) = (X_{i1}(t), \ldots, X_{iq}(t))^\top \in \mathcal{C}\big([0,T]^q\big)$ with $q \geq 1$. The known functional finite population mean is denoted by $\boldsymbol{\mu}_X(t) = N^{-1}\sum_{i\in U_N} \boldsymbol{X}_i(t)$.

The main goal is to obtain a design consistent estimator for the unknown functional finite population mean, $\mu_Y(t) = N^{-1}\sum_{i\in U_N} Y_i(t)$, based on the calibration method. The idea consists in modify the basic sampling design weights, $d_i = \pi_i^{-1}$, of the unbiased functional Horvitz-Thompson estimator defined by $\hat{\mu}_Y^{\mathrm{HT}}(t) = N^{-1}\sum_{i\in a} d_i Y_i(t)$, for new more efficient weights $w_i > 0$ incorporating the auxiliary information. These weights must to be sufficiently close to d_i's according to some dissimilarity distance function $\mathcal{D}_a(w,d)$ on $\mathbb{R}_+^n$, and satisfying the set of calibration constraints

$$N^{-1}\sum_{i\in a} w_i \boldsymbol{X}_i(t) = \boldsymbol{\mu}_X(t).$$

The functional estimator for $\mu_Y(t)$ based on the calibration weights is expressed by the linear weighted estimator $\hat{\mu}_Y(t) = N^{-1}\sum_{i\in a} w_i Y_i(t)$. Different calibration estimators can be obtained depending on the chosen distance function (Deville and Särndal [30]). However, it is well known that, in the finite dimensional setting, all of calibration estimators are asymptotically equivalent to the one obtained through the use of the popular chi-square distance function $\mathcal{D}_a(w,d) = \sum_{i\in a}(w_i - d_i)^2/2d_i q_i$, where q_i is an individual given positive weight uncorrelated with d_i.

Assuming a point-wise multiple linear regression model (Ramsay and Silverman [79]), $Y_i(t) = \boldsymbol{X}_i(t)^\top \boldsymbol{\beta}(t) + \varepsilon_i(t)$, where $\varepsilon_i(t)$ is the ith zero-mean measurement functional error independent of $\boldsymbol{X}_i(t)$ with variance structure given by a diagonal matrix with elements $1/q_i$ unrelated to d_i, then the estimator for $\mu_Y(t)$ from the restricted minimization problem can be expressed as

$$\hat{\mu}_Y(t) = \hat{\mu}_Y^{\mathrm{HT}}(t) + \big\{\boldsymbol{\mu}_X(t) - \hat{\boldsymbol{\mu}}_X^{\mathrm{HT}}(t)\big\}^\top \widehat{\boldsymbol{\beta}}(t),$$

where $\hat{\boldsymbol{\mu}}_X^{\mathrm{HT}}(t) = \sum_{i\in a} d_i \boldsymbol{X}_i(t)$ denotes the Horvitz-Thompson estimator for the functional vector $\boldsymbol{X}(t)$, and $\widehat{\boldsymbol{\beta}}(t) = \big\{\sum_{i\in a} d_i q_i \boldsymbol{X}_i(t)\boldsymbol{X}_i(t)^\top\big\}^{-1}\sum_{i\in a} d_i q_i \boldsymbol{X}_i(t) Y_i(t)$ is the weighted estimator of the functional coefficient vector $\boldsymbol{\beta}(t)$, whose uniqueness relies on the existence of the inverse of the matrix $\sum_{i\in a} d_i q_i \boldsymbol{X}_i(t)\boldsymbol{X}_i(t)^\top$ for all t.

The calibration weights can be generalized allowing functional calibration weights $w_i(t)$ which can be obtained from the minimization of the generalized chi-square distance $\mathcal{D}_a^*(w,d)$, expressed below, subject to the functional calibration restriction

$$N^{-1}\sum_{i\in a} w_i(t)\boldsymbol{X}_i(t) = \boldsymbol{\mu}_X(t). \tag{4.1}$$

The existence of functional calibration weights is stated in the next theorem, which is a straightforward generalization of the finite dimensional results of Deville and Särndal [30].

Theorem 4.1. *Assume the existence of a functional vector* $\boldsymbol{w}(t) = (w_1(t), \dots, w_n(t))^\top$ *such that* (4.1) *holds, and the inverse of the matrix* $\sum_{i\in a} d_i q_i\,(t)\,\boldsymbol{X}_i(t)\boldsymbol{X}_i(t)^\top$. *Then, for a fixed* $t \in [0,T]$, $\hat{\boldsymbol{w}}(t)$ *minimizes over* $\mathcal{C}([0,T]^n)$ *the generalized chi-square distance*

$$\mathcal{D}_a^*(w,d) = \sum_{i\in a} \frac{\big(w_i(t)-d_i\big)^2}{2d_i q_i(t)}$$

subject to (4.1), *where the functional calibration weight* $\hat{w}_i(t)$ *for all* $i \in a$ *is given by*

$$\hat{w}_i(t) = d_i \left[1 + q_i(t)\left\{\boldsymbol{\mu}_X(t) - \hat{\boldsymbol{\mu}}_X^{\mathrm{HT}}(t)\right\}^\top \left\{\sum_{i\in a} d_i q_i(t)\boldsymbol{X}_i(t)\boldsymbol{X}_i(t)^\top\right\}^{-1} \boldsymbol{X}_i(t)\right].$$

Note that, for this generalized setting, the functional calibration estimator for $\mu_Y(t)$ is expressed by

$$\hat{\mu}_Y(t) = N^{-1}\sum_{i\in a} \hat{w}_i(t) Y_i(t) = \hat{\mu}_Y^{\mathrm{HT}}(t) + \left\{\boldsymbol{\mu}_X(t) - \hat{\boldsymbol{\mu}}_X^{\mathrm{HT}}(t)\right\}^\top \widehat{\boldsymbol{\beta}}(t),$$

where

$$\widehat{\boldsymbol{\beta}}(t) = \left\{\sum_{i\in a} d_i q_i(t)\boldsymbol{X}_i(t)\boldsymbol{X}_i(t)^\top\right\}^{-1} \sum_{i\in a} d_i q_i(t)\boldsymbol{X}_i(t) Y_i(t),$$

provided the inverse of the matrix $\sum_{i\in a} d_i q_i(t)\boldsymbol{X}_i(t)\boldsymbol{X}_i(t)^\top$ exists for all t.

4.3 Maximum entropy on the mean for survey sampling

Let $(\widetilde{\mathcal{X}}, \mathcal{F})$ be an arbitrary measurable space over which we want to search for an σ-finite positive measure μ. The maximum entropy on the mean principle provides an efficient way of getting an estimator for some linear functional $\mu_{\widetilde{Y}}(t) = \int_{\widetilde{\mathcal{X}}} \widetilde{Y}(t)\mathrm{d}\mu$ satisfying a known q-vector of functionals $\int_{\widetilde{\mathcal{X}}} \widetilde{\boldsymbol{X}}(t)\mathrm{d}\mu = \boldsymbol{\mu}_X(t)$, where $\widetilde{Y}(t)\colon \widetilde{\mathcal{X}} \to \mathcal{C}([0,T])$ and $\widetilde{\boldsymbol{X}}(t)\colon \widetilde{\mathcal{X}} \to \mathcal{C}([0,T]^q)$.

A natural unbiased and consistent estimator of $\mu_{\widetilde{Y}}(t)$ is the empirical functional mean $\hat{\mu}_{\widetilde{Y}}(t) = \int_{\mathcal{X}} \widetilde{Y}(t)\mathrm{d}\mu_n = n^{-1}\sum_{i\in a}\widetilde{Y}_i(t)$, where $\mu_n = n^{-1}\sum_{i\in a}\delta_{T_i}$ is the corresponding empirical distribution with $T_1, \dots, T_n$ an observed random sample from μ. Despite properties of this estimator, it may not have the smallest variance in this kind of framework. Therefore, incorporating prior functional auxiliary information the variance of an asymptotically unbiased functional estimator can be reduced applying the maximum entropy on the mean principle (Gamboa et al. [43]).

The philosophy of the principle consists in to enhance $\hat{\mu}_{\widetilde{Y}}(t)$ considering the maximum entropy on the mean functional estimator

$$\hat{\mu}_{\widetilde{Y}}^{\mathrm{MEM}}(t) = \int_{\mathcal{X}} \widetilde{Y}(t) \mathrm{d}\hat{\mu}_n^{\mathrm{MEM}} = n^{-1} \sum_{i \in a} \hat{p}_i(t) \widetilde{Y}_i(t), \quad \text{for all } t \in [0, T],$$

where $\hat{\mu}_n^{\mathrm{MEM}} = n^{-1} \sum_{i \in a} \hat{p}_i(t) \delta_{T_i}$ is a weighted version of the empirical distribution μ_n, with $\hat{\boldsymbol{p}}(t) = (\hat{p}_1(t), \ldots, \hat{p}_n(t))^\top$ given by the expectation of the independent n-dimensional stochastic process $\boldsymbol{P}(t) = (P_1(t), \ldots, P_n(t))^\top$ drawn from a posterior finite distribution ν^*, $\hat{\boldsymbol{p}}(t) = \mathbb{E}_{\nu^*}[\boldsymbol{P}(t)]$ for all $t \in [0, T]$, where ν^* must to be close to a prior distribution υ, which transmits the information that $\hat{\mu}_n^{\mathrm{MEM}}$ must to be sufficiently close to μ_n.

Therefore, the maximum entropy on the mean principle focuses on reconstructing the posterior measure ν^* that maximizes the entropy, over the convex set of all probability measures, $S(\nu \parallel \upsilon) = -D(\nu \parallel \upsilon)$ subject to the linear functional constraint holds in mean,

$$\mathbb{E}_{\nu^*}\left[n^{-1} \sum_{i \in a} P_i(t) \widetilde{\boldsymbol{X}}_i(t) \right] = \boldsymbol{\mu}_X(t), \qquad \forall t \in [0, T].$$

We recall that $D(\nu \parallel \upsilon)$ is the I-divergence or relative divergence or Kullback-Leibler information divergence between a feasible finite measure ν with respect to a given prior measure υ (see for details, e.g., Csiszár [24]) defined by

$$D(\nu \parallel \upsilon) = \begin{cases} \int_\Omega \log\left(\frac{\mathrm{d}\nu}{\mathrm{d}\upsilon}\right) \mathrm{d}\nu - \nu(\Omega) + 1 & \text{if } \nu \ll \upsilon \\ +\infty & \text{otherwise.} \end{cases}$$

To establish the connection between calibration and maximum entropy on the mean approaches the following notation, based on Gamboa et al. [43], is adopted $\widetilde{Y}_i(t) = N^{-1} n d_i Y_i(t)$, $\widetilde{\boldsymbol{X}}_i(t) = N^{-1} n d_i \boldsymbol{X}_i(t)$ and $p_i(t) = \pi_i w_i(t)$, such that the functional Horvitz-Thompson estimator of $\mu_Y(t)$ and the functional calibration constrain (4.1) can be, respectively, expressed as

$$\hat{\mu}_Y^{\mathrm{HT}}(t) = N^{-1} \sum_{i \in a} d_i Y_i(t) = n^{-1} \sum_{i \in a} \widetilde{Y}_i(t)$$

and

$$n^{-1} \sum_{i \in a} p_i(t) \widetilde{\boldsymbol{X}}_i(t) = N^{-1} \sum_{i \in a} w_i(t) \boldsymbol{X}_i(t) = \boldsymbol{\mu}_X(t), \qquad \forall t \in [0, T].$$

Hence, the functional calibration estimation problem follows the structure of the maximum entropy on the mean principle, where the corresponding estimator is defined by

$$\hat{\mu}_Y^{\mathrm{MEM}}(t) = n^{-1}\sum_{i\in a}\hat{p}_i(t)\widetilde{Y}_i(t) = N^{-1}\sum_{i\in a}\hat{w}_i(t)Y_i(t).$$

The functional calibration weighting vector $\hat{\boldsymbol{p}}(t)$ with coordinates $\hat{p}_i(t) = \pi_i\hat{w}_i(t)$ for $i \in a$, is the expectation of the n-dimensional stochastic process $\boldsymbol{P}(t)$ with coordinates $P_i(t) = \pi_i W_i(t)$, drawn from ν^*,

$$\hat{\boldsymbol{p}}(t) = \mathbb{E}_{\nu^*}\left[\boldsymbol{P}(t)\right], \qquad \forall t \in [0, T]\,,$$

where the posterior measure $\nu^* = \otimes_{i\in a}\nu_i^*$ (by the independence of P_i's) maximizes the entropy $S(\cdot \parallel \upsilon)$ subject to the calibration constraint is fulfilled in mean,

$$\mathbb{E}_{\nu^*}\left[n^{-1}\sum_{i\in a}P_i(t)\widetilde{\boldsymbol{X}}_i(t)\right] = \mathbb{E}_{\nu^*}\left[N^{-1}\sum_{i\in a}W_i(t)\boldsymbol{X}_i(t)\right] = \boldsymbol{\mu}_X(t), \quad \forall t \in [0, T]\,.$$

Note that as $p_i(t) = \pi_i w_i(t)$ and $\hat{w}_i(t)$ must to be sufficiently close to d_i, then the $\hat{p}_i(t)$ must be close enough to 1 for each $i \in a$.

4.3.1 Reconstruction of the measure ν^*

For simplicity and without loss generality we assume that $T = 1$. The posterior distribution ν^* can be reconstructed adopting the random measure approach for infinite-dimensional inverse problems explained in detail by Gzyl and Velásquez [49]. To do this, we express the calibration constraint (4.1) as an infinite-dimensional linear inverse problem writing $w_i(t)$ as

$$w_i(t) = \int_0^1 K(s,t)\varpi_i\left(s\right)\mathrm{d}s + d_i \quad \text{for each } i \in a,$$

where $K(s,t)$ is a known continuous, real-valued and bounded kernel function and $\varpi_i = \mathbb{E}_\nu\left[\mathcal{W}_i\left(s\right)\right]$, where $\mathcal{W}$ is a stochastic process. Note that, as $p_i(t) = \pi_i w_i(t)$ then $p_i(t) = \pi_i\int_0^1 K(s,t)\varpi_i\left(s\right)\mathrm{d}s + 1$.

Hence, the infinite-dimensional inverse problem, which takes the form of a Fredholm integral equation of the first kind, is

$$\begin{aligned}\mathbb{E}_\nu\left[\mathcal{K}\mathcal{W}\right] &= \mathbb{E}_\nu\left\{\sum_{i\in a}\left[\int_0^1 K(s,t)\mathrm{d}\mathcal{W}_i\left(s\right) + d_i\right]\boldsymbol{X}_i(t)\right\}\\ &= \int_0^1\sum_{i\in a}K(s,t)\boldsymbol{X}_i(t)\varpi_i\left(s\right)\mathrm{d}s + \sum_{i\in a}d_i\boldsymbol{X}_i(t) \qquad (4.2)\\ &= N\boldsymbol{\mu}_X(t), \qquad t \in [0,1]\,.\end{aligned}$$

To obtain the functions $\varpi_i^*(s)$ that solve the integral equation $\mathbb{E}_\nu[\mathcal{K}\mathcal{W}] = N\boldsymbol{\mu}_X(t)$, the random measure approach adopted considers $\varpi_i(s)$ as a density of a measure $\varpi_i(s)\,\mathrm{d}s$, $i \in a$. Under this setting, we define the random measure $\mathcal{W}_i(a,b] = \mathcal{W}_i(b) - \mathcal{W}_i(a)$ for $(a,b] \subset [0,1]$ such that $\mathrm{d}\mathbb{E}_\nu\{\mathcal{W}_i(0,s]\} = \varpi_i(s)\mathrm{d}s$ for each $i \in a$. The next theorem ensures the existence of the posterior distribution ν^* to obtain the functions $\varpi_i^*(s)$ depending on the assumed prior distribution υ.

Theorem 4.2. *Let υ be a prior positive probability measure, $\boldsymbol{\lambda} = \boldsymbol{\lambda}(t)$ a measure in the class of continuous measures on $[0,1]$, $\mathcal{M}(C[0,1]^q)$, and $\mathcal{V} = \{\nu \ll \upsilon \colon Z_\upsilon(\boldsymbol{\lambda}) < +\infty\}$ a nonempty open class, where $Z_\upsilon(\boldsymbol{\lambda}) = \mathbb{E}_\upsilon[\exp\{\langle\boldsymbol{\lambda}, \mathcal{K}\mathcal{W}\rangle\}]$, with*

$$\langle\boldsymbol{\lambda}, \mathcal{K}\mathcal{W}\rangle = \int_0^1 \boldsymbol{\lambda}^\top(\mathrm{d}t)\left(\int_0^1 \sum_{i\in a} K(s,t)\boldsymbol{X}_i(t)\mathrm{d}\mathcal{W}_i(s) + \sum_{i\in a} d_i\boldsymbol{X}_i(t)\right). \tag{4.3}$$

Then there exists a unique probability measure

$$\nu^* = \arg\max_{\nu\in\mathcal{V}} S(\nu \parallel \upsilon),$$

subject to $\mathbb{E}_\nu[\mathcal{K}\mathcal{W}] = N\boldsymbol{\mu}_X(t)$, which is achieved at

$$\mathrm{d}\nu^*/\mathrm{d}\upsilon = Z_\upsilon^{-1}(\boldsymbol{\lambda}^*)\exp\{\langle\boldsymbol{\lambda}^*, \mathcal{K}\mathcal{W}\rangle\},$$

where $\boldsymbol{\lambda}^*(t)$ minimizes the functional

$$H_\upsilon(\boldsymbol{\lambda}) = \log Z_\upsilon(\boldsymbol{\lambda}) - \langle\boldsymbol{\lambda}, N\boldsymbol{\mu}_X\rangle.$$

Based on the Theorem 4.2, we will carry out the reconstruction of ν, assuming the centered Gaussian and compound Poisson random measures as prior measures, in order to estimate the respective functional calibration weights $\hat{w}_i(t)$, $i \in a$. The estimates are given by the following two Lemmas.

Lemma 4.1. *Let υ be a prior centered stationary Gaussian measure on the measurable space $\big(\mathcal{C}([0,1]), \mathcal{B}(\mathcal{C}([0,1]))\big)$, and $\boldsymbol{\lambda} = \boldsymbol{\lambda}(t) \in \mathcal{M}(C[0,1]^q)$. Then, $\hat{w}_i(t) = \int_0^1 K(s,t)\varpi^*(s)\mathrm{d}s + d_i$ $i \in a$, where*

$$\varpi^*(s) = \sum_{i'\in a}\int_0^1 K(s,t')\boldsymbol{X}_{i'}^\top(t')\boldsymbol{\lambda}^*(\mathrm{d}t').$$

Lemma 4.2. *Let $\mathcal{W}_i(s) = \sum_{k=1}^{N(s)} \xi_{ik}$ be a compound Poisson process, where $N(s)$ is a homogeneous Poisson process on $[0,1]$ with intensity parameter $\gamma > 0$, and ξ_{ik}, $k \geq 1$ are independent and identically distributed real-valued random variables for each $i \in a$ with distribution u on $\mathbb{R}$ satisfying $u(\{0\}) = 0$, and independent of $N(s)$. Then, $\hat{w}_i(t) = \int_0^1 K(s,t)\varpi_i^*(s)\mathrm{d}s + d_i$ $i \in a$, where*

$$\varpi_i^*(s) = \int_{\mathbb{R}} \xi_i \exp\left\{\sum_{i\in a}\int_0^1 K(s,t)\xi_i\boldsymbol{X}_i^\top(t)\boldsymbol{\lambda}^*(\mathrm{d}t)\right\} u(\mathrm{d}\xi_i).$$

4.4 Approximation of the maximum entropy on the mean functional estimator

To approximate the functional calibration weights and the functional maximum entropy on the mean estimator for the finite population mean of $Y(t)$ with the assumed prior measure, an Euler discretization scheme is used. Consider a partition of $(s,t) \in [0,1]^2$ in J and L equidistant fixed points, $(j-1)/J < s_j \leq j/J$, $j = 1, \dots, J$, $(l-1)/L < t_l \leq l/L$, $l = 1, \dots, L$, respectively. For the corresponding prior measures, the approximations for functions $Z_v(\boldsymbol{\lambda})$, $H_v(\boldsymbol{\lambda})$ and $\boldsymbol{\lambda}^*(t)$ are based on the respective results found in the Appendix.

4.4.1 Centered Gaussian measure

For a prior centered Gaussian measure, the approximations of the linear moment calibration constraint (4.2) and the inner product $\langle \boldsymbol{\lambda}, \mathcal{K}\mathcal{W} \rangle$ are, respectively, given by

$$\mathbb{E}_\nu \left[\sum_{j=1}^{J} \sum_{i \in a} K(s_j, t_l) \Delta\mathcal{W}_i(s_j) \boldsymbol{X}_i(t_l) + \sum_{i \in a} d_i \boldsymbol{X}_i(t_l) \right] = N \boldsymbol{\mu}_X(t_l)$$

and

$$\begin{aligned} &\frac{1}{L} \sum_{l=1}^{L} \boldsymbol{\lambda}^\top(t_l) \sum_{j=1}^{J} \sum_{i \in a} K(s_j, t_l) \Delta\mathcal{W}_i(s_j) \boldsymbol{X}_i(t_l) + \frac{1}{L} \sum_{l=1}^{L} \boldsymbol{\lambda}^\top(t_l) \sum_{i \in a} d_i \boldsymbol{X}_i(t_l) \\ &= \frac{1}{L} \sum_{j=1}^{J} \sum_{i \in a} \sum_{l=1}^{L} K(s_j, t_l) \Delta\mathcal{W}_i(s_j) \boldsymbol{\lambda}^\top(t_l) \boldsymbol{X}_i(t_l) + \frac{1}{L} \sum_{i \in a} d_i \sum_{l=1}^{L} \boldsymbol{\lambda}^\top(t_l) \boldsymbol{X}_i(t_l), \end{aligned}$$

where $\Delta\mathcal{W}_i(s_j) = \mathcal{W}_i(s_j) - \mathcal{W}_i(s_{j-1})$ is the discrete version of $\mathrm{d}\mathcal{W}_i(s)$ for $i \in a$.

Therefore, we have that $Z_v(\boldsymbol{\lambda})$ is approximated at the grid (see equation (4.6) of the proof of Lemma 1 in the Appendix) by

$$\begin{aligned} &\mathbb{E}_v \left[\exp \left\{ \frac{1}{L} \sum_{i \in a} d_i \sum_{l=1}^{L} \boldsymbol{\lambda}^\top(t_l) \boldsymbol{X}_i(t_l) + \frac{1}{L} \sum_{j=1}^{J} \sum_{i \in a} \sum_{l=1}^{L} K(s_j, t_l) \boldsymbol{\lambda}^\top(t_l) \boldsymbol{X}_i(t_l) \Delta\mathcal{W}_i(s_j) \right\} \right] \\ &= \exp \left\{ \frac{1}{L} \sum_{i \in a} d_i \sum_{l=1}^{L} \boldsymbol{\lambda}^\top(t_l) \boldsymbol{X}_i(t_l) + \sum_{j=1}^{J} \frac{1}{2J} \left(\frac{1}{L} \sum_{i \in a} \sum_{l=1}^{L} K(s_j, t_l) \boldsymbol{\lambda}^\top(t_l) \boldsymbol{X}_i(t_l) \right)^2 \right\} \\ &= \exp \left\{ \frac{1}{L} \sum_{i \in a} d_i \sum_{l=1}^{L} \boldsymbol{\lambda}^\top(t_l) \boldsymbol{X}_i(t_l) \right\} \prod_{j=1}^{J} \exp \left\{ \frac{1}{2J} \sum_{i \in a} \sum_{i' \in a} h_i(s_j) h_{i'}(s_j) \right\} \\ &= \exp \left\{ \frac{1}{L} \sum_{i \in a} d_i \sum_{l=1}^{L} \boldsymbol{\lambda}^\top(t_l) \boldsymbol{X}_i(t_l) \right\} \prod_{j=1}^{J} z_i \left(h_i(s_j) \right), \end{aligned}$$

where $h_i(s_j) = L^{-1}\sum_{l=1}^{L} K(s_j, t_l)\boldsymbol{\lambda}^\top(t_l)\boldsymbol{X}_i(t_l)$, $i \in a$, $j = 1, \dots J$, and $l = 1, \dots L$.

Now, the finite dimensional maxentropic solution for $\varpi_i(s_j)$ for each $i \in a$ is approximated by (see Gzyl and Velásquez [49])

$$\begin{aligned}\varpi_i^*(s_j) &= \left.\frac{\mathrm{d}\log z_i\left(h_i(s_j)\right)}{\mathrm{d}(2J)^{-1}h_i(s_j)}\right|_{h_i(s_j)=\mathcal{K}\boldsymbol{\lambda}^*}\\ &= \left.\sum_{i'\in a} h_{i'}(s_j)\right|_{h_i(s_j)=\mathcal{K}\boldsymbol{\lambda}^*}\\ &= \frac{1}{L}\sum_{l=1}^{L}\sum_{i'\in a} K(s_j, t_l')\boldsymbol{\lambda}^{*\top}(t_l')\boldsymbol{X}_{i'}(t_l'),\end{aligned} \tag{4.4}$$

where the finite dimensional version of $\boldsymbol{\lambda}^*(t_l')$, $(l-1)/L < t_l \le l/L$, $l = 1, \dots, L$, is the minimizer of $H_v(\boldsymbol{\lambda})$, whose approximation (see equation (4.7) of the proof of Lemma 1 in the Appendix) is

$$\begin{aligned}&\frac{1}{2}\sum_{l=1}^{L}\sum_{l=1}^{L}\boldsymbol{\lambda}^\top(t_l)\left(\frac{1}{JL^2}\sum_{j=1}^{J}K(s_j, t_l)K(s_j, t_l')\sum_{i\in a}\sum_{i'\in a}\boldsymbol{X}_i(t_l)\boldsymbol{X}_{i'}^\top(t_l')\right)\boldsymbol{\lambda}(t_l')\\ &\quad + \frac{1}{L}\sum_{l=1}^{L}\left(\sum_{i\in a} d_i\boldsymbol{X}_i^\top(t_l) - N\boldsymbol{\mu}_X^\top(t_l)\right)\boldsymbol{\lambda}(t_l).\end{aligned}$$

The first order condition (see equation(4.8)) associated to this minimization problem is

$$\begin{aligned}&\frac{1}{JL^2}\sum_{j=1}^{J}\sum_{l=1}^{L}K(s_j, t_l)K(s_j, t_l')\sum_{i\in a}\sum_{i'\in a}\boldsymbol{X}_i(t_l)\boldsymbol{X}_{i'}^\top(t_l')\boldsymbol{\lambda}^*(t_l')\\ &\quad + \frac{1}{L}\left(N\boldsymbol{\mu}_X(t_l) - \sum_{i\in a} d_i\boldsymbol{X}_i(t_l)\right) = \mathbf{0},\end{aligned}$$

whose solution $\boldsymbol{\lambda}^*(t_l')$ is given by

$$\begin{aligned}\boldsymbol{\lambda}^*(t_l') &= \left(\frac{1}{JL^2}\sum_{j=1}^{J}\sum_{l=1}^{L}K(s_j, t_l)K(s_j, t_l')\sum_{i\in a}\sum_{i'\in a}\boldsymbol{X}_i(t_l)\boldsymbol{X}_{i'}^\top(t_l')\right)^{-1}\\ &\quad\times\frac{1}{L}\left(N\boldsymbol{\mu}_X(t_l) - \sum_{i\in a} d_i\boldsymbol{X}_i(t_l)\right)\\ &= JL\left(\sum_{j=1}^{J}\sum_{l=1}^{L}K(s_j, t_l)K(s_j, t_l')\sum_{i\in a}\sum_{i'\in a}\boldsymbol{X}_i(t_l)\boldsymbol{X}_{i'}^\top(t_l')\right)^{-1}\\ &\quad\times\left(N\boldsymbol{\mu}_X(t_l) - \sum_{i\in a} d_i\boldsymbol{X}_i(t_l)\right).\end{aligned}$$

Finally, the approximation of the finite dimensional solution of $\hat{w}_i(t)$ is

$$\hat{w}_i(t_l) = \frac{1}{J}\sum_{j=1}^{J} K(s_j, t_l)\varpi_i^*(s_j) + d_i,$$

where $\varpi_i^*(s_j)$ es given by the equation (4.4).

4.4.2 Compound Poisson measure

Based on equations (4.9) and (4.10) of the proof of Lemma 2 in the Appendix, the approximation of $Z_v(\boldsymbol{\lambda})$ is given by

$$\begin{aligned}
&\mathbb{E}_v\left[\exp\left\{\langle g(s_j), \mathrm{d}\mathcal{W}_i\rangle + \left\langle \boldsymbol{\lambda}, \sum_{i\in a} d_i \boldsymbol{X}_i(t_l)\right\rangle\right\}\right]\\
&= \exp\left\{\left\langle \boldsymbol{\lambda}, \sum_{i\in a} d_i \boldsymbol{X}_i(t_l)\right\rangle\right\}\mathbb{E}_v\left[\exp\left\{\langle g(s_j), \mathrm{d}\mathcal{W}_i\rangle\right\}\right]\\
&= \exp\left\{\left\langle \boldsymbol{\lambda}, \sum_{i\in a} d_i \boldsymbol{X}_i(t_l)\right\rangle\right\}\prod_{j=1}^{J}\mathbb{E}_v\left[\exp\left\{g\left(s_j\right) m_i\left(\left(s_{j-1}, s_j\right]\right)\right\}\right]\\
&= \exp\left\{\left\langle \boldsymbol{\lambda}, \sum_{i\in a} d_i \boldsymbol{X}_i(t_l)\right\rangle\right\}\prod_{j=1}^{J}\exp\left\{\mathbb{E}_v\left[\exp\left\{g\left(s_j\right)\xi_i\right\}\right]\right\}\\
&= \exp\left\{\left\langle \boldsymbol{\lambda}, \sum_{i\in a} d_i \boldsymbol{X}_i(t_l)\right\rangle\right\}\prod_{j=1}^{J}\exp\left\{\frac{\gamma}{J}\int_{\mathbb{R}}\left(\exp\left\{g\left(s_j\right)\xi_i\right\} - 1\right)u\left(d\xi_i\right)\right\}\\
&= \exp\left\{\left\langle \boldsymbol{\lambda}, \sum_{i\in a} d_i \boldsymbol{X}_i(t_l)\right\rangle\right\}\\
&\quad\times\prod_{j=1}^{J}\exp\left\{\frac{\gamma}{J}\int_{\mathbb{R}}\left(\exp\left\{\frac{1}{L}\sum_{i\in a}\xi_i\sum_{l=1}^{L} K(s_j, t_l)\boldsymbol{\lambda}^\top(t_l)\boldsymbol{X}_i(t_l)\right\} - 1\right)u\left(\mathrm{d}\xi_i\right)\right\}\\
&= \exp\left\{\left\langle \boldsymbol{\lambda}, \sum_{i\in a} d_i \boldsymbol{X}_i(t_l)\right\rangle\right\}\prod_{j=1}^{J}\exp\left\{\frac{\gamma}{J}\int_{\mathbb{R}}\left(\exp\left\{\sum_{i\in a}\xi_i h_i(s_j)\right\} - 1\right)u\left(\mathrm{d}\xi_i\right)\right\}\\
&= \exp\left\{\left\langle \boldsymbol{\lambda}, \sum_{i\in a} d_i \boldsymbol{X}_i(t_l)\right\rangle\right\}\prod_{j=1}^{J} z_i\left(h_i(s_j)\right), \qquad i \in a,
\end{aligned}$$

where $h_i(s_j) = L^{-1}\sum_{l=1}^{L} K(s_j, t_l)\boldsymbol{\lambda}^\top(t_l)\boldsymbol{X}_i(t)$ for all $i \in a$ and $j = 1, \ldots J$, with $\left\langle \boldsymbol{\lambda}, \sum_{i\in a} d_i \boldsymbol{X}_i(t_l)\right\rangle = L^{-1}\sum_{i\in a} d_i \sum_{l=1}^{L}\boldsymbol{\lambda}^\top(t_l)\boldsymbol{X}_i(t_l)$.

The approximated maxentropic solution for $\varpi_i(s_j)$ for each $i \in a$ is

$$\begin{aligned}\varpi_i^*(s_j) &= \left.\frac{\mathrm{d}\log z_i\left(h_i(s_j)\right)}{\mathrm{d}h_i(s_j)}\right|_{h_i(s_j)=\mathcal{K}\boldsymbol{\lambda}^*} \\ &= \left.\frac{\gamma}{J}\int_{\mathbb{R}} \xi_i \exp\left\{\sum_{i\in a} \xi_i h_i(s_j)\right\} u\left(\mathrm{d}\xi_i\right)\right|_{h_i(s_j)=\mathcal{K}\boldsymbol{\lambda}^*} \\ &= \frac{\gamma}{J}\int_{\mathbb{R}} \xi_i \exp\left\{\frac{1}{L}\sum_{i\in a}\sum_{l=1}^{L} K(s_j,t_l)\xi_i \boldsymbol{X}_i^\top(t_l)\boldsymbol{\lambda}^*(t_l)\right\} u\left(\mathrm{d}\xi_i\right),\end{aligned} \tag{4.5}$$

where the finite dimensional version of $\boldsymbol{\lambda}^*(t_l)$, is the minimizer of $H_v(\boldsymbol{\lambda})$, whose approximation, by the equation (4.11) of the proof of Lemma 2 in the Appendix, is

$$\begin{aligned}H_v(\boldsymbol{\lambda}) &= \log Z_v(\boldsymbol{\lambda}) - \langle \boldsymbol{\lambda}, N\boldsymbol{\mu}_X\rangle \\ &= \frac{\gamma}{J}\sum_{j=1}^{J}\int_{\mathbb{R}}\left(\exp\left\{\frac{1}{L}\sum_{i\in a}\sum_{l=1}^{L} K(s_j,t_l)\xi_i\boldsymbol{X}_i^\top(t_l)\boldsymbol{\lambda}(t_l)\right\} - 1\right) u\left(\mathrm{d}\xi_i\right) \\ &\quad + \frac{1}{L}\sum_{l=1}^{L}\left(\sum_{i\in a} d_i\boldsymbol{X}_i^\top(t_l) - N\boldsymbol{\mu}_X^\top(t_l)\right)\boldsymbol{\lambda}(t_l)\end{aligned}$$

The corresponding equation for $\boldsymbol{\lambda}^*(t_l)$ that minimizes $H_v(\boldsymbol{\lambda})$ is given by the nonlinear system of equations (see equation (4.12) in the Appendix)

$$\sum_{i\in a}\left[\frac{1}{J}\sum_{j=1}^{J} K(s_j,t_l)\left(\gamma L\int_{\mathbb{R}} \xi_i \exp\left\{\frac{1}{L}\sum_{i\in a}\sum_{l=1}^{L} K(s_j,t_l)\xi_i\boldsymbol{X}_i^\top(t_l)\boldsymbol{\lambda}^*(t_l)\right\} u\left(\mathrm{d}\xi_i\right)\right) + d_i\right] \\ \times \boldsymbol{X}_i(t_l) = N\boldsymbol{\mu}_X(t_l).$$

Finally, as in the Gaussian measure case, the finite dimensional solution of $\hat{w}_i(t)$ is approximated by $\hat{w}_i(t_l) = J^{-1}\sum_{j=1}^{J} K(s_j,t_l)\varpi_i^*(s_j) + d_i$ with $\varpi_i^*(s_j)$ given by the equation (4.5).

4.5 Simulation study

We shall illustrate through a simple simulation study the performance of results obtained in the above section. Considering a finite population U_N of size $N = 1000$, we generate a functional random variable $Y_i(t)$ by the point-wise multiple linear regression model (see for instance Ramsay and Silverman [79], Horváth and Kokoszka [51] or Zhang and Chen [115])

$$Y_i(t) = \alpha(t) + \boldsymbol{X}_i(t)^\top\boldsymbol{\beta}(t) + \varepsilon_i(t), \qquad i \in U_N,$$

where $\alpha(t) = 1.2 + 2.3\cos(2\pi t) + 4.2\sin(2\pi t)$, $\boldsymbol{\beta}(t) = (\beta_1(t), \beta_2(t))^\top$ with $\beta_1(t) = \cos(10t)$ and $\beta_2(t) = t\sin(15t)$, $\boldsymbol{X}_i(t) = (X_{i1}(t), X_{i2}(t))^\top$, and $\varepsilon_i(t) \sim \mathcal{N}(0, \sigma_\varepsilon^2(1+t))$ with $\sigma_\varepsilon^2 = 0.1$, and independent of $\boldsymbol{X}_i(t)$. The auxiliary functional covariates are defined by $X_{i1}(t) = \mathcal{U}_{i1} + f_1(t)$ with $f_1(t) = 3\sin(3\pi t + 3)$, and $X_{i2}(t) = \mathcal{U}_{i2} + f_2(t)$ with $f_2(t) = -\cos(\pi t)$, where $\mathcal{U}_{i1}$ and $\mathcal{U}_{i2}$ are independent and, respectively, i.i.d. uniform random variables on the intervals $[-1, 1.3]$ and $[-0.5, 0.5]$. The design time points for $t \in [0,1]$ and $s \in [0,1]$ are $t_j = j/J$, $j = 1, \ldots, J$ and $s_l = l/L$, $l = 1, \ldots, L$, with $J = 50$ and $L = 80$.

The Figures 4.1 and 4.2 show, respectively, the simulated finite population auxiliary functional covariates and functional responses for each $i \in U_N$, and the respective finite population functional means, $\boldsymbol{\mu}_X(t) = (\mu_{X_1}(t), \mu_{X_2}(t))^\top$ and $\mu_Y(t) = N^{-1}\sum_{i \in U_N} Y_i(t)$. Assuming a uniform fixed-size sampling design we drawn a sample $a \in U_N$ of $n = 0.12N$ elements without replacement. For the kernel function we assumed a Gaussian one, $K(t,s) = \exp\left\{-|t-s|^2/2\sigma^2\right\}$ with $\sigma^2 = 0.5$. The random variables ξ_i for the compound Poisson case are assumed i.i.d. uniform on the interval $[-1, 1]$, and $\gamma = 1$. To solve the nonlinear system of equations for $\boldsymbol{\lambda}^*(t_l)$ in the compound Poisson case, we used the R-package BB (see Varadhan [102] and Varadhan and Gilbert [103]).

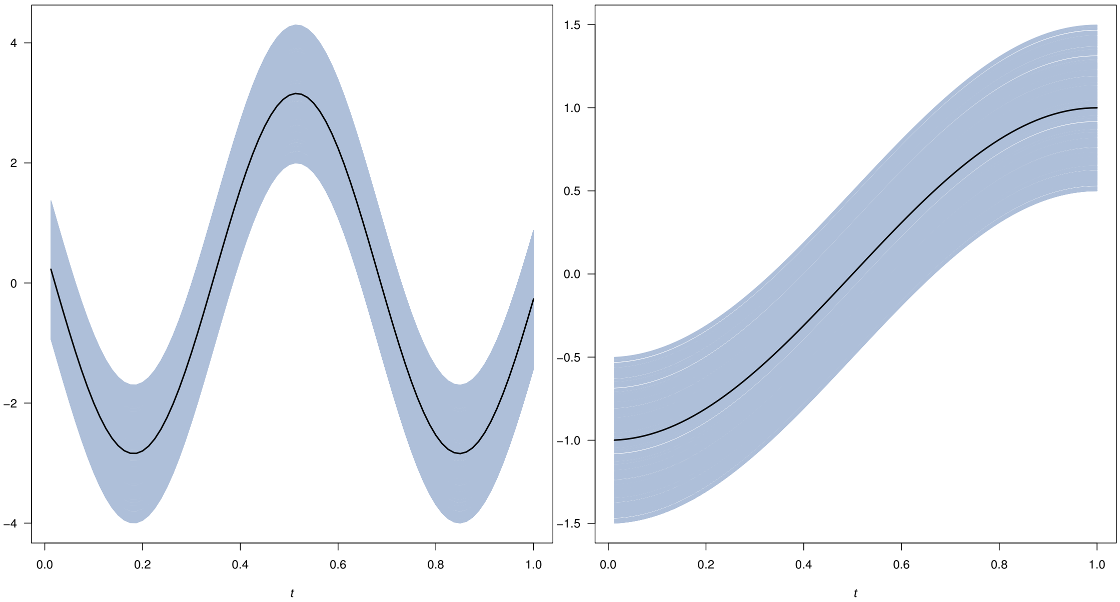

Figure 4.1: Population auxiliary functional variables (gray lines), $X_{i1}(t)$ (on left) and $X_{i2}(t)$ (on right). Functional finite population means, $\mu_{X_1}(t)$ and $\mu_{X_2}(t)$ (bold solid line).

The graphical comparisons of the estimators for a random selected repetition are illustrated in the Figure 4.2. The Figure shows, in general, a good performance, specially for the estimator assuming the Gaussian measure. The principal differences with respect to the theoretical functional finite population mean are localized on the edges, particularly on the left edge. The Horvitz-Thompson estimator, in both cases, has a little departure localized around the deep valley. However our estimator has not this departure. A nice feature of the functional calibration method is that permits to check graphically how well the estimator satisfies the calibration constraints for each covariate, $N^{-1}\sum_{i\in a}\hat{w}_i(t)\boldsymbol{X}_i(t)=\boldsymbol{\mu}_X(t)$. This is illustrated in the Figure 4.3.

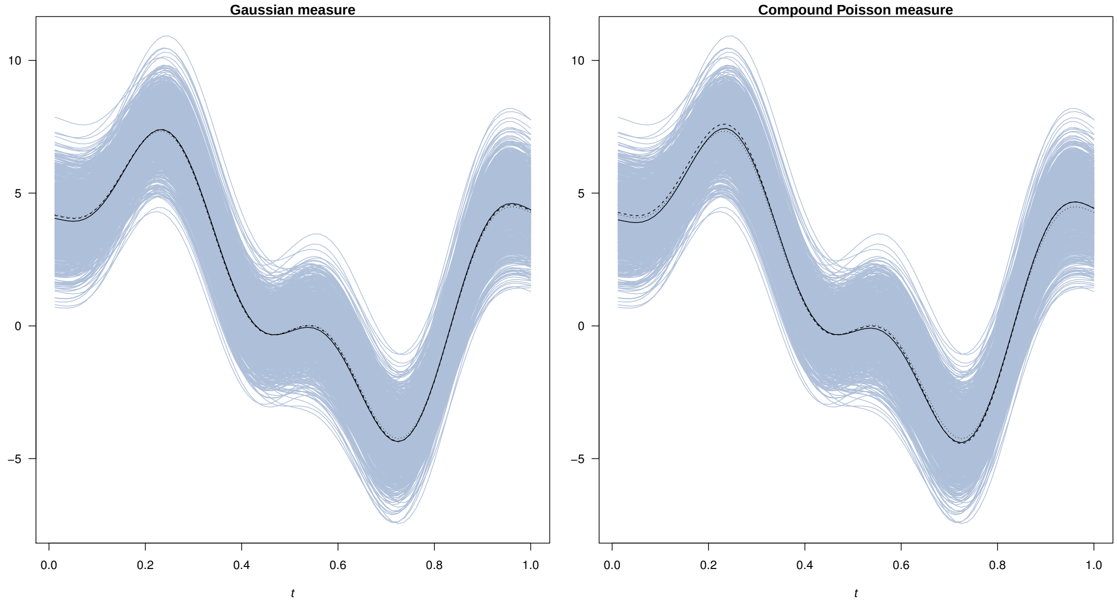

Figure 4.2: Population survey functions $Y_i(t)$ (gray lines), finite population mean $\mu_Y(t)$ (solid line), and the functional Horvitz-Thompson (dotted line) and maxentropic (dashed line) functional estimators.

To evaluate the performance of the maximum entropic functional calibration estimator, $\hat{\mu}_Y^{\mathrm{MEM}}(t)$, assuming the Gaussian and compound Poisson prior measures, we calculated its empirical bias–variance decomposition of the mean square errors and compare it with the functional Horvitz-Thompson estimator $\hat{\mu}_Y^{\mathrm{HT}}(t)$. The simulation study was conducted with 100 repetitions. In Table 4.1 we can see that, with respect to the Horvitz-Thompson estimator, the maximum entropic estimator has smaller variance and mean square error for both prior measures, particularly for the Gaussian prior. Although the Horvitz-Thompson estimator has smaller bias squared, the differences are not significant. Also, the small value for the bias confirm the unbiasedness of the functional maximum entropy on the mean and Horvitz-Thompson estimators.

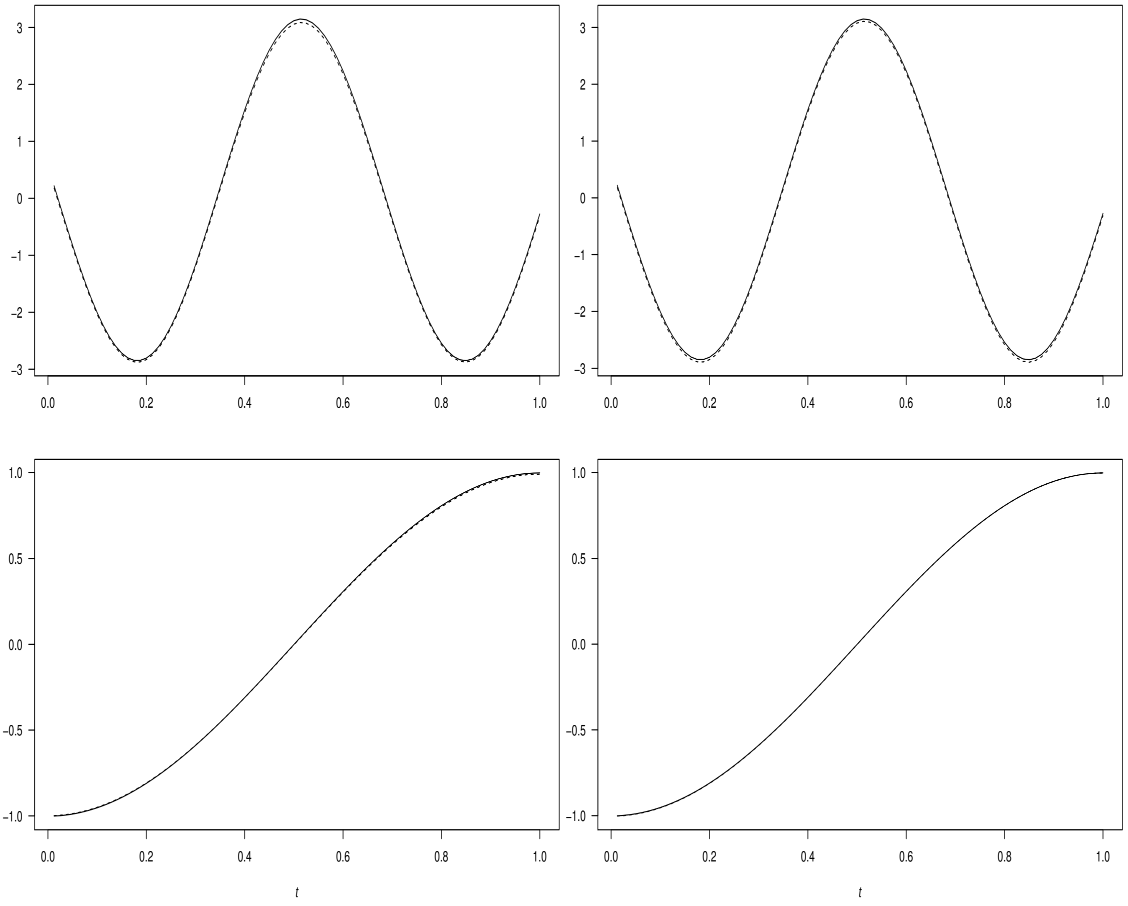

Figure 4.3: Functional calibration constraint (4.1) for Gaussian (on left) and compound Poisson (on right) measures. $\boldsymbol{\mu}_X(t)$ (solid line), $N^{-1}\sum_{i\in a}\hat{w}_i(t)X_i(t)$ (dashed line).

Table 4.1: Bias-variance decomposition of MSE

Functional estimator	MSE	Bias^2	Variance
Horvitz-Thompson	0.2391	0.0005	0.2386
Maximum entropy on the mean (Gaussian)	0.2001	0.0006	0.1995
Maximum entropy on the mean (Poisson)	0.2333	0.0084	0.2249

4.6 Concluding remarks

In this chapter we have proposed an extension to the problem of obtaining an estimator for the finite population mean of a survey variable incorporating complete auxiliary information under an infinite-dimensional setting. Considering that both the survey and the set of auxiliary variables are functions, the respective functional calibration constraint is expressed as an infinite-dimensional linear inverse problem, whose solution offers the functional survey weights of the calibration estimator. The solution of the problem is conducted by mean the maximum entropy on the mean principle, which is a powerful probabilistic-based regularization method to solve constrained linear inverse problems. Here we assume a centered Gaussian and compound Poisson random measures as prior measures to obtain the functional calibration weights. However, other random measures can be considered also.

The simulations study results show that the proposed functional calibration estimator improves its accuracy compared with the Horvitz-Thompson estimator. In the simulations, both the functional survey and auxiliary variables where assumed with amplitude variations (variation in the y-axis) only. More complex extensions allowing both amplitude and phase (variation in the x-axis) variations are possible.

Finally, a further interesting extension of the functional calibration estimation problem under the maximum entropy on the mean approach can be conducted following the idea of model-calibration proposed by Wu and Sitter [110], Wu [109] and Montanari and Ranalli [72]. This may be accomplished considering a nonparametric functional regression $Y_i(t) = \mu\left\{\boldsymbol{X}_i(t)\right\} + \varepsilon_i(t)$, $i \in U_N$, $t \in \big([0,T]$ to model the relation between the functional survey variable and the set of functional auxiliary covariates in order to allows a more effective use of the functional auxiliary information.

Appendix

Proof of Theorem 4.1. The Lagrangian function associated to the restricted minimization problem is

$$L_a(\boldsymbol{w},\boldsymbol{\lambda}) = \mathcal{D}_a^*(w,d) + \boldsymbol{\lambda}^\top(t)\left(\boldsymbol{\mu}_X(t) - N^{-1}\sum_{i\in a} w_i(t)\boldsymbol{X}_i(t)\right),$$

where $\boldsymbol{\lambda}(t)$ is the corresponding functional Lagrange multiplier vector. The first order conditions are

$$\frac{w_i(t) - d_i}{d_i q_i\,(t)} - \boldsymbol{\lambda}(t)^\top \boldsymbol{X}_i(t) = 0, \qquad i \in a$$

which can be expressed as

$$w_i(t) = d_i\left[1 + q_i(t)\boldsymbol{\lambda}(t)^\top \boldsymbol{X}_i(t)\right], \qquad i \in a$$

where, its uniqueness is guaranteed by the continuous differentiability of $\mathcal{D}_a^*(w,d)$ with respect to $w_i(t)$ for all $i \in a$, and by its strictly convexity.

From the functional calibration restriction (4.1) and by the existence assumption on the inverse of the matrix $\sum_{i\in a} d_i q_i(t)\boldsymbol{X}_i(t)\boldsymbol{X}_i(t)^\top$ for all t, the Lagrange functional multiplier vector is determined by

$$\hat{\boldsymbol{\lambda}}(t) = \left(\sum_{i\in a} d_i q_i(t)\boldsymbol{X}_i(t)\boldsymbol{X}_i(t)^\top\right)^{-1} \left(\boldsymbol{\mu}_X(t) - \hat{\boldsymbol{\mu}}_X^{\mathrm{HT}}(t)\right).$$

Finally, replacing $\hat{\boldsymbol{\lambda}}(t)$ into the first order conditions, the calibration functional estimator $\hat{w}_i(t)$ of the Theorem is obtained. ∎

Proof of Theorem 4.2. Csiszár [25, Theorem 3, page 775]. ∎

Proof of Lemma 4.1. According to Theorem 4.2, the maximum of the entropy $S(\nu \parallel \upsilon)$ over the class $\mathcal{V} = \{\nu \ll \upsilon \colon Z_\upsilon(\boldsymbol{\lambda}) < \infty\}$ subject to the linear moment calibration constraint $\mathbb{E}_\upsilon[\mathcal{K}\mathcal{W}] = N\boldsymbol{\mu}_X(t)$ is attained at $\mathrm{d}\nu^*/\mathrm{d}\upsilon = Z_\upsilon^{-1}(\boldsymbol{\lambda}^*)\exp\{\langle\boldsymbol{\lambda}^*, \mathcal{K}\mathcal{W}\rangle\}$, where

$$\begin{aligned} Z_\upsilon(\boldsymbol{\lambda}) &= \exp\left\{\mathbb{E}_\upsilon[\langle\boldsymbol{\lambda}, \mathcal{K}\mathcal{W}\rangle] + \frac{1}{2}\mathrm{Var}_\upsilon[\langle\boldsymbol{\lambda}, \mathcal{K}\mathcal{W}\rangle]\right\} \\ &= \exp\left\{\sum_{i\in a} d_i \int_0^1 \boldsymbol{\lambda}^\top(\mathrm{d}t)\boldsymbol{X}_i(t) + \frac{1}{2}\int_0^1 \left(\sum_{i\in a}\int_0^1 K(s,t)\boldsymbol{\lambda}^\top(\mathrm{d}t)\boldsymbol{X}_i(t)\right)^2 \mathrm{d}s\right\}, \end{aligned} \tag{4.6}$$

owing to that $\mathbb{E}_\upsilon[\mathrm{d}\mathcal{W}_i(s)] = 0$, and $\mathrm{Var}_\upsilon[\mathrm{d}\mathcal{W}_i(s)] = \mathrm{d}s$, $i \in a$.

Now we proceed with the problem of finding $\boldsymbol{\lambda}^*(\mathrm{d}t) \in \mathcal{M}_b(C[0,1]^q)$, where $\mathcal{M}_b$ is the class of bounded continuous measures, such that minimizes

$$\begin{aligned} H_\upsilon(\boldsymbol{\lambda}) &= \frac{1}{2}\int_0^1 \left(\sum_{i\in a}\int_0^1 K(s,t)\boldsymbol{\lambda}^\top(\mathrm{d}t)\boldsymbol{X}_i(t)\right)\left(\sum_{i'\in a}\int_0^1 K(s,t')\boldsymbol{\lambda}^\top(\mathrm{d}t')\boldsymbol{X}_{i'}(t')\right)\mathrm{d}s \\ &\quad + \int_0^1 \boldsymbol{\lambda}^\top(\mathrm{d}t)\left(\sum_{i\in a} d_i\boldsymbol{X}_i(t) - N\boldsymbol{\mu}_X(t)\right) \\ &= \frac{1}{2}\sum_{i\in a}\sum_{i'\in a}\int_0^1\int_0^1\int_0^1 K(s,t)K(s,t')\boldsymbol{\lambda}^\top(\mathrm{d}t)\boldsymbol{X}_i(t)\boldsymbol{X}_{i'}^\top(t')\boldsymbol{\lambda}(\mathrm{d}t')\mathrm{d}s \\ &\quad + \int_0^1 \boldsymbol{\lambda}^\top(\mathrm{d}t)\left(\sum_{i\in a} d_i\boldsymbol{X}_i(t) - N\boldsymbol{\mu}_X(t)\right). \end{aligned} \tag{4.7}$$

The corresponding equation for $\boldsymbol{\lambda}^*(\mathrm{d}t)$ that minimizes $H_v(\boldsymbol{\lambda})$ is given by

$$\sum_{i\in a}\sum_{i'\in a}\int_0^1\int_0^1 K(s,t)K(s,t')\boldsymbol{X}_i(t)\boldsymbol{X}_{i'}^\top(t')\boldsymbol{\lambda}^*(\mathrm{d}t')\mathrm{d}s+\sum_{i\in a}d_i\boldsymbol{X}_i(t)=N\boldsymbol{\mu}_X(t), \quad (4.8)$$

which can be rewritten as

$$\sum_{i\in a}\left[\int_0^1 K(s,t)\left(\sum_{i'\in a}\int_0^1 K(s,t')\boldsymbol{X}_{i'}^\top(t')\boldsymbol{\lambda}^*(\mathrm{d}t')\right)\mathrm{d}s+d_i\right]\boldsymbol{X}_i(t)=N\boldsymbol{\mu}_X(t),$$

obtaining, by the moment calibration constraint (4.2), the Lemma's result. ■

Proof of Lemma 4.2. For each $i\in a$, define a random variable $m_i\left((a,b]\right)$ for $(a,b]\subset[0,1]$,

$$m_i\left((a,b]\right)\triangleq\mathcal{W}_i(b)-\mathcal{W}_i(a)=\sum_{k=N(a)+1}^{N(b)}\xi_{ik}.$$

By the Lévy-Khintchine formula for Lévy processes, the moment generating function of the n-dimensional compound Poisson process $\boldsymbol{\mathcal{W}}(s)$ is given by

$$\mathbb{E}_v\left[\exp\left\{\langle\boldsymbol{\alpha},\boldsymbol{\mathcal{W}}(s)\rangle\right\}\right]=\exp\left\{s\gamma\int_{\mathbb{R}^n}\left(e^{\langle\boldsymbol{\alpha},\boldsymbol{\xi}_k\rangle}-1\right)u\left(\mathrm{d}\boldsymbol{\xi}_k\right)\right\},\qquad\boldsymbol{\alpha}\in\mathbb{R}^n,$$

where $\boldsymbol{\xi}_k=(\xi_{1k},\ldots,\xi_{nk})^\top$. This formula can be generalized for a continuous function $g(s)$ from $[0,1]$ to $\mathbb{R}$ and defining $\langle g(s),\mathcal{W}_i\rangle=\int_0^1 g(s)\mathrm{d}\mathcal{W}_i(s)$ for each $i\in a$, which is approximated by $\sum_{j=1}^J g\left(s_{j-1}\right)m_i\left((s_{j-1},s_j]\right)$, with $s_j=j/J$, $j=1,\ldots,J$. Thus, by the independence of $m_i\left((a,b]\right)$, we have that for all $i\in a$

$$\begin{aligned}\mathbb{E}_v\left[\exp\left\{\langle g(s),\mathrm{d}\mathcal{W}_i\rangle\right\}\right]&=\lim_{J\to\infty}\prod_{j=1}^J\mathbb{E}_v\left[\exp\left\{g\left(s_{j-1}\right)m_i\left((s_{j-1},s_j]\right)\right\}\right]\\&=\lim_{J\to\infty}\prod_{j=1}^J\exp\left\{\mathbb{E}_v\left[\exp\left\{g\left(s_{j-1}\right)\xi_i\right\}\right]\right\}\\&=\lim_{J\to\infty}\prod_{j=1}^J\exp\left\{\frac{\gamma}{J}\int_{\mathbb{R}}\left(\exp\left\{g\left(s_{j-1}\right)\xi_i\right\}-1\right)u\left(\mathrm{d}\xi_i\right)\right\}\\&=\exp\left\{\gamma\int_0^1\mathrm{d}s\int_{\mathbb{R}}\left(\exp\left\{g\left(s\right)\xi_i\right\}-1\right)u\left(\mathrm{d}\xi_i\right)\right\}.\end{aligned}\quad(4.9)$$

Now, by the Theorem 4.2, the maximum of the entropy S over the class $\mathcal{V}$ subject to $\mathbb{E}_v\left[\mathcal{K}\mathcal{W}\right]=N\boldsymbol{\mu}_X(t)$ is achieved at $\mathrm{d}\nu^*/\mathrm{d}v=Z_v^{-1}(\boldsymbol{\lambda}^*)\exp\left\{\langle\boldsymbol{\lambda}^*,\mathcal{K}\mathcal{W}\rangle\right\}$ with

$$\begin{aligned}\langle \boldsymbol{\lambda}, \mathcal{K}\mathcal{W}\rangle &= \int_0^1 \boldsymbol{\lambda}^\top(\mathrm{d}t)\int_0^1 \sum_{i\in a} K(s,t)\boldsymbol{X}_i(t)\mathrm{d}\mathcal{W}_i(s) + \int_0^1 \boldsymbol{\lambda}^\top(\mathrm{d}t)\sum_{i\in a} d_i\boldsymbol{X}_i(t)\\ &= \langle g(s), \mathcal{W}_i\rangle + \left\langle \boldsymbol{\lambda}, \sum_{i\in a} d_i\boldsymbol{X}_i(t)\right\rangle,\end{aligned}$$

where $g(s) = \int_0^1 \boldsymbol{\lambda}^\top(\mathrm{d}t)\sum_{i\in a} K(s,t)\boldsymbol{X}_i(t)$.

Therefore,

$$\begin{aligned}Z_v(\boldsymbol{\lambda}) &= \exp\left\{\gamma\int_0^1 \mathrm{d}s\int_{\mathbb{R}} \left(\exp\left\{g\left(s\right)\xi_i\right\} - 1\right)u\left(\mathrm{d}\xi_i\right)\right\}\exp\left\{\left\langle \boldsymbol{\lambda}, \sum_{i\in a} d_i\boldsymbol{X}_i(t)\right\rangle\right\}\\ &= \exp\left\{\gamma\int_0^1 \mathrm{d}s\int_{\mathbb{R}} \left(\exp\left\{g\left(s\right)\xi_i\right\} - 1\right)u\left(\mathrm{d}\xi_i\right) + \left\langle \boldsymbol{\lambda}, \sum_{i\in a} d_i\boldsymbol{X}_i(t)\right\rangle\right\}\end{aligned} \tag{4.10}$$

Finally, as in the proof of Lemma 4.1, the problem is concentrated to find $\boldsymbol{\lambda}^*(t)$ such that minimizes

$$\begin{aligned}H_v(\boldsymbol{\lambda}) = \gamma\int_0^1 \mathrm{d}s\int_{\mathbb{R}} &\left(\exp\left\{\int_0^1 \boldsymbol{\lambda}^\top(\mathrm{d}t)\sum_{i\in a} K(s,t)\xi_i\boldsymbol{X}_i(t)\right\} - 1\right)u\left(\mathrm{d}\xi_i\right)\\ &+ \int_0^1 \boldsymbol{\lambda}^\top(\mathrm{d}t)\left(\sum_{i\in a} d_i\boldsymbol{X}_i(t) - N\boldsymbol{\mu}_X(t)\right).\end{aligned} \tag{4.11}$$

The corresponding equation for $\boldsymbol{\lambda}^*(\mathrm{d}t)$ that minimizes $H_v(\boldsymbol{\lambda})$ is given by

$$\begin{aligned}\sum_{i\in a}\left[\int_0^1 K(s,t)\left(\int_{\mathbb{R}} \xi_i\exp\left\{\sum_{i\in a}\int_0^1 K(s,t)\xi_i\boldsymbol{X}_i^\top(t)\boldsymbol{\lambda}^*(\mathrm{d}t)\right\}u\left(\mathrm{d}\xi_i\right)\right)\mathrm{d}s + d_i\right]\\ \times\boldsymbol{X}_i(t) = N\boldsymbol{\mu}_X(t),\end{aligned} \tag{4.12}$$

obtaining, by the moment calibration constraint (4.2), the Lemma's result. ■

References

[1] M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[2] B. Arnold, N. Balakrishnan, and H. Nagaraja. *A First Course in Order Statistics*, volume 54. Classics in Applied Mathematics, SIAM, 2008. Philadelphia.

[3] A. Arribas-Gil and J. Romo. Robust depth-based estimation in the time warping model. *Biostatistics*, 13:398–414, 2012.

[4] M. Balasubramanian and E. L. Schwartz. The isomap algorithm and topological stability. *Science*, 295, 2002.

[5] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15:1373–1396, 2003.

[6] M. Bernstein, V. De Silva, J. C. Langford, and J. B. Tenenbaum. Graph approximations to geodesics on embedded manifolds. Technical report, Stanford University, 2000. Available at http://isomap.stanford.edu/BdSLT.pdf.

[7] R. Bhattacharya and V. Patrangenaru. Large sample theory of intrinsic and extrinsic sample means on manifolds. I. *The Annals of Statistics*, 31:1–29, 2003.

[8] R. Bhattacharya and V. Patrangenaru. Large sample theory of intrinsic and extrinsic sample means on manifolds. II. *The Annals of Statistics*, 33: 1225–1259, 2005.

[9] J. Bigot and S. Gadat. A deconvolution approach to estimation of a common shape in a shifted curves model. *The Annals of Statistics*, 38:2422–2464, 2010.

[10] J. Bigot, J.-M. Loubes, and M. Vimond. Semiparametric estimation of shifts on compact Lie groups for image registration. *Probability Theory and Related Fields*, 152:425–473, 2012.

[11] E. Boissard, T. Le Gouic, and J.-M. Loubes. Distribution's template estimate with Wasserstein metrics. 2011. ArXiv:1111.5927. Available at http://arxiv.org/pdf/1111.5927v1.pdf.

[12] B. M. Bolstad, R. A. Irizarry, M. Åstrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

[13] D. Bosq. *Linear Processes in Function Spaces: Theory and Applications.* Number 149 in Lecture Notes in Statistics. Springer-Verlag, 2000.

[14] D. Bosq and D. Blanke. *Inference and Prediction in Large Dimensions.* Wiley Series in Probability and Statistics. Wiley, 2007.

[15] J. Breidt, F and J. D. Opsomer. Local polynomial regression estimators in survey sampling. *The Annals of Statistics*, 28:1026–1053, 2000.

[16] H. Cardot and E. Josserand. Horvitz-Thompson estimators for functional data: asymptotic confidence bands and optimal allocation for stratified sampling. *Biometrika*, 98:107–118, 2011.

[17] H. Cardot, D. Degras, and E. Josserand. Confidence bands for Horvitz-Thompson estimators using sampled noisy functional data. 2011. ArXiv:1105.2135v1. Available at http://arxiv.org/abs/1105.2135v1.

[18] I. Castillo and J.-M. Loubes. Estimation of the distribution of random shifts deformation. *Mathematical Methods of Statistics*, 18:21–42, 2009.

[19] L. Cayton. Algorithms for manifold learning. Technical report, University of California, San Diego, 2005.

[20] D. Chen and H. G. Müller. Nonlinear manifold representations for functional data. *The Annals of Statistics*, 40:1–29, 2012.

[21] R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21:5–30, 2006.

[22] T. Cox and M. Cox. *Multidimensional Scaling*, volume 88 of *Monographs on Statistics and Applied Probability.* Chapman & Hall, 2nd edition, 2000.

[23] G. Csárdi and T. Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. http://igraph.sf.net.

[24] I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3:146–158, 1975.

[25] I. Csiszár. Sanov property, generalized I-projection and a conditional limit theorem. *The Annals of Probability*, 12:768–793, 1984.

[26] D. Dacunha-Castelle and F. Gamboa. Maximum d'entropie et problème des moments. *Annales de l'Institut Henri Poincaré. Series B, Probabilités et Statistiques*, 26:567–596, 1990.

[27] F. N. David and N. L. Johnson. Statistical treatment of censored data part I. Fundamental formulae. *Biometrika*, 41(1/2):228–240, 1954.

[28] H. A. David and H. N. Nagaraja. *Order Statistics.* Wiley, 3rd edition, 2003. New Jersey.

[29] V. de Silva and J. B. Tenenbaum. Unsupervised learning of curved manifolds. In D. Denison, M. Hansen, C. Holmes, B. Mallick, and B. Yu, editors, *Nonlinear estimation and classification*, volume 171 of *Lecture Notes in Statistics*, pages 453–465. Springer-Verlag, New York, 2003.

[30] J. C. Deville and C. E. Särndal. Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87:376–382, 1992.

[31] C. Dimeglio, S. Gallón, J.-M. Loubes, and E. Maza. A robust algorithm for template curve estimation based on manifold embedding. *HAL: hal-00786338*, 2013. http://hal.archives-ouvertes.fr/hal-00786338, Submitted.

[32] M. P. do Carmo. *Riemannian Geometry.* Mathematics: Theory & Applications. Birkhäuser Boston Inc., Boston, MA, 1992. Translated from the second Portuguese edition by Francis Flaherty.

[33] D. Donoho and C. Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 100:5591–5596, 2003.

[34] S. Dudoit and Y. H. Yang. Bioconductor R packages for exploratory analysis and normalization of cDNA microarray data. In G. Parmigiani, E. S. Garrett, R. A. Irizarry, and S. L. Zeger, editors, *The Analysis of Gene Expression Data: Methods and Software*, Statistics for Biology and Health, pages 73–101, New York, 2003. Springer.

[35] J. F. Dupuy, J. M. Loubes, and E. Maza. Non parametric estimation of the estructural expectation of a stochastic increasing function. *Statistics and Computing*, 21:121–136, 2011.

[36] H. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*, volume 375 of *Mathematics and Its Applications.* Kluwer Academic Publishers, Dordrecht, 2000.

[37] F. Ferraty and P. Vieu. *Nonparametric Functional Data Analysis: Theory and Practice.* Springer Series in Statistics. Springer–Verlag, 2006.

[38] W. A. Fuller. *Sampling Statistics.* Wiley Series in Survey Methodology. Wiley, 2009.

[39] S. Gallón, J.-M. Loubes, and E. Maza. Statistical properties of the quantile normalization method for density curve alignment. *Mathematical Biosciences*, 242:129–142, 2013.

[40] F. Gamboa. *Méthode du maximum d'entropie sur la moyenne et applications.* PhD thesis, Université de Paris-Sud, Orsay, 1989.

[41] F. Gamboa and E. Gassiat. Bayesian methods and maximum entropy for ill-posed inverse problems. *The Annals of Statistics*, 25:328–350, 1997.

[42] F. Gamboa, J. Loubes, and E. Maza. Semi-parametric estimation of shits. *Electronic Journal of Statistics*, 1:616–640, 2007.

[43] F. Gamboa, J.-M. Loubes, and P. Rochet. Maximum entropy estimation for survey sampling. *Journal of Statistical Planning and Inference*, 141:305–317, 2011.

[44] T. Gasser and A. Kneip. Searching for structure in curve sample. *Journal of the American Statistical Association*, 90:1179–1188, 1995.

[45] D. Gervini and T. Gasser. Nonparametric maximum likelihood estimation of the structural mean of a sample of curves. *Biometrika*, 95:801–820, 2005.

[46] A. Golan and H. Gzyl. An entropic estimator for linear inverse problems. *Entropy*, 14:892–923, 2012.

[47] N. Gozlan. *Principe conditionnel de Gibbs pour des contraintes fines approchées et Inégalités de Transport.* PhD thesis, Université Paris X, Nanterre, France, 2005.

[48] H. Gzyl. Maxentropic reconstruction of some probability distributions. *Studies in Applied Mathematics*, 105:235–243, 2000.

[49] H. Gzyl and Y. Velásquez. *Linear Inverse Problems: the Maximum Entropy Connection*, volume 83 of *Series on Advances in Mathematics for Applied Sciences.* World Scientific, 2011.

[50] T. Harms and P. Duchesne. On calibration estimation for quantiles. *Survey Methodology*, 32:37–52, 2006.

[51] L. Horváth and P. Kokoszka. *Inference for Functional Data with Applications.* Springer Series in Statistics. Springer, 2012.

[52] D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685, 1953.

[53] R. Irizarry, B. Hobbs, F. Collin, Y. Beazer-Barclay, K. Antonellis, U. Scherf, and T. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003.

[54] A. Izenman. *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning.* Springer, 2008.

[55] G. James. Curve alignment by moments. *The Annals of Aplied Statistics*, 1: 480–501, 2007.

[56] J. Jost. *Riemannian Geometry and Geometric Analysis.* Springer, 2011.

[57] A. Kneip and T. Gasser. Convergence and consistency results for self-modelling regression. *The Annals of Statistics*, 16:82–112, 1988.

[58] A. Kneip and T. Gasser. Statistical tools to analyze data representing a sample of curves. *The Annals of Statistics*, 20:1266–1305, 1992.

[59] A. Kneip and J. Ramsay. Combining registration and fitting for funstional models. *Journal of the American Statistical Association*, 103:1155–1165, 2008.

[60] A. Kneip and J. Utikal. Inference for density families using functional principal component analysis. *Journal of the American Statistical Association*, 96:519–542, 2001.

[61] A. Kneip, X. Li, X. MacGibbon, and J. Ramsay. Curve registration by local regression. *Canadian Journal of Statistics*, 28:19–29, 2000.

[62] R. Koenker. The median is the message: Toward the Fréchet mean. *Journal de la Société Française de Statistiques*, 147:61–64, 2006.

[63] M. Kovaĉević. Calibration estimation of cumulative distribution and quantile functions from survey data. *Proceedings of the Survey Methods Section, Statistical Society of Canada*, 47:139–144, 1997.

[64] R. Kress. *Linear Integral Equations*, volume 82 of *Applied Mathematical Sciences.* Springer-Verlag, New York, 2nd edition, 1999.

[65] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete 3. Folge. A Series of Modern Surveys in Mathematics.* Springer, 1991. Berlin.

[66] J. A. Lee and M. Verleysen. *Nonlinear Dimensionality Reduction.* Information Science and Statistics. Springer, New York, 2007.

[67] X. Liu and H. G. Müller. Functional convex averaging and synchronization for time-warped random curves. *Journal of the American Statistical Association*, 99:687–699, 2004.

[68] S. López-Pintado and J. Romo. On the concept of depth for functional data. *Journal of the American Statistical Association*, 104:718–734, 2009.

[69] P. Maréchal. On the principle of maximum entropy as a methodology for solving linear inverse problems. In B. Grigelionis, V. P. J. Kubilius, H. Pragarauskas, R. Ruszkis, and V. Statulevicius, editors, *Probability Theory and Mathematical Statistics*, Proceedings of the Seventh Vilnius Conference, pages 481–492, Vilnius, Lithuania, 1999. VPS/TEV.

[70] E. Maza. Estimation de l'espérance structurelle d'une fonction aléatoire. *Comptes Rendus Mathématique. Académie des Sciences. Paris*, 343, 2006.

[71] A. Mohammad-Djafari. A comparison of two approaches: Maximum entropy on the mean (MEM) and Bayesian estimation (BAYES) for inverse problems. In M. Sears, V. Nedeljkovic, N. E. Pendock, and S. Sibisi, editors, *Maximum Entropy and Bayesian Methods*, pages 77–91, Berg-en-Dal, South Africa, 1996. Kluwer Academic Publishers.

[72] G. E. Montanari and M. G. Ranalli. Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association*, 100:1429–1442, 2005.

[73] J. Navaza. On the maximum entropy estimate of electron density function. *Acta Crystallographica*, A41:232–244, 1985.

[74] J. Navaza. The use of non-local constraints in maximum-entropy electron density reconstruction. *Acta Crystallographica*, A42:212–223, 1986.

[75] X. Pennec. Intrinsic statistics on Riemannian manifolds: basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25: 127–154, 2006.

[76] H. G. Ramsay, J. and S. Graves. *Functional Data Analysis with R and MATLAB.* Springer, 2009.

[77] J. O. Ramsay and X. Li. Curve registration. *Journal of the Royal Statistical Society. Series B*, 60:351–363, 1998.

[78] J. O. Ramsay and B. W. Silverman. *Applied Functional Data Analysis: Methods and Case Studies.* Springer-Verlag, 2002.

[79] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis.* Springer Series in Statistics. Springer-Verlag, 2nd edition, 2005.

[80] J. N. K. Rao, J. G. Kovar, and H. J. Mantel. On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika*, 77:365–375, 1990.

[81] B. Rønn. Nonparametric maximum likelihood estimation for shifted curves. *Journal of the Royal Statistical Society. Series B*, 63:243–259, 2001.

[82] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[83] M. Rueda, S. Martínez, H. Martínez, and A. Arcos. Estimation of the distribution function with calibration methods. *Journal of Statistical Planning and Inference*, 137:435–448, 2007.

[84] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26:43–49, 1978.

[85] C. E. Särndal. The calibration approach in survey theory and practice. *Survey Methodology*, 33:99–119, 2007.

[86] C. E. Särndal, B. Swensson, and J. Wretman. *Model Assisted Survey Sampling.* Springer Series in Statistics. Springer, 1992.

[87] R. Serfling. *Approximation Theorems of Mathematical Statistics.* Wiley, 1980. New York.

[88] B. W. Silverman. Incorporating parametric effects into functional principal components analysis. *Journal of Royal Statistical Society. Series B*, 57:673–689, 1995.

[89] S. Singh. Generalized calibration approach for estimating variance in survey sampling. *Annals of the Institute of Statistical Mathematics*, 53:404–417, 2001.

[90] C. G. Small. *The Statistical Theory of Shape.* Springer Series in Statistics. Springer, New York, 1996.

[91] G. K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments.

[92] G. K. Smyth. `limma`: linear models for microarray data. In R. Gentleman, V. Carey, W. Huber, R. Irizarry, and S. Dudoit, editors, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pages 397–420, New York, 2005. Springer.

[93] G. K. Smyth and T. P. Speed. Normalization of cDNA microarray data. *Methods*, 31(4):265–273, 2003.

[94] G. K. Smyth, M. Ritchie, N. Thorne, J. Wettenhall, and W. Shi. *limma: Linear Models for Microarray Data User's Guide.* 2012. Software manual available from http://www.bioconductor.org/.

[95] R. Tang and H.-G. Müller. Pairwise curve synchronization for functional data. *Biometrika*, 95:875–889, 2008.

[96] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[97] A. Théberge. Extensions of calibration estimators in survey sampling. *Journal of the American Statistical Association*, 94:635–644, 1999.

[98] T. Trigano, U. Isserles, and Y. Ritov. Semiparametric curve alignment and shift density estimation for biological data. *IEEE Transactions on Signal Processing*, 59:1970–1984, 2011.

[99] A. Tsybakov. *Introduction to Nonparametric Estimation.* Springer Series in Statistics. Springer, 2008.

[100] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes.* Springer, 1996. New York.

[101] S. Vantini. On the definition of phase and amplitude variability in functional data analysis. *Test*, 21:676–696, 2012.

[102] R. Varadhan. R-package BB: Solving and optimizing large-scale nonlinear systems. 2012. Available at http://cran.r-project.org.

[103] R. Varadhan and P. Gilbert. BB: An R package for solving a large system of nonlinear equations and for optimizing a high-dimensional nonlinear objective function. *J. Stat. Softw.*, 32, 2009. Available at http://www.jstatsoft.org/v32/i04/.

[104] C. Villani. *Optimal Transport: Old and New*, volume 338 of *A Series of Comprehensive Studies in Mathematics.* Springer-Verlag, Berlin, 2009.

[105] H. Wang, N. Schauer, B. Usadel, P. Frasse, M. Zouine, M. Hernould, A. Latché, J.-C. Pech, A. Fernie, and B. M. Regulatory features underlying pollination-dependent and -independent tomato fruit set revealed by transcript and primary metabolite profiling. *The Plant Cell*, 21(5):1428–1452, 2009.

[106] K. Wang and T. Gasser. Alignment of curves by dynamic time warping. *The Annals of Statistics*, 25:1251–1276, 1997.

[107] K. Wang and T. Gasser. Synchronizing sample curves nonparametrically. *The Annals of Statistics*, 27:439–460, 1999.

[108] D. B. West. *Introduction to Graph Theory.* Pearson Education, Inc., Delhi, second edition, 2001.

[109] C. Wu. Optimal calibration estimators in survey sampling. *Biometrika*, 90: 937–951, 2003.

[110] C. Wu and R. R. Sitter. A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96:185–193, 2001.

[111] L. Yang. *Médianes de mesures de probabilité dans les variétés riemanniennes et applications à la détection de cibles radar*. PhD thesis, Université de Poitiers, Poitiers, France, 2012.

[112] Y. H. Yang and A. C. Paquet. Preprocessing two-color spotted arrays. In R. Gentleman, V. Carey, W. Huber, R. Irizarry, and S. Dudoit, editors, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pages 49–69, New York, 2005. Springer.

[113] Y. H. Yang and T. P. Speed. Design and analysis of comparative microarray experiments. In T. P. Speed, editor, *Statistical Analysis of Gene Expression Microarray Data*, pages 35–91, Boca Raton, 2003. Chapman & Hall/CRC.

[114] Y. H. Yang and N. P. Thorne. Normalization for two-color cDNA microarray data. In D. R. Goldstein, editor, *Science and Statistics: A Festschrift for Terry Speed*, volume 40, pages 403–418, New York, 2003. IMS Lecture Notes.

[115] J.-T. Zhang and J. Chen. Statistical inferences for functional data. *The Annals of Statistics*, 35:1052–1079, 2007.

[116] Z. Zhang and H. Zha. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM Journal of Scientific Computing*, 26:313–338, 2004.

Résumé

L'une des principales difficultés de l'analyse des données fonctionnelles consiste à extraire un motif commun qui synthétise l'information contenue par toutes les fonctions de l'échantillon. Le Chapitre 2 examine le problème d'identification d'une fonction qui représente le motif commun en supposant que les données appartiennent à une variété ou en sont suffisamment proches, d'une variété non linéaire de basse dimension intrinsèque munie d'une structure géométrique inconnue et incluse dans un espace de grande dimension. Sous cette hypothèse, un approximation de la distance géodésique est proposé basé sur une version modifiée de l'algorithme Isomap. Cette approximation est utilisée pour calculer la fonction médiane empirique de Fréchet correspondante. Cela fournit un estimateur intrinsèque robuste de la forme commune.

Le Chapitre 3 étudie les propriétés asymptotiques de la méthode de normalisation quantile développée par Bolstad, et al. (2003) qui est devenue l'une des méthodes les plus populaires pour aligner des courbes de densité en analyse de données de microarrays en bioinformatique. Les propriétés sont démontrées considérant la méthode comme un cas particulier de la procédure de la moyenne structurelle pour l'alignement des courbes proposée par Dupuy, Loubes and Maza (2011). Toutefois, la méthode échoue dans certains cas. Ainsi, nous proposons une nouvelle méthode, pour faire face à ce problème. Cette méthode utilise l'algorithme développée dans le Chapitre 2.

Dans le Chapitre 4, nous étendons le problème d'estimation de calage pour la moyenne d'une population finie de la variable de sondage dans un cadre de données fonctionnelles. Nous considérons le problème de l'estimation des poids de sondage fonctionnel à travers le principe du maximum d'entropie sur la moyenne -MEM-. En particulier, l'estimation par calage est considérée comme un problème inverse linéaire de dimension infinie suivant la structure de l'approche du MEM. Nous donnons un résultat précis d'estimation des poids de calage fonctionnels pour deux types de mesures aléatoires a priori: la measure Gaussienne centrée et la measure de Poisson généralisée.

Abstract

One of the main difficulties in functional data analysis is the extraction of a meaningful common pattern that summarizes the information conveyed by all functions in the sample. The problem of finding a meaningful template function that represents this pattern is considered in Chapter 2 assuming that the functional data lie on an intrinsically low-dimensional smooth manifold with an unknown underlying geometric structure embedding in a high-dimensional space. Under this setting, an approximation of the geodesic distance is developed based on a robust version of the Isomap algorithm. This approximation is used to compute the corresponding empirical Fréchet median function, which provides a robust intrinsic estimator of the template.

The Chapter 3 investigates the asymptotic properties of the quantile normalization method by Bolstad, et al. (2003) which is one of the most popular methods to align density curves in microarray data analysis. The properties are proved by considering the method as a particular case of the structural mean curve alignment procedure by Dupuy, Loubes and Maza (2011). However, the method fails in some case of mixtures, and a new methodology to cope with this issue is proposed via the algorithm developed in Chapter 2.

Finally, the problem of calibration estimation for the finite population mean of a survey variable under a functional data framework is studied in Chapter 4. The functional calibration sampling weights of the estimator are obtained by matching the calibration estimation problem with the maximum entropy on the mean -MEM- principle. In particular, the calibration estimation is viewed as an infinite-dimensional linear inverse problem following the structure of the MEM approach. A precise theoretical setting is given and the estimation of functional calibration weights assuming, as prior measures, the centered Gaussian and compound Poisson random measures is carried out.