Learning sparse spline-based shape models

Laure Amate

UNIVERSITÉ DE NICE - SOPHIA ANTIPOLIS

Ecole doctorale STIC

Sciences et Technologies de l'Information et de la Communication

THÈSE

pour obtenir le titre de

Docteur en Sciences

de l'Université de Nice - Sophia Antipolis

Mention : Automatique, Traitement du Signal et des Images

présentée et soutenue par

Laure AMATE

Apprentissage de modèles de formes parcimonieux basés sur les représentations splines

---

Thèse dirigée par: Pr. Tarek HAMEL et Dr. Maria-João RENDAS

soutenue le 10 décembre 2009

M. M. Figueiredo	Professeur	IST (Lisbonne, Portugal)	Rapporteur
M. R. Garello	Professeur	ENST Bretagne (Brest, France)	Rapporteur
M. J. Blanc-Talon		DGA/MRIS (France)	Membre du jury
M. I. Jermyn	Chercheur	INRIA (Sophia Antipolis, France)	Membre du jury
M. C. Giurcaneanu	Chercheur	TUT (Tampere, Finlande)	Membre invité
M. T. Hamel	Professeur	UNS (Nice-Sophia Antipolis, France)	Directeur de thèse
Mme. M.J. Rendas	Chercheur	I3S (Sophia Antipolis, France)	Encadrant de thèse

Thanks

The work presented in this manuscript would not have been the same without the supervision of João Rendas. She is the first person I would like to thank for believing in me, supporting me and introducing me in the research world. I also address many thanks to the members of my jury for agreeing to review my work and for the kindness and the relevance of their remarks: Prof. Mario Figueiredo, Prof. René Garello, Mr. Ian Jermyn, Mr. Jacques Blanc-Talon and Mr. Ciprian Gircaneanu.

The achievement of this thesis would not have been possible without the support of Laboratoire I3S and I am particularly grateful to Viviane who always has the answers to my countless questions. My everyday (and night) work has been brightened by all phd students, internship students and permanent researchers in the lab who provide a pleasant environment.

Of course the major part of my thanks are going to my family and friends without whom nothing would have been possible. My parents and my sister were, are and will be, always involved in my success, and, for a few years now, Ronald is a constant support and always gives good advice. I will not name all my friends but they will recognize themselves, those from the South of France, those from the lab and those knowing me for a longer time.

I also would like to thank the first person who believed in me and introduced me to João, Mr. Frédéric Dabe from GESMA, and the French Department of Defense (DGA) for their financial support.

Finally I just give you my personal leitmotif:

Start by doing what is necessary then do what is possible and you will realize the impossible.

Saint François d'Assises
Laure Amate
February 2nd, 2010.

Contents

List of Figures

List of Tables

GLOSSARY

AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
BMS	Bayesian Model Selection
EM	Expectation Maximization
LS	Least Squares
MAP	Maximum A Posteriori
MC	Monte Carlo
MCEM	Monte Carlo Expectation Maximization
MCMC	Markov Chain Monte-Carlo
MDL	Minimum Description Length
MEM	Metropolis Expectation Maximization
MH	Metropolis Hastings
ML	Maximum Likelihood
pp	piecewise polynomial
RJMCMC	Reversible Jump Markov Chain Monte-Carlo
SA	Simulated Annealing
SAEM	Stochastic Approximation Expectation Maximization
SEM	Stochastic Expectation Maximization

Chapter 1

Introduction

In many contexts it is important to be able to find compact representations of the collective morphological properties of a set of objects. This is the case, for instance, of autonomous robotic platforms operating in natural environments that must use the perceptual properties of the objects present in their workspace to correctly plan their motion, navigate and execute their mission. This thesis is a contribution to the definition of formalisms and methods for automatic identification of such models. The shapes we want to characterize are closed curves in an ambient Euclidean space, corresponding to contours of objects detected in the scene.

We begin with the formal definition of the notion of shape as classes of equivalence with respect to groups of basic geometric operators, introducing two distinct approaches that have been used in the literature: i) the discrete theory of shape introduced in the seminal work of Kendall [30] and ii) the definition of space of continuous closed curves recently proposed by Klassen and Srivastava [31]. We briefly characterize their topology and geometry, and point out their relative advantages and limitations from the point of view of the specific goals of our work. The discrete theory, admitting the existence of a finite number of recognizable landmarks, provides in an obvious manner a compact representation. Its strong advantage is that the shape space is finite dimensional, its topology is known, and use of complex algebra enables analytical treatment in the important case of two dimensional shapes, for which shape space is a smooth manifold embedded in an ambient Euclidean space. Most importantly, the definition of probability distributions corresponding to a set of observed shapes can be addressed, by considering the local mappings of neighborhoods of this shape space to their tangent vector spaces. However, its use can lead to unstable results when the definition of these landmarks needs to be automated, as we consider here. The continuous theory of shapes, presented for closed curves in $\mathbb{R}^2$ in [31], provides a more fundamental approach, but leads to shape spaces of infinite dimension, lacking the parsimony of the discrete representation.

This space of continuous shapes is an infinite dimensional Riemannian manifold whose finite dimensional orthogonal spaces are known, and study of its topology can be assessed using differential geometry tools. It presents two main drawbacks with respect to our goals: it is subject to numerical instabilities when building discrete approximations of the relevant elements and operators, and, more fundamentally, we are not aware of attempts to define probability distributions in this infinite dimensional shape space.

Based on the remarks above, we combine in our work the advantages of both approaches. We base shape description on spline representations of curves: piece-wise continuous polynomials defined by sets of knots (limits of the polynomial pieces) and control points (the coefficients of the polynomials). Our shapes are thus continuous, at the same time benefiting from the flexibility (we consider a varying number of free knots, and thus our model is dense in the set of continuous curves) and the sparsity of the spline representation. The compact model of the properties of a set of shapes that we learn from a given data are parametric probability distributions defined over the space of parameters of the spline representation, inducing in this manner a probability distribution in the dense subset of all continuous closed curves. In the rest of the manuscript, we first study the simpler problem of fitting free-knots splines of varying complexity to a single observed curve, to address, in a subsequent step, the estimation of probability distributions over the spline parameters that will represent a set of curves.

Real observations are affected by noise, and exact representation of all samples of a given curve would lead to overly complex models. One must then find a balance between the parsimony of the representation and its fidelity to the observations, looking for minimal spline representations with good fitting properties. This trade-off is a well known characteristic of model identification using nested families of increasing dimension, and is an active research topic addressed by many authors. After presenting an overview of methods previously proposed in the literature, pointing out some inconsistencies in previously published references, we single out a two-step approach which is formally sound and matches our specific requirements. It splits the identification, simulating a reversible jump Markov chain to select the complexity of the model (number of knots in the case of splines) followed by a simulated annealing algorithm to estimate its parameters. We investigate the link between Kendall's shape space and spline representations when we take the spline control points as landmarks. Two important questions arise: in which measure our approach provides, as a by-product, a method for automatic landmark identification, and, how much discriminative and descriptive power do we gain by using the more complex continuous representation.

After having discussed modeling of a single curve, we consider the more complex problem of modeling a set of objects with similar morphological characteristics. We equate the problem to finding the statistical distribution of the parameters of the spline representation, modeling the knots and

control points as unobserved variables.The identified distribution is the maximizer of a marginal likelihood criterion, and we propose a new Expectation-Maximization algorithm to optimize it. Because we may want to treat a large number of curves observed sequentially, we adapt to our problem an iterative (on-line) version of the EM algorithm recently proposed in the literature, where one data curve is processed at each iteration. For the choice of statistical distributions that we consider, both the expectation and the maximization steps must resort to numerical approximations. We rely on sequential Monte Carlo simulations to compute the expected value at each EM iteration, leading to a stochastic/on-line variant of the EM algorithm that, as far as we know, is implemented here for the first time.

Several illustrating examples are shown throughout the presentation, so that no separate chapter of manuscript is reserved to numerical results. In the last chapter, we recall the main novel aspects and contributions of the thesis, and discuss possible directions for future work. In particular, further work is required concerning the important problem of reducing the numerical complexity of the algorithms presented in the thesis to make them suitable for real robotic applications. More fundamentally, extension of the stochastic/on-line EM method to mixtures of exponentials would enable a joint estimation of all (structural and continuous) model parameters. Finally, we return to the motivating framework of the thesis and conclude with a discussion of how this shape modeling framework can be integrated into navigation and mapping systems of autonomous mobile platforms.

Chapter 2

The theory of shape

2.1 Introduction

Many applications require to distinguish objects with respect to their morphology: color, texture or geometry, for example. In our case, dealing with side-scan sonar images of the ocean bottom, we are interested in characterizing the objects' appearance: their sizes, locations, orientations and shapes. While the first three characteristics are mainly determined by the observer's point of view, the shape of large scale objects lying on the sea floor is fundamentally an intrinsic property of the objects themselves. For this reason, we want to be able to clearly separate the description of shape from other factors, and derive representations allowing the direct comparison of two shapes. Since our observations are uncertain, our aim is not only to identify a single possible instance of a convenient mathematical definition of the shape of an object, but rather to do this statistically, i.e., to be able to define probability measures over the set of possible shapes. This is particularly important, for compactness and robustness reasons, if we are interested in collectively describing the shape of a collection of observed objects.

We thus focus our attention on the definition of shape given by Kendall [29] as "what is left when the differences which can be attributed to translations, rotations, and dilatations have been quotiented out". We distinguish two major approaches that share this point of view that shapes are orbits under the action of certain classes (groups) of operators. The first one is the original theory developed by Kendall in [30] and represents an object as a finite set of landmarks: for example, the vertices of a polygon or the characteristic points on monkey skulls. We refer to this approach as the *discrete theory* of shape, and present its basic definitions in the first section of this chapter. In the second section, we present an alternative approach, in which shapes are infinite-dimensional continuous curves describing object contours. This theory has been initiated by Klassen and Srivastava in [31]. In contrast with the discrete theory, we refer to this theory as the *continuous*

theory of shapes. We outline their approach and focus on the special case of two dimensional closed curves which is the relevant case in the context of this thesis.

2.2 Discrete shape space (Kendall)

Kendall's theory, developed in [29], relies on the postulate that any change in position, size or orientation of an object does not change the global perception we have about its shape. An object is represented by a set of distinguishable points called landmarks that characterize it: the vertices for polygonal shapes or the tips of the fingers for human hands for example. This set is called the configuration of an object and noted Z. It is represented by the configuration matrix $Z \in \mathbb{R}^{m\times k}$ whose columns are the k landmarks $z_i = \in (z_{i1}\, z_{i2} \cdots z_{im})^T \mathbb{R}^m$, where $(u)^T$ denotes the transpose of u, and $Z = (z_1\, z_2 \cdots z_k)$.

The shape associated to Z is then the orbit of Z with respect to the transformations that leave unchanged its shape: translation, rotation and scaling. Let $[Z]$ denote the shape of Z.

Let $SO(m)$ denotes the special orthogonal group in $\mathbb{R}^m$, i.e. the group of rotation matrices.

Definition (2.1): Shape equality

Let $[Z_1]$ and $[Z_2]$ be the shapes associated to two configurations Z_1 and Z_2. We say that $[Z_1]$ and $[Z_2]$ are identical and write: $[Z_1] \overset{s}{=} [Z_2]$ if and only if $\exists\ \alpha \in \mathbb{R}^*_+$, $R \in SO(m)$ and $\mathbf{t} \in \mathbb{R}^m$ that map Z_2 into Z_1:

$$[Z_1] \overset{s}{=} [Z_2] \Leftrightarrow Z_1 = \alpha R Z_2 + \mathbf{t}\ . \tag{2.1}$$

It can be shown that $\overset{s}{=}$ is an equivalence relation (reflexive, symmetric and transitive) that partitions the configuration space into equivalence classes.

2.2.1 The preshape space

For notational simplicity, let introduce the $\mathrm{Vec}\,(\cdot)$ operator:

Definition (2.2): $\mathrm{Vec}\,(A)$

Let $A = [A_{ij}]$, $i = 1 \cdots m$, $j = 1 \cdots k$ be a $\mathbb{R}^{m\times k}$ matrix, $\mathrm{Vec}\,(A) \in \mathbb{R}^{mk}$ is the column vector that stacks the columns of A one under the other:

$$[\mathrm{Vec}\,(A)]_{(j-1)m+i} = A_{ij}\ .$$

The study of the set of shapes, or the equivalence classes, is usually not done directly in the configuration space $\mathbb{R}^{m\times k}$. We present below definitions that are classically used to analyze its geometry.

Definition (2.3): Centered representation

Let $Z \in \mathbb{R}^{m\times k}$ be a configuration. We denote by Z_c its centered representation:

$$Z_c \triangleq Z - z^c \mathbf{1}^k \ , \tag{2.2}$$

where $\mathbf{1}^k$ is the k-dimensional row vector with unit components, and $z^c = (z_1^c\, z_2^c \cdots z_m^c)^T \in \mathbb{R}^m$ is the barycenter of Z: $z^c = \dfrac{1}{k}\displaystyle\sum_{i=1}^{k} z_i$.

The application $Z \to Z_c$ that maps each point of $\mathbb{R}^{m\times k}$ in the orthogonal subspace of $\mathbf{1}^k$ is invariant with respect to translation.

Definition (2.4): Frobenius norm

We denote by $||Z||$ the Frobenius norm of $Z \in \mathbb{R}^{m\times k}$ or equivalently, the Euclidean norm in $\mathbb{R}^{mk}$ of $\mathrm{Vec}\,(Z)$:

$$||Z|| \triangleq \sqrt{\mathrm{Tr}\,(Z^T Z)} = \sqrt{(\mathrm{Vec}\,(Z))^T\, \mathrm{Vec}\,(Z)} \ . \tag{2.3}$$

Definition (2.5): Preshape

Let $Z \in \mathbb{R}^{m\times k}$ be a configuration. The preshape of Z, denoted by $\tilde{Z}$, is the centered and normalized version of Z:

$$\tilde{Z} \triangleq \frac{Z_c}{||Z_c||} \ . \tag{2.4}$$

The application $Z_c \to \tilde{Z}$ is invariant to scaling. Figure 2.1 illustrates the computation of the preshape of a planar object.

Definition (2.6): Preshape space $\mathcal{S}_m^k$

We denote by $\mathcal{S}_m^k$ the preshape space, the set of all preshapes obtained from objects with k landmarks in $\mathbb{R}^m$:

$$\mathcal{S}_m^k \triangleq \left\{ Z \in \mathbb{R}^{k\times m}\,;\, z_c = \left(\mathbf{0}^k\right)^T \text{ and } ||Z|| = 1 \right\} \ , \tag{2.5}$$

where $\mathbf{0}^k$ is the column vector with k zeros.

From this definition, $\mathcal{S}_m^k$ is the hypersphere of unit radius in $\mathbb{R}^{m(k-1)}$. It is a $(m(k-1)-1)$- dimensional non-linear variety. The preshape space is

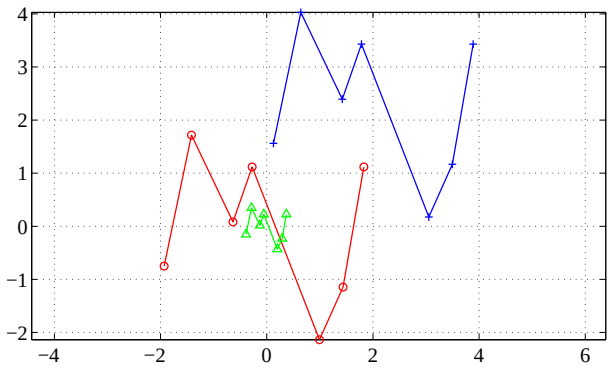

Figure 2.1: Illustration of preshape determination: Z (in blue) is the configuration, Z_c (in red) is the centered configuration and $\tilde{Z}$ (in green) is the preshape.

thus a mathematical object with a simple and well-known structure.

Three distinct distances have been defined in the preshape space, enabling the direct comparison of two preshapes $\tilde{Z}_1$ and $\tilde{Z}_2$ (see Figure 2.2 for a geometric interpretation):

- the scaled distance, noted D_s, which is an extrinsic distance defined outside $\mathcal{S}_m^k$ by scaling one of the preshapes (Definition 2.7),
- two intrinsic distances:
  - the chordal distance D_P (Definition 2.8),
  - the great-circle distance D (Definition 2.9).

Definition (2.7): Scaled distance

The scaled distance D_s is the minimum Euclidean distance between $\tilde{Z}_1$ and the line through the origin containing $\tilde{Z}_2$ (its scaled version):

$$D_s(\tilde{Z}_1, \tilde{Z}_2) = \inf_{\alpha \in \mathbb{R}_*^+} ||\tilde{Z}_2 - \alpha \tilde{Z}_1|| \ , \tag{2.6}$$

with $||\cdot||$ the Frobenius norm (definition 2.4) and $\mathbb{R}_*^+$, the set of non-null positive real numbers.

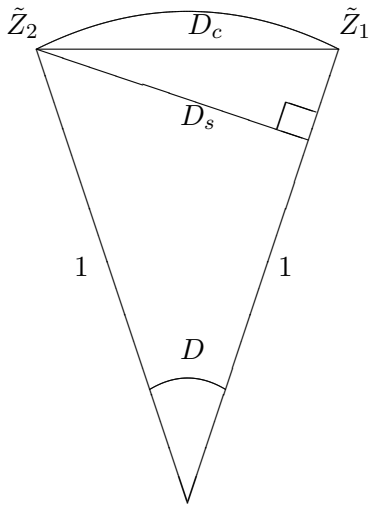

Figure 2.2: Illustration of the scaled distance D_s, the chordal distance D_c and the great-circle distance D, in the preshape space.

Definition (2.8): Chordal distance

The chordal distance D_c is the Euclidean distance between the preshapes $\tilde{Z}_1$ and $\tilde{Z}_2$:

$$D_c(\tilde{Z}_1, \tilde{Z}_2) = ||\tilde{Z}_2 - \tilde{Z}_1|| \ , \tag{2.7}$$

with $||\cdot||$ the Frobenius norm (definition 2.4).

Definition (2.9): Great-circle distance

The great-circle distance D is the length of the geodesic in $\mathcal{S}_m^k$ between $\tilde{Z}_1$ and $\tilde{Z}_2$:

$$D(\tilde{Z}_1, \tilde{Z}_2) = \arg\left(\tilde{Z}_2, \tilde{Z}_1\right) \ , \tag{2.8}$$

where $\arg(\tilde{Z}_2, \tilde{Z}_1)$ denotes the angle between $\tilde{Z}_1$ and $\tilde{Z}_2$.

Note that:

- $D_s \in [0, 1]$ and $D_s = \sin D$,
- $D_c \in [0, 2]$ and $D_c = 2 \sin \frac{D}{2}$.
- $D \in [0, \pi]$ and $D(\tilde{Z}_1, \tilde{Z}_2) = \arccos \mathrm{Tr}\left(\tilde{Z}_1^T \tilde{Z}_2\right)$,

We thus have that geodesics of the preshape spaces $\mathcal{S}_m^k$ are the geodesics of spheres, the great circles.

2.2.2 The shape space

It follows from definitions 2.1 and 2.6 that the shape $[Z]$ of a configuration Z is:

$$[Z] \triangleq \left\{ R\tilde{Z} \ : \ \tilde{Z} \in \mathcal{S}_m^k \, , \, R \in SO(m) \right\} \ . \tag{2.9}$$

Definition (2.10): Shape space Σ_m^k

Let $\mathcal{S}_m^k$ be the preshape space of the configuration space $\mathbb{R}^{m \times k}$. The shape space, noted Σ_m^k, is the set of all shapes of objects of k landmarks in $\mathbb{R}^m$. Σ_m^k is the quotient space of $\mathcal{S}_m^k$:

$$\Sigma_m^k \triangleq \mathcal{S}_m^k / SO(m) \ . \tag{2.10}$$

Construction of the shape space is summarized in Figure 2.3.

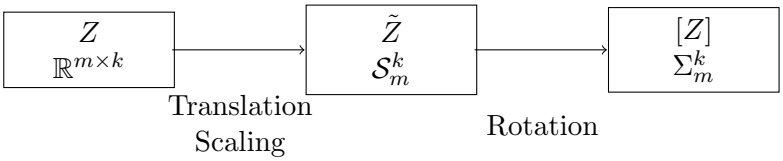

Figure 2.3: Summary of the shape representation process and the corresponding spaces.

The shape space is a $\left(m(k-1) - 1 - \dfrac{m(m-1)}{2} \right)$- dimensional non-linear space [30]. Figure 2.4 illustrates the shape space as the set of orbits of $\mathcal{S}_m^k$ under the action of $SO(m)$. Note that this figure does not correspond to any real shape space: when $m = 2$ and $k = 2$, $\mathcal{S}_2^2$ is a one-dimensional sphere in $\mathbb{R}^2$, a point, and the minimal non-trivial shape space is obtained for $m = 2$ and $k = 3$, implying that $\mathcal{S}_2^3$ is a three-dimensional sphere in $\mathbb{R}^4$.

In [30], Kendall studies the topology of generic shape spaces Σ_m^k, showing that:

- Σ_1^2 is a 2-point space and thus a zero-dimensional sphere of unit radius.
- Σ_2^3 (the space of planar triangles) is a metric 2-dimensional sphere in $\mathbb{R}^3$ of radius $1/2$, largely studied in the litterature [7, 34, 29, 18, 30, 45].
- Σ_1^k is equivalent to the $(k-2)$- dimensional sphere of unit radius, noted $\boldsymbol{S}^{k-2}(1)$.
- Σ_2^k is equivalent to the $(k-2)$- dimensional complex projective space of curvature 4, noted $\mathbb{C}P^{k-2}(4)$.
- Σ_{k-1}^k, $k \geq 3$, are topological spheres but not metric spheres studied in [30].

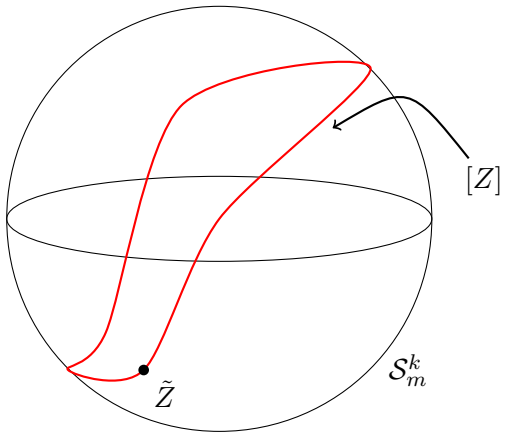

Figure 2.4: Illustration of the preshape space as an hypersphere and shapes as orbits.

- Σ_m^k, $k \leq m$ are overdimensioned spaces. They are topologically equivalent to hemispheres, the halved version of Σ_{k-1}^k [30].
- Σ_m^k, $k \geq m+2$ and $m \geq 3$ are not spheres, nor manifolds, as they have singularity sets [30]. Note that outside the singularity set, the shape space is a differential manifold that can be endowed with a Riemannian metric and where geodesics can be defined.

In [30], three distances in Σ_m^k are defined almost everywhere (outside the singularity sets) as the distances inherited from the distances in $\mathcal{S}_m^k$ with definitions 2.7, 2.8 and 2.9.

Definition (2.11): Distances in Σ_m^k

Let $D(\cdot,\cdot)$ be a distance in $\mathcal{S}_m^k$. Then,

$$d([Z_1],[Z_2]) = \inf_{R \in SO(m)} D(\tilde{Z}_1, R\tilde{Z}_1) \ , \tag{2.11}$$

is a distance in Σ_m^k between two shapes $[Z_1]$ and $[Z_2]$, obtained by searching the minimum distance between the orbits of the preshapes $\tilde{Z}_1$ and $\tilde{Z}_2$.

We will use the following notations:

- d_F is called the full Procrustes distance inherited from D_s (Definition 2.7),
- d_P is called the partial Procrustes distance inherited from D_c (Definition 2.8),

- d is called the Procrustes distance inherited from D (definition 2.9).

The determination of geodesics in Σ_m^k is a difficult problem treated in [30].

2.2.3 Planar shapes

In the context of this thesis, we are interested in characterizing shapes of planar objects i.e. $m = 2$ and Σ_2^k, which, as the authors of [30] show, is a Riemannian manifold.

In the special case of planar objects, we can exploit the complex representation of the plane: $z_i \in \mathbb{C}$, $Z = (z_1, z_2, \cdots z_k) \in \mathbb{C}^k$, where $\mathbb{C}$ is the complex representation of $\mathbb{R}^2$, is possible. The Frobenius norm is: $||Z|| = \sqrt{ZZ^H}$ where Z^H denotes the conjugate transpose of Z, and the rotations are equivalent to multiplication by unit modulus complexes $e^{i\theta}$ ($i^2 = -1$), $\theta \in [0, 2\pi[$, so that shapes are the equivalence classes:

$$[Z] = \left\{ e^{i\theta}\tilde{Z},\ \tilde{Z} \in \mathcal{S}_2^k\,,\ \theta \in [0, 2\pi[\right\}\ . \tag{2.12}$$

All distances given in definition 2.11 have analytical expressions in Σ_2^k:

- $d_F([Z_1], [Z_2]) = 1 - |\tilde{Z}_1 \tilde{Z}_2^H|$,
- $d_P([Z_1], [Z_2]) = \sqrt{2}\left(1 - |\tilde{Z}_1 \tilde{Z}_2^H|\right)$,
- $d([Z_1], [Z_2]) = \arccos |\tilde{Z}_1 \tilde{Z}_2^H|$.

2.2.4 Extension to closed shapes

Kendall's original theory considers sets of distinguishable points. In many applications, these points are samples of closed curves $c(t), t \in T \subset \mathbb{R}^+$ such that $z_k = c(\varphi(t_k))$. The parametrization $\varphi(t)$ of curves is not fixed and there is no guarantee of a systematic labeling of the points as there is generally no natural origin for curves.

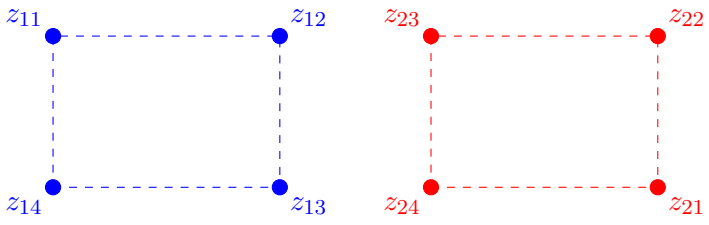

Figure 2.5: Two objects with the same shape and different labeling of their landmarks.

Figure 2.5 shows two closed objects with the same shape but with different labeling of their landmarks. Any of the distances introduced before between their shapes is non-null, $d = \pi/2$ for example, as they are in different orbits. The two objects will have the same shape if we additionally impose invariance with respect to labeling. An additional invariance with respect to parameterization $\varphi(t)$ should be imposed to recover the intuitive notion of shape, such as, fixed chord-length for the set of points or regular sampling of the curve. This invariance, not considered here in the discrete case, is imposed for shape of continuous curves in section 2.3. However since points correspond to samples taken along the curve such that: $k_1 > k_2$ implies $t_{k_1} > t_{k_2}$, we consider only relabeling that preserve their order.

We give a new definition of shape, denoted by (Z), as the equivalence class of $[Z]$ with respect to the group noted $\mathcal{P}$, the set of shift and reflective operators:

Definition (2.12): Shift and reflective group

Let $\mathcal{P}$ be the subset of the permutation group generated by 2 matrices $\Delta_1 \in \mathbb{R}^{k\times k}$ and $\Delta_2 \in \mathbb{R}^{k\times k}$ such that any matrix $P \in \mathcal{P} \in \mathbb{R}^{k\times k}$ can be written:

$$P = (\Delta_1)^r (\Delta_2)^p \quad ,$$

$p \in \{1, 2 \dots k\}$ and $r \in \{0, 1\}$.

$$\Delta_1 = \begin{bmatrix} 0 & \cdots & \cdots & 0 & 1 \\ \vdots & \iddots & \iddots & \iddots & 0 \\ \vdots & \iddots & \iddots & \iddots & vdots \\ 0 & \iddots & \iddots & \iddots & \vdots \\ 1 & 0 & \cdots & \cdots & 0 \end{bmatrix} \text{ and } \Delta_2 = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \\ 1 & \ddots & \ddots & \ddots & 0 \\ 0 & 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 1 & 0 \end{bmatrix}.$$

Note that these operators act upon the labels and not the points and form a cyclic group of order k with the finite set $\{1 \cdots k\}$.

Definition (2.13): Circular shape

Let Z_1 and Z_2 be two configurations in $\mathbb{R}^{m\times k}$. We say that Z_1 and Z_2 have the same circular shape and write $(Z_1) \overset{circ}{=} (Z_2)$ if and only if $\exists\, \alpha \in \mathbb{R}^+_*$, $R \in SO(m)$, $\mathbf{t} \in \mathbb{R}^m$ and $P \in \mathcal{P}$:

$$(Z_1) \overset{circ}{=} (Z_2) \Leftrightarrow Z_1 = \alpha R Z_2 P + \mathbf{t} \ .$$

Let Z be a configuration. (Z) is the equivalence class:

$$(Z) = \left\{ R\tilde{Z}P \,:\, \tilde{Z} \in \mathcal{S}^k_m \,,\, P \in \mathcal{P},\, R \in SO(m) \right\} \ . \tag{2.13}$$

Definition (2.14): Circular shape space

We denote by Σ_m^{*k} the circular shape space, the set of circular shapes in $\mathbb{R}_m^k$. It is the quotient of the shape space defined in 2.10 by the shift group:

$$\Sigma_2^{*k} = \Sigma_2^k / \mathcal{P} \ . \tag{2.14}$$

As quotient spaces, Σ_m^{*k} inherit distances defined in Σ_2^k. Let d be a distance defined in Σ_m^k. Then, $\mathcal{D}$:

$$\mathcal{D}((Z_1),(Z_2)) = \min_{P\in\mathcal{P}} d([Z_1],[Z_2\,P]) \ , \tag{2.15}$$

is a distance between two circular shapes in Σ_m^{*k}.

When $\mathcal{D}$ is derived from d_F or d_P, i.e. when d in the previous equation is given by equation (2.11) and either equation (2.6) or eqaution (2.7), its determination is a version of the general *Procrustes problem*, more particularly, the two sided orthogonal Procrustes problem [43], as it is a double minimization problem, for example with d_P:

$$\mathcal{D}_P\left((Z_1),(Z_2)\right) = \min_{P\in\mathcal{P}} d_P([Z_1],[Z_2]) = \min_{R\in SO(m),P\in\mathcal{P}} ||\tilde{Z}_1 - R\tilde{Z}_2 P|| \ , \tag{2.16}$$

where $\tilde{Z}_1$ and $\tilde{Z}_2 \in \mathcal{S}_m^k$ are any two preshapes representing (Z_1) and (Z_2).

When no special restrictions are imposed on Z_1 and Z_2 (such has their rank for example), there exists no closed form expression for $\mathcal{D}$ and the minimization must be solved it numerically.

To conclude this brief review on the discrete shape theory, we must note that preshape spaces have a well-known geometry allowing very simple definition for distances and, in the planar case (that is the focus of this thesis), analytical expressions exist which are easy to handle. However, the theory suffers a major basic drawback that is the selection of landmarks and their one-to-one correspondence, detailed in the concluding section of this chapter 2.5.

2.3 Continuous shape space (Srivastava)

As we have seen, the definition of the discrete shape space presented in the previous section goes through the identification of the preshape space, that collects selected representatives of each class of equivalence with respect to two groups of operators: translation and scaling. The shape space is finally

built as the quotient of this space with respect to the actions of $SO(m)$, the group of rotations in $\mathbb{R}^m$.

The approaches that stem from the work introduced in [31] follow the same construction even if considering distinct configuration spaces $\mathcal{L}_c$, the set of mathematical objects that contains all possible instances of shapes. First a selection function $Q(\cdot)$ is defined such that the corresponding many-to-one mapping of $\mathcal{L}_c$, noted $\mathcal{C}$, is invariant with respect to the action of a given set of groups $\mathcal{G}$:

$$\begin{aligned} Q: \quad \mathcal{L}_c &\rightarrow \mathcal{C} \subset \mathcal{L}_c \\ C &\rightarrow Q(C) = s \end{aligned} \tag{2.17}$$

The sets $\{C : Q(C) = s\}$ are equivalence classes with respect to $\mathcal{G}$, i.e. if $s = Q(C_1) = Q(C_2)$, then, $\exists g \in \mathcal{G}$: $C_1 = g(C_2)$. The preshape space $\mathcal{C}$ is thus the set of possible values of $Q(C) \in \mathcal{L}_c$.

Let $\mathbb{L}^2$ be the Hilbert space of functions from $[0, 2\pi]$ to $\mathbb{R}$ with inner product $\langle f, g \rangle_{\mathbb{L}^2}$:

$$\langle f, g \rangle_{\mathbb{L}^2} = \int_0^{2\pi} f(t)g(t)dt \ , \tag{2.18}$$

and associated norm $\|f\|_{\mathbb{L}^2} = \sqrt{\langle f, f \rangle_{\mathbb{L}^2}}$.

Let $\mathbb{L}^2(\mathbb{R}^2)$ be the set of continuous and differentiable functions defined on $[0, 2\pi]$ with values in $\mathbb{R}^2$. Let $f(t) = (f_x(t), f_y(t))$ and $g(t) = (g_x(t), g_y(t))$ be two elements of $\mathbb{L}^2(\mathbb{R}^2)$, the inner product is:

$$\langle f, g \rangle_{\mathbb{L}^2(\mathbb{R}^2)} = \int_0^{2\pi} \langle f(t), g(t) \rangle_E \, dt \ , \tag{2.19}$$

where, $\langle f(t), g(t) \rangle_E = f_x(t)g_x(t) + f_y(t)g_y(t)$ is the standard Euclidean inner product in $\mathbb{R}^2$.

The following two configurations have been proposed as a basis for the definition of shape spaces:

- the angle function with respect to arc-length parametrization [31], noted $\theta(t) \in \mathbb{L}^2$,
- the square-root velocity function [25], noted $q(t) \in \mathbb{L}^2(\mathbb{R}^2)$.

Some other functions have been proposed [31, 37] that we do not expose here as they are extensions of the former two.

Definition (2.15): Velocity vector

The velocity vector $\alpha(t)$ of the curve $c(t)$ is:

$$\alpha(t) = \frac{dc(t)}{dt} \ . \tag{2.20}$$

Definition (2.16): Curve length

Let $c(t)$ be a curve. The length of $c(t)$, noted $L(c)$ is:

$$L(c) = \int_0^{2\pi} \sqrt{\langle \alpha(t), \alpha(t) \rangle} dt \ . \tag{2.21}$$

When the parametrization of $c(t)$ is fixed to arc-length (unit speed), as it is the case for the first proposed configuration, $|\alpha(t)| = 1$. Then:

Definition (2.17): Angle function

$\theta(t) \in \mathbb{L}^2$ is the angle between the x-axis and the velocity vector $\alpha(t)$ such that:

$$\alpha(t) = e^{\Phi(t)} e^{i\theta(t)} \quad , \quad t \in [0, 2\pi] \ . \tag{2.22}$$

where $\Phi(t)$ is the log-velocity function:

$$\Phi(t) = \log |\alpha(t)| \ .$$

If the parametrization is fixed to arc-length, or unit speed, then $|\alpha(t)| = 1$ and, for all $t \in [0, 2\pi]$, $e^{\Phi(t)} = 1$ and $\Phi(t) = 0$. In this case, $\theta(t)$ is a representation of $c(t)$.

Definition (2.18): Square-root velocity function

The square-root velocity function, $q(t) \in \mathbb{L}^2(\mathbb{R}^2)$ by definition is:

$$q(t) = \frac{\alpha(t)}{\sqrt{||\alpha(t)||_E}} \quad , \quad t \in [0, 2\pi] \ , \tag{2.23}$$

with $\alpha(t)$ given in equation (2.20).

2.3.1 The preshape space

As differentiable functions, the two configuration spaces $\mathbb{L}^2$ and $\mathbb{L}^2(\mathbb{R}^2)$ are invariant with respect to translation of the original curves. Corresponding preshape spaces are obtained considering the action of selected groups:

- For $\theta(t)$, $\mathcal{G}$ is the group of rotations, and the selected representative, the preshape, is such that

$$\mathrm{E}[\theta(t)] = \frac{1}{2\pi} \int_0^{2\pi} \theta(t) dt = \pi \ .$$

Then, the preshape space $\mathcal{C}_\theta$ is:

$$\begin{aligned}\mathcal{C}_\theta &= \{\theta(t) = \arg(\alpha(t)) : \\ &\quad c(t) \text{ is arc-length parametrized}, L(c(t)) = 2\pi, \mathrm{E}[\theta(t)] = \pi\} \ .\end{aligned}$$

- For $q(t)$, $\mathcal{G}$ is the group of scalings, and the selected representative, the preshape, is such that

$$\int_0^{2\pi} ||q(t)||^2 dt = 1 \ ,$$

Then the preshape space $\mathcal{C}_q$ is:

$$\mathcal{C}_q = \left\{ q(t) = \frac{\dot{c}(t)}{\sqrt{||\dot{c}(t)||}} : \mathrm{E}\left[||q(t)||^2\right] = 1 \right\} \ . \tag{2.24}$$

Let $c(t)$ be a curve. We note $\tilde{c}(t)$ its preshape (either with θ or q configuration space. In figure 2.6, we give, on the left, an example of a closed curve $c(t)$ (seahorse) and, on the right, its preshape $\theta(t)$.

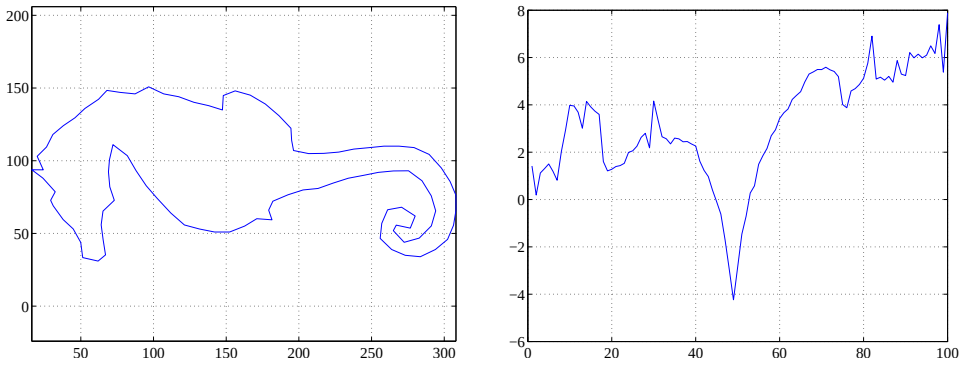

Figure 2.6: Example of the angle function (right) for the contour of the object on the left.

In [31, 25], authors define differentiable mappings from the configuration space $\mathcal{L}_c$ to $\mathbb{R}$ that define preshape spaces:

$$\mathcal{F} : \mathcal{L}_c \rightarrow \mathbb{R} \tag{2.25}$$

$$\mathcal{F}(C) = a \ . \tag{2.26}$$

Then, the preshape space $\mathcal{C} = \mathcal{F}^{-1}(a)$. They conclude that both preshape spaces $\mathcal{C}_\theta$ and $\mathcal{C}_q$ are infinite-dimensional "Riemannian" manifolds and they determine the tangent and normal spaces to the preshape space. Indeed, from the computation of the directional derivative of the mapping $\mathcal{F}$, it follows that the normal space is a one-dimensional space with inner product

inherited from $\mathcal{L}_c$, and the tangent its infinite-dimensional complement in $\mathcal{L}_c$. With definition 2.24, $\mathcal{C}_q$ is an infinite-dimensional unit radius hypersphere with well characterized geometry and for which geodesics are great-circles. However there is no analytical expression for geodesics in $\mathcal{C}_\theta$ and in [31] authors propose a numerical method to find them.

Definition (2.19): Distance in $\mathcal{S}$

Distance D between the preshapes of two continuous curves is the length of the geodesic between them and it is numerically computed.

2.3.2 The shape space

Let $c(t)$ be a curve. The shape of $c(t)$, noted $[c(t)]$, is the orbit of the preshape $\tilde{c}(t)$ with respect to some group operators.

- Let $\chi_{x.f}(t) = f(t-x) + x$, $x \in [0, 2\pi]$ be a function that changes the initial point of $f(t)$ defined on $[0, 2\pi]$, preserving its mean value, and let $\mathcal{X}_f$ be the set of such functions for $x \in [0, 2\pi]$.
- Let Γ be the set of orientation preserving diffeomorphisms of $[0, 2\pi]$ (equivalent to actions on the circle noted $\mathbb{S}^1$), or re-parametrization group. For $\gamma \in \Gamma$ and $q(t) \in \mathbb{L}^2(\mathbb{R}^2)$, $\tilde{q}(t) = \sqrt{\dot{\gamma}(t)}\,(q \circ \gamma(t))$.
- Let $SO(2)$ be the special orthogonal group of rotations in $\mathbb{R}^2$.

Definition (2.20): Shape space

Shape space, noted $\mathcal{S}$, is the quotient space of the preshape space $\mathcal{C}$ under the action of groups operators:

- $\mathcal{S}_\theta = \mathcal{C}_\theta / \mathcal{X}_\theta$,
- $\mathcal{S}_q = \mathcal{C}_q / \left(\mathcal{X}_q \times \Gamma \times SO(2)\right)$.

In [31, 26], the authors determine that the shape space is an infinite-dimensional Riemannian manifold that inherits the distance definition from those defined in the preshape space (definition 2.19).

Definition (2.21): Distance in $\mathcal{S}$

The distance d between two shapes $[c_1(t)]$ and $[c_2(t)]$ is defined as the minimal distance between orbits of preshapes $\tilde{c}_1(t)$ and $\tilde{c}_2(t)$:

$$d([c_1(t)], [c_2(t)]) = \min_{P \in \mathcal{P}} D(\tilde{c}_1(t), P \cdot \tilde{c}_2(t)) \ , \tag{2.27}$$

where $\mathcal{P}$ is the group of actions that form the equivalence class given in definition 2.20 ($\mathcal{P} = \mathcal{X}_\theta$ for $\theta(t)$ representation and $\mathcal{P} = \mathcal{X}_q \times \Gamma \times SO(2)$ for q representation), and D is the distance in $\mathcal{S}$ (definition 2.19).

2.3.3 Closed shapes

The representation of closed curves requires an additional closure condition. For the two representations introduced previously, these conditions are:

- for $\theta(t)$, $\int_0^{2\pi} e^{i\theta(t)}dt = 0$,
- for $q(t)$, $\int_0^{2\pi} ||q(t)||_E q(t)dt = 0$.

The closure condition (in fact two conditions in the plane) must be added in the mapping $\mathcal{F}$ given in equation 2.26: $\mathcal{F} : \mathcal{L}_c \to \mathbb{R}^3$. Then, both preshape spaces are infinite-dimensional Riemannian manifolds with tangent spaces that are the complement of three-dimensional normal spaces for whose bases are known.

The closure condition is a nonlinear constraint and $\mathcal{C}_q$ is thus no longer a sphere. For both representations, geodesics and distances in the preshape spaces have no analytical expressions and in [31, 26], authors propose numerical methods to determine them.

2.4 Statistical shape spaces

We now focus on the construction of statistical shape spaces by first focusing on the definition of statistics such as mean and variance, and then summarizing current efforts on the definition of probability distributions in shape spaces.

2.4.1 Statistics in shape space

Definition (2.22): Fréchet mean

Let $(S, dist)$ be a metric space with probability measure w. The Fréchet mean $\bar{S}_F$ is the set of global minimizers of $f(x)$:

$$f(x) = \int_S dist^2(x, y)dw(y) \quad x \in S \ ,$$

i.e. it is the set $\{([\mu]_{dist} \in S : f([\mu]_{dist}) \leq f(x),\ x \in S\}$.

Let $\{X_i\}_{i=1}^n , X_i \in S$ be independent and identically distributed elements of S. Then, we obtain the sample Fréchet mean or Karcher mean of $\{X_i\}_{i=1}^n$ using in the previous equation the empirical distribution, as the local minimizers of $F(x)$:

$$F(x) = \frac{1}{N}\sum_{i=1}^{N} dist^2(x, X_i) \quad x \in S \ . \tag{2.28}$$

Definition (2.23): Fréchet variation

Let μ denotes a minimizer of $f(x)$. Then, the variation, noted V_μ, of the Fréchet mean is $V_\mu = f(\mu)$.

Continuous and discrete mean shape can be defined with the corresponding distances definitions (2.11), (2.27).

For discrete planar shape in Σ_2^k, the Fréchet means and their variations are studied in [6], with the full Procrustes distance d_F. The mean associated to the full Procrustes distance, noted $[\mu]_F$, has an analytical expression. An example of this mean shape is given in Figure 2.7 with, in the left, 10 objects, and, in the right, their preshapes and their mean shape $[\mu]_F$. Note that the mean shape is not rotationally aligned with the preshapes as it is the mean of orbits, invariant to rotation, and not the mean of preshapes.

Definition (2.24): Full Procrustes mean shape

Let $\left\{\tilde{Z}_i\right\}_{i=1}^n$ be a set of preshapes defined in equation (2.4). Let $[\mu]_F$ be the Fréchet mean shape of the set. Let S be the matrix $S = \sum_{i=1}^N \tilde{Z}_i^H \tilde{Z}_i$, and let λ and ν be its largest eigenvalue and its corresponding eigenvector. Then, $[\mu]_F = \nu$.

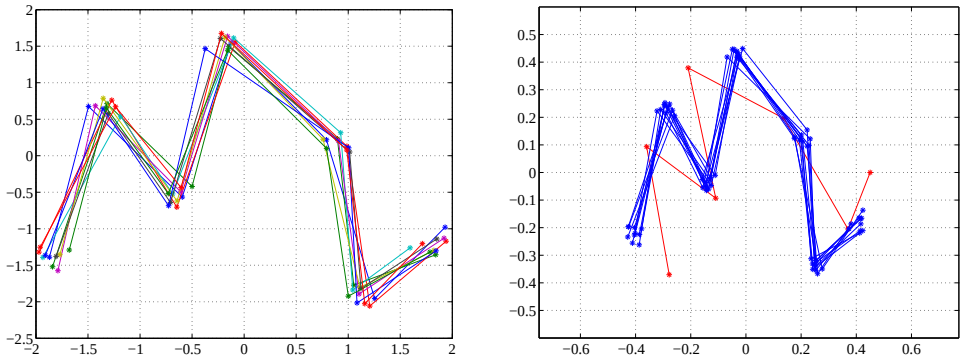

Figure 2.7: Example of a set of curves (left) and, in the right, their preshapes (blue) and the Karcher mean shape $[\mu_F]$(red).

Uniqueness of the Fréchet mean in Σ_2^k is proven in [6]. They also propose nonparametric tests to compare mean and variations of two sample sets of shapes. We do not further explore these properties here.

For continuous shape spaces, as their is no analytical expression for the distance, the Fréchet mean must be numerically computed.

2.4.2 Statistical shape models

We now want to go further in the statistical analysis and define distributions in shape spaces. In the literature there are different approach to the definition of shape distributions.

We begin by introducing an approach for discrete representations of objects, that consists in defining distributions in the configuration space $\mathbb{R}^{m\times k}$ and transposing them in the shape space. Authors in [30, 45] and in [34] give some of them:

- [30, 45] transpose the distribution of k labeled iid (independently identically distributed) isotropic Gaussian distributions in $\mathbb{R}^m$, to its induced shape measure in Σ_m^k, outside the singularity set (that has measure zero). Note that for $m = 2$, this shape measure is the uniform measure on Σ_2^k.
- [30, 45] transpose the distribution of k labeled iid non-isotropic Gaussian distributions in $\mathbb{R}^2$, to its induced shape measure in Σ_2^k,
- [30] transpose the complex normal distributions in $\mathbb{C}$, to its induced shape measure in Σ_2^k,
- [34] transpose k independently distributed bivariate Gaussian with Bookstein's coordinates to its induced shape measure in Σ_2^k.

Note that these approaches do not directly define distribution in the shape space but consider a distribution in the configuration space and transform it to the shape space. It is only possible to do so for a reduced number of well-known distributions.

We also introduce the point distribution model (PDM), presented in [11]:

Definition (2.25):

Let $\{x_i\}_{i=1}^n$ be a set of aligned objects, then each object is described using a typical object $\bar{x}$ with a typical variation:

$$x_i = \bar{x} + \sum_{j=1}^{p} h_{ij} v_j \quad . \tag{2.29}$$

A standard tool for computing the modes of the variations $\{v_j\}_{j=1}^p$ in the set of objects is the principal component analysis of the data set.

This theory is the basis of lots of studies. One major issue in PDM determination is the way landmarks are selected and aligned. [48, 13] propose methods for automatic selection of landmarks and [5] proposes to use equally spaced control points of a spline as the landmarks. In [23],the authors propose a multimodal PDM for tracking applications.

In this part on PDM, we intentionally did not mention shape but rather refer to mention objects. Indeed the method does not consider the global shape of the objects but their variations inside an observed set. Thus, even when some restrictions are applied to objects (such as, size, position, ...), the model is still in the configuration space and not in the shape space.

A second type of approach for shape distribution definition is when defining distributions in the preshape space and then transporting them in the shape space. Considering the discrete theory, this method requires to define distributions on hyperspheres. [18] and [40] present some distributions in the preshape space:

- the complex Bingham distribution [18] in $\mathcal{S}_2^k$, denoted $\mathbb{C}B_{k-2}(A)$, where A is a matrix determining the mode of the distribution and its concentration around the mode. Note that $\mathbb{C}B_{k-2}(A)$ is invariant with respect to rotations and is equivalent to a zero-mean complex multivariate normal distribution conditioned to have unit norm: if $w \sim \mathbb{C}N(0, \Sigma)$,[1]

$$w|||w|| = 1 \sim \mathbb{C}B_{k-2}(-\Sigma^{-1}) \ .$$

- the complex Watson distribution [18] in $\mathcal{S}_2^k$, denoted $\mathbb{C}W_{k-2}(\mu, \xi)$, where μ and ξ determine its mode and concentration. Note that $\mathbb{C}W_{k-2}(\mu, \xi)$ is a special case of $\mathbb{C}B_{k-2}(A)$ when A has just two distinct eigenvalues.

- in [40], Pennec propose a generalization of the Gaussian distribution for manifolds, and particularly for spheres. This could be applied for the definition of Gaussian mixtures on the preshape space for example.

The last approach introduced in this thesis for the definition of shape distributions is those that are defined in the tangent space of the preshape space or of the shape space. In general, considering geometry on Riemannian manifolds, the projection on the tangent space is defined with the exponential mapping.

[1] $\mathbb{C}N(0, \Sigma)$ denotes the complex gaussian distribution with null mean and covariance matrix Σ.

Definition (2.26): Exponential map

Let $\mathcal{M}$ be a manifold and μ, a point in $\mathcal{M}$. Let $T_\mu(\mathcal{M})$ be the tangent space to $\mathcal{M}$ at μ and v a vector of $T_\mu(\mathcal{M})$. Let $g_{\mu,v}$ be the geodesic in $\mathcal{M}$ going through μ in the direction v.
The exponential map maps each vector $v \in T_\mu(\mathcal{M})$ to a point in $\mathcal{M}$ attained in a unit time along $g_{\mu,v}$:

$$\begin{aligned} T_\mu(\mathcal{M}) &\rightarrow \mathcal{M} \\ v &\rightarrow exp_\mu(v) = g_{\mu,v}(1) \ . \end{aligned}$$

The inverse mapping, from the manifold to the tangent space, is called the logarithmic map. So the projection detailed above (2.30) can be seen as the inverse of exponential mapping with geodesic on the preshape space (hypersphere).

So defining distributions on the tangent space to shape space and then, transporting them onto the shape space with the logarithmic mapping, we can define distributions on the shape space.

Let $T_{[\mu]}(\Sigma_2^k)$ be the tangent space to Σ_2^k at the shape $[\mu]$. $T_{[\mu]}(\Sigma_2^k)$ is an Euclidean approximation of the shape space in the neighborhood of $[\mu]$.

The vector $v \in T_{[\mu]}(\Sigma_2^k)$ corresponding to the shape $[Z] \neq [\mu]$ is its orthogonal projection in the tangent space. $\hat{Z}$, the icon of $[Z]$ that minimizes the distance with $[\mu]$, is first selected:

$$\hat{Z} = \arg \min_{Z \in [Z]} dist(Z, [\mu]) \ .$$

where $dist(\cdot)$ is one of the three distances defined above. Then, $\hat{Z}$ is projected on $T_{[\mu]}(\Sigma_2^k)$ so that:

$$v = \left(I_k - \mu^H \mu\right) \hat{Z} \ , \tag{2.30}$$

with μ the preshape of $[\mu]$ closest to $\hat{Z}$. Note that v is a complex vector. The inverse projection from v to Z, an icon of the shape $[Z]$, is given by:

$$Z = \sqrt{(1 - vv^H)}\, \mu + v \ . \tag{2.31}$$

Note that (2.31) and (2.30) correspond respectively to the exponential and logarithm mappings, generally defined above.

In the case of continuous shape spaces, infinite-dimension Riemannian manifolds, one must resort to approximations for distributions definition. The tangent space to a shape space is also an infinite dimensional space. Thus exponential mapping is not defined and [31, 25] propose to restrict to a finite-dimensional subspace of the tangent space (with Fourier descriptors for example in [31]). Then, one can define the exponential map 2.26 to

transport distributions from the tangent space to the shape space, that is a Riemannian manifold and benefits from the work described in [40].

2.5 Conclusion and discussion

2.5.1 Discussion on Kendall's approach

Kendall's shape spaces are very powerful in the sense that they are manifolds of fixed dimensions. The main advantage of the discrete shape theory is its ease of manipulation, especially in the case of planar shapes where we have analytical expressions for distances, mean shape, and where its Riemannian structure allows to use the tangent space as an Euclidean approximation to the shape space.

The major drawback comes from the definition of an object as a set of remarkable points. Indeed they are usually manually selected and thus their selection highly varies with the perception of the selector itself. Some authors proposed to automate the selection [13, 5] but the problem due to the matching of points between distinct objects still remains. We can also easily define two sets of points that have the same shape but does not correspond to the same object. A solution may be to use sample points of a curve to define its shape. Then, the sampling scheme must be tight enough to capture the global shape but this goes paired with an increased computational cost.

Practically, to be able to compare the shapes of two distinct objects, their configurations must lie in the same space. This corresponds to establish that they have the same number of landmarks. But when trying to compare the shapes of a fish and a triangle it may be difficult to use the same number of key points to determine their shapes and thus it is meaningless to try to match them.

To avoid these major problems due to the discretization of the definition of the shape, we have seen that some authors propose continuous approaches to the shape definition.

2.5.2 Discussion on this continuous approach

The continuous approach to the definition of shape is an appealing theory as it overcomes the main drawback of Kendall's theory: the selection of landmarks. Continuous shape spaces are infinite dimensional Riemannian manifolds and Riemannian geometry tools allow the definition of some major elements such as the distance between two shapes, the geodesic on the shape space and the mean of a set of shapes. A problem with this approach is that

their is no analytical expressions available for these tools and one must use numerical methods for their determination.

The main drawback arises as the shape space are non-linear infinite-dimensional spaces. So they are not probability spaces and to perform statistical analysis and modeling, one must resort to approximations. The Riemannian structure allows to use the tangent space to shape space and the exponential mapping between the two spaces. However the tangent space is still infinite-dimensional and to define standard statistics and probability density functions we must approximate it with a finite-dimensional vector space with, for example, Fourier descriptors or wavelet bases.

The representation of the shape of a curve is a continuous function and generally not a sparse representation. In the continuous theory there is no requirement about the sparsity or the smoothness of the representation and, for numerical computations, there is no discussion on the choice of the sampling scheme of curves.

We note that this theory is a burning issue in the shape theory domain and authors of papers quoted above are still working on this theory, especially on the statistical point of view. Besides A. Srivastava and E. Klassen kindly provides to us a pre-press version of their book "Statistical analysis of shapes of curves and surfaces". This book mainly refers to the articles we cited along this section and generalizes the theory for open curves in $\mathbb{R}^n$, and was a grateful help for us in the preparation of this chapter.

Chapter 3

Spline modeling

3.1 Introduction

Splines are piecewise polynomials functions. They form a dense set in the space of continuous and differentiable functions [14], in the sense that, given a function f in this set, we can find a spline that gets as close as we want to f. In this chapter, we focus on the problem of representing contours with splines. A spline is completely determined by a set of parameters and thus is an efficient and sparse representation of a contour.

In the first section, we introduce spaces of spline functions, noted $\boldsymbol{S}$, giving some definitions and notations used in this chapter. The traditional problems involving splines are briefly reviewed and we focus, in the third section, on the regression problem where some conditions are imposed on the complexity of the spline representation such as its order (the degree of the polynomials) or the number of polynomials in the piecewise representation. We dedicate the fourth section to free-knots spline modeling that is a special case of the regression problem where the number of polynomials is a parameter of the modeling problem. We present this problem as the identification of a model using nested families of increasing dimension. We review the solutions available in the literature and we recommend an approach that first selects the complexity of the model and then estimates the parameters of the model. The last section is dedicated to results of spline modeling for simulated and real contours.

3.2 Spline generalities

We begin this chapter on spline modeling with some basic definitions and properties of splines. A more detailed presentation can be found in [14].

3.2.1 Piecewise polynomials and splines

Definition (3.1): Piecewise polynomials of order m

Let $f(t)$ be a piecewise polynomial (pp) function of order m (degree $m-1$) with $l+1$ breakpoints $\tau = (\tau_i)_{i=1}^{l+1}$:

$$f(t) = P_i(t) \text{ if } \tau_i \leq t < \tau_{i+1} \, , i = 1 \cdots l \ . \tag{3.1}$$

For each $i = 1 \cdots l$, $P_i(t)$ is a polynomial of order m on its interval.

The function $f(t)$ is defined on $I = [\tau_1 \ \tau_{l+1}]$. Let $\Pi_{m,\tau}$ be the set of all pp functions $f(t)$ of order m with breaks τ.

In definition (3.1) no indication is given about the continuity and derivability of $f(t)$ at the break points. Let $\nu = (\nu_i)_{i=2}^{l}$ be a vector of positive integers, $\nu_i \leq m$, where each ν_i denotes the number of continuity conditions enforced at the corresponding break:

- $\nu_i = 0$ means that there is no condition at τ_i,
- $\nu_i = 1$ means that $f(t)$ is continuous at τ_i and
- $\nu_i = r$, $r > 1$, means that $f(t)$ is r^{th}-differentiable at τ_i.

Let $\Pi_{m,\tau,\nu}$ be the set of all functions of $\Pi_{m,\tau}$ satisfying these conditions.

From τ and ν, define the vector ξ of $(k+m)$ elements, called the knot vector, such that:

- $\xi_1 \leq \xi_2 \cdots \leq \xi_m \leq \tau_1$,
- $\tau_{l+1} \leq \xi_{k+1} \leq \xi_{k+m}$,
- and τ_i is repeated $m - \nu_i$ times in ξ for $i = 2 \cdots l$.

The set of functions $f(t)$ in $\Pi_{m,\tau,\nu}$, also noted $\boldsymbol{S}_{m,\xi}$, is the set of spline functions of order m with knot vector ξ. $\boldsymbol{S}_{m,\xi}$ admits a basis consisting of B-spline functions (or *basis splines*) commonly used for representing spline functions as they are easily defined and computed by means of the recursive equation (3.2). Figure 3.1 shows an example of a B-spline basis with knot vector $\xi = (0, 0, 0, 0, 0.3, 0.4, 0.7, 0.8, 1, 1, 1, 1)$.

Definition (3.2): B-splines

The j^{th} B-spline function, $j = 1 \dots k$, is:

$$b_j^m(t;\xi) = \omega_{j,m}(t) b_j^{m-1}(t;\xi) + (1 - \omega_{j+1,m}(t)) b_{j+1}^{m-1}(t;\xi) \ , \qquad (3.2)$$

with,

$$\omega_{j,m}(t) = \begin{cases} \frac{t-\xi_j}{\xi_{j+m-1}-\xi_j} & \text{if } \xi_{j+m-1} > \xi_j \\ 0 & \text{otherwise,} \end{cases}$$

$$b_j^0(t;\xi) = \begin{cases} 1 & \text{if } \xi_j \le t < \xi_{j+1} \\ 0 & \text{elsewhere.} \end{cases}$$

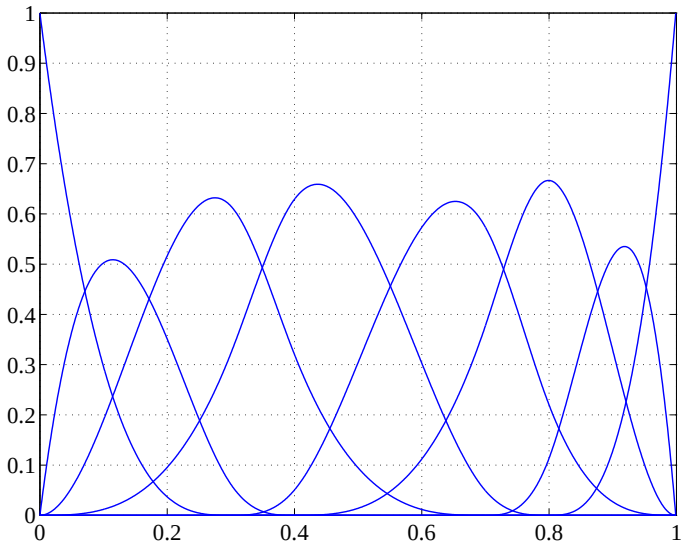

Figure 3.1: Example of a cubic B-spline basis.

Definition (3.3): B-spline representation

Any spline function $s(t) \in \boldsymbol{S}_{m,\xi}$, $t \in [a\ b]$ is a linear combination of the B-splines:

$$s(t) = \sum_{j=1}^{k} \beta_j \, b_j^m(t;\xi) \quad \text{with} \quad \xi_m = a \text{ and } \xi_{k+1} = b \ . \qquad (3.3)$$

Figure 3.2 gives an example of a spline curve, a two dimensional spline function, with $\xi = (0, 0, 0, 0, 0.3, 0.4, 0.7, 0.8, 1, 1, 1, 1)$.

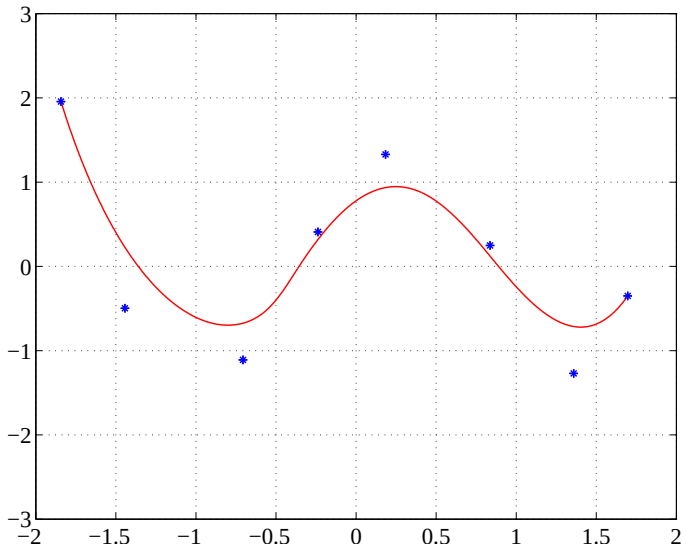

Figure 3.2: Example of cubic splines (in red) with β points in blue.

We now define some spline spaces that will be useful for the problem presented later.

Definition (3.4): Sets of spline functions

Let $\boldsymbol{S}_m^k$ be the set of splines of order m with fixed number of knots k:

$$\boldsymbol{S}_m^k = \left\{ \boldsymbol{S}_{m,\xi}; \xi = (\xi_i)_{i=1}^k \right\} \ .$$

Let $\boldsymbol{S}_m$ be the set of splines of order m:

$$\boldsymbol{S}_m = \bigcup_{k=m+1}^{\infty} \boldsymbol{S}_m^k \ .$$

3.2.2 Brief review of traditional problem solved with spline functions

In the literature, a large amount of research is involved with spline functions. In this section, some of them are briefly exposed and discussed.

Interpolating spline.

A traditional problem aims at finding a spline that interpolates a set of observed data. Such splines are called *i*nterpolating splines. Let $(y_i)_{i=1}^N$ be a set of data points. Then $s(t)$ is an interpolating spline if and only if $\exists$

sampling times $(t_i)_{i=1}^N$

$$y_i = s(t_i) \quad i = 1 \cdots N, \tag{3.4}$$

where $(t_i)_{i=1}^N$ are sampling times.

This interpolating problem may be solved by imposing continuity conditions at the break points leading to a system of equations developed in [14]. An example of an interpolating cubic spline for a set of points $\{y_i\}_{i=1}^N$ is given in figure 3.3.

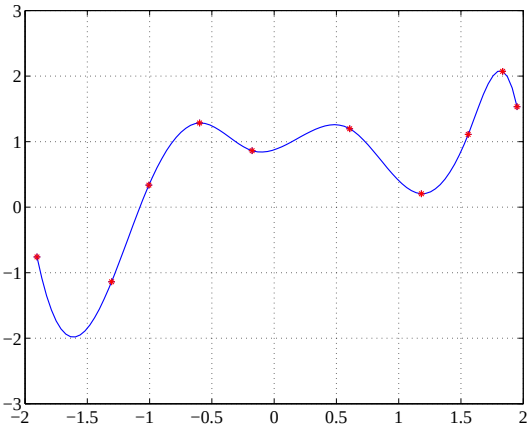

Figure 3.3: Example of interpolating spline (in blue) with $(y_i)_{i=1}^N$ (in red).

Regression spline.

In the previous example, the spline function exactly fits the data points. For noisy data, splines are used as approximating, denoising functions:

$$y_i = s(t_i) + \epsilon_i \quad i = 1 \cdots N \ , \tag{3.5}$$

where $(\epsilon_i)_{i=1}^N$ are unknown noise samples.

The problem reduces to finding the spline function that minimizes a function of the error. Let $||\epsilon||$ be the L^2-norm of the error also called residual sum of squares (RSS).

$$\mathrm{RSS}(s) = \frac{1}{N} \sum_{i=1}^{N} \left(y_i - s(t_i)\right)^2 \ . \tag{3.6}$$

The regression spline is the spline that minimizes RSS(s).

Identifying a spline is equivalent to determining its parameters β and ξ, a problem known as *regression*. Figure 3.4 shows an example of such a spline where the knot vector is fixed to $\xi = (0, 0, 0, 0, 0.3, 0.4, 0.7, 0.8, 1, 1, 1, 1)$. The estimation of β is a linear problem detailed later in section 3.3.1.

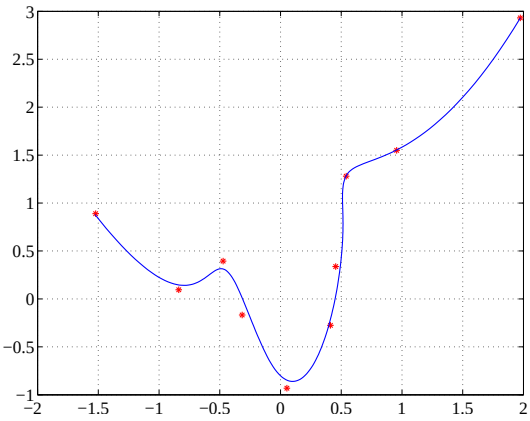

Figure 3.4: Example of approximating spline with noisy data in red and the spline in blue.

Smoothing spline.

Considering the regression spline problem, one may impose a smoothness condition on the spline function resulting in minimizing criterion SC which is the sum of an error term and a smoothing term:

$$\mathrm{SC}(s) = \frac{1}{N}\sum_{i=1}^{N}(y_i - s(t_i))^2 + \lambda \int_0^1 \left(s^{(p)}(u)\right)^2 du \ , \tag{3.7}$$

where $s^{(p)}(u)$ denotes the p-th derivative of $s(\cdot)$.

The parameter λ controls the degree of smoothness of the solution. The automatic selection of λ is still a current issue in spline estimation, and has been studied e.g. in [12, 47]. Figure 3.5 shows examples of smoothing splines with different values for λ: $\lambda_1 < \lambda_2 < \lambda_3$.

With this example we can see that the least smooth spline (λ_1) is the closest to the data and when $\lambda = 0$ the problem reduces to the regression spline presented above. If the smoothness of the spline is large, then it does not necessarily capture the details of the data (underfitting), but when the smoothness is near zero, it may model the noise present in the data (overfitting).

3.3 Curve modeling with spline

In the context of this thesis, the contour of an object is a continuous curve and we want to find a spline curve that best represent the data. Because we are dealing with noisy data, we focus on the regression problem and

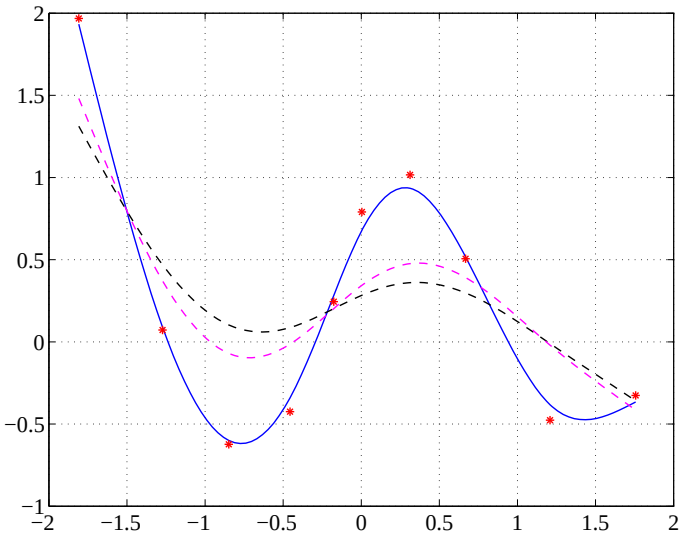

Figure 3.5: Example of smoothing splines with λ_1 (blue), λ_2 (magenta) and λ_3 (black), with data in red.

the identification of a spline with its parameters β and ξ that are a good representation of the contour.

We assume that cubic splines (order fixed to $m = 4$) are sufficiently smooth while being flexible enough to represent a wide variety of observed curves.

In this section, we focus on the modeling problem where a (cubic) spline must "best" fit the data that are noisy samples of a 2D curve. $s(t) \in \boldsymbol{S}_m$ (definition 3.4) is a function defined from $I = [a, b]$ to $\mathbb{R}^2$. The two coordinates in the plan are spline functions of the same order and with the same knot vector: they just differ by their control points.

$$\begin{aligned} s(t) &= (s_X(t), s_Y(t)) \ , \\ s_X(t) &= \sum_{j=1}^{k} \beta_{jX} b_j(t;\xi) \ , \\ s_Y(t) &= \sum_{j=1}^{k} \beta_{jY} b_j(t;\xi) \ . \end{aligned}$$

We use the complex representation of the plan so that $s(t)$ is a complex function: $s(t) = \sum_{j=1}^{k} \beta_j b_j(t;\xi)$, $(\beta_j = \beta_{jX} + i\beta_{jY})_{j=1}^{k} \in \mathbb{C}^k$. Determining a spline curve is equivalent to estimating the parameter $\theta = (\xi, \beta)$.

Modeling equation given in (3.5) is given in its matrix representation as:

$$Y = B_\xi \beta + \varepsilon \ , \tag{3.8}$$

where $Y = (y_1 \cdots y_N)^T \in \mathbb{C}^N$ is the data vector, $\varepsilon = (\varepsilon_1 \cdots \varepsilon_N)^T \in \mathbb{C}^N$ is the noise vector, B_ξ is the spline design matrix depending on the knot vector ξ with generic element $[B_\xi]_{i,j} = b_j(t_i;\xi) \in \mathbb{R}$ defined in equation (3.2), and $\beta = (\beta_1 \cdots \beta_k)^T \in \mathbb{C}^k$ is the vector of control points.

3.3.1 Regression splines with fixed knot vector ξ

When the knot vector ξ is fixed, the design matrix B_ξ is constant and the problem reduces to the estimation of the control points which are linear parameters in (3.8) and $s(t) \in \boldsymbol{S}_{m,\xi}$.

The least-squares (LS) criterion aims at minimizing the L^2-norm of the error and has an analytical solution:

$$\hat{\beta}_{LS} = \arg\min_{\beta \in \mathbb{C}^k} ||Y - B_\xi \beta||^2 \ , \tag{3.9}$$

$$\hat{\beta}_{LS} = \left(B_\xi^T B_\xi\right)^{-1} B_\xi^T Y \ . \tag{3.10}$$

The same solution is observed when applying the Maximum Likelihood (ML) criterion, assuming a Gaussian model for the error, $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_N)$:

$$\hat{\beta}_{ML} = \arg\max_{\beta \in \mathbb{C}^k} p(Y|\beta, \sigma^2) \ . \tag{3.11}$$

Figure 3.6 shows two examples of spline curve with fixed knot vector and its control points estimated from the sampled data for different number of knots. This example illustrates that we can get a spline arbitrarily close to the observed data when augmenting the number of knots. It also shows that these two splines do not capture the details in the tail of the seahorse. To do so, we must either add some knots to the representation, either adapt the knots position so that less knots are assigned to the part of the head and more are assigned to the tail.

As noted in [27], a sufficient condition for $\left(B_\xi^T B_\xi\right)$ to be invertible, so that the solution is identifiable, is that the matrix B_ξ has rank k. To ensure this, one may check if the data samples are dense with respect to the knots vector in the sense that there is at least one data sample between two knots. In practice this condition is usually verified as we want to find a representation with smaller dimension than the size of the full data set.

If the knot vector is unknown, the estimation problem is nonlinear and its solution requires numerical methods. The knot vector is composed of two spline parameters: its dimension $k \in \mathbb{N}$ ($\mathbb{N}$ is the set of positive integers), and the locations of the k knots in I. We emphasize this point by changing the notation: $\xi \to \xi^k = (\xi_1 \cdots \xi_k)$. Since once the knot vector is fixed, the problem is linear, the estimates of the control points have an analytical

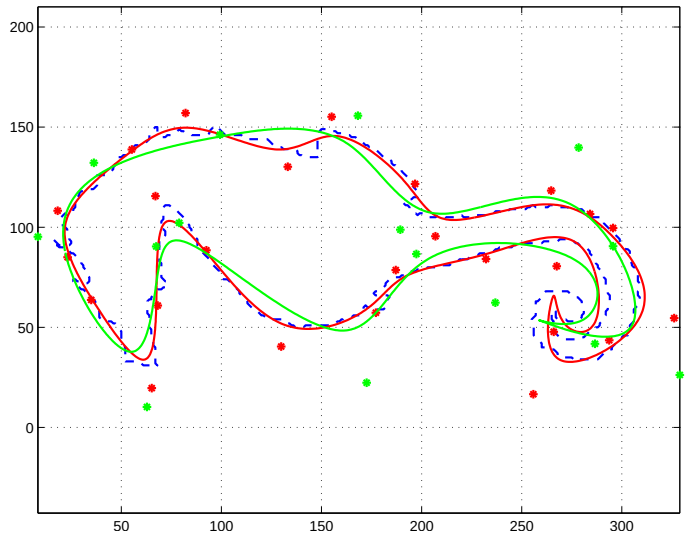

Figure 3.6: Example of spline curve estimation for data in blue: with 15 equispaced knots and β (green), with 25 equispaced knots and β (red).

expression, (3.10). We focus on the major difficulty: the estimation of the knots.

The knot estimation is a major problem when interested in determining the most sparse spline model for the data. Indeed we expect that fewer well placed knots will yield a spline more efficient to represent the data than fixed knots spline.

3.3.2 Regression splines with fixed number of knots

In this section, the number of knots k is fixed and their locations are free. The solution of the regression problem $s(t) \in \boldsymbol{S}_m^k$. Note that B-splines no longer form a basis $\boldsymbol{S}_m^k$. The estimation problem is known as the "free knots" spline problem. The space of all possible knot vector is the simplex in k dimensions with $I = [a, b]$.

Definition (3.5): Knot space

Let S_I^k be the space of all ξ^k configurations on I:

$$S_I^k = \left\{ \xi^k : a \leq \xi_1 \leq \xi_2 \cdots \leq \xi_k \leq b \right\} \quad , \tag{3.12}$$

A LS approach to this nonlinear problem is proposed in [27], applying a Gauss-Newton algorithm with Marquardt modification [35] to search for the minimum error with a parameter that is a transformation of the knot vector. This approach suffers from drawbacks traditional on nonlinear es-

timation problems: sensitivity to initialization, local convergence, and a specific drawback, resulting from freeing the knots, that is the existence in S_I^k of many stationary points of the LS function. This last point is called the "lethargy" problem in [27].

Another approach to the estimation of knot locations is to consider a statistical model for the noise in (3.8): $\varepsilon_i \sim \mathcal{N}\left(0, \sigma^2 \mathbf{I}_N\right)$, $\mathbf{I}_N$ being the $N \times N$ identity matrix. The noise variance σ^2 is unknown and introduced as an additional unknown parameter in the estimation problem. A solution can be found applying a Maximum Likelihood (ML) criterion with parameter $\theta = (\beta, \xi, \sigma^2)$.

Definition (3.6): Maximum Likelihood criterion

The Maximum Likelihood (ML) criterion aims at determining the estimate of parameter $\theta \in \Theta$ that maximizes the criterion:

$$\hat{\theta}_{ML} = \arg\max_{\theta \in \Theta} p(Y|\theta) \ . \tag{3.13}$$

For the spline fitting problem, the ML criterion leads to:

$$\hat{\sigma^2}_{ML}|\xi = \frac{Y^H(\mathbf{I}_N - B_{\xi^k}\left(B_{\xi^k}^T B_{\xi^k}\right)^{-1} B_{\xi^k}^T)Y}{2N} \ , \tag{3.14}$$

$$\hat{\beta}_{ML}|\xi = \left(B_\xi^T B_\xi\right)^{-1} B_\xi^T Y \ , \tag{3.15}$$

$$\hat{\xi}^k = \arg\max_{\xi^k \in \Delta_I^k} p(Y|\xi^k, \hat{\beta}_{ML}^k, \hat{\sigma^2}_{ML}) \ . \tag{3.16}$$

We must resort to numerical methods to find the estimates of the knots by solving the maximization problem. For example one can built a Markov Chain that simulates knot vectors from $p(Y|\xi^k, \hat{\beta}_{ML}^k, \hat{\sigma^2}_{ML})$ and, using a Simulated Annealing (SA) algorithm, the chain will converge to the maximizers of the function. Details about the SA algorithm can be found in [39]. In figure 3.7, we can see an application of this numerical method with ML criterion and a SA algorithm, where the number of knots has been fixed to $k = 15$.

Other criteria can be used to estimate the knot vector. For example, if we have some a priori knowledge on the parameters, such as their distributions, we can maximize the posterior distribution of the parameters with a similar numerical method.

3.3.3 Regression splines with fixed location of knots

If we now fix the location of the knots but not their number, the problem is to choose a correct model complexity: if the number of knots is too small,

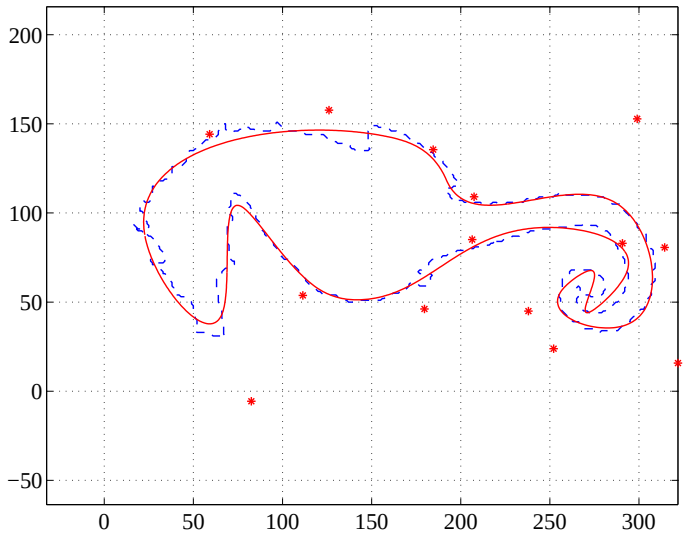

Figure 3.7: Example of the estimation of a spline curve and its control points (in red) for data in blue with $k = 15$.

the spline curve will not be a good representation of the data (underfitting) and if the number of knots is too high, the spline curve will try to fit the noise in the observed data (overfitting). We give an example for both situations in figure 3.6 where we choose knots that are equispaced in I: $\xi_{i+1} - \xi i = c$, for $i = 1 \cdots (k-1)$, where c is a constant in $\mathbb{R}_*^+$.

The determination of the model complexity k, relates to the model selection problems largely studied in the literature (e.g. [44, 28, 8]). Within the statistical approach exposed above in section 3.3.2 with the Gaussian noise model, we use a classical method: the penalized likelihood criterion.

Definition (3.7): Penalized ML criterion

$$\hat{\theta}_{penML} = \arg\max_{\theta \in \Theta} \left(\log p(Y|\theta) - \mathcal{P}(k) \right) \quad , \tag{3.17}$$

where $\mathcal{P}(k)$ is a penalty term that increases with k.

The criterion is a trade-off between under- and over-fitting of the resulting spline curve. Standard penalizing term are:

- Akaike Information Criterion (AIC), $\mathcal{P}_{AIC}(k) = k$,
- Bayesian Information Criterion (BIC), $\mathcal{P}_{BIC}(k) = \frac{k}{2} \ln(N)$,
- Minimum Description Length (MDL).

These penalties have been studied in [19, 1]. The choice of the penalty term is further explored in the next section where we study the likelihood approaches for the free-knot spline modeling. Figure 3.8 shows an example

of spline identification with penalized ML criterion with the BIC penalty term.

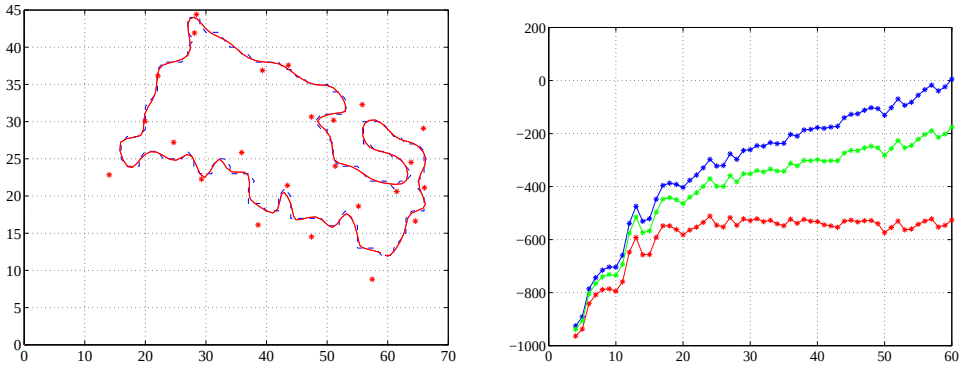

Figure 3.8: Example of spline curves estimation (left) (in red) for data in blue with equispaced knots vector: $k =$. Penalized likelihood criterion (right): no penalty (in blue), BIC (in red), AIC (in green).

3.4 Free-knots spline

Spline modeling without fixing the dimension of the spline curve (i.e. the number of knots) nor the locations of the knots is a difficult problem that has been addressed by many authors [27, 16, 33, 17]. This section highlights the major difficulties. The discussion presented here is not restricted to spline modeling but can be extended to other problems of model fitting. We thus use general notations and we come back to spline modeling at the end of the section.

3.4.1 Model structure

Free-knots spline modeling is a particular case of sparse model fitting where the model is the set of spline functions. In this domain, a convenient way of adjusting the model dimension to the data complexity is to consider a set $\mathcal{M}$ of candidate models that is the union of nested families of parametric models $\mathcal{M}_k$:

$$\mathcal{M} = \bigcup_{k \in \mathcal{K}} \mathcal{M}_k \quad ; \quad \mathcal{K} = \{k_{\min} \cdots k_{\max}\} \tag{3.18}$$

$$\mathcal{M}_k = \{p(\cdot|\theta),\, \theta \in \Theta_k\} \subset \mathcal{M}_{k+1} \ . \tag{3.19}$$

In this setting, models are probabilistic parametric functions assumed generally to be Gaussian distribution for the noise.

We note that the integer k is no longer the number of knots in the spline representation but an index directly related to the complexity, or number of degrees of freedom, of each family of models $\mathcal{M}_k$: if $k' > k$, the complexity of models in $\mathcal{M}_{k'}$ is higher than the complexity of those in $\mathcal{M}_k$.

The overall parameter space Θ of $\mathcal{M}$ is simply

$$\Theta = \bigcup_{k \in \mathcal{K}} \Theta_k \ . \tag{3.20}$$

Even when each Θ_k is a vector space, Θ is not a vector space, but a union of vector spaces.

We turn now to the definition of criteria to estimate the parameter $\theta \in \Theta$. Two common choices for parameter estimation are Maximum Likelihood (ML) and Bayesian. We detail them below in our particular setting.

3.4.2 Likelihood approaches

Maximizing the log-likelihood of the parameter θ yields

$$\hat{\theta}_{ML} = \arg \max_{\theta \in \Theta} \log p(Z|\theta) \ . \tag{3.21}$$

As noted previously (section 3.3.1), under Gaussian models, this is equivalent to minimizing the L^2-norm of the error between the data and the samples of the spline model. As it tries to reduce the error term, the ML criterion will always choose the most complex model to fit the data: $\hat{k}_{ML} = k_{max}$. This is a well-known drawback of ML estimation in nested families of models.

To correct the overfitting behavior of ML, many authors proposed the addition of corrective terms that favor selection of models using less parameters:

$$\hat{\theta}_{PL} = \arg \max_{k \mathcal{K} \, ; \, \theta \in \Theta} \left(\log p(Z|\theta) + \mathcal{P}(k) \right) \ , \tag{3.22}$$

where $\mathcal{P}(k)$ is a decreasing function of k. These techniques are generally known by the name of "penalized likelihood" methods, and have received three distinct justifications.

- they are sometimes dictated by the desire to impose some regularity characteristics on the solution, requiring in this case prior knowledge on its characteristics which is not necessarily available,
- they have also been justified as particular cases of MAP (Maximum a Posteriori) estimation for special selections of the prior $p(\theta)$, e.g. in [4, 8]. We will see below in section 3.4.3 that there is a fundamental flaw in this approach,

- a more generic approach says that they are derived on the basis of asymptotic arguments (when the data set is infinite) which relate them to the models' Bayesian marginal posterior $Pr(\mathcal{M}_k|Z)$ [28]. This is the case for AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion) [44] and MDL (Minimum Description Length) [41]. They cannot offer any guarantee of optimality for finite data sets.

3.4.3 Bayesian approaches

Bayesian approaches regularize the identification problem through the definition of a prior over Θ, and optimize the expected value of some function of the estimation error under the posterior distribution over Θ. As noted above, generic justifications of penalized likelihood link it to Bayesian estimates: either to the model posterior $p(\theta|Z)$ or to the marginal model posteriors $Pr(\mathcal{M}_k|Z)$. The two most popular Bayesian criteria are the MMSE (Minimum Mean Square Error) and the MAP (Maximum A Posteriori).

For the union-type models like (3.18)-(3.19)-(3.20), the MMSE criterion is meaningless, since the associated cost function $(\theta - \hat{\theta})^2$ is undefined when the complete parameter space is not a vector space. On the contrary, the 0/1 cost function of the MAP criterion is well defined, and should enable determination of a unique model among the set of candidate models. When Θ has the simple structure of a vector space, this criterion leads to

$$\hat{\theta}_{MAP} = \arg\max_{\theta\in\Theta} p(\theta|Z) \ . \tag{3.23}$$

Definition of distributions which are the basic entities manipulated by Bayesian techniques (e.g. prior or posterior distributions) must be done with care. When parametric probabilistic families over each Θ_k are known, as we assume here, an intuitive way is to use a mixture-like approach, writing posterior densities over Θ as

$$p(\theta|Z) = \sum_{k\in\mathcal{K}} p_k(\theta|Z), \qquad \theta \in \Theta \ , \tag{3.24}$$

where each $p_k(\theta|Z)$ is the Radon-Nikodym derivative of an un-normalized measure with respect to (w.r.t) the invariant measure over Θ_k: if $\theta \notin \Theta_k$ then $p_k(\theta|Z) = 0$, and

$$1 \geq \int_{\Theta_k} p_k(\theta|Z)d\theta \equiv Pr(\mathcal{M}_k|Z) \ . \tag{3.25}$$

A proper density $\nu_k(\theta|Z)$ over Θ_k is obtained by normalization:

$$\nu_k(\theta|Z) \triangleq \frac{p_k(\theta|Z)}{\int_{\Theta_k} p_k(\theta|Z)d\theta} \equiv p(\theta|\mathcal{M}_k, Z) \quad , \theta \in \Theta_k \ , \tag{3.26}$$

where we stressed the meaning of ν_k as resulting from conditioning on $\theta \in \Theta_k$. Note that these "local" densities (as well as their un-normalized versions) are defined w.r.t distinct reference measures, the invariant measures over each Θ_k, and are thus not directly comparable.

Surprisingly, several authors have proposed estimators based on a direct transposition of the MAP criterion (3.23) to the model structure considered herein [3, 8]:

$$\begin{aligned} \hat{\theta}_{nMAP} &= \arg\max_{\theta\in\Theta} p(\theta|Z) = \arg\max_{k\in\mathcal{K}} \max_{\theta\in\Theta_k} p_k(\theta|Z) \\ &= \arg\max_{k\in\mathcal{K}} p_k(\hat{\theta}_k|Z) \ , \\ \hat{\theta}_k &= \arg\max_{\theta\in\Theta_k} p_k(\theta|Z) \end{aligned} \tag{3.27}$$

We designate them by "naïve MAP" estimators.

As we pointed out before, the un-normalized densities $p_k(\theta|Z)$ defined over each Θ_k are not defined with respect to the same measures. This criterion, that abusively compares them directly, may lead to estimates with pathological behavior as demonstrated by the example below.

3.4.4 Pathological behaviour of the "naive" MAP estimator

Let $Y = [y_1, \ldots, y_n] \in \mathbb{R}^n$ be the observation vector, and consider a model with just two families: $\mathcal{M} = \mathcal{M}_1 \cup \mathcal{M}_2$, where

$$\mathcal{M}_k = \left\{ \mathcal{N}\left(f_k(\cdot|\theta), \sigma^2 \mathbf{I}_n\right), \theta \in \Theta_k \right\}, k = 1, 2 \ ,$$

with $\mathcal{N}(\mu, \Sigma)$ the normal density with mean μ and covariance matrix Σ.

Above, $f_k(\cdot|\theta)$ is a k-piecewise linear signal, see Figure 3.9, such that the parameters of the models are $\theta_1 = [P^0, \sigma^2]$ and $\theta_2 = [P^1, P^2, \sigma^2]$, where $\{P^i = (P^i_x, P^i_y) \in \mathbb{R}^2\}_{i=0}^2$ are the break point coordinates.

We use a factored prior distribution over both parameter spaces: $P^i_x \propto \mathcal{U}([0\, L])$, $P^i_y \propto \mathcal{N}(0, \Sigma^2), i = 0, 1, 2$, and $\sigma^2 \propto \mathcal{U}(\mathbb{R}^+)$[1]. To minimize the impact of the prior over the breakpoints we set $\Sigma^2 \gg 1$. Models are equiprobable: $\pi(\mathcal{M}_1) = \pi(\mathcal{M}_2)$.

Apparently, the prior chosen is uninformative and does not express preference for $\mathcal{M}_1$. One would expect that the ML overfitting behaviour would not be corrected and that $\mathcal{M}_2$ will always be selected.

[1] $X \propto p$ indicates that random variable X is drawn according to p and $\mathcal{U}(A)$ denotes the uniform distribution over set A.

In this example, parameters are set at: $P^1 = (250, 10), P^2 = (800, 6)$, $\Sigma = 10^{16}$, $X = \{10m\}_{m=0}^{100}$ and $L = 1000$, see figure 3.10. Table 3.1 displays estimates of the error probability $\Pr\left(\hat{\theta} \in \Theta_1 | \theta \in \mathcal{M}_2\right)$ for several values of the noise variance σ^2, obtained over 200 Monte Carlo runs. We see that even for small values of the noise variance the error probability is very high: the "naive MAP" estimator is biased toward the **simplest** model $\mathcal{M}_1$, even if the data clearly shows the existence of 3 different slopes. The ML overfitting behaviour is not observed and there is a bias of opposite sense that exhibits a strong preference for the simpler model $\mathcal{M}_1$.

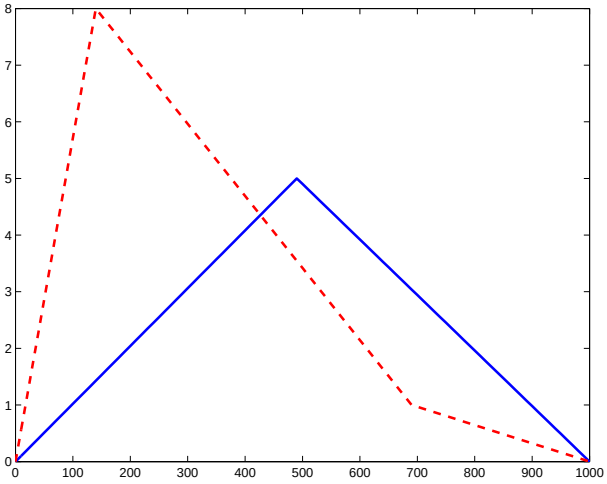

Figure 3.9: $f(\cdot|\theta_1)$ ($-$ blue), $f(\cdot|\theta_2)$ ($--$ red).

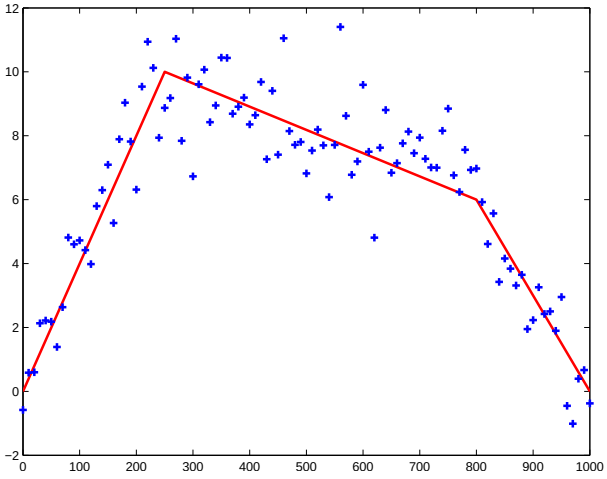

Figure 3.10: $Z \propto p(\cdot) \in \mathcal{M}_2, \quad \sigma^2 = 1.2$.

σ^2	0.6	1	1.2	1.6	2	4
$Pr(\tilde{\theta} \in \mathcal{M}_1 \vert \mathcal{M}_2)$	0	0.25	0.61	0.945	0.985	1

Table 3.1: $Pr(\tilde{\theta} \in \mathcal{M}_1 \vert \mathcal{M}_2)$, $\sigma^2 = (0.6, 1, 1.2, 1.6, 2, 4)$.

We now show that this preference for the simplest models is actually an artifact caused by the comparison of densities defined with respect to distinct measures. Let $\rho_i^2 = ||Y - f_i(X;\theta_i)||^2$ be the residuals for model $\mathcal{M}_i$, $i = (1,2)$. Criterion (3.27) leads to a model selection rule $r \underset{\mathcal{M}_2}{\overset{\mathcal{M}_1}{\lessgtr}} 1$ where r is the ratio

$$\begin{aligned} r &= \frac{\pi(\mathcal{M}_2)\, p_2(\theta_2|Y)}{\pi(\mathcal{M}_1)\, p_1(\theta_1|Y)} \\ &= \frac{\left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{\rho_2^2}{2\sigma^2}\right) L^{-2} \left(2\pi\Sigma^2\right)^{-1} \exp\left(-\frac{(P_y^1)^2+(P_y^2)^2}{2\Sigma^2}\right)}{(2\pi\sigma^2)^{-n/2} \exp\left(-\frac{\rho_1^2}{2\sigma^2}\right) L^{-1} \left(2\pi\Sigma^2\right)^{-1/2} \exp\left(-\frac{(P_y^0)^2}{2\Sigma^2}\right)} \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\Sigma L} \exp\left(\frac{\rho_1^2 - \rho_2^2}{2\sigma^2}\right) \exp\left(-\frac{(P_y^1)^2 + (P_y^2)^2 - (P_y^0)^2}{2\Sigma^2}\right) . \end{aligned}$$

If $\Sigma^2 \gg 1$, such that $\exp\left(-\frac{(P_y^1)^2+(P_y^2)^2-(P_y^0)^2}{2\Sigma^2}\right) \simeq 1$, then

$$\rho_1^2 - \rho_2^2 \underset{\mathcal{M}_2}{\overset{\mathcal{M}_1}{\lessgtr}} 2\sigma^2 \log\left(\Sigma L \sqrt{2\pi}\right) = \gamma.$$

Note that ΣL is the ratio of the normalizing constants of the prior densities over Θ_1 and Θ_2, that increases with the number of breakpoints, then larger for $\mathcal{M}_2$. As the previous equation shows, the decision region for $\mathcal{M}_1$ increases monotonically with ΣL, explaining why the apparently uninformative prior leads to a strong bias in favor of $\mathcal{M}_1$. This is the opposite behaviour of the ML criterion drawback.

This biased behaviour is entirely due to the fact that we are comparing the "densities" $p_k(\cdot|Z)$ defined with respect to measures μ_k (the Lebesgue measure in both cases) over spaces of distinct dimensions ($d_1 = 3$, $d_2 = 5$). Depending on the priors chosen, this may bias the decision, in an unclear manner, in favor of simpler or more complex models. We stress that these remarks do not concern penalized likelihood methodologies globally, but only their interpretation as Bayesian MAP estimators. Indeed penalized likelihood function is defined in the data space , with respect to to the same measure.

3.4.5 Bayesian model selection

We address now the second justification of penalized likelihood, that relates it to Bayesian Model Selection (BMS) which does not attempt at directly

selecting the model set *and* the parameter value, but instead starts by selecting the model $\mathcal{M}_k$ first using the posterior probabilities $Pr(\mathcal{M}_k|Z)$, $k \in \mathcal{K}$ [49]:

$$\hat{k} = \arg\max_{k\in\mathcal{K}} \int_{\Theta_k} p(\theta|Z)\ d\theta\ . \tag{3.28}$$

Determination of these posteriors requires specification of a priori distribution $\pi(\theta)$ defined over Θ. If $\pi(\theta)$ is of the form (3.24), $\forall k \in \mathcal{K}$, we have

$$Pr(\mathcal{M}_k|Z) = \frac{p(Z|\mathcal{M}_k)\pi(\mathcal{M}_k)}{\sum_{j\in\mathcal{K}} p(Z|\mathcal{M}_j)\pi(\mathcal{M}_j)}\ , \tag{3.29}$$

$$Pr(Z|\mathcal{M}_k) = \int_{\Theta_k} p(Z|\mathcal{M}_k,\theta_k)\,\pi(\theta_k|\mathcal{M}_k)d\theta_k\ . \tag{3.30}$$

We refer to [28] for discussion about the prior selection.

Once the model has been determined, the parameter $\theta \in \Theta_{\hat{k}}$ can be estimated using one of the standard statistical estimation criteria (MMSE, MAP,...). This (sound) estimation approach selects k by comparing the total posterior probability mass accumulated over $\mathcal{M}_k$. One may question whether this marginal approach can guarantee the aptitude of the elements of $\mathcal{M}_{\hat{k}}$, alone, to fit the data well. We will see below that in all the examples considered the models selected by BMS have fitting properties close to those obtained by penalized likelihood criteria.

More surprisingly, numerical studies presented below in section 3.5.5 show that their fitting performance in "signal-in-noise" problems is similar to direct MMSE "signal" estimation, which belongs to a set much richer class of models than $\mathcal{M}$.

3.4.6 Consistent MAP estimator

We present now a numerical implementation of the "Two-step MAP estimator",

$$(i)\quad \hat{k} = \arg\max_{k\in\mathcal{K}} Pr(\mathcal{M}_k|Z)\ , \tag{3.31}$$

$$(ii)\quad \hat{\theta} = \arg\max_{\theta\in\Theta_{\hat{k}}} p(\theta|Z,\mathcal{M}_{\hat{k}})\ . \tag{3.32}$$

that uses BMS to select $\mathcal{M}_k$ and MAP to identify a $\theta \in \Theta_{\hat{k}}$. In this manner, optimization is done in each step using commensurable score functions: the (discrete) posterior distribution over the families of models $\mathcal{M}_k$ in *(i)*, and a regular posterior density over $\Theta_{\hat{k}}$, with respect to a selected base measure, in *(ii)*. Note that [32] proposes a similar idea for ML estimation, selecting first the $\mathcal{M}_k$ using the BMS criterion and computing the ML estimates of the $\theta \in \Theta_k$ in a second step.

We here consider a prior π of the form (3.24),

$$\pi_\theta(\theta) = \sum_{k=k_{\min}}^{k_{\max}} \pi(\theta|\mathcal{M}_k)\ Pr(\mathcal{M}_k), \qquad \theta \in \Theta\ .$$

For most problems the posterior probabilities $Pr(\mathcal{M}_k|Z)$ have no closed-form and their maximum must be determined numerically. We obtain estimates $\hat{Pr}(\mathcal{M}_k|Z)$ by sampling from $p(\theta|Z), \theta \in \Theta$ using Reversible Jump Markov Chain Monte Carlo (RJMCMC) [24] and computing the total mass of each $\mathcal{M}_k$ as the corresponding marginals.

RJMCMC uses a proposal distribution $q(\theta'|\theta)$, $\theta' \in \Theta_{k'}$, $\theta \in \Theta_k$, where θ (resp. θ') is the current (resp. candidate) state of the chain, that is a mixture of basic transition distributions (birth, death or change) moving across neighboring families of models $\mathcal{M}_k$, $\mathcal{M}_{k-1}$ and $\mathcal{M}_{k+1}$. To ensure chain reversibility (and thus convergence in distribution to the target distribution $p(\theta|Z)$) the acceptance function of the chain is [24] $\alpha_{RJ}(\theta, \theta') = \min\{1, r_{RJ}\}$, with

$$r_{RJ} = \frac{p(\theta'|Z)}{p(\theta|Z)} \frac{q(\theta'|\theta)}{q(\theta|\theta')} J(\theta', \theta)\ , \tag{3.33}$$

where $J(\theta', \theta)$ is the Jacobian of the mapping from θ to θ'.

In our application the proposal is a mixture of basic transition distributions: birth (b) of a new knot, death (d) of a knot or change (c) of a knot location. This allows the chain to "jump" from one model $\mathcal{M}_k$ to its neighbors $\mathcal{M}_{k-1}$ and $\mathcal{M}_{k+1}$. Details about the implementation of RJMCMC are given below in section 3.5.2.

Finally, the family $\mathcal{M}_{\hat{k}}$ is chosen using the RJMCMC samples $\left(\theta^{(i)}_{k^{(i)}}\right)_{i=1}^{M} \propto p(\theta|Z)$ in the following manner

$$\hat{k} = \arg\max_k \hat{Pr}(\mathcal{M}_k|Z), \qquad \hat{Pr}(\mathcal{M}_k|Z) = \frac{M_k}{M}, \tag{3.34}$$

where $M_k = \#\left\{\left(\theta^{(i)}_{k^{(i)}}\right)_{i=1}^{M} : k^{(i)} = k\right\}$ and $\#\ A$ is the cardinality of set A.

Once the family $\mathcal{M}_{\hat{k}}$ has been determined, parameter estimation is done for $\theta \in \Theta_{\hat{k}}$. Again, there is, in general, no analytical solution, and we must resort to a numerical method to find the model with the maximal posterior density $p_{\hat{k}}(\theta|Z)$. A common choice for approximating the solution of this optimization problem is Simulated Annealing (SA), with an acceptance probability α_{SA}:

$$\alpha_{SA}(\theta'_{\hat{k}}, \theta_{\hat{k}}, T_i) = \min\left\{1; \left(\frac{p(\theta'_{\hat{k}}|Z, \mathcal{M}_{\hat{k}})}{p(\theta_{\hat{k}}|Z, \mathcal{M}_{\hat{k}})}\right)^{\frac{1}{T_i}}\right\}, \tag{3.35}$$

where $\theta_{\hat{k}}$ (resp. $\theta'_{\hat{k}}$) is the current (resp. candidate) state, and T_i is the chain temperature that must decrease according to a convenient cooling scheme. We refer the interested reader to [39] for details on SA.

3.5 Experimental results

In this section, we compare BIC to the two-step BMS/MAP semi-parametric identification described in section 3.4.6 for curve modeling with free-knot (cubic) splines. We begin with a brief description of the model and of numerical issues related to its optimization, presenting our comparative study in a second step.

We assume that the observations follow a normal model

$$p(Z|\theta, \mathcal{M}_k) = \mathcal{N}\left(f(t;\theta), \sigma^2 I\right), \quad f(t;\theta) = \sum_{i=1}^{k} \beta^i b_i(t, \xi_k) \ . \tag{3.36}$$

where k is the number of knots, $b_i(t, \xi_k)$ is the i^{th} B-Spline function, $\xi_k \in [0,1]^k$ is the (ordered) knots vector, and $\beta_k \in \mathbb{R}^{2k}$ is the vector of control points. We refer to [14] for details about splines. The parameter vector of $\mathcal{M}_k$ is $\theta_k = (\xi_k, \beta_k, \sigma^2)$.

We use the factored prior already proposed for this problem in [17], except for the prior over k which we consider uniform, hence establishing not a preference for simpler models:

$$\pi(\theta_k) = \pi(\beta_k|\mathcal{M}_k, \xi_k, \sigma^2)\, \pi(\xi_k|\mathcal{M}_k)\, \pi(\mathcal{M}_k)\, \pi(\sigma^2).$$

- $\pi(\mathcal{M}_k) = \mathcal{U}(k \in [k_{\min}, k_{\max}])$,
- $\pi(\xi_k|\mathcal{M}_k) \sim \mathcal{U}([0\ 1]^k)$,
- $\pi(\beta_k|\mathcal{M}_k, \xi_k, \sigma^2) = \mathcal{N}\left(0, \sigma^2 N (B^T B)^{-1}\right)$ where $B = B_{\hat{k},\xi}$ is the spline design matrix with entries $b_i(t, \xi_k)$,
- $\pi(\sigma^2) = 1/\sigma^2$.

3.5.1 Numerical issues for BIC

Once ξ determined, maximization of the likelihood allows analytic determination of the estimates of β_k and σ^2 for a fixed model order k. Using the reduced likelihood at this estimated values as the target distribution of the SA algorithm, we can find the ML estimate of the knot vector ξ_k. Temperature T_i is initialized at $T_0 = 50$ and decreases every 500 iterations by a factor of 5. The maximum number of iterations is fixed to 10000. We thus obtain

the ML estimate of the parameter vector (with k fixed). Adding the BIC penalty term defined in section 3.3.3 to the likelihood and maximizing the sum with respect to k allows determination of the BIC-penalized estimate of k.

3.5.2 Numerical issues for RJMCMC

As described in section 3.4.6, we first identify $\mathcal{M}_{\hat{k}}$, using RJMCMC to estimate $Pr(\mathcal{M}_k|Z)$. The prior given in (3.5) allows analytical expressions for MAP estimate of β_k and σ^2. Details about this are given in Appendix A. We thus show that k and ξ_k are the structural parameters of the model and the target distribution of RJMCMC reduces to

$$\mathcal{P}(\mathcal{M}_k, \xi|Z) = p(\mathcal{M}_k, \xi|Z)p(\hat{\sigma^2}_{k,\xi}|Z, \mathcal{M}_k, \xi)p(\hat{\beta}_{\hat{k},\xi}|Z, \mathcal{M}_k, \xi, \hat{\sigma^2}_{k,\xi}). \quad (3.37)$$

The proposal distribution $q(\theta'|\theta)$ reduces to $q(k', \xi'|k, \xi)$ and is taken as a mixture of basic transition laws that allow "jumps" between families of models: birth (b), death (d) and change (c) of a knot point in the knot vector ξ_k.

$$q(\theta'|\theta) = m(k^{cand}|k^{curr})p(\xi^{cand}|\xi^{curr}, k^{curr}) \ , \quad (3.38)$$

where $m(\cdot|\cdot)$ is the probability associated with each type of move and $p(\xi^{cand}|\xi^{curr}, k^{curr})$ is the probability to select the candidate knot vector.

We choose:

$$m(k^{cand}|k^{curr}) = \begin{cases} c \min\left\{1, \frac{\pi(\mathcal{M}_{k^{cand}})}{\pi(\mathcal{M}_{k^{curr}})}\right\} \\ \qquad \text{if } k^{cand} = k^{curr} - 1 \text{ or } k^{cand} = k^{curr} + 1, \\ 1 - m(k^{curr} + 1|k^{curr}) - m(k^{curr} - 1|k^{curr}) \\ \qquad \text{if } k^{cand} = k^{curr}, \\ 0 \quad \text{otherwise.} \end{cases} \quad (3.39)$$

The three types of move, addition, deletion or change of a knot location, completely determine the chain of the RJMCMC algorithm. The probability to choose a move depends on the parameter c in equation (3.39). We set $c = 0.2$ so that exploring a subspace $\mathcal{M}_k$ by changing a knot location is preferred on jumping to another subspace by either adding or deleting a knot.

Leaves us to define $p(\xi^{cand}|\xi^{curr}, k^{curr})$ and the Jacobian computation:

- When "change" move is selected, we change a knot of the current knot vector by choosing randomly $\xi_j \in \xi^k$ and we replace it by ξ_{j*}, a random knot in ξ_j neighborhood (between the previous ξ_{j-1} and the next knot ξ_{j+1} of the vector). Then, we have $k^{cand} = k^{curr}$, $J(s^{curr}, s^{cand}) = 1$

(there is no change of space dimension) and the selection function becomes:

$$q(\xi^{cand}|\xi^{curr}, k^{curr}) = \frac{1}{k^{curr} - 2m} \times \frac{1}{\xi_{j+1} - \xi_{j-1}}.$$

- For the "birth" move, we add a knot to the current knot vector by choosing a location ξ_{j*}, randomly in [0 1] and we place it in the knots vector. Then $k^{cand} = k^{curr} + 1$ and the selection function becomes:

$$\begin{aligned} q(\xi^{cand}|\xi^{curr}, k^{curr}) &= 1 \ , \\ J(s^{curr}, s^{cand}) &= \left| \frac{\partial(\xi^{k+1})}{\partial(\xi^k, \xi_{j*})} \right| = 1 \ . \end{aligned}$$

- For the "death" move, we delete a knot to the current knot vector by choosing a location ξ_{j*}, randomly in the knot vector. Then $k^{cand} = k^{curr} - 1$ and the selection function becomes:

$$\begin{aligned} q(\xi^{cand}|\xi^{curr}, k^{curr}) &= \frac{1}{k^{curr} - 2m} \ , \\ J(s^{curr}, s^{cand}) &= \left| \frac{\partial(\xi^k)}{\partial(\xi^{k-1}, \xi_{j*})} \right| = 1 \ . \end{aligned}$$

We performe $M = 10000$ iterations of the RJMCMC algorithm.

3.5.3 Numerical issues for SA

Fixing $k = \hat{k}$, the second step identifies the parameters of $\mathcal{M}_{\hat{k}}$ by maximizing the "local posterior density" $p(\theta|Z, \mathcal{M}_{\hat{k}}), \theta \in \Theta_{\hat{k}}$. Maximization with respect to $(\beta_{\hat{k}}, \sigma^2)$ can again be found analytically, see appendix A.

A SA algorithm is run with $\mathcal{P}(\xi_{\hat{k}}|Z, \mathcal{M}_{\hat{k}})$ as the score function, producing a sequence of values of $\xi_{\hat{k}}$ that converge in distribution to its maximum, completely identifying a single model amongst $\mathcal{M}$. We performe $L = 2000$ SA iterations. Temperature T_i is initialized at $T_0 = 0.02$, and is halved every 500 iterations.

3.5.4 BMS-MAP and BIC on simulated data

We first compare the two methods BIC and BMS/MAP on simulated data, with the goal of exposing the asymptotic nature of the BIC criterion, which only for very large data sets is an approximation of BMS, inheriting the problems of "naive Bayes" under its interpretation as a posterior computed for a particular "prior."

We simulated data from a spline model with $k = 13$ knots and $\sigma^2 = 0.04$ (see Fig. 3.11), and considered two observation sets: $\mathcal{D}_1$ with $N = 51$ data

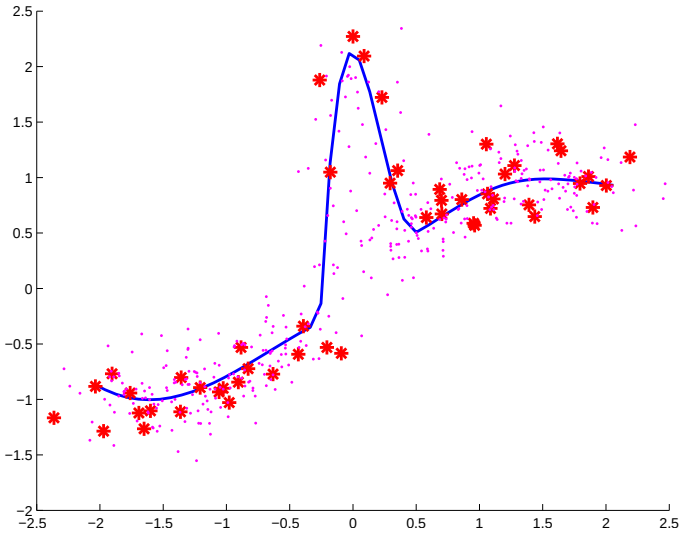

Figure 3.11: Spline model (-) and data sets $\mathcal{D}_1$ (*) and $\mathcal{D}_2$ ($\cdot$).

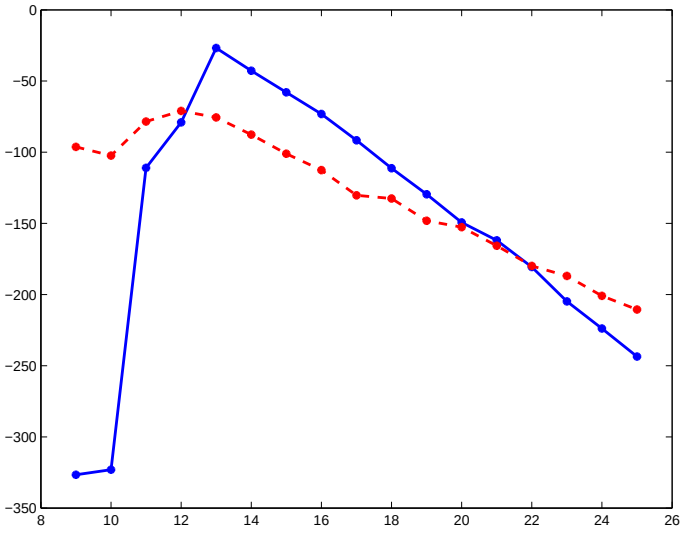

Figure 3.12: BIC criterion for $\mathcal{D}_1$ ($--$) and $\mathcal{D}_2$ ($-$) depending on the model order in abscissa.

points and $\mathcal{D}_2$ with $N = 401$ data points, see Figure 3.11. Figures 3.12 and 3.13 summarizes the comparison of the two methods.

For the shorter data set $\mathcal{D}_1$ BIC systematically chooses a 'wrong' model order $k = 12$ (Fig. 3.12), underestimating the data complexity, while BMS/MAP correctly identifies the true value $k = 13$ (Fig. 3.13). For the larger data set $\mathcal{D}_2$ both criteria choose the same (and correct) model order, confirming that only asymptotically BIC yields an unbiased estimate of the model com-

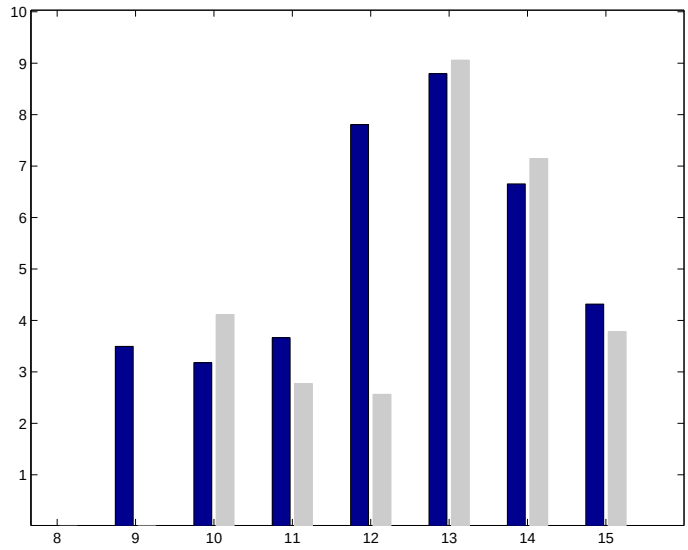

Figure 3.13: Log-posterior probability of models for $\mathcal{D}_1$ (black) and $\mathcal{D}_2$ (gray) with model order in abscissa.

	$\mathcal{D}_1$	$\mathcal{D}_2$
BIC	0.0597	0.0848
BMS/MAP	0.0620	0.0845

Table 3.2: Average (over 50 runs) of the mean square error between data and models.

plexity, while the two-step numerical estimator BMS/MAP has an unbiased behavior for all values of N. Table 3.2 shows the mean square error of the models identified by the two criteria, averaged over the 50 MC runs. We note here that while producing overall similar error figures, for the larger data set $\mathcal{D}_2$ the two models have indeed virtually identical residual error, while for $\mathcal{D}_1$ BIC yields a slightly smaller error, revealing its close relation to ML.

3.5.5 BMS-MAP and BARS on real data

BARS (Bayesian Adaptive Regression Splines) is a spline fitting method introduced in [17]. It allows to find a function that best fit the data points, this function being the average over splines obtained with a RJMCMC. It is thus no longer a sparse representation and we compare our BMS-MAP method to BARS. As a mixture of splines, BARS solution should provide a better fit than BMS-MAP with a single spline. We compare the results by

computing the mean square error, MSE:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left\{Z_i - \hat{f}(t_i, \hat{\theta})\right\}^2.$$

We present 2 examples:

- the first one is one of the *sin-exp* function used by DiMatteo in [17] and
- the second one is a benthic contour between a region of posidonia and a region of sand ripples extracted from a seabed Sonar image.

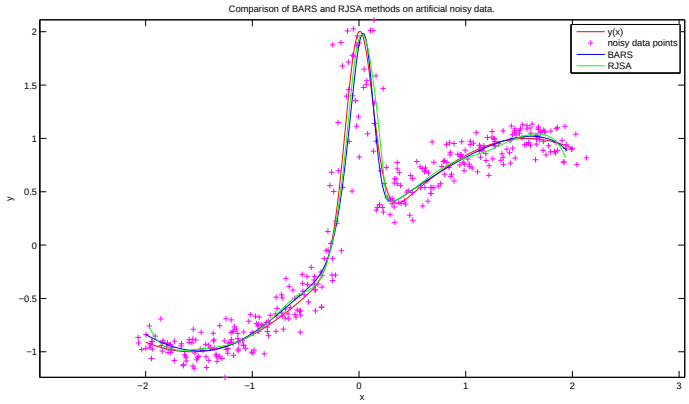

Figure 3.14: The red line is the *sin-exp* function and the magenta + are the noisy data points. The blue line is the result of the BARS method and the green line is the BMS-MAP result.

With figure 3.14, we can see that both methods produce the same visual results. Figure 3.15 displays the original contour of the second example and results of the BARS and the BMS-MAP methods.

We compare the results with their respective MSE in Table 3.5.5. For these 2 examples, MSE are in the same range and they are both equally good approximation of the original data. In addition to the results accuracy (shared with BARS), we can compare the resulting number of knots of BMS-MAP and the histogram of k for BARS samples. Figure 3.5.5 displays histograms of k for the first set of data (function+noise) on the left and the second set (contour) on the right. For the first set of data, the mode of the posterior of k with the BARS method is located for $k = 5$, while we obtain $\hat{k} = 9$ with BMS-MAP. For the second set of data, the mode of the posterior of k with the BARS method is located for $k = 10$ while we obtain $\hat{k} = 8$ with BMS-MAP. These results show that BMS-MAP method yield to similar accuracy than BARS but with a sparse model.

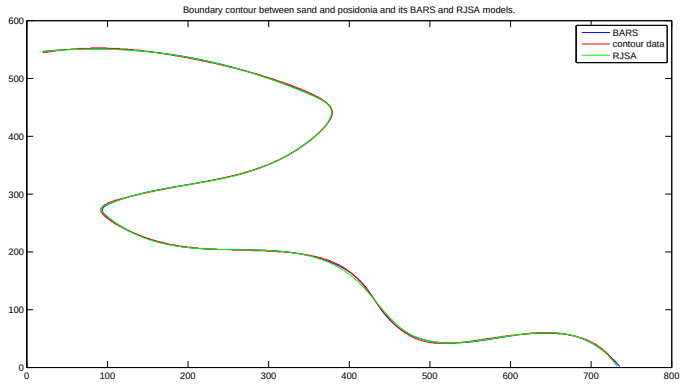

Figure 3.15: The red line is the detected boundary. The blue line is the result of the BARS method and the green line is the BMS method result.

MSE	$y(x)$	contour
BARS	0.0181	2.0
BMS-MAP	0.0174 ($\hat{k} = 9$)	4.3($\hat{k} = 8$)

Table 3.3: MSE for BARS and BMS-MAP are in the same range for these 2 examples.

3.5.6 BMS-MAP on real data

Posidonia clutter contour.

The contour of posidonia clutter example used to compare with BMS-MAP and BIC is shown in figure 3.17, together with the control points estimated in the 50 runs and an example of the resulting spline model. Figure 3.18 displays the knot vector in $[0, 1]$ for the 50 runs (y axis). We can see that there exists a small variability in knots estimation due to the complexity of the right part of the contour, while the knots corresponding to the left part, less complex, have a more stable identification. We can see this also in the 50 spline models shown in figure 3.19.

Sand ripple contour.

We consider here a sand ripple contour extracted from of side-scan sonar image. Figure 3.20 show results of 10 runs of BMS-MAP with the estimated knot vectors on the right and one example of histogram on the left. We note first that BMS always selects $\hat{k} = 9$ and that the knot estimate is very stable due to the simplicity of the contour. Figure 3.21 show the model given by the 10 runs and the contour itself with similar conclusions.

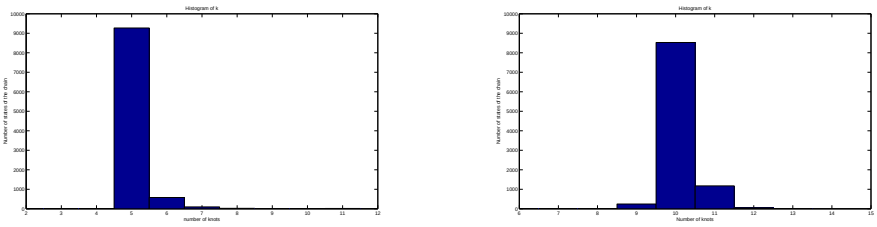

Figure 3.16: These figures are histograms of the number of knots for the first set of data (left) and for the contour data (right).

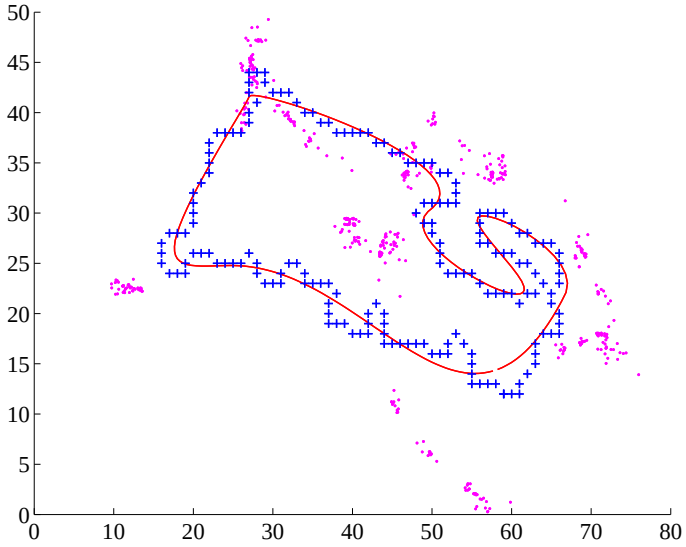

Figure 3.17: Posidonia clutter (blue data points) with the estimated spline model (red curve) and the 50 control points vector estimated (magenta).

Seahorse contour.

This example consider a more complex contour: a seahorse (it is a contour from a data base of fishes kindly provided to us by Prof. Srivastava). Figure 3.22 show one histogram resulting from BMS and selecting $\hat{k} = 25$. We run BMS-MAP 10 times and it always selects the same model complexity. Estimated knot vectors of the 10 runs are shown in the right of figure 3.22. We can see that even with this complex contour (compared to the sand ripple above), the resulting spline model is stable both for the knot vector and for the curve (see figure 3.23. A closer look at the variabilty between models is given in figure 3.24.

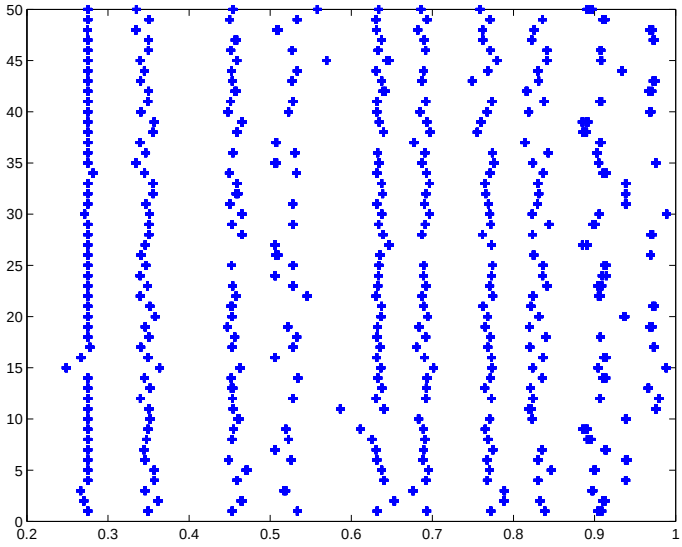

Figure 3.18: Posidonia clutter: estimated knots for the 50 runs in the y-axis.

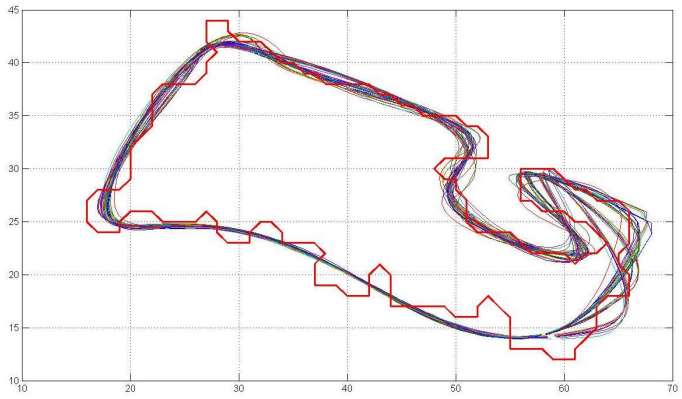

Figure 3.19: Posidonia clutter: the spline model estimated for the 50 runs.

3.6 Conclusion

In this chapter, we show that free-knot spline is a tool for curve modeling that allow to adjust the complexity of the data to the complexity of the spline model. We also stress a major misunderstanding of the MAP criterion when identification is performed in nested families of models. We thus recommend a two-steps approach: BMS-MAP. We show with some examples that BMS-MAP identify a sparse free-knots spline model with good fitting properties and thus is the representation corresponding to our requirements.

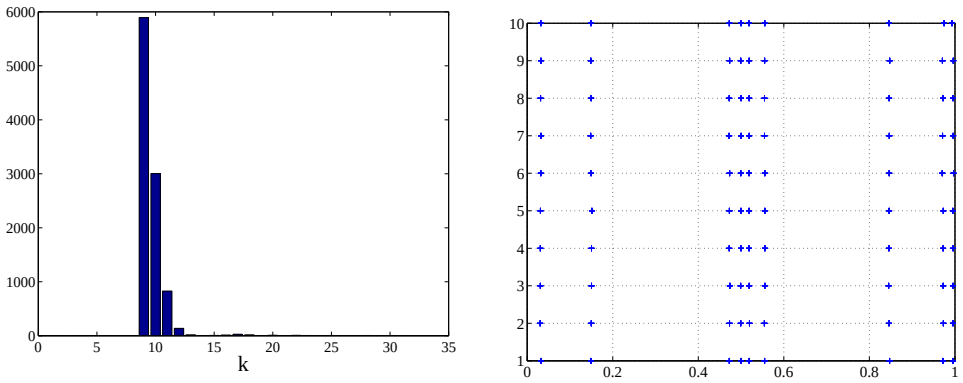

Figure 3.20: Sand ripple.

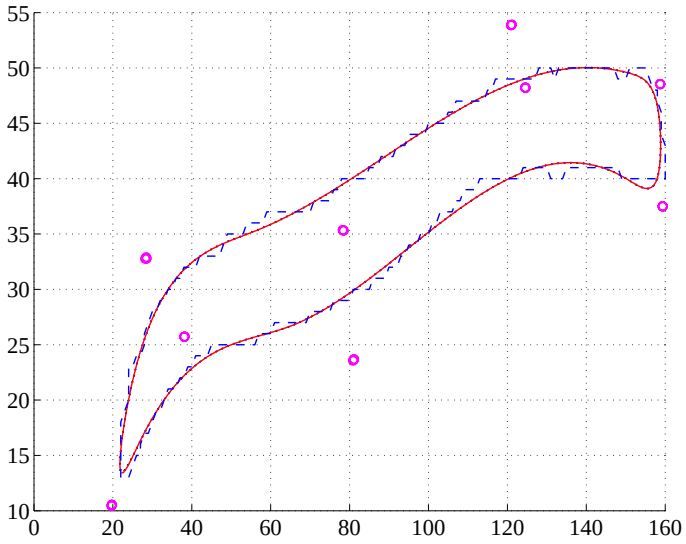

Figure 3.21: Sand ripple.

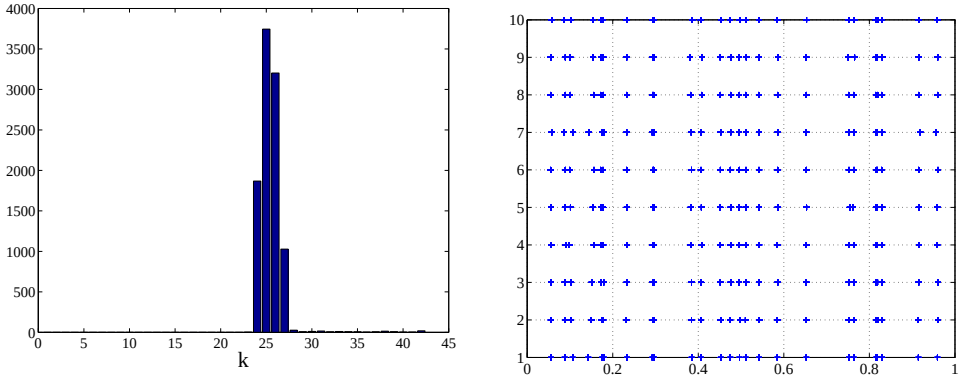

Figure 3.22: Seahorse contour with an example of BMS histogram (left) and results of knot vectors (right) over 10 runs of BMS-MAP (y axis).

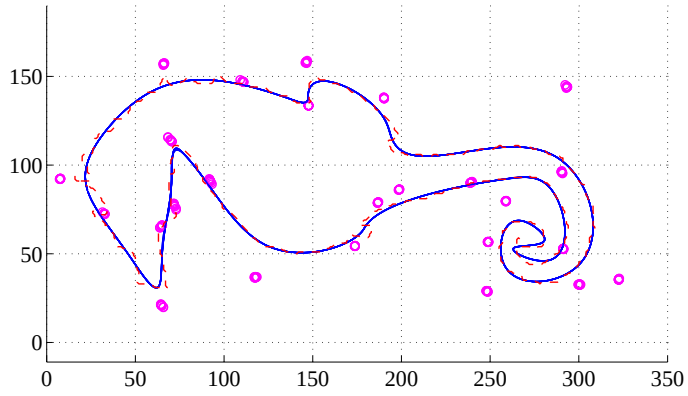

Figure 3.23: Seahorse contour: 10 BMS-MAP models with data in red, control points in magenta and splines in blue.

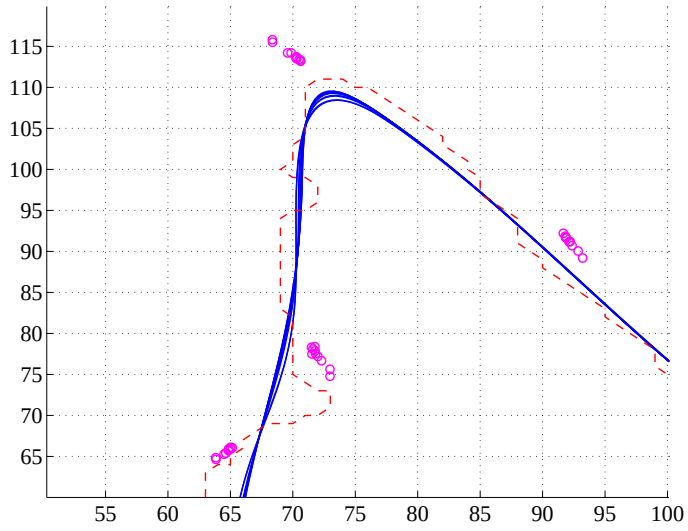

Figure 3.24: Seahorse contour: zoom of figure 3.23 over the space below the head of the sehorse.

Chapter 4

Collective spline modeling

4.1 Motivation and problem formulation

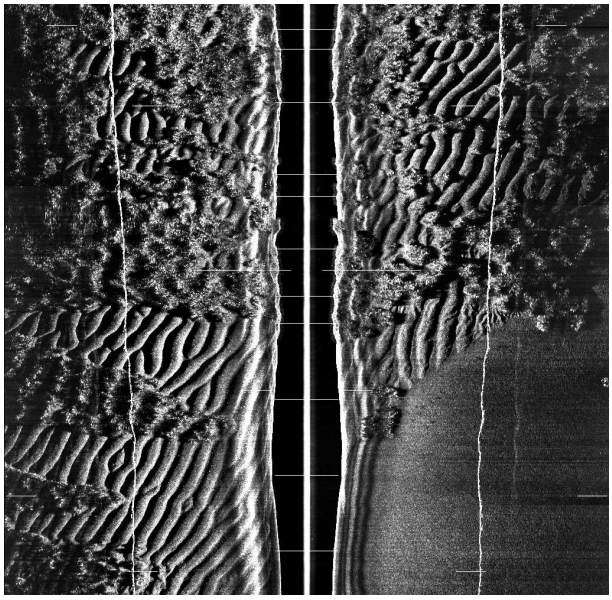

Figure 4.1: Example of side-scan sonar image.

Autonomous underwater robots, operating in natural environment, rely in their environment of objects present in the area to correctly plan their trajectory. In [42], the transition zones between distinct seabed regions are proposed as the features that must be used for safe navigation. The identified regions are those with different habitats such as sand ripples areas, posidonia area, ... An example of side-scan sonar image is given in figure 4.1 with sand ripples in the bottom left, flat sand in the bottom right, posidonia in the upper left and mixing of posidonia and sand ripples in the upper right. The goal is thus to characterize these distinct regions and to

characterize the different elements constituting the regions. In this chapter, we want to determine a model for a set of closed curves: the contours of objects (sand ripples, posidonia clutters, ...).

Let $\mathbb{L}_c$ be the space of continuous functions from $[0,1]$ to $\mathbb{R}^2$. Let $C = \{c_i(t)\}_{i=1}^n$ be a set of continuous planar closed curves: $c_i(t) \in \mathbb{L}_c$, $i = 1 \cdots n$. We choose a parametric family of models such that:

$$c_i(t) \sim \mu(\gamma) \ ,$$

with $\mu(\cdot)$ a probability measure in the curve space $\mathbb{L}_c$ and $\gamma \in \mathcal{G}$, the parameter vector.

Note that $c_i(t)$ is a closed curve that is defined with respect to a fixed origin $t = 0$. To model the set C, we must first align the curves with respect to a reference curve (the first one for example) so that they all have the same origin. This is a common normalization when modeling curves [36, 11]. Abusing the notation, we keep $C = \{c_i(t)\}_{i=1} n$ for aligned version of the curves.

We adopt a mixture model for the curves:

$$p(C) = \sum_{k \in \mathcal{K}} \omega_k p_k(C) \quad \sum_{k \in \mathcal{K}} \omega_k = 1 \ , \tag{4.1}$$

where $p_k(C)$ denotes the probability density function in $\boldsymbol{S}_m^k$ (definition 3.4.

Model determination for both the mixture weights and the curves is quite difficult and in this chapter, we restrict to simpler problem where each curve is represented by a single spline with fixed complexity. Thus the weights ω_k correspond to the ratio of the number of curves in C with complexity k, with the total number of curves:

$$\omega_k = \frac{n_k}{n} \ ,$$

where $n_k = \#\left\{c_j; \hat{k}_j = k, \ j = 1 \cdots n\right\}$. Note that $\hat{k}_j$ can be determined with a Bayesian model selection criterion such as the one defined in chapter 3. Then C is the union of subsets C_k of curves with complexity k. We are now left with the determination of a model for subset C_k. To simplify notation, we do not specify the subscript k in the following and C is a set of observed curves with fixed complexity spline model, and n is their number.

Based on section 3.3, $c_i(t)$ is modeled by a spline curve with control points β and knots ξ. We recall here the curve modeling with spline such that:

$$c_i(t) = s_i(t) + \epsilon_i(t) \quad i = 1 \cdots n \ ,$$

where $s_i(t)$ is the spline model with knots ξ^i and control points β^i, and $\epsilon^i(t)$ is the approximation noise function. The corresponding matrix notation where Z^i is the column vector of observed samples of $c_i(t)$ is:

$$Z_i = B^i\beta^i + \varepsilon^i \quad i = 1 \cdots n \ , \tag{4.2}$$

where B^i is the spline design matrix with knots ξ^i and ε^i is the noise vector for both the approximation and the observation noises. We choose a fixed order spline, $m = 4$ (cubic splines), with fixed number of knots k. We thus restrict the model space to $\boldsymbol{S}_m^k$ given in definition 3.4.

We assume a Gaussian model for the noise in the curve samples in equation (4.2):

$$\varepsilon^i \sim \mathcal{N}\left(0, \sigma^2\mathbf{I}_{N_i}\right) \ , \tag{4.3}$$

where $\sigma^2\mathbf{I}_{N_i}$ assumes independent realizations of the noise with similar variance σ^2.

The closed spline curve depends on knots intervals, noted Δ_j^i, as they are defined circularly. As the curves are aligned we propose to fix the first knot position ξ_1^i so that we fix the parametrization of the curve. We choose to set the first knot position to $t = 0$, in other words, $\xi_1^i = 0$.

Definition (4.1): Knot intervals

The knot intervals vector, noted $\Delta^i = \left(\Delta_1^i\Delta_2^i \cdots \Delta_k^i\right)$, is such that:

$$\Delta_j^i = \xi_{j+1}^i - \xi_j^i, \quad j = 1 \cdots (k-1) \text{ and } \Delta_k^i = 1 - \xi_k^i \ . \tag{4.4}$$

Definition (4.2): $(k-1)$-dimensional simplex

The simplex in $(k-1)$ dimensions is the subset of $\mathbb{R}^k$ such that:

$$\mathcal{D}^{(k-1)} = \left\{ (\Delta_1, \cdots, \Delta_k) \in \mathbb{R}^k | \sum_{j=1}^{k} \Delta_j = 1, \, \Delta_j > 0 \, \forall j = 1 \cdots k \right\} \ . \tag{4.5}$$

As $\xi_1^i = 0$ is fixed, $\left\{\Delta_j^i\right\}_{j=1}^k$ uniquely defines a knot vector and determining a model for ξ^i is equivalent to determining a model for Δ^i. Δ^i is an element of the $(k-1)$-dimensional simplex, $\mathcal{D}^{k-1}$ (definition 4.2), and defines a partition of the unity: i.e. $\sum_{j=1}^k \Delta_j^i = 1$. Intervals can thus be interpreted as determining a probability law and we choose a parametric model for Δ^i: a Dirichlet distribution with unknown parameter $\alpha = (\alpha_1 \cdots \alpha_k)$ ($\alpha_j > 0$, $j = 1 \cdots k$). Since there is a one-to-one correspondence between intervals and knots, we make abuse of notation and use ξ instead of Δ in the following.

Definition (4.3): Dirichlet distribution

The Dirichlet distribution is

$$\mathcal{D}(\xi^i|\alpha) = \frac{1}{B(\alpha)} \prod_{j=1}^{k} \left(\Delta_j^i\right)^{\alpha_j - 1} \quad , \tag{4.6}$$

where $B(\alpha) = \frac{\prod_{j=1}^{k} \Gamma(\alpha_j)}{\Gamma\left(\sum_{j=1}^{k} \alpha_j\right)}$ is the normalizing constant, and $\Gamma(\cdot)$ is the standard Gamma function.
The mode of $\mathcal{D}(\xi^i|\alpha)$ is $\left(\frac{\alpha_1}{\bar{\alpha}}, \frac{\alpha_2}{\bar{\alpha}}, \cdots, \frac{\alpha_k}{\bar{\alpha}}\right)$ with $\bar{\alpha} = \sum_{j=1}^{k} \alpha_j$.

Note that the parameter α determines the variability of the distribution:

- $\alpha = (1 \cdots 1)$ defines a uniform distribution over $\mathcal{D}^{k-1}$,
- $\alpha_j \to 0$ defines a distribution with large probability centered around zero and one. Thus, if ξ_{j-1} is fixed, ξ_j is in its close neighborhood which tends to produce a discontinuity in the spline curve,
- large values for $\{\alpha_j\}_{j=1}^{k}$ defines a highly peaked distribution (small variance) around its mode.

For the control points, we choose a Gaussian model with unknown mean vector noted μ_0 and a covariance matrix noted Σ^i. Similarly to [17], we choose $\Sigma^i = \sigma^2 N_i \left(B^{iT} B^i\right)^{-1}$. It is the inverse of the averaged Fisher information matrix evaluated for the N_i samples of the curve c_i and thus reflects the amount of knowledge about the control points given by the observed samples. The averaged Fisher Information matrix is:

$$\bar{\mathcal{I}}(\beta^i) = -\frac{1}{N_i} \frac{\partial^2 \log p(Z^i|w_i)}{\partial \beta^i \, \partial \beta^{iH}} = (N_i \sigma^2)^{-1} B^{iT} B^i \quad . \tag{4.7}$$

So the parameter vector of our model is:

$$\gamma = \left(\mu_0, \alpha, \sigma^2\right) \quad , \tag{4.8}$$

and the parameter space is:

$$\mathcal{G} = \mathbb{C}^k \times \mathcal{D}^{k-1} \times \mathbb{R}_*^+ \quad . \tag{4.9}$$

We can interpret the model parameters from the point of view of shapes:

- μ_0 may be the mean shape of the control points $\left\{\beta^i\right\}_{i=1}^{n}$, determining the coarse mean shape of curves,
- α allows to determine variability in knot vector $\left\{\xi^i\right\}_{i=1}^{n}$ and thus refines the shape of curves; moreover the mode $\left(\frac{\alpha_1}{\bar{\alpha}}, \frac{\alpha_2}{\bar{\alpha}}, \cdots, \frac{\alpha_k}{\bar{\alpha}}\right)$ of the Dirichlet distribution can be seen as the mean knot vector.

To identify the model we choose a criterion that maximizes the likelihood of the data and:

$$\hat{\gamma} = \arg\max_{\gamma \in \mathcal{G}} p(C|\gamma) \ . \tag{4.10}$$

Note that we could choose a Bayesian criterion by specifying a prior on $\gamma \in \mathcal{G}$ so as to maximize the posterior $p(\gamma|C)$ for example.

Figure 4.2 shows the hierarchical model we adopt and the important expressions of the model are:

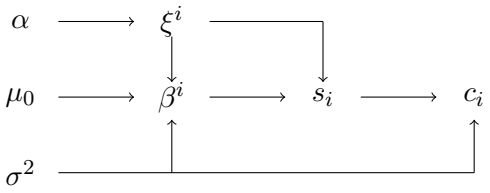

Figure 4.2: Hierarchical design of our curve model showing the dependencies.

Noise model:

$$p(Z^i|\beta^i, \xi^i, \gamma) = \left(2\pi\sigma^2\right)^{-N_i} \exp\left(-\frac{(Z^i - B^i\beta^i)^H (Z^i - B^i\beta^i)}{2\sigma^2}\right) \ , \tag{4.11}$$

Spline parameters models:

$$\begin{aligned} \xi^i, \beta^i|\gamma &\sim Dirichlet(\xi^i|\alpha) \times \mathcal{N}\left(\beta^i|\mu_0, \Sigma^i\right) \\ p(\xi^i, \beta^i|\gamma) &= \frac{1}{B(\alpha)} \prod_{j=1}^{k} (\Delta_j^i)^{(\alpha_j - 1)} \times (2\pi\sigma^2 N_i)^{-k} |B^{iT} B^i| \times \\ &\quad \exp\left(-\frac{(\beta^i - \mu_0)^H B^{iT} B^i (\beta^i - \mu_0)}{2\sigma^2 N_i}\right) \ . \end{aligned} \tag{4.12}$$

Definition (4.4): Curved exponential family density

Density $f(x;\gamma)$ belongs to the curved exponential family if and only if $\exists$ functions, $h(x)$, $S(x)$, $\Psi(\gamma)$ and $\Phi(\gamma)$ such that:

$$f(x;\theta) = h(x) \exp\left\{-\Psi(\gamma) + \langle S(x), \Phi(\gamma)\rangle\right\} \ , \tag{4.13}$$

where $< \cdot, \cdot >$ denotes the standard Euclidean inner product in $\mathbb{R}^d$, d being the dimension of $\Phi(\gamma)$. Note that $d \geq \dim(\gamma)$. $S(x)$ is called the sufficient statistics.

Let $w_i = \left(\beta^i, \xi^i\right)$. Note that the model $p(Z^i, w_i|\gamma) = p(Z^i|w_i, \gamma)p(w_i|\gamma)$ belongs to the exponential family with:

$$p(Z^i, w_i|\gamma) = h(Z^i, w_i) \exp\left(-\Psi(\gamma) + \left\langle S(Z^i, w_i), \Phi(\gamma)\right\rangle\right) \ , \tag{4.14}$$

with

$$h(Z^i, w_i) = N_i|B^{iT}B^i| \ ,$$

$$\Psi(\gamma) = (N_i + k)\log\left(2\pi\sigma^2\right) + \log B(\alpha) \ , \tag{4.15}$$

$$S(Z^i, w_i) = \begin{bmatrix} ln\Delta^i \\ (Z^i - B^i\beta^i)^H(Z^i - B^i\beta^i) + \frac{\beta^{iH}B^{iT}B^i\beta^i}{N_i} \\ \Re\left(\frac{B^{iT}B^i}{N_i}\beta^i\right) \\ \Im\left(\frac{B^{iT}B^i}{N_i}\beta^i\right) \\ \frac{\mathrm{Vec}\left(B^{iT}B^i\right)}{N_i} \end{bmatrix} \ , \tag{4.16}$$

$$\Phi(\theta) = \begin{bmatrix} \alpha - 1 \\ -\frac{1}{2\sigma^2} \\ \Re\left(\frac{\mu_0}{\sigma^2}\right) \\ \Im\left(\frac{\mu_0}{\sigma^2}\right) \\ -\frac{\mathrm{Vec}\left(\mu_0^*\mu_0^T\right)}{2\sigma^2} \end{bmatrix} \ . \tag{4.17}$$

$S(Z^i, w_i)$ is the sufficient statistics of the model. $\Re(\cdot)$ (resp. $\Im(\cdot)$) denotes the real (resp. imaginary) part of its argument. $\mathrm{Vec}\,(A)$ is defined in definition 2.2.

4.2 Model identification

With the model detailed above, we want to identify a model for the set of curves C. But if n, the number of curves, is large, the identification of the model may be computationally expensive. As we may observe new curves that will be added to the set along the identification, the modeling problem should be performed sequentially so that parameters γ are updated with each new observation. This can be summarized in the schematic algorithm:

Initialization: Set $i = 1$,

Update: $c_i(t)$ is observed and:

$$\hat{\gamma}^{(i)} = \arg\max_{\gamma \in \mathcal{G}} p(\{c_j\}_{j=1}^{i} | \gamma) \ ,$$

Iteration: loop until the set of observed curves is completely processed.

4.2.1 Unwanted parameters

For the identification of the model detailed in Figure 4.2, we distinguish two different approaches:

- a straightforward approach consists in first estimating the hidden parameters β^i and ξ^i for all $i = 1 \cdots n$, followed by an estimation of parameter γ,
- a marginal likelihood approach where hidden parameters are marginalized out.

The first approach would require to process each curve at a time to find its spline model following one of the methods described in chapter 3. Once all spline parameters are estimated, another estimation should be done to identify the model and $\gamma \in \mathcal{G}$. Then, for $i = 1 \cdots n$:

$$\begin{aligned} \left(\hat{\beta}^i, \hat{\xi}^i\right) &= \arg \max_{\beta \in \mathbb{C}^k, \xi \in \mathcal{D}^{k-1}} p(Z^i | \beta, \xi, \gamma) \ , \\ \hat{\gamma} &= \arg\max_{\gamma \in \mathcal{G}} p(\left\{\hat{\beta}^i, \hat{\xi}^i\right\}_{i=1}^{n} | \gamma) \ , \end{aligned}$$

This method is computationally demanding as each iteration corresponds to spline modeling in $\boldsymbol{S}_m^k$.

The second approach consists in integrating the likelihood with respect to hidden variables $\left(\beta^i, \xi^i\right)$ which corresponds to averaging over them:

$$\hat{\gamma} = \arg\max_{\gamma \in \mathcal{G}} \int_{\mathcal{W}^i} p(Z^i, w_i | \gamma) dw_i \ . \tag{4.18}$$

Generally the integration and the maximization are untractable and we must use numerical methods to identify $\hat{\gamma}$. The Expectation-Maximization (EM) method is one of them.

4.2.2 EM algorithm

In this section we briefly review the EM algorithm and some of its variants we use in the next section for the collective spline modeling problem.

Note that for simplicity we conserve the notation given above for generalities about EM algorithm: Z is the observed data set, W are the unavailable variables and γ is the model parameter. The data likelihood is $p(Z|\gamma)$ and the complete data likelihood, taking into account the unobserved variables, is $p(Z, W|\gamma)$.

The EM algorithm [15] has been developed to identify models with missing variables with a marginalization approach. This algorithm has proved to be useful in a multitude of problems such has the identification of Gaussian mixture densities (e.g. [20]) where both the parameters of the distributions and weights of the mixture are unknown. It is a two steps iterative process where the observed data likelihood is increased at each iteration guaranteeing the (local) convergence of the algorithm as it reaches a bound of the function.

Expectation (E) step: the expectation of the log-likelihood of the complete data is computed, noted $Q(\gamma|\gamma^{(t)})$ where $\gamma^{(t)}$ is the parameter estimated at the previous iteration:

$$Q(\gamma|\gamma^{(t)}) = \mathrm{E}_W \left[\log p(Z, W|\gamma)|Z, \gamma^{(t)}\right] \ , \tag{4.19}$$

where E denotes the expectation operator and t denotes the iteration number.

Maximization (M) step: the maximizer of $Q(\cdot|\gamma^{(t)})$ is evaluated and becomes the estimate of the parameter at iteration $(t+1)$, $\gamma^{(t+1)}$:

$$\gamma^{(t+1)} = \arg\max_{\gamma \in \mathcal{G}} Q(\gamma|\gamma^{(t)}) \ . \tag{4.20}$$

Loop until: the process is iterated until a convergence criterion is met such as no major change in the parameter estimates.

When the complete data likelihood $p(Z, W|\gamma)$ belongs to the exponential family, as it is our case, the (E) step in EM algorithm reduces to:

$$\bar{s}(Z, \gamma) = \mathrm{E}[S(Z, W)|Z, \gamma] \ , \tag{4.21}$$

$S(Z, W)$ given in equation (4.16).

This standard algorithm suffers some major drawbacks:

- the estimate depends on initialization $\gamma^{(0)}$ of the algorithm as the algorithm converges to a local maximum of $p(Z|\gamma)$, not to the global maximum,
- the convergence may be very slow.

The EM algorithm is useful when both the computations of the expected value in (4.19) and the maximum in (4.20) are analytically tractable. This is not the case in general, although for simple problem such as the identification of mixture of Gaussian densities [15]. In all other cases one must resort on numerical approximations. Some authors propose variants of EM algorithm:

- Monte-Carlo EM allows to approximate the expectation in (4.19),
- Metropolis EM allows to find numerically the maximum in (4.20).

Monte-Carlo EM.

When $Q(\gamma, \gamma^{(t)})$ has no closed form, an alternative is to use an empirical approximation. Wei and Tanner in [50] propose to use a Monte-Carlo method to approximate it, called MCEM (Monte-Carlo EM):

Monte-Carlo step:

$$W^m \sim p(W|\gamma^{(t)}, Z), \quad \text{for } m = 1 \cdots M \ , \tag{4.22}$$

(E) step:

$$Q(\gamma, \gamma^{(t)}) \approx \frac{1}{M} \sum_{m=1}^{M} \log p(W^m, Z|\gamma) \ . \tag{4.23}$$

Then, the standard maximization step is performed.

The sampling step of equation (4.22) is in general not an easy task and one must use numerical method to sample from the conditionnal posterior, such as Metropolis-Hastings (MH) method.

Note that if $M = 1$, then MCEM is equivalent to a method presented in [10], the Stochastic EM (SEM). If $M \to +\infty$, then, by the law of large numbers, MCEM is equivalent to EM. The quality of the approximation depends on the number of samples, M, and the authors propose to use an increasing scheme for M. Then, MCEM can be interpreted as a Simulated Annealing algorithm with M equivalent to the inverse of the temperature.

In [21], author proves the convergence of MCEM to a local maxima in the case of complete likelihood function being an element of the exponential family. They also give some results on the rate of convergence and recommend to use a polynomial scheme for the increase of M. We do not explore this point here, but we retain their conclusions and:

$$M_t = t^\rho \quad \rho > 1 \tag{4.24}$$

In practice, they recommend to use: $M_t = M_0 + t^{1.2}$.

In our case, the computation of the expectation is only partly untractable: $W = (\beta, \xi)$ and the computation of the expectation with respect to to β is

tractable (details are in appendix B), while the computation with respect to ξ is untractable. We thus require to use a MCEM approach to approximate the expectation in the (E) step. As our complete data likelihood belongs to the exponential family, based on results in [21], our MCEM application converges to a local maximum of the observed data likelihood.

Metropolis EM.

To deal with models where both (E) and (M) steps are untractable, [22] proposes an approach called Metropolis EM (MEM) based on MCEM with a Metropolis-type step. While MCEM allows to compute an approximation of $Q(\gamma, \gamma^{(t)})$, the Metropolis-type step is building a Markov chain with the parameters estimated at each iteration:

MC step: First step follows MCEM detailed in (4.23) and draw M samples $\{W_m\}_{m=1}^{M}$,

Metropolis update: A parameter γ' is simulated from a proposal $\mathcal{P}_t(\gamma^{(t-1)}, \gamma)$, and is acceptated as parameter estimate with probability $\mathcal{A}_t$:

$$\begin{aligned} \mathcal{A}_t(\gamma^{(t-1)}, \gamma', M) &= \min\left\{1, r_t(\gamma^{(t-1)}, \gamma', M)\right\} \ , \\ r_t(\gamma^{(t-1)}, \gamma', M) &= \exp\left(M(Q(\gamma', \gamma^{(t-1)}) - Q(\gamma^{(t-1)}, \gamma^{(t-1)}))\right) \ . \end{aligned}$$

The Metropolis step allows updates of the parameter γ that do not increase the likelihood function and thus this algorithm may escape from local minima. As detailed in [22], the MEM algorithm produces a sequence of parameters which distribution converges to the set of global maxima of the marginal likelihood, under the condition that the number of samples, M, increases with t.

In our case, the maximization is untractable but as our parameter vector γ is not very complex (small vector size: $3k + 1$, and standard parameter space $\mathbb{C}^k, \mathcal{D}^{k-1}, \mathbb{R}_*^+$), and as the maximization can be separated into two independent maximizations (detailed in appendix B), we prefer to adopt a Gradient approach for the maximization step.

Sequential EM.

In all EM procedures exposed above, we note that the observed data are entirely used at each iteration.As we discuss in the introduction of this section, it would be preferred to update the estimate of the parameter sequentially as the observations (curves) are acquired.

In the literature, [46] proposes to update the parameter estimate after each new observation, the iteration being thus equivalent to the number of processed data. Let $c_i(t)$ be the curve observed at time i with its samples Z^i. Let $\mathbf{S}\left(Z^i, \gamma\right) = \frac{\partial \log p(Z^i|\gamma)}{\partial \gamma_i}$, $\gamma = (\gamma_1 \cdots \gamma_p)$ be the score vector.

Each parameter update involves a stochastic gradient approach requiring the computation of the Fisher Information Matrix $\mathcal{I}^{-1}(\gamma)$ corresponding to one observation:

$$\mathcal{I}^{-1}(\gamma) = \mathrm{E}_\gamma \left[\mathbf{S}\left(Z^i, \gamma\right) \mathbf{S}^T \left(Z^i, \gamma\right) \right] \ .$$

Then,

$$\tilde{\gamma}^{(i)} = \tilde{\gamma}^{(i-1)} + \left(i\mathcal{I}^{-1}(\gamma^{(i)}) \right)^{-1} \mathbf{S}\left(Z^i, \tilde{\gamma}^{(i-1)}\right) \ . \tag{4.25}$$

This approach is a general sequential setting and its use in EM algorithm is only one of its applications. This method is computationally expensive as each iteration require the computation of the inverse of the matrix $\mathcal{I}(\gamma)$. The major drawback of the recursive approach is that the estimated parameter may be outside the definition space: $\tilde{\gamma}^{(t)} \notin \mathcal{G}$. For example, when $\mathcal{G}$ is not a vector space, imposing for example some constraints on the parameters, the parameter update in equation (4.25) may leads to a parameter that do not fit the constaints and thus outside the parameter space. Moreover, the evaluation of $\mathcal{I}(\gamma)$ is not possible here as we do not have access to the marginal likelihood $p(Z|\gamma)$.

The online EM proposed in [9] follows the idea of sequentially processing the data and overcomes the drawbacks of Titterington's recursive approach. Instead of treating the (M) step as the one to update at each new data, the sequential feature arises by the mean of stochastically approximating the expectation value in the traditional (E) step. This is the online (E) step:

$$\tilde{Q}_{i+1}(\gamma) = \tilde{Q}_i(\gamma) + \eta_{i+1} \left(\mathrm{E}\left[\log p(Z^{(i+1)}, w_{(i+1)} | \gamma^{(i)}, Z^{(i+1)}) \right] - \tilde{Q}_i(\gamma) \right) \ , \tag{4.26}$$

where $\gamma^{(i)}$ is the parameter estimated at the previous step i.

In the special case of exponential family, the online (E) step 4.26 reduces to:

$$\hat{s}_{i+1} = \hat{s}_i + \eta_{i+1} \left(\bar{s}(Z_{i+1}, \hat{\gamma}_i) - \hat{s}_i \right) \ . \tag{4.27}$$

The choice of the step size η_i is discussed in [9] and the authors recommend to use $\eta_i = \eta_0 i^{-\kappa}$, $\kappa \in]1/2, 1[$ and $\eta_0 \in [0, 1]$. In practice, they recommend to put $\eta_0 = 1$ and $\kappa = 0.6$.

This method is thus more related to the incremental version of EM introduced in [38]. The online EM has the advantage of not having to compute the inverse Fisher Information Matrix of the sequential approach. The maximization step in equation (4.20) is then performed within the defined parameter space. If it is not tractable then we must use a numerical approach to identify the maximizer.

4.3 Monte-Carlo Online EM

We propose in this section a new variant of the EM algorithm, the Monte-Carlo Online EM, as the solution of the problem exposed at the beginning of this chapter: collectively modeling a set of curves with splines. Numerical results of this method are given in the next section of this chapter.

Firstly, MC online EM algorithm processes the data curves sequentially. Let $\hat{\gamma}_i$ be the parameter estimated at iteration (i). We describe here the $(i+1)$-th iteration of the algorithm corresponding to the curve $c_{i+1}(t)$ and its sampled version Z^{i+1}:

Computation of $\bar{s}(Z^{i+1}, \hat{\gamma}_i)$:

$$\begin{aligned} \bar{s}(Z^{i+1}, \hat{\gamma}_i) &= \mathrm{E}_{w_{i+1}}\left[S(Z^{i+1}, w_{i+1}) | Z^{i+1}, \hat{\gamma}_i\right] \ , \\ &= \int_{\mathcal{W}_{i+1}} S(Z^{i+1}, w_{i+1}) p(w_{i+1} | Z^{i+1}, \hat{\gamma}_i) dw_{i+1} \ . \end{aligned} \tag{4.28}$$

As the integration is untractable, we use a Monte-Carlo method to approximate $\bar{s}(Z^{i+1}, \hat{\gamma}_i)$:

- Let $M^i = M_0 + i^{1.2}$ (equation (4.24)),
- Let $\left\{w_{i+1}^j\right\}_{j=1}^{M^i}$ be M^i samples of $p(w_{i+1}|Z^{i+1}, \hat{\gamma}_i)$,
- Then,

$$\bar{s}(Z^{i+1}, \hat{\gamma}_i) \approx \frac{1}{M^i} \sum_{j=1}^{M^i} S(Z^{i+1}, w_{i+1}^j) \ .$$

Online step:

$$\hat{s}_{i+1} = \hat{s}_i + \eta_i \left(\bar{s}(Z^{i+1}, \hat{\gamma}_i) - \hat{s}_i\right) \ , \tag{4.29}$$

with $\eta_i = i^{-0.6}$.

Maximization step: Let $\ell(s, \gamma) \triangleq -\Psi(\gamma) + \langle s, \Phi(\gamma) \rangle$, $\Psi(\cdot)$ and $\Phi(\cdot)$ given in equations (4.15) and (4.17). Then,

$$\hat{\gamma}_{i+1} = \arg\max_{\gamma \in \mathcal{G}} \ell(\hat{s}_{i+1}, \gamma) \ . \tag{4.30}$$

The maximization is performed numerically with a gradient approach.

To simplify the notation, we remove the index corresponding to the iteration, $i+1$ and i, in the next section where the computation of $\bar{s}(Z^{i+1}, \hat{\gamma}_i)$ is detailed.

4.3.1 Computation of $\bar{s}(Z,\gamma)$

We consider the general problem of computing $\bar{s}(Z,\gamma) = \mathrm{E}_w\left[S(Z,w)|Z,\gamma\right]$.

$$\begin{aligned}
\bar{s}(Z,\gamma) &= \int_{\mathcal{W}} S(Z,w)p(w|Z,\gamma)dw \ , \\
&= \frac{1}{p(Z|\gamma)}\int_{\mathcal{W}} S(Z,w)p(Z|w,\gamma)p(w|\gamma)dw \ , \\
&= \frac{1}{p(Z|\gamma)}\iint_{\mathbb{C}^k\times\mathcal{D}^{k-1}} S(Z,\beta,\xi)p(Z|\beta,\xi,\gamma)p(\beta|\gamma,\xi)p(\xi|\gamma)d\beta d\xi \ .
\end{aligned}$$

The first integral with respect to β is tractable (see appendix B) and leads to:

$$\bar{s}(Z,\gamma) = \begin{bmatrix}
\int_{\mathcal{D}^{k-1}} \log\Delta^i q(\xi|Z,\gamma)d\xi \ , \\
Z^HZ + k\sigma^2 - 2\Re\left(\int_{\mathcal{D}^{k-1}} Z^HB\varphi q(\xi|Z,\gamma)d\xi\right) + \cdots \\
\quad \cdots + \frac{N+1}{N}\int_{\mathcal{D}^{k-1}} \varphi^HB^TB\varphi q(\xi|Z,\gamma)d\xi \ , \\
\Re\left(\int_{\mathcal{D}^{k-1}} \frac{B^TB}{N}\varphi q(\xi|Z,\gamma)d\xi\right) \ , \\
\Im\left(\int_{\mathcal{D}^{k-1}} \frac{B^TB}{N}\varphi q(\xi|Z,\gamma)d\xi\right) \ , \\
\int_{\mathcal{D}^{k-1}} \mathrm{Vec}\left(B^TB\right) q(\xi|Z,\gamma)d\xi \ ,
\end{bmatrix} \tag{4.31}$$

where

- $q(\xi|Z,\gamma) = \frac{\exp\left[-\frac{T(Z,\xi,\gamma)}{2\sigma^2}\right]p(\xi|\gamma)}{\int \exp\left[-\frac{T(Z,\xi,\gamma)}{2\sigma^2}\right]p(\xi|\gamma)d\xi}$
- $T(Z,\xi,\gamma) = Z^HZ + \frac{1}{N}\mu_0B^TB\mu_0 - \frac{N}{N+1}\left(Z+\frac{B\mu_0}{N}\right)^H B(B^TB)^{-1}B^T\left(Z+\frac{B\mu_0}{N}\right)$
- $\varphi = \frac{N}{N+1}\left(B^TB\right)^{-1}B^T(Z+\frac{B\mu_0}{N})$ . .

Note that $\bar{s}(Z,\gamma) \in \mathbb{R}^{k^2+3k+1}$.

For our spline model, there is no analytical expression for $\bar{s}(Z,\gamma)$: the integral with respect to $\xi \in \mathcal{D}^{k-1}$ is untractable. We thus use a Monte-Carlo approach to approximate it, following the MCEM approach. Let I denotes any one of integrals listed above in equation (4.31) and let $f(\xi)$ be the corresponding function such that

$$I = \int_{\mathcal{D}^{k-1}} f(\xi)q(\xi|Z,\gamma)d\xi$$

Then, using a sampling scheme, we approximate I, such that:

$$I \approx \frac{1}{M}\sum_{j=1}^{M} f(\xi_j)\,, \quad \xi_j \sim q(\xi|Z,\gamma) \quad j = 1\cdots M \ . \tag{4.32}$$

4.3.2 Sampling from $q(\xi|Z,\gamma)$

Sampling from $q(\xi|Z,\gamma)$ is not straightforward and we use a Metropolis-Hastings (MH) algorithm with proposal density $P(\xi'|\xi_p)$ to obtain samples $\xi_j,\ j = 1\cdots M$. Let ξ_p be the state of the chain at time p and ξ' the candidate state simulated from a known probability $P(\xi'|\xi_p)$.

Initialization: Set ξ_0 and $p = 0$,

Draw a state: $\xi' \sim P(\xi'|\xi_p)$,

Accept the new state: ξ' is accepted as the next state of the Markov chain with probability $r(\xi', \xi_p)$ given below in equation (4.33),

Iterate: loop until a stop criterion is reached such as the number of samples for example.

MH builds a Markov Chain over ξ which distribution converges toward the target distribution $q(\xi|Z,\gamma)$. Detailed balance condition for the convergence is respected by using an acceptance ratio $r(\xi', \xi_p)$ such that:

$$r(\xi', \xi_p) = \frac{q(\xi'|Z,\gamma)}{q(\xi_p|Z,\gamma)} \frac{P(\xi_p|\xi')}{P(\xi'|\xi_p)} \ , \tag{4.33}$$

In practice, we begin the chain with a burn-in period: a predefined number of states that allow the chain to converge to the target distribution. We choose 1000 iterations for the burn-in.

The choice of the proposal is crucial for the mixing of the chain. We consider, for the simulations, different proposals:

- $P_1(\xi'|\xi_p)$: we change randomly one knot in its neighborhood following a triangular distribution,
- $P_2(\xi'|\xi_p)$: $\xi' \sim Dir(\xi; \alpha^0 = [11\cdots 1])$, corresponding to uniform distribution in the simplex ,
- $P_3(\xi'|\xi_p)$: $\xi' \sim Dir(\xi; \alpha^i)$, using the estimated parameter α^i at iteration i of the online EM algorithm.

In the next section, we give results for these 3 different proposals.

4.3.3 Numerical method for maximization

Once $\hat{s}_i$ is computed, γ is updated by maximizing the log-likelihood function, $\ell(\hat{s}_i, \gamma)$, (4.30).

The maximization has no analytical solution and we use numerical method, a standard gradient algorithm, to perform it. However the estimation of the

Dirichlet parameter α is independent of other parameters estimation and we can separate the maximization into two maximizations:

$$\hat{\alpha} = \arg \max_{\alpha \in \left(\mathbb{R}_*^+\right)^k} \left(-\log B(\alpha) + \sum_{j=1}^{k} (\alpha_j - 1) \log \Delta_j \right) \quad ,$$

$$\begin{aligned} \left(\hat{\mu}_0, \hat{\sigma^2}\right) \quad = \quad & \arg \max_{\mu_0 \in \mathbb{C}^k, \sigma^2 \in \mathbb{R}_*^+} \left(-(N+k) \log(2\pi\sigma^2) - \frac{\mu_0^H B^T B \mu_0}{2N\sigma^2} + \cdots \right. \\ & \left. \cdots - \frac{1}{2\sigma^2} \left((Z - B\beta)^H (Z - B\beta) + \frac{\beta^H B^T B \beta}{N} \right) + \Re \left(\frac{\mu_0^H B^T B \beta}{\sigma^2} \right) \right) \quad . \end{aligned}$$

4.4 Results

In this section, we show results obtained on curves simulated from a cubic spline model with $k = 12$ knots. Figure 4.3 shows 20 curves amongst the 200 simulated from the model with μ_0, the red crosses,
$\alpha = [3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27]$ and $\sigma^2 = 2.310^{-3}$.

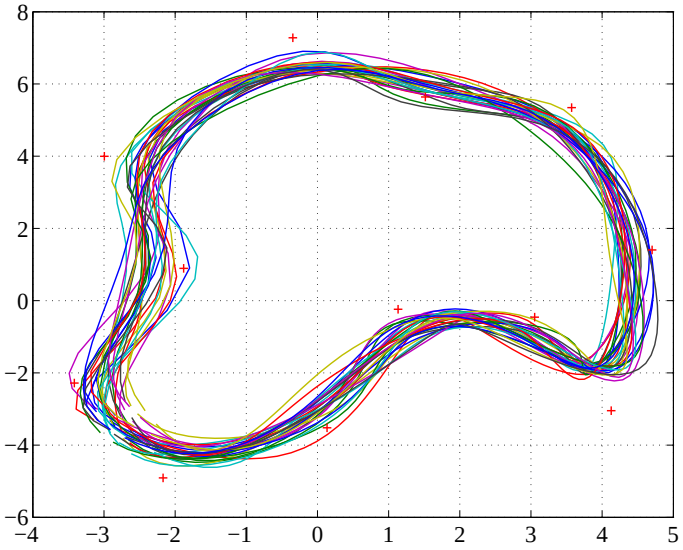

Figure 4.3: Subset of the 200 curves simulated from .

We now test different proposals presented in section 4.3.2. Results are shown in figures 4.4 and 4.5: in the first figure, we can see that the estimated μ_0 are similar, in the second, we can see that α is better estimated using

the proposal P_1 while for the other proposals the Markov chain is not mixed enough to obtain a good estimation of α. But, when looking at the bottom figure, the mode of the estimated Dirichlet distribution is estimated correctly for all the proposals.

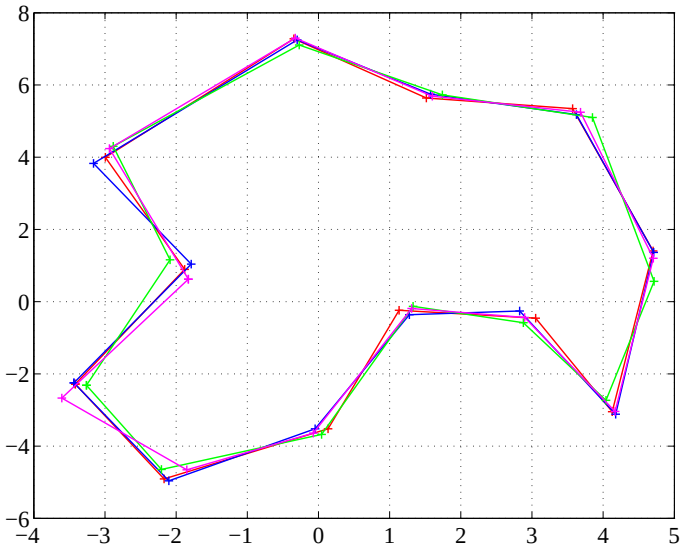

Figure 4.4: μ_0 estimated from 3 different proposals: real in red, estimated μ_0, P_1 proposal in magenta, P_2 proposal in green and P_3 in blue.

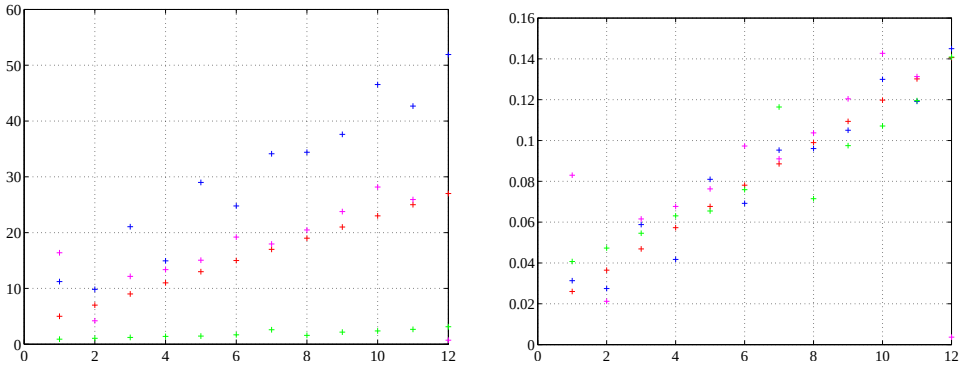

Figure 4.5: α (top) and $\frac{\alpha}{\bar{\alpha}}$ (bottom) estimated from 3 different proposals: real in red, P_1 proposal in magenta, P_2 proposal in green and P_3 in blue.

Figure 4.6 show that the estimated value of the noise covariance σ^2 is away from the exact value. However, we see that P_1 and P_3 proposals converge to the estimated value while for P_2, there is no convergence. It is due to the lack of mixing of the Markov Chain in the MC step of the algorithm.

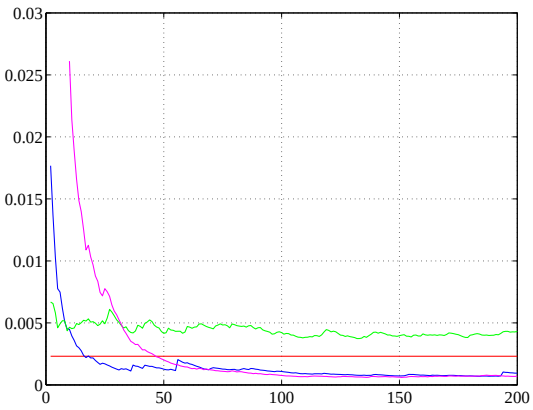

Figure 4.6: σ^2 estimated from 3 different proposals: real in red, P_1 proposal in magenta, P_2 proposal in green and P_3 in blue.

We conclude that the first proposal has the best properties for MC online EM algorithm.

We then test on the same set of simulated curves the influence due to the initialization of the algorithm. We know that EM algorithm converges locally and thus the identified model depends on the initialization. We perform 22 Monte Carlo runs of the algorithm with P_1 as proposal and with different initializations. We choose a random knot vector sampled from Dirichlet distribution with parameter $[1, 1, \cdots, 1]$. Then, $mu_0^{(0)}$ and $\sigma^{2(0)}$ are initialized with the maximum likelihood estimates, and $\alpha^{(0)}$ is set to $[1, 1 \cdots 1]$.

Figure 4.7 show the maximum log-likelihood of the 22 runs with lots of local maxima and one global maxima for 18-th run. The estimated parameters of the run corresponding to this maxima are very close from the real model (see figure 4.8) while some of the estimated parameters of runs with local maxima are rotated with respect to to real model. The remaining runs, 9 out of the 22, are corresponding to local maxima that are not close from the model.

We now take the parameters estimated at the run with the maximum of the log-likelihood, and we simulate some curves with this model and with a ba estimated model to compare them qualitatively with those of the data set. We can see in figure 4.9 that the correctly estimated model is a coherent model with the data set (left) while the badly estimated model does not produce curves similar to the data set (right).

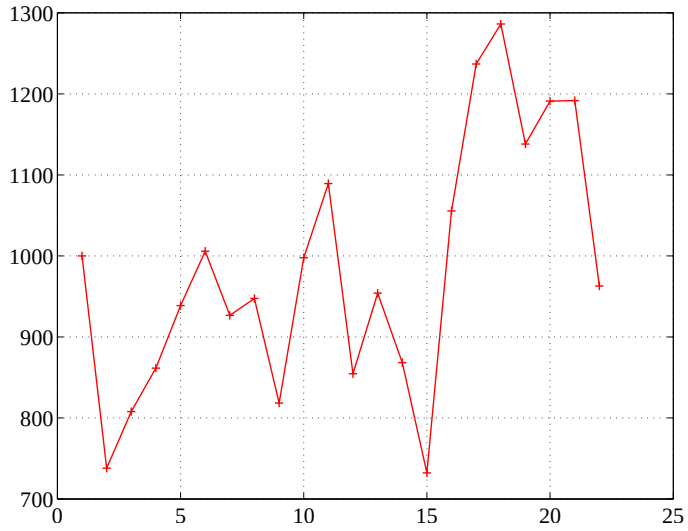

Figure 4.7: Maximum log-likelihood with the index of the run in abcissae.

4.4.1 Real data

In this section, we present the results of modeling a set of real data curves: 30 contours of sand ripples extracted manually from side-scan images which forms are characteristic of the sand ripples contours. For MC online EM, we choose the first proposal that demonstrates the better results on simulated data. We show here the dependence of the identified model on the complexity of the spline, or equivalently with k. We present 3 identified models with the maximum log-likelihood over the Monte-Carlo runs: γ_1 for $k = 8$ over 50 runs, γ_2 for $k = 15$ over 26 runs and γ_3 for $k = 25$ over 25 runs.

Figure 4.10 show the identified μ_0 in each case. We can see that the more complex model ($k = 25$) is not identifying a correct model. It is due to the fact that our data set is only composed of 30 curves which is too small to be able to identify $3k + 1 = 76$ parameters. In figure 4.11, we show the identified model for α. We can see that coefficients are closed from one, meaning that the variability in curves is high and thus the Dirichlet distribution is almost uniform in the simplex.

4.5 Conclusion

In this chapter, we proposed a new variant of the EM algorithm that sequentially identify a collective spline model of a set of curves. We show that when applied to simulated data, the model is coherent with the data. With real data curves, one must take care of adjusting the dimension of the model

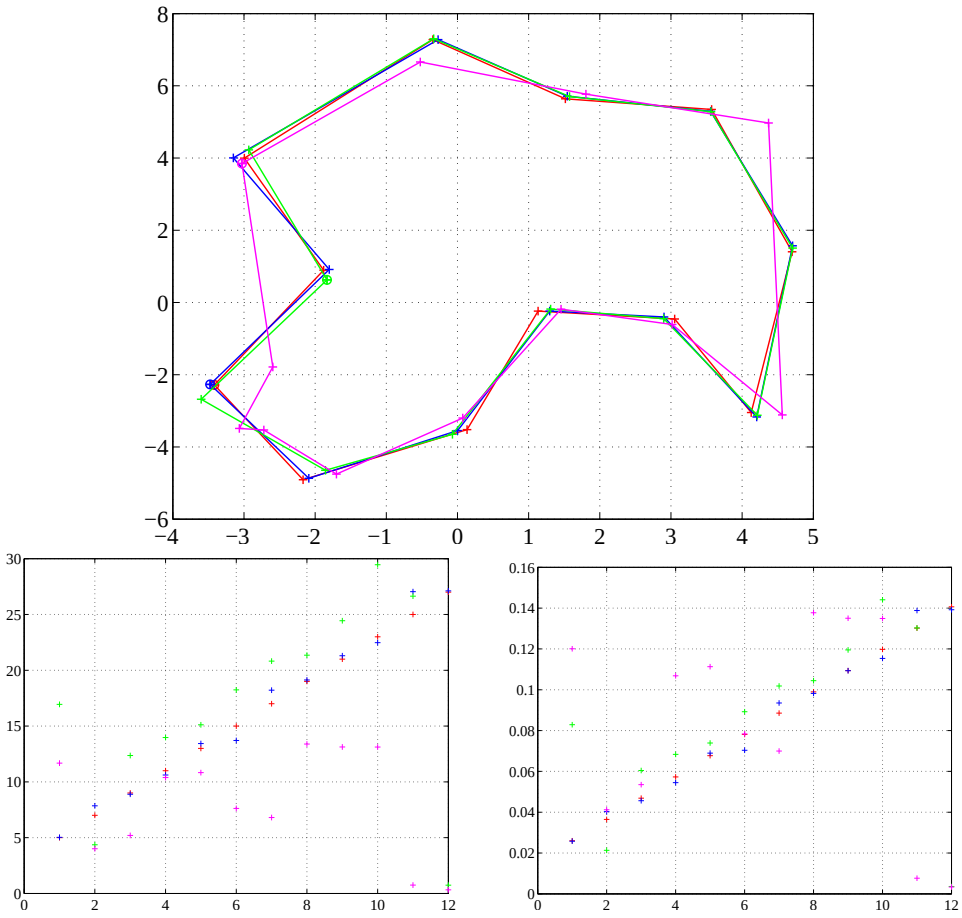

Figure 4.8: Parameters estimated for the global maxima (in blue) and for some local maxima (in green), with the real model (in red): μ_0 on the top, α in the bottom left and $\frac{\alpha}{\bar{\alpha}}$ in the bottom right.

with the number of available curves. In the context of sequentially acquiring the curves, we can update the model at each observation and thus stabilize the identified parameters.

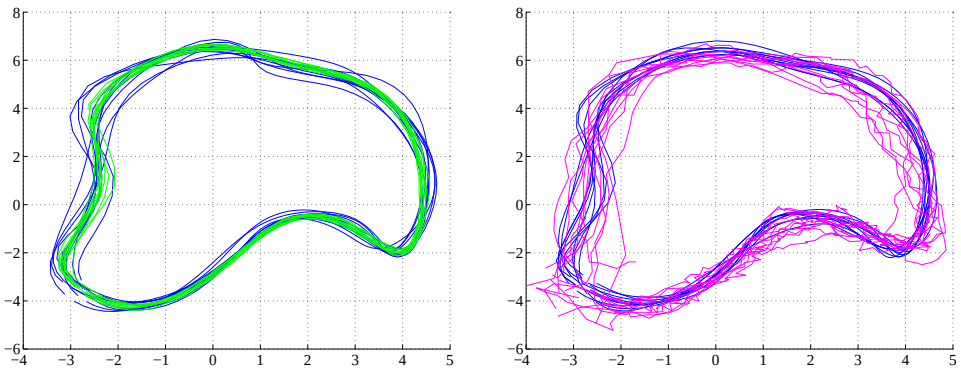

Figure 4.9: Simulated curves from the model estimated with the maximum log-likelihood (left, in green) and curves extracted of the data set (in blue) and with a bad estimation (right, in magenta).

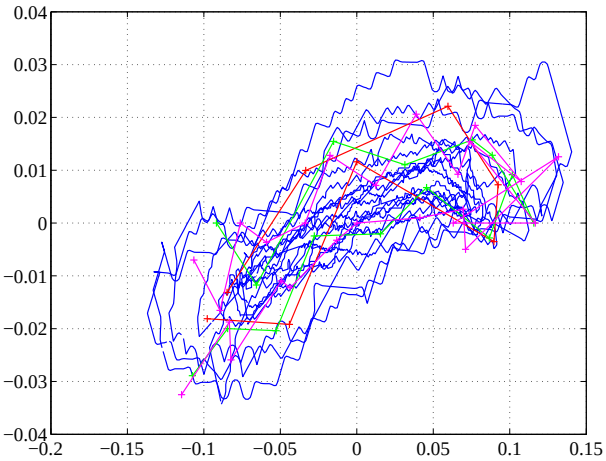

Figure 4.10: μ_{01} (red), μ_{02} (green) and μ_{03} (magenta) with some sand ripples contours (blue).

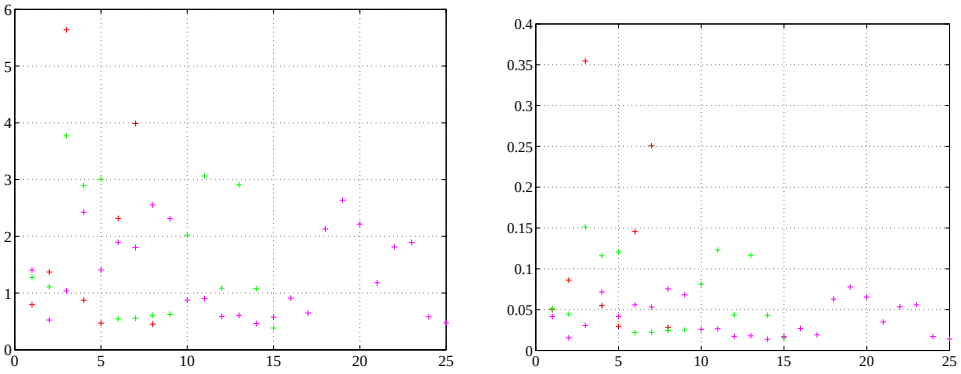

Figure 4.11: α (left) and $\frac{\alpha}{\bar{\alpha}}$ (right) of γ_1 (red), γ_2(green) and γ_3 (magenta).

Chapter 5

Spline-based shape model

5.1 Introduction

The major advantage justifying our choice of spline representations to characterize shapes is that it has the power of continuous descriptions while still leading to sparse representations through its parameters β and ξ. As several authors have already suggested, the identification of spline parameters, in particular of control points, can be interpreted as an automatic way to select the landmarks of the discrete shape theory presented in chapter 2. This idea is for instance presented in [5] considering the special case of fixed equispaced knots selecting the remarkable points of the object as the identified control points. In this chapter, we study how we can base the definition of a shape space in spline models.

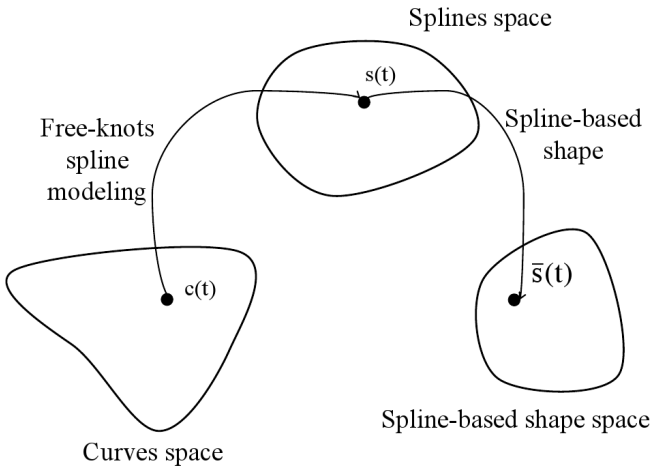

Figure 5.1: Illustration of spline-based shape identification of contours.

Figure 5.1 illustrates the scheme we follow for the shape ($\bar{s}(t)$) definition

of a curve $s(t)$, passing through the spline model $s(t)$(chapter3).

We study the definition of shapes of free-knots spline representations of curves using the basic Kendall's shape definition: "what is left after removing information of position, size and orientation of an object". Since our objects are continuous curves, we also add the parametrization invariance as it is customary in the continuous theories (see chapter 2). In the first section, we study on spline representations what is the result of imposing invariances with respect to a number of operating groups, and we show that, contrary to what happens in both shape theories presented in chapter 2 – where the shape space has no simpler representation than a set of equivalence classes – we can here select a unique spline as a representative of each shape, i.e. of an entire class of equivalence with respect to the operating groups considered.

In the second section, we briefly compare Kendall's shapes and spline-based shapes, focusing on the comparison of distances between shapes defined in each theory.

5.2 Spline-based shape definition

In this chapter, as we already did in previous sections, we use the complex representation of points in the plane. Let $s(t) = (s_X(t), s_Y(t))$ be a spline curve from $[0,1]$ to $\mathbb{R}^2 \equiv \mathbb{C}$. Let $s(t) \in \boldsymbol{S}_m^k$ be a cubic spline with parameters (β, ξ), $\beta \in \mathbb{C}^k$ and $\xi \in [0,1]^k$.

We maintain the shape definition used for continuous curves presented in chapter 2, restriciting it to splines:

Definition (5.1): Spline-based shape

Two splines $s_1(t) \in \boldsymbol{S}_m^k$ and $s_2(t) \in \boldsymbol{S}_m^k$ have identical shapes $[s_1(t)]$ and $[s_2(t)]$, i.e. $[s_1(t)] = [s_2(t)]$, if and only if $\exists\ \alpha \in \mathbb{R}_*^+$, $R \in SO(2)$, $\mathbf{t} \in \mathbb{R}^2$ and $\gamma(t) \in \Gamma$ such that:

$$s_1(t) = \alpha R s_2(\gamma(t)) + \mathbf{t} \ ,$$

where Γ is the parametrization group defined in section 2.3.2.

5.2.1 Translation and scaling invariances

We study now the impact of these shape invariances on the spline parameters.

We begin by fixing the position of $s(t)$ in the plane, so that the representation becomes invariant with respect to translation. We will denote by

$s_c(t)$ the centered version of $s(t)$, so that:

$$\int_0^1 s_c(t)dt = 0 \ . \tag{5.1}$$

Let $\mathcal{B}_i = \int_0^1 b_i(t;\xi)dt$ and $\mathcal{B} = [\mathcal{B}_1 \cdots \mathcal{B}_k]$.

$$\begin{aligned}\int_0^1 s_c(t)dt &= \int_0^1 \sum_{i=1}^k \beta_i^c b_i(t;\xi)dt = \sum_{i=1}^k \beta_i^c \int_0^1 b_i(t;\xi)dt \ , \\ &= \sum_{i=1}^k \beta_i^c \mathcal{B}_i = \mathcal{B}\beta^c = 0 \ .\end{aligned}$$

Thus, the control points of the centered spline are orthogonal to the hyperplane $\mathcal{B}$. Note that this is not equivalent to centering the control points, i.e. $\sum_{i=1}^k \beta_{iX}^c \neq 0$, unless $\mathcal{B} = b\mathbf{1}$, which happens, for instance, when ξ is uniform, and the $b_i(t;\xi)$ are shifted versions of the same function.

We can now impose invariance with respect to scale to the centered spline representation $s_c(t)$ by fixing its size. We may consider two distinct definitions of the size of a curve:

- unit length, given in definition 2.16 and recalled here:

$$L_1^2[s_c] = \int_0^1 \left\langle \frac{ds_c(t)}{dt}, \frac{ds_c(t)}{dt} \right\rangle dt = 1 \ , \tag{5.2}$$

- 'unit-'ball" size:

$$L_2^2[s_c] = \int_0^1 ||s_c(t)||^2 dt = 1 \ , \tag{5.3}$$

We denote by $G \in \mathbb{R}^{k\times k}$ the Gramian matrix of the B-splines and by $H \in \mathbb{R}^{k\times k}$ the Gramian matrix of the derivatives of B-splines:

$$G_{ij} = \int_0^1 b_i(t;\xi)b_j(t;\xi)dt \quad i,j = 1\cdots k \ , \tag{5.4}$$

$$H_{ij} = \int_0^1 \frac{db_i(t;\xi)}{dt}\frac{db_j(t;\xi)}{dt}dt \quad i,j = 1\cdots k \ . \tag{5.5}$$

Then,

$$\begin{aligned}L_1^2[s_c] &= \int_0^1 \sum_{i=1}^k \sum_{j=1}^k \left\langle \beta_i^c \frac{db_i(t;\xi)}{dt}, \beta_j^c \frac{db_j(t;\xi)}{dt} \right\rangle dt \ , \\ L_1^2[s_c] &= \beta^{cH} H \beta^c \ . \end{aligned} \tag{5.6}$$

and

$$\begin{aligned} L_2^2[s_c] &= \int_0^1 \sum_{i=1}^{k}\sum_{j=1}^{k} \left\langle \beta_i^c b_i(t;\xi), \beta_j^c b_i(t;\xi) \right\rangle dt \ , \\ L_2^2[s_c] &= \beta^{cH} G \beta^c \ . \end{aligned} \tag{5.7}$$

We see that both the curve length and the "ball" length are Frobenius norms of the vectors of control points under convenient metrics defined by the B-spline basis. Let $L_n[s_c], n = 1, 2$ be the size definitions above. The corresponding unit size representations $\tilde{s}_n(t)$ are

$$\tilde{s}_n(t) = \frac{s_c(t)}{L_n[s_c]} \quad n = 1, 2 \ , \tag{5.8}$$

The mapping from $s(t)$ to the centered and normalized spline $\tilde{s}_n(t)$ is a many-to-one application that allows to select a unique representation of all translated and scaled representations of a spline curve. Imposing these conditions on the spline leads to spline representation that is invariant to translation and scaling.

For fixed k and ξ, the space of control points of the corresponding centered and normalized spline functions, noted $\tilde{S}_m^{\xi}$, is thus an ellipsoid:

$$\tilde{S}_m^k = \left\{ s(t) \in \boldsymbol{S}_m^k : \int_0^1 s(t)dt = 0 \text{ and } L_n[s] = 1 \right\} \ . \tag{5.9}$$

5.2.2 Origin and rotational invariances

We now consider the invariance with respect to the origin of the curve parametrization. Indeed, any closed curve has identical shape when changing its origin.For closed spline curves, we choose to fix the origin on a knot point so that changing the origin corresponds to a circular shift of parameters β and ξ. We select the knot vector ordered so that intervals between knots, $\Delta_i = \xi_{i+1} - \xi_i,\ i = 1 \cdots (k-1)$ and $\Delta_k = 1 - \xi_k + \xi_1$, are following the lexicographic order:

$$\begin{aligned} T &= \arg\max \{\Delta_i, \quad i = 1 \cdots k\} \ , \\ \check{\Delta}_i &= \Delta_{(T+i-1) \bmod k} \quad i = 1 \cdots k \ . \end{aligned}$$

And we fix the origin t_0 of the curve so that $\check{\xi}_1 = t_0$.

We note $\check{s}(t)$ the new spline representation with parameters $(\check{\beta}, \check{\xi})$ that are circularly shifted version of (β, ξ). Figure 5.2 illustrates this selection.

The mapping from $\tilde{s}(t)$ to $\check{s}(t)$ is thus invariant to the choice of the curve parametrization.

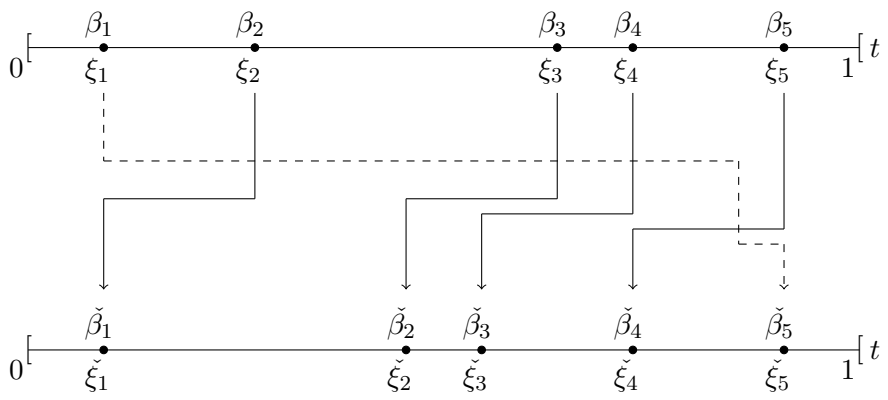

Figure 5.2: Illustration of the fixed lexicographic parametrization of splines: the original parametrization (β, ξ) is above and the circularly transformed parametrization $(\check{\beta}, \check{\xi})$ is below.

Kendall's landmarks are distinguishable points, i.e. their labeling is defined by the selection mechanism, on the basis of relevant criteria for the problem at hand. In our case, as the enforcement of permutation invariance discussed in section 2.2.4 demonstrates, all points are in principle exchangeable (points are just points). The coupling between control points and knots allow us to select a canonical labeling of the control points, the one that leads to a knot vector whose intervals are sorted by lexicographic order.

We remark here that this choice of ordered knot vector is impossible for some problematic cases: the splines that present a periodicity between their knots for which uniform splines with equispaced knots is a special case. However we consider here that this does not occur in practice as the spline model of an arbitrary curve identified with approaches detailed in chapter 3, does not have this very particular setting. Indeed, the set of such splines in $\boldsymbol{S}_m^k$ are a set of null measure in $\boldsymbol{S}_m^k$.

Note that when considering open curves, the origin of the curve is naturally fixed and there is no meaning in considering this invariance.

We now consider the orbits of $\check{s}_n(t)$ under the group of rotations in $\mathbb{R}^2$. We fix the orientation of the spline by fixing the origin on the positive x-axis so that $\check{s}(t_0) \in R^+$. This arbitrary convention allows to select a unique representative of the equivalence class with respect to rotation. We expect that once t_0 and the knots have been fixed as explained before, in a manner that is driven by the characteristics of the shape itself, fixing the orientation in this manner is sufficient to align curves and thus to distinguish spline shapes.

We can question this choice as this fixed orientation may not minimize

the difference between two shapes in the sense that it does not lead to the minimum $\mathbb{L}^2$-distance between their corresponding equivalence classes. However this is also the case for Kendall's definition of distances, for example the full Procrustes distance considers explicitly minimization with respect to scaling and in this sense needs to consider elements outside the set of normalized representations. We hope that in practice this selection will not significantly impact the distance between curves determined through their spline-based shape representations: if they are close, the distance should be small (near zero), if they are highly dissimilar, the distance should be high.

Any $\mathbb{L}^2$-induced distance in the spline-based shape space will be an upper bound on the true $\mathbb{L}^2$ distance between the corresponding equivalence classes. The problem may thus be that two close shapes will not be properly aligned, resulting in a spline-based shape distance that is larger than it should be.

We note $\hat{s}(t)$ the rotated version of $\check{s}(t)$. The mapping from $\check{s}(t)$ to $\hat{s}(t)$ is a many-to-one mapping that is invariant to rotations.

5.2.3 Invariant spline parametrization

In the previous sections we considered the standard shape invariances that Kendall used to define a shape: translation, scaling, rotation, and we added the invariance with respect to the definition of origin. We now consider the invariance with respect to the choice of parametrization of the curve. Figure 5.3 shows the same closed curve with different parametrizations: the spline with non-uniform knots $s(t)$ and the uniform spline $s(\gamma(t))$ where $\gamma(t)$ is given in figure 5.4.

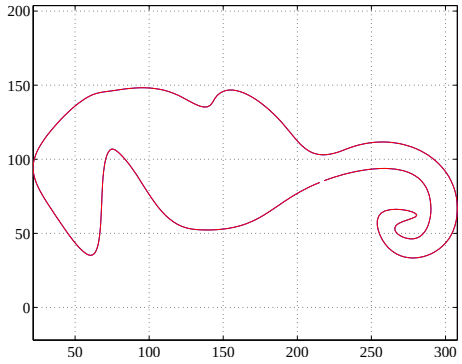

Figure 5.3: Closed spline curve with two distinct parametrizations corresponding to uniform (in red) and non-uniform (in blue) knot vector.

Curve re-parametrization directly affects the knot positions and, obviously indirectly, also the control points of the spline description. So we

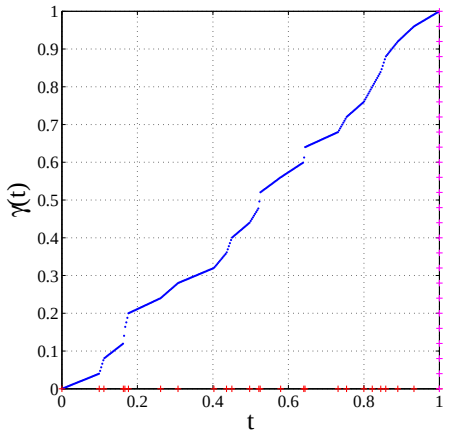

Figure 5.4: Parametrization function with the initial knot (in red) on the x-axis and the uniform knot vector (in magenta) on the y-axis.

would like to be insensitive with respect to the particular observed knot vector. We do this by imposing uniform knots vector $\bar{\xi}$ so that, $\Delta_i = \frac{1}{k}$ and $\bar{\xi}_i = (i-1)/k$ for $i = 1 \cdots k$. This is equivalent to a piecewise linear re-parametrization of the original curve:

$$\gamma : t \in [\xi_i, \xi_{i+1}] \rightarrow u = a_i t + b_i \in [(i-1)/k, i/k] \ , \tag{5.10}$$

where $a_i \in R_*^+$ and $b_i \in \mathbb{R}$.

Figure 5.5 illustrates the re-parametrization for a particular ξ. This re-parametrization also affects the control points and we will denote it by $\left(\bar{\beta}, \bar{\xi}\right)$, and designate it by the "matched curve parametrization". The detailed computation of $\bar{\beta}$ is given in appendix C for the particular cases of $m = 1, 2, 3$. Similar formulas can be obtained for higher values of m.

Note that this is not the usual re-parametrization group consider in shape theory which elements are commonly considered strictly increasing continuous differentiable functions.

Being linear, the re-parametrization γ does not change the degree of the spline. However, being piecewise, the differentiability of $s(u) = s(\gamma(t))$ at the knots is generally lost. For example for ξ_i, $i = 1 \cdots k$:

$$\left.\frac{d(s \circ \gamma(t))}{dt}\right|_{\xi_i - \delta} = \left.\frac{d\gamma(t)}{dt}\frac{ds(u)}{du}\right|_{\xi_i - \delta} = a_{i-1} \left.\frac{ds(u)}{du}\right|_{(i-1)/k - \delta} \ ,$$

and,

$$\left.\frac{d(s \circ \gamma(t))}{dt}\right|_{\xi_i + \delta} = \left.\frac{d\gamma(t)}{dt}\frac{ds(u)}{du}\right|_{\xi_i + \delta} = a_i \left.\frac{ds(u)}{du}\right|_{(i-1)/k + \delta} \ ,$$

Assuming that the spline $s(t)$ is a continuous and differentiable function, then $s(u)$ is continuous as $\gamma(t)$ is continuous but it will not be differentiable

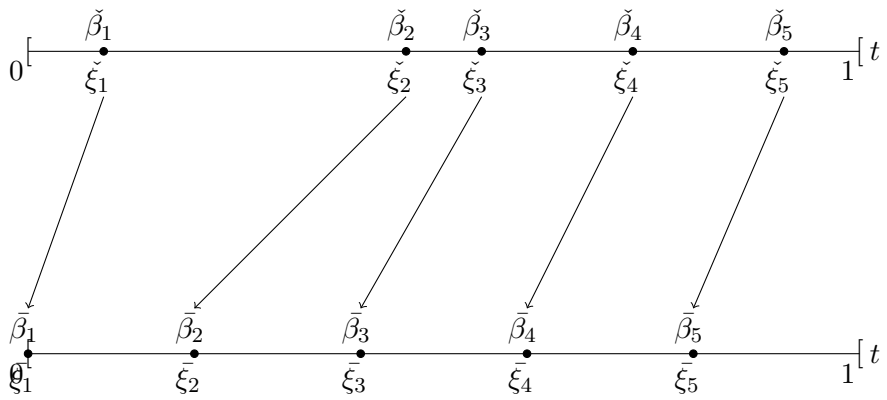

Figure 5.5: Illustration of the re-parametrization of splines to obtain the matched curve parametrization.

if $\gamma(t)$ is not. To cope with this non-differentiability at the knot points, we must use multiple knots in the matched curve representations. For splines of order m, knots are repeated $(m-1)$ times in the knot vector. For example for cubic splines:

$$\bar{\xi} = [0, 0, 0, 1/k, 1/k, 1/k, \cdots (k-1)/k, (k-1)/k, (k-1)/k] \quad . \tag{5.11}$$

Note that this matched curve representation transfers complexity from the knot sequence (which is now always uniform) to the control points (which now belong to a space of higher dimensionality).

With the matched curve parametrization, we expect to reduce variability of spline representation inside the equivalence class with respect to parametrizations. Finally $\bar{s}(t)$ is the spline-based shape of $s(t)$. Figure 5.6 shows a diagram for the spline-based shape selection. Note that contrary to other shape theories, the spline shape is a unique representation of the complete equivalence classes.

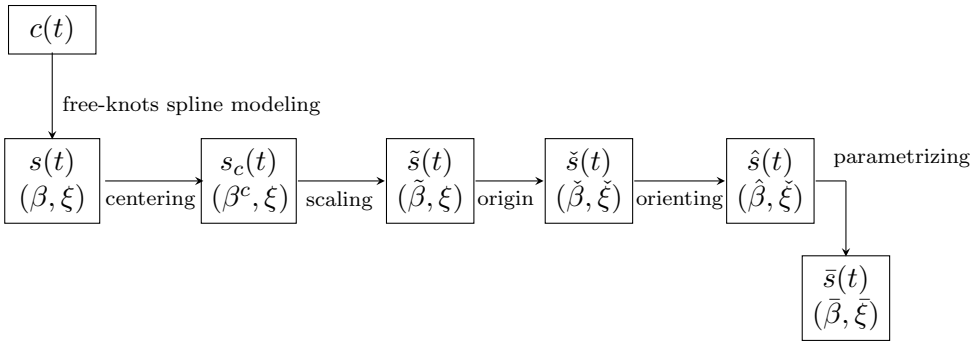

Figure 5.6: Diagram showing the spline shape identification.

The set of spline shape functions, noted $\bar{\boldsymbol{S}}_m$ is the union of spline shape functions with k knots $\bar{\boldsymbol{S}}_m^k$:

$$\bar{\boldsymbol{S}}_m^k = \left\{ s(t) \in \boldsymbol{S}_m : \begin{array}{l} L[s] = 1, \\ \text{fixed origin and orientation} \\ s(t) \text{ is matched curve parametrized} \end{array} \right\}, \quad (5.12)$$

$$\bar{\boldsymbol{S}}_m = \bigcup_k \bar{\boldsymbol{S}}_m^k \ . \quad (5.13)$$

Note that due the re-parametrization of the spline, $\bar{s}(t)$ is not, in general, a centered curve and $\int_0^1 s(t)dt \neq 0$.

5.2.4 Spline-shape distance definition

$\bar{\boldsymbol{S}}_m$ is a subset of the set of continuous functions $\mathbb{L}^2$ from $[0,1]$ to $\mathbb{R}^2$. It thus inherits the $\mathbb{L}^2$ inner product given by:

$$\langle f, g \rangle = \int_0^1 \langle f(t), g(t) \rangle_E \, dt \ .$$

Let $\bar{s}_1(t) \in \bar{\boldsymbol{S}}_m^{k_1}$ and $\bar{s}_2(t) \in \bar{\boldsymbol{S}}_m^{k_2}$ be two spline curves with respective control points vector β^1 and β^2. Note that due to the matched curve parametrization, β^i, $i = 1, 2$, is no longer an element of $\mathbb{C}^{k_i}$ but is an element of $\mathbb{C}^{(m-1)k_i}$.

$\left(\bar{\boldsymbol{S}}_m, d\right)$ is a metric space if we define distance d such that:

$$d^2(\bar{s}_1, \bar{s}_2) = \int_0^1 (\bar{s}_2(t) - \bar{s}_1(t))^H (\bar{s}_2(t) - \bar{s}_1(t)) dt \ . \quad (5.14)$$

Let W_{k_1,k_2} be the $(k_1 \times k_2)$ real matrix such that,

$$[W_{k_1,k_2}]_{i,j} = \int_0^1 b_i(t; \xi_{k_1}) b_j(t; \xi_{k_2}) dt \ .$$

The distance (5.14) can be expressed in terms of the spline parameters and matrix W_{k_1,k_2}:

$$d^2(\bar{s}_1, \bar{s}_2) = 2\left(1 - \Re\left(\beta^{2H} W_{k_1,k_2} \beta^1\right)\right) \ . \quad (5.15)$$

Note that when $k_1 = k_2 = k$, $W_{k,k} = G$ the Gramian matrix of B-splines, and the mapping from s to $G^{1/2}\beta$ is an isometry from $\bar{\boldsymbol{S}}_m^k$ to $\mathbb{C}^{(k-1)m}$ with the standard norm.

5.3 Spline-based shape and the discrete theory of shape

In this section we briefly study the link between Kendall's discrete shape theory and the spline-based shape theory sketched in the previous section, by taking the uniquely identified control points $\bar{\beta}$ as the landmarks of Kendall's theory.

Let $s(t) \in \bar{\boldsymbol{S}}_m^k$ and (β, ξ) be its spline parameters [1]. Remember that $s(t)$ is a uniform spline and $\beta \in \mathbb{C}^{(m-1)k}$.

We begin by studying the shape of the set of control points β under Kendall's theory. The control points β are not centered and not normalized (in Kendall's meaning): $\sum_{i=1}^{2k} \beta_i \neq 0$ and $\beta^H \beta \neq 1$.

Let $[\beta]$ denotes the Kendall's shape of the configuration β. In this case, Kendall's shape space is $\Sigma_2^{(m-1)k}$ and distances between shapes are given in section 2.2.3.

We now develop the partial Procrustes distance between two control points configurations β^1 and β^2 obtained with the matched curve parametrization of their corresponding spline curves. Note that we use the notations introduced in section 2.2.

$$d_P([\beta^1], [\beta^2]) = \sqrt{2}\left(1 - |\tilde{\beta}^1 \tilde{\beta}^{2H}|\right) \quad .$$

We now develop the distance between spline curves with similar complexity, defined in (5.15), and:

$$d(\bar{s}_1, \bar{s}_2) \quad = \quad \sqrt{2}\sqrt{(1 - |\beta^{2H} G \beta^1| \cos\theta)} \ ,$$

where θ is the argument of $\beta^{2H} G \beta^1$.

The angle θ has the sense of the rotation between $G^{1/2}\beta^1$ and $G^{1/2}\beta^2$. When $\theta \approx 0$, then $d^2 = \sqrt{2} d_P$. This suggests that our apprehension in section 5.2.3 that distance between spline-based shapes is maintained is correct.

5.4 Conclusion

This chapter reports our preliminary work on the definition of spline-based shape space for closed curves. We propose to use the free-knots spline model identified in chapter 3 as the representative of the contour of an object and

[1] Note that we adopt the simpler notation β for $\bar{\beta}$ of the previous section

to define its shape. From equivalence classes with respect to group actions we select a unique representative spline so that the spline-based shape of the curve is normalized, with fixed origin and matched curve parametrized. We also define a distance between two spline-based shapes and discussed its relation with Kendall's partial Procrustes distance.

In this chapter we succeeded in proposing a spline-based shape space. We put forward a tentative definition of a matched spline representation that enabled us to move all the complexity of the curve representation to the representation of uniform splines.

This thesis does not fully establish the properties of this representation which we expect to address in detail in future work.

Chapter 6

Conclusion

6.1 Contributions

This thesis is a contribution to the definition of formal representations of the shape of objects as "what is left after removing information of position, scale and orientation" following the seminal definition of Kendall. Our objects are the contours of identified distinct regions such as sand ripples or posidonia clutters in the special case of underwater images. Chapter 2 presents two theories reported in the literature that define shape as equivalence classes with respect to some groups action: translations, scales, rotations and parametrization. Both the discrete and the continuous definitions of shapes allow the determination of planar shape spaces that are Riemannian manifolds, and relevant geometric tools are defined such as distances between shapes, geodesics, and mean shape. These approaches suffer from major drawbacks: the selection of landmarks for the discrete case and the infinite dimension of continuous shape spaces. In both cases, the definition of statistical shape models is not an easy task and requires approximations procedures, for example, using tangent spaces or projections on finite-dimensional spaces.

We propose here to combine both approaches by using spline curves to represent the shape of the contours. We begin by studying the curve modeling problem with splines in chapter 3 and we introduce the free-knots spline model as the one to use to obtain a sparse representation of a contour with good fitting properties. We formulate the identification of a spline curve as a problem of selecting a model amongst nested families of models. Our review of the solutions proposed in the literature (with maximum likelihood or Bayesian criteria) reveals a major flaw when applying the MAP criterion to the identification problem that may lead to paradoxical behaviors. The two steps approach proposed in section 3.4.6 is the procedure that can be safely used in this framework. This work has been published in [2].

We can now identify a spline model for any arbitrary contour. We propose in chapter 5 to use the spline parameters to express the shape of the contour. We address the basic issues involved on building spline-based shape theory, studying shape invariances when imposed on spline curves and showing that we can select a spline representative in each equivalence class so that its spline-based shape is uniquely defined. We also define a distance in spline-based shape spaces as the distance inherited from the one in the set of continuous curves.

Chapter 4 presents a method for the determination of a probabilistic spline-based model for a set of curves. We propose to use a hierarchical parametric model and to consider the spline parameters as unobserved variables of the modeling problem. Then, the identification of the model is performed using a variant of the expectation-maximization algorithm that, as far as we know, is proposed here for the first time: the Monte-Carlo Online EM. This method sequentially processes the observed curves updating the estimates of the model parameters with each new observation and, as some steps cannot be analytically performed, resorts to Monte-Carlo approximations.

6.2 Perspectives

During this thesis we identified some major axis that would require future developments. We enumerate here some of them.

The results presented in chapter 4 indicate that the Monte-Carlo Online EM algorithm may converge to local maxima of the likelihood function. This is a well-known EM problem. Some studies have been published on the convergence of MCEM algorithm [21] and on the convergence of the online EM [9], in the special case of exponential family. The results on these references do not allow simple characterization of the convergence of MC online EM algorithm. The formal establishment of the properties of the algorithm is an important issue that requires further work.

The work presented in chapter 5 is both innovative and very preliminary. The geometry of the spline-based shape spaces needs to be studied so that it becomes a useful tool to define geodesics between curves, distance (as the length of the geodesic), mean shape of a set of curves and also distributions of shapes. Addressing these issues will reveal the real pertinence of the

notion that is put forward here.

A key stone of our construction is the matched representation that maps a free-knots spline to a given uniform spline. The intuition behind the definition of this representation is that the knots of a free-knots spline capture, in a non-linear manner, the most distinctive elements of the morphology of a curve. Our representation moves this information from the non-linear part of the spline representation to its linear parameters.

We did not have the opportunity to study the properties of the resulting shape. This definitely needs further work.

Links between the spline-based shape and the discrete and continuous shapes require further explorations so that bridges can be build between the different shape theories. It would be also interesting to compare the shape distances given by each theory to extract similarities and/or dissimilarities that may be due to imposed invariances.

Appendix A

Details for RJMCMC implementation

When $\mathcal{M}_k$ and ξ are fixed, the MAP estimation of β and σ^2 can be done analytically.

Posteriors are given by:

$$\begin{aligned} p(\beta|Z,\mathcal{M}_k,\xi,\sigma^2) &= (2\pi\sigma^2)^{-(k-m)}\left(\frac{N+1}{N}\right)^{(k-m)}|B_{k,\xi}^T B_{k,\xi}|\cdots \\ &\quad \exp\left(-\frac{1}{2\sigma^2}\frac{N+1}{N}(\beta-\varphi_{k,\xi})^H B_{k,\xi}^T B_{k,\xi}(\beta-\varphi_{k,\xi})\right), \\ p(\sigma^2|Z,\mathcal{M}_k,\xi) &= \frac{(\sigma^2)^{-(N+1)}}{\Gamma(N)}\left(\frac{Z^H\alpha_{k,\xi}Z}{2}\right)^N \exp\left(-\frac{Z^H\alpha_{k,\xi}Z}{2\sigma^2}\right), \end{aligned}$$

with

$$\begin{aligned} \alpha_{k,\xi} &= I_N - \frac{N}{N+1}B_{k,\xi}(B_{k,\xi}^T B_{k,\xi})^{-1}B_{k,\xi}^T \ , \\ \varphi_{k,\xi} &= \frac{N}{N+1}\left(B_{k,\xi}^T B_{k,\xi}\right)^{-1}B_{k,\xi}^T Z \ . \end{aligned}$$

The MAP estimate are:

$$\hat{\beta}_{k,\xi} = \varphi_{k,\xi} \ , \tag{A.1}$$

$$\hat{\sigma^2}_{k,\xi} = \frac{Z^H\alpha_{k,\xi}Z}{2(N+k-m+1)} \ . \tag{A.2}$$

We introduce a new notation:

$$\mathcal{P}(\mathcal{M}_k,\xi|Z) = p(\mathcal{M}_k,\xi|Z)p(\hat{\sigma^2}_{k,\xi}|Z,\mathcal{M}_k,\xi)p(\hat{\beta}_{\hat{k},\xi}|Z,\mathcal{M}_k,\xi,\hat{\sigma^2}_{k,\xi}).$$

Then, k and ξ are the structural variables of our model and we call $\mathcal{P}(\hat{k},\xi|Z)$ the reduced posterior that is the target of the RJMCMC.

Appendix B

Details of Monte-Carlo Online EM algorithm

In this appendix, we give the detailed expression involved in MC online EM algorithm.

The marginal likelihood is:

$$\begin{aligned} p(Z|\gamma) &= \int_W p(Z,w|\gamma)dw \ , \\ &= \int_W p(Z|w,\gamma)p(w|\gamma)dw \ , \\ &= \iint_{\mathcal{C}^k\times\mathcal{D}^{k-1}} p(Z|\gamma,\beta,\xi)p(\beta,\xi|\gamma)d\beta d\xi \ . \end{aligned}$$

Given models in equation (4.11) and (4.12), we can perform the integration with respect to β. This leads to:

$$p(Z|\gamma) = (2\pi\sigma^2)^{-N}(N+1)^{-k}\int_{\mathcal{D}^{k-1}} \exp\left[-\frac{T(Z,\xi,\gamma)}{2\sigma^2}\right] p(\xi|\gamma)d\xi \ ,$$

with

$$T(Z,\xi,\beta) = Z^H Z + \frac{\mu_0 B^T B\mu_0}{N} - \frac{N}{N+1}\left(Z+\frac{B\mu_0}{N}\right)^H B(B^TB)^{-1}B^T\left(Z+\frac{B\mu_0}{N}\right) \ .$$

Then, we can express $\bar{s}(Z,w)$ (equation (4.28)):

$$\begin{aligned} \bar{s}(Z,\gamma) &= \int_{\mathcal{W}} S(Z,w)p(w|Z,\gamma)dw \ , \\ &= \frac{1}{p(Z|\gamma)}\int_{\mathcal{W}} S(Z,w)p(Z,w|\gamma)dw \ , \\ &= \frac{1}{p(Z|\gamma)}\iint_{\mathbb{C}^k\times\mathcal{D}^{k-1}} S(Z,w)p(Z|\beta,\xi,\gamma)p(\beta|\gamma,\xi)p(\xi|\gamma)d\beta d\xi \ . \end{aligned}$$

We note $q(\xi|Z,\gamma) = \frac{\exp\left[-\frac{T(Z,\xi,\gamma)}{2\sigma^2}\right]p(\xi|\gamma)}{\int_{\mathcal{D}^{k-1}}\exp\left[-\frac{T(Z,\xi,\gamma)}{2\sigma^2}\right]p(\xi|\gamma)d\xi}$. Then,

$$\bar{s}(Z,\gamma) = \int_{\mathcal{D}^{k-1}} \tilde{s}(Z,\xi,\gamma)q(\xi|Z,\gamma)d\xi \ ,$$

where $\tilde{s}(Z,\xi,\gamma)$ is such that:

$$\bar{s}(Z,\gamma) = \begin{bmatrix} \int_{\mathcal{D}^{k-1}} \log\Delta^i q(\xi|Z,\gamma)d\xi \ , \\ Z^HZ + k\sigma^2 - 2\Re\left(\int_{\mathcal{S}^k} Z^HB\varphi q(\xi|Z,\gamma)d\xi\right) + \cdots \\ \quad \cdots + \frac{N+1}{N}\int_{\mathcal{D}^{k-1}} \varphi^HB^TB\varphi q(\xi|Z,\gamma)d\xi \ , \\ \Re\left(\int_{\mathcal{D}^{k-1}} \frac{B^TB}{N}\varphi q(\xi|Z,\gamma)d\xi\right) \ , \\ \Im\left(\int_{\mathcal{D}^{k-1}} \frac{B^TB}{N}\varphi q(\xi|Z,\gamma)d\xi\right) \ , \\ \int_{\mathcal{D}^{k-1}} \mathrm{Vec}\left(B^TB\right) q(\xi|Z,\gamma)d\xi \ , \end{bmatrix}$$

with $\varphi = \frac{N}{N+1}\left(B^TB\right)^{-1}B^T(Z + \frac{B\mu_0}{N})$.

Appendix C

Details of the matched curve parametrization

In this appendix, we detail the computation of the B-splines from the knot vector ξ to the uniform knot vector $\bar{\xi}$. Each interval $[\xi_i, \xi_{i+1}]$ corresponds to $[\bar{\xi}_i, \bar{\xi}_{i+1}]$ with $\bar{\xi}_i = \frac{i-1}{k}$, for $i = 1 \cdots k$.

On interval $[\xi_i, \xi_{i+1}]$, let $u_i = \gamma(t)$ be the re-parametrization such that:

$$u_i = a_i t + b_i \ , \quad i = 1 \cdots k \ .$$

Considering the one order B-splines $b_i^0(t, \xi)$ and $b_i^0(u, \bar{\xi})$, the re-parametrization is evident:

$$b_i^0(t, \xi) = \begin{cases} 1 & , \quad t \in [\xi_i, \xi_{i+1}] \ , \\ 0 & , \quad t \text{ elsewhere} \ . \end{cases}$$

$$b_i^0(u, \bar{\xi}) = \begin{cases} 1 & , \quad u \in [\frac{i-1}{k}, \frac{i}{k}] \ , \\ 0 & , \quad u \text{ elsewhere} \ . \end{cases}$$

We develop the computation for two order B-splines:

$$b_i^1(t, \xi) = \begin{cases} \frac{t-\xi_i}{\xi_{i+1}-\xi_i} & , \quad t \in [\xi_i, \xi_{i+1}] \ , \\ \frac{\xi_{i+2}-t}{\xi_{i+2}-\xi_{i+1}} & , \quad t \in [\xi_{i+1}, \xi_{i+2}] \ , \\ 0 & , \quad t \text{ elsewhere} \ . \end{cases}$$

Then, for $t \in [\xi_i, \xi_{i+1}]$, $t = \frac{u-b_i}{a_i}$, and for $t \in [\xi_{i+1}, \xi_{i+2}]$, $t = \frac{u-b_{i+1}}{a_{i+1}}$. We thus obtain:

$$b_i^1(u, \xi) = \begin{cases} \frac{u-(a_i\xi_i+b_i)}{(a_i\xi_{i+1}+b_i-(a_i\xi_i+b_i)} & , \quad u \in [\frac{i-1}{k}, \frac{i}{k}] \ , \\ \frac{(a_{i+1}\xi_{i+2}+b_{i+1})-u}{a_{i+1}\xi_{i+2}+b_{i+1}-(a_{i+1}\xi_{i+1}+b_{i+1})} & , \quad u \in [\frac{i}{k}, \frac{i+1}{k}] \ , \\ 0 & , \quad u \text{ elsewhere} \ . \end{cases}$$

It follows that,

$$b_i^1(u,\bar{\xi}) = \begin{cases} \frac{u-\frac{i-1}{k}}{\frac{1}{k}} & , \quad u \in [\frac{i-1}{k}, \frac{i}{k}] \ , \\ \frac{\frac{i+1}{k}-u}{\frac{1}{k}} & , \quad u \in [\frac{i}{k}, \frac{i+1}{k}] \ , \\ 0 & , \quad u \text{ elsewhere} \ . \end{cases}$$

From this last computation, we can see that the re-parametrization only affects the knot vector. This is not the case for higher order B-splines. We develop below the computation for the three order B-spline $b_i^2(t)$ to show that the re-parametrization can be done analytically. We will not develop it for higher order but the computation is similar.

$$b_i^2(t,\xi) = \begin{cases} \frac{t-\xi_i}{\xi_{i+2}-\xi_i}\frac{t-\xi_i}{\xi_{i+1}-\xi_i} & , \quad t \in [\xi_i, \xi_{i+1}] \ , \\ \frac{t-\xi_i}{\xi_{i+2}-\xi_i}\frac{\xi_{i+2}-t}{\xi_{i+2}-\xi_{i+1}} + \frac{\xi_{i+3}-t}{\xi_{i+3}-\xi_{i+1}}\frac{t-\xi_{i+1}}{\xi_{i+2}-\xi_{i+1}} & , \quad t \in [\xi_{i+1}, \xi_{i+2}] \ , \\ \frac{\xi_{i+3}-t}{\xi_{i+3}-\xi_{i+1}}\frac{\xi_{i+3}-t}{\xi_{i+3}-\xi_{i+2}} & , \quad t \in [\xi_{i+2}, \xi_{i+3}] \ , \\ 0 & , \quad t \text{ elsewhere} \ . \end{cases}$$

The re-parametrization induces some equalities for $i = 1 \cdots k$:

$$\begin{aligned} \xi_{i+1} - \xi_i &= \frac{1}{ka_i} \ , \\ \xi_{i+2} - \xi_i &= \frac{1}{k}\left(\frac{1}{a_{i+1}} + \frac{1}{a_i}\right) \ . \end{aligned}$$

Then,

$$b_i^2(u,\xi) = \begin{cases} \frac{2}{1+a_i/a_{i+1}}\frac{u-(i-1)/k}{2/k}\frac{u-(i-1)/k}{1/k} & , \quad u \in [(i-1)/k, i/k] \ , \\ \frac{2}{1+a_{i+1}/a_i}\frac{u-(i-1)/k}{2/k}\frac{(i+1)/k-u}{1/k} + & \\ \quad + \frac{2}{1+a_{i+1}/ai+2}\frac{(i+2)/k-u}{2/k}\frac{u-i/k}{1/k} + & \\ \quad + \frac{a_{i+1}-a_i}{a_{i+1}+a_i}\frac{(i+1)/k-u}{1/k} + \frac{a_{i+1}-a_{i+2}}{a_{i+1}+a_{i+2}}\frac{u-i/k}{1/k} & , \quad u \in [i/k, (i+1)/k] \ , \\ \frac{2}{1+a_{i+2}/a_{i+1}}\frac{(i+2)/k-u}{2/k}\frac{(i+2)/k-u}{1/k} & , \quad u \in [(i+1)/k, (i+2)/k] \ , \\ 0 & , \quad u \text{ elsewhere} \ . \end{cases}$$

Bibliography

[1] L. Amate and M.J. Rendas. A contribution to the problem of mapping seabed transition zones. In *Proc. of Oceans'06, Boston, USA*, September 2006.

[2] L. Amate and M.J. Rendas. Sparse model fitting in nested families: Bayesian approach vs. penalized likelihood. In *Proc. Eusipco 2009, Glasgow, UK*, 2009.

[3] Christophe Andrieu, Nando de Freitas, and Arnaud Doucet. Reversible jump mcmc simulated annealing for neural networks. In *UAI '00: Proc. of the 16th Conference on Uncertainty in Artificial Intelligence*, pages 11–18, San Francisco, CA, USA, 2000.

[4] Christophe Andrieu, Nando De Freitas, and Arnaud Doucet. Robust full bayesian learning for radial basis networks. *Neural Comput.*, 13(10):2359–2407, 2001.

[5] A.M. Baumberg and D.C. Hogg. Learning flexible models from image sequences. In *In European Conference on Computer Vision*, pages 299–308. Springer-Verlag, 1993.

[6] Abhishek Bhattacharya and Rabi Bhattacharya. Nonparametric statistics on manifolds with applications to shape spaces. *IMS collections - Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, 3:282301, 2008.

[7] F.L. Bookstein. Size and shape spaces for landmark data in two dimensions. *Statictical Science*, 1(2):181–242, 1986.

[8] S. P. Brooks, N. Friel, and R. King. Classical model selection via simulated annealing. *Journal Of The Royal Statistical Society Series B*, 65(2):503–520, 2003.

[9] Olivier Cappé and Eric Moulines. Online expectation maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B*, 71, 2009.

[10] Gilles Celeux, Didier Chauveau, and Jean Diebolt. On stochastic versions of the em algorithm. Technical report, INRIA, 1995.

[11] Timothy F. Cootes, Christopher J. Taylor, David H. Cooper, and Jim Graham. Active shape models: Their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

[12] P. Craven and G. Wahba. Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.*, 31:377–403, 1979.

[13] Rhodri H. Davies, Carole J. Twining, Tim F. Cootes, John C. Waterton, and Chris J. Taylor. A minimum description length approach to statistical shape modeling. *IEEE Transactions on Medical Imaging*, 21:525–537, 2002.

[14] C. DeBoor. *A Practical Guide to Splines*. Springer, 1978.

[15] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum-likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society B*, 39:1–38, 1977.

[16] D. G. T. Denison, B. K. Mallick, and A. F. M. Smith. Automatic bayesian curve fitting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60:333–350(18), 1998.

[17] I. DiMatteo, C.R. Genovese, and R.E. Kass. Bayesian curve-fitting with free-knot splines. *Biometrika*, 88:1055–1071, 2001.

[18] I.L. Dryden and K.V. Mardia. *Statistical Shape Analysis*. John Wiley and sons, 1998.

[19] M. Figueiredo, J. Leito, and A. Jain. Unsupervised contour representation and estimation using b-splines and a minimum description length criterion. *IEEE Transactions on Image Processing*, 9(6):1075–1087, June 2000.

[20] M. A. T. Figueiredo and A. K. Jain. Unsupervised learning of finite mixture models. *IEEE Trans. Pattern Anal. Mach. Intell*, 24(3):381–396, March 2002.

[21] Gersende Fort and Eric Moulines. Convergence of the monte carlo expectation maximization for curved exponential families. *Annals of Statistics*, 31(4):1220–1259, 2003.

[22] Carlo Gaetan and jian-Feng Yao. A multiple-imputation metropolis version of the em algorithm. *Biometrika*, 90(3):643–654, September 2003.

[23] Jan Giebel and Dariu Gavrila. Multimodal shape tracking with point distribution models. In *Proceedings of the 24th DAGM Symposium on Pattern Recognition*, pages 1–8, London, UK, 2002. Springer-Verlag.

[24] Peter J. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82:711–732, 1995.

[25] S. Joshi, E. Klassen, A. Srivastava, and I. H. Jermyn. A novel representation for riemannian analysis of elastic curves in r^n. In *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, Minneapolis, USA, June 2007.

[26] S. Joshi, E. Klassen, A. Srivastava, and I. H. Jermyn. Removing shape-preserving transformations in square-root elastic (sre) framework for shape analysis of curves. In *Proc. Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR)*, Ezhou, China, August 2007.

[27] David L. B. Jupp. Approximation to data by splines with free knots. *SIAM Journal on Numerical Analysis*, 15(2):328–343, April 1978.

[28] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995.

[29] D. G. Kendall. Shape manifolds, procrustean metrics and complex projective space. *Bull. London Math. Soc.*, pages 81–121, 1984.

[30] D.G. Kendall, D. Barden, T.K. Carne, and H. Le. *Shape and Shape Theory*. John Wiley and sons, 1999.

[31] Eric Klassen, Anuj Srivastava, and Washington Mio. Analysis of planar shapes using geodesic paths on shape spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26:372–383, 2004.

[32] Mark Kliger and Joseph M. Francos. Map model order selection rule for 2-d sinusoids in white noise. *IEEE Transactions on Signal Processing*, 53(7):2563–2575, 2005.

[33] Mary J. Lindstrom. Penalized estimation of free-knot splines. *Journal of Computational and Graphical Statistics*, 8(2):333–352, 1999.

[34] K.V. Mardia and I.L. Dryden. Shape distributions for landmark data. *Adv. Appl. Prob.*, 21:742–755, 1989.

[35] Donald W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal on Applied Mathematics*, 11(2):431–441, 1963.

[36] Jorge S. Marques and Arnaldo J. Abrantes. Shape alignment – optimal initial point and pose estimation. *Pattern Recognition Letters*, 18(1):49–53, 1997.

[37] Washington Mio, Anuj Srivastava, and Shantanu Joshi. On shape of plane elastic curves. *Int. J. Comput. Vision*, 73(3):307–324, 2007.

[38] Radford M. Neal and Geoffrey E. Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. *Learning in graphical models*, pages 355–368, 1999.

[39] R. H. J. M. Otten and L. P. P. P. van Ginneken. *The Annealing algorithm*. Kluwer Academic, 1989.

[40] Xavier Pennec. Probabilities and statistics on Riemannian manifolds: Basic tools for geometric measurements. In *Proc. of Nonlinear Signal and Image Processing (NSIP'99)*, volume 1, pages 194–198, June 20-23, Antalya, Turkey, 1999.

[41] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

[42] S. Rolfes. *Stochastic geometry : an approach to featureless perception-based robot navigation*. PhD thesis, University of Nice-Sophia Antipolis, December 2002.

[43] Peter Schnemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, March 1966.

[44] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[45] Christopher G. Small. *The statistical theory of shape*. Springer series in statistics, 1996.

[46] D. M. Titterington. Recursive parameter estimation using incomplete data. *Journal Of The Royal Statistical Society Series B*, 46(2):257–267, 1984.

[47] G. Wahba. *Spline models for observational data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1990.

[48] K.N. Walker, T.F. Cootes, and C.J. Taylor. Automatically building appearance models from image sequences using salient features. *Image and Vision Computing*, 20(5-6):435–440, April 2002.

[49] Larry Wasserman. Bayesian model selection and model averaging. *J Math Psychol*, 44(1):92–107, March 2000.

[50] Greg C. G. Wei and Martin A. Tanner. A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411):699–704, 1990.