UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

Fakultät für
Elektrotechnik,
Informatik und
Mathematik

Extensions of statistical shape models for medical imaging and computer vision

Von der Fakultät für Elektrotechnik, Informatik und Mathematik
der Universität Paderborn

zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften (Dr.-Ing.)

genehmigte Dissertation
von

M. E. Alma Eguizabal Aguado

Erster Gutachter:	Prof. Dr. Peter Schreier
Zweiter Gutachter:	Prof. Dr.-Ing. Reinhold Häb-Umbach

Tag der mündlichen Prüfung: 16.03.2020

Paderborn 2020

Diss. EIM-E/350

Propongo un brindis preventivo.
Por si acaso todo sale bien.

—*El Kanka*

Affiliations

Signal and System Theory Group
Dept. of Electrical Engineering and Information Technology (EIM-E)
Universität Paderborn

Abstract

Statistical Shape Models (SSM) are beneficial in image processing and computer vision. SSM account for the variability of the shape from the objects that are present in the images. There are different reasons why an object may vary its shape from observation to observation. For instance, in medical imaging the anatomy changes from patient to patient. SSM are information *a priori* about the shape of objects, and are used as constraints by image processing and machine learning approaches. Given prior knowledge of the position of some meaningful points belonging to the contour of the object, i.e., the landmarks, a supervised segmentation method can find these landmarks in new images. Point Distribution Models (PDM) are landmark-based SSM whose parameters are learned from training samples, and are promising approaches to solve segmentation problems. PDM are linear models, and therefore provide efficiency, make the estimation procedures simple and have an intuitive interpretation.

There are different issues to be addressed in SSM, and hence improve their segmentation accuracy. To begin with, registering the PDM landmarks in new images requires good image contrast. Therefore, in challenging medical imaging modalities, such as fluoroscopic X-ray, the state-of-the-art PDM-based segmentation algorithms do not provide enough quality. Additionally, heuristics are frequently present in SSM. One example is the model-order selection of the PDM, which is often chosen with a "rule-of-thumb" percentage of variance. *Ad hoc* designs complicate the generalization, and the lack of optimization makes SSM lose potential. Moreover, when training the PDM, we require a collection of sampled points (the landmarks) of the training contours. These points are very often manually selected, and based only on experience or subjective criteria.

The goal in this work is to study how to overcome these issues and enhance the segmentation performance of the SSM, and more particularly the PDM. Firstly, we have studied ways to add robustness to the segmentation algorithms and at the same time keep their simplicity. Secondly, we have designed an information-theoretic model-selection technique that provides optimized PDM with good generalization. Finally, we have considered how to register the training boundaries without the need of the corresponding manual landmarks using Dynamic Time Warping.

Zusammenfassung

Statistische Formmodelle (Statistical Shape Models/ SSMs) sind für die Bildverarbeitung von Vorteil und berücksichtigen die Formvariabilität von Objekten, die in den Bildern vorhanden sind. Es gibt verschiedene Gründe, warum ein Objekt seine Form von Beobachtung zu Beobachtung verändern kann. Beispielsweise ändert sich in der medizinischen Bildgebung die Anatomie von Patient zu Patient. Statistische Formmodelle beinhalten a priori Informationen über die Form von Objekten und werden von Bildverarbeitungs- und maschinellen Lernansätzen als Einschränkungen verwendet. Bei vorheriger Kenntnis der Position einiger bestimmter Punkte, die zur Kontur des Objekts gehören, den sogenannten Landmarken, kann ein überwachtes Segmentierungsverfahren diese Landmarken in neuen Bildern finden. Punktverteilungsmodelle (Point Distribution Models/ PDMs) sind auf Landmarken basierende statistische Formmodelle, deren Parameter aus Trainingsmustern erzeugt werden und vielversprechende Ansätze zur Lösung von Segmentierungsproblemen darstellen. Punktverteilungsmodelle (PDMs) sind lineare Modelle und daher effizient anzuwenden, vereinfachen die Schätzverfahren und sind intuitiv zu interpretieren.

In statistischen Formmodellen müssen verschiedene Probleme behoben werden, um die Segmentierungsgenauigkeit zu verbessern. Um die PDM-Landmarken in neuen Bildern zu identifizieren, sind zunächst ein guter Kontrast und eine gute Bildqualität erforderlich. Daher bieten die modernen PDM-basierten Segmentierungsalgorithmen bei anspruchsvollen medizinischen Bildgebungsmodalitäten, wie fluoroskopischem Röntgen, nicht genügend Genauigkeit. Zusätzlich werden häufig Heuristiken und Ad-hoc-Regeln verwendet. Ein Beispiel ist die Auswahl der Modellordnung des PDMs, die häufig mit einer Faustformel basierend auf der Varianz erfolgt. Darüber hinaus wird für das PDM-Training eine Sammlung von Stichprobenpunkten (Landmarken) der Trainingskonturen benötigt. Diese Punkte werden sehr oft manuell ausgewählt und basieren nur auf Erfahrungen bzw. subjektiven Kriterien.

Ziel dieser Arbeit ist es, zu untersuchen, wie diese Probleme behoben und die Segmentierungsleistung des SSMs und insbesondere des PDMs verbessert werden kann. Um die Robustheit der Segmentierungsalgorithmen zu erhöhen, wurden verschiedene Möglichkeiten untersucht. Des Weiteren wurde eine informationstheoretische Modellauswahlmethode entwickelt, die in einem optimierten PDM resultiert. Abschließend haben wir uns überlegt, wie Sie die Landmarken zen mithilfe von Dynamic Time Warping registrieren können, ohne die entsprechenden manuellen Orientierungspunkte zu benötigen.

Acknowledgements

The start of this section is an overused sentence, but it is completely honest: this thesis would have not been possible without the support, advice, help, patience, hard work and time of so many people. They absolutely deserve this section.

First of all, I would like to express my deepest gratitude to my advisor and *Doktorvater* Prof. Peter Schreier. Six years ago he gave me the opportunity to work in his group, where I became the *pioneer* PhD student. Peter has worked on transmitting his wisdom to me, professionally and personally, with lots of patience and understanding. I am very happy to have learnt from him many of my best working strengths of today. Also, I am truly thankful for the enriching Industry-Academia doctoral path that Peter prepared. First, I joined his industrial project with Stryker during two years. At that time I worked together with Dr. Ulrich Hoffmann and Arno Blau, to whom I am very thankful as well. Uli and Arno guided me along my very first steps in industrial research. The second phase of this doctoral path was academic: during four more years I could develop this thesis and my teaching skills. It has been overall a very rewarding career experience. Furthermore, during these six years Peter has provided all necessary funding, including *Boney* and the amazing chance to attend two summer schools to strengthen my background.

I also want to thank the SST members and ex-members for being good colleagues and friends. Especially David, who is a professional model to me, and at the same time a good friend. David always had time to discuss, and gave me great ideas to my thesis. Also, my friends Yang and Wen, who gave me support and very good advice, from close by and from far away. The SST group would not be the same without the enormous heart of Annaleinchenlein. She deserves more than a thousand thanks. Anna has been my family in Paderborn, and helped me survive every little (and big) issue that one encounters abroad. Thank you Annalein for your *Buenos Aires* and your precious time. Furthermore, one of the very interesting opportunities of being an SST member is to meet Peter's distinguished and talented visitors. I would like to thank very much Prof. Louis Scharf (my *Doktorgrossvater*) for his great discussions about UME, and his tips for a successful career. Also, I thank Prof. Barry Quinn, whose discussions have pushed forward part of the contributions of this thesis. A special thanks to Nacho for his interesting discussions about manifolds (and also to Steven, who I visited instead, but gave me so great hints about DTW). And many thanks to Prof. Tülay Adali for being such a great model to me, as well as for her wise advice.

Me voy a permitir el lujo de seguir en español. Esta tesis no habría sido posible sin Olga. La confianza en mí misma que me permite lanzarme a por todas y asumir mis fracasos se la debo a ella. Gracias a Olga conocí el mundo de la investigación.

Mis primeros sueños sobre la carrera investigadora fueron con ella. Al igual que, por supuesto, mis primeros éxitos. Olga estaba ahí cuando fui por primera vez coautora de un *paper* (en el CSIC con César y Andreu, a los que agradezco su ilusión y su pasión por la Ciencia, que me inspiraron muchísimo). Olga me guió para escribir mi primera contribución de conferencia (y todas las que vinieron después), y siempre hizo todo lo posible para que yo pudiera asistir a presentar mis trabajos y disfrutar de ello. Gracias a Olga publiqué mi primer *paper* de *journal*, porque ella creía en mí, tiraba de mí si hacía falta, y buscaba mis éxitos. Juntas conseguimos una FPU, dos becas del SPIE, y un largo etcétera. Olga me mandó a Dartmouth (mil gracias a Ashley y a Brian) y a Munich (un millón de gracias a Bea y a Zsuzsi). Siempre pensando en lo que era bueno para mi carrera investigadora. Todas esas experiencias son la base de mi presente y mi futuro profesional. Gracias Olga por despertar en mí a una investigadora curiosa, y por todo tu apoyo durante estos años.

Mi *personal coach* durante este tiempo ha sido mi mejor amiga Marina (Dra. Torres Trimállez). Pasando cada una por su tesis hemos compartido consejos de supervivivencia 24/7, y su apoyo ha sido imprescindible. Por supuesto no olvido a Javi, que también desde Santander siempre ha escuchado mis lloros, y celebrado mis éxitos. Y muchísimas gracias a los dos por vuestro apoyo cuando decidí irme a Alemania. Vuestros mensajes de ánimo aún seguían en mi pared cuando terminaba esta tesis. Paderborn ha sido un lugar mucho más familiar gracias a los grandes amigos que he conocido, y que voy a echar muchísimo de menos al finalizar esta tesis. Ellos también han escuchado mis penas doctorales, me han dado buenísimos consejos, y siempre me han ayudado en todo lo que he necesitado. Mil gracias Soraya y Roland, Caco y Bettina, Ali y Jens.

Mención especial para la familia. Porque son el pilar donde se sostiene todo. Muchísimas gracias a mis padres Ignacio y Dori, y mi hermano Jose, por creer en mi (cliché lo sé, pero es totalmente cierto). Por apoyarme en esta aventura en Alemania. Por estar tan orgullosos de mis éxitos, y por ayudarme a levantarme tras los fracasos. Incluso por los *speech*, porque en el fondo sé que son necesarios. A mis tíos, mis primos y mi abuelo, por estar pendientes de mi tesis, y siempre celebrando mis avances. A mis suegros Pedro y Pili por preocuparse por mi cuando he estado lejos, y por desearme suerte. Lo mismo que mis cuñados Connie, Pedro y Mario. En especial Connie, que siempre ha estado pendiente por el *Whatsapp* de mis avances y me ha dado fantásticos consejos. Y mil gracias a mi sobrina Nicole por su amor incondicional y su alegría, que me han aportado una energía preciosa y muy valiosa este tiempo. Y por último, mi mención más especial va para mi chico, mi amigo, mi compañero de trabajo, mi compañero de carrera, mi compañero de piso, y mi compañero de vida: Christian. Esta tesis se ha terminado sólo porque tú sabías que yo podía hacerlo.

Alma Eguizabal

Paderborn, July 2019

Prólogo

Una tesis es un camino, no una senda perdida (Holzwege), pero menos aún algo predecible, con un comienzo y un final feliz. Esta investigación ha sido una tentativa, rigurosa, bien fundada pero no cerrada, en la que no se adivina el final más que en un horizonte. Solo nos queda avanzar sabiendo que si los resultados parecen felices, y lo parecen, el horizonte se irá acercando, y si no lo son tanto... se habrán de desechar ciertas estrategias y abrir, además, otros caminos. Porque la aventura del conocimiento, si sabe ser realista y por tanto discreta, agradecerá los resultados sean estos los que fueren. No avanzará el conocimiento menos en un caso que en el otro.

Esta tesis se ha encuadrado en el campo de la salud. No necesita, por tanto, como pensaba Ortega -con ese prurito de la filosofía de situarse un poco por encima de la ciencias y también de la ingeniería, - una dirección previa, unos valores. La dirección la ha puesto la autora y no puede ser más noble: la vida, la salud. Este trabajo, como decíamos, abre nuevos caminos en sus aciertos y señala vías cerradas, no transitables. Avanza siempre. Parafraseando a Breton, podríamos decir que esta tesis que se presenta a continuación es el comienzo y solo el comienzo.

José Ignacio Eguizabal

Notation and Acronyms

Used Notation

a	Scalar (lowercase)
$\mathbf{a}$	Column vector (lowercase boldface)
$\mathbf{A}$	Matrix (uppercase boldface)
$\mathbf{A}^{-1}$	Inverse of square matrix $\mathbf{A}$
$\mathbf{A}^{T}$	Transpose of matrix $\mathbf{A}$
$\mathbf{A}^{H}$	Conjugate transpose (Hermitian) of matrix $\mathbf{A}$
$\mathrm{Tr}(\mathbf{A})$	Trace of matrix $\mathbf{A}$
$\mathrm{Re}(a)$, $\mathrm{Im}(a)$	Real and imaginary part of a
$\lvert a \rvert$	Absolute value of a
$\arg(a)$	Phase of a
$\exp(a)$	The natural exponential function of a
$\log(a)$	The natural logaritmus function of a
$\lVert \mathbf{a} \rVert$	ℓ^2-norm of vector $\mathbf{a}$
$\lvert \mathbf{A} \rvert$	Determinant of square matrix $\mathbf{A}$
$\lVert \mathbf{A} \rVert_F$	Frobenius norm of matrix $\mathbf{A}$
$\mathrm{blkdiag}(\mathbf{A}_1, \ldots, \mathbf{A}_N)$	Block-diagonal matrix with matrices $\mathbf{A}_1, \ldots, \mathbf{A}_N$ along the main diagonal
$\mathrm{diag}(a_1, \ldots, a_N)$	Diagonal matrix with elements $a_1, \ldots, a_N$ along the main diagonal
$\mathrm{vec}(\mathbf{A})$	Column vector comprised of a vertical concatenation of the columns of matrix $\mathbf{A}$
$[\mathbf{A}_1, \ldots, \mathbf{A}_N]$	Horizontal concatenation of matrices $\mathbf{A}_1, \ldots, \mathbf{A}_N$
$[\mathbf{A}_1; \ldots; \mathbf{A}_N]$	Vertical concatenation of matrices $\mathbf{A}_1, \ldots, \mathbf{A}_N$
$\mathbf{A} \succeq \mathbf{B}$	Matrix $\mathbf{A} - \mathbf{B}$ is positive semidefinite
$a[n] * b[n]$	Linear convolution of sequences $a[n]$ and $b[n]$
$\mathbf{A} \otimes \mathbf{B}$	Kronecker product of matrices $\mathbf{A}$ and $\mathbf{B}$
$\mathbf{A}^{\star}$	Optimal solution of an optimization problem in the matrix variable $\mathbf{A}$
$\nabla_{\mathbf{A}} f(\mathbf{A})$	Derivative of $f(\mathbf{A})$ with respect to $\mathbf{A}$
$\mathbf{I}_N$	$N \times N$ identity matrix (subscript is omitted when the dimension is self-evident)
$\mathbf{0}_N$	$N \times 1$ zero vector (subscript is omitted when the dimension is self-evident)

$\mathbf{1}_N$	$N \times 1$ ones vector (subscript is omitted when the dimension is self-evident)
$\mathbb{R}^{M \times N}$	Space of $M \times N$ real matrices
$\mathbb{C}^{M \times N}$	Space of $M \times N$ complex matrices
$O(N)$	Orthogonal group of dimension N
$SO(N)$	Special orthogonal group of dimension N
$\mathbb{C}P^N$	Complex projective space of dimension N
$\mathbb{S}^N$	Hypersphere of dimension N
$\mathbb{T}$	The scale/translation-normalized space
$\mathbb{R}^+$	Real positive scalars
$d_p(\mathbf{a}, \mathbf{b})$	Procrustes distance between $\mathbf{a}$ and $\mathbf{b}$
$d_s(\mathbf{a}, \mathbf{b})$	Shape distance between $\mathbf{a}$ and $\mathbf{b}$
$g(\mathbf{a}, \mathbf{b})$	Geodesic distance between the pre-shapes $\mathbf{a}$ and $\mathbf{b}$
$\sim$	Distributed as
$\mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Multivariate complex Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Multivariate real Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$\mathcal{X}^2_\ell$	Chi-squared distribution with ℓ degrees of freedom
$\Gamma(a)$	Gamma function of scalar a
$\mathrm{E}[X]$	Mathematical expectation of random variable X
σ_x	Standard deviation of random variable X
$p(X)$	Probability density function of random variable X
$P(X \leq x)$	Probability of random variable X being equal to or lower than x
$\mathcal{L}(x)$	Likelihood of parameter x
$GP(\mathbf{X}, \boldsymbol{\Gamma})$	Gaussian Process with mean deformation $\mathbf{X}$ and covariance function (kernel) $\boldsymbol{\Gamma}$

Acronyms

AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
CAS	Computer-Assisted Surgery
CLM	Constrained Local Models
CPD	Coherent Point Drift
CT	Computed Tomography
DNN	Deep Neural Network
DTW	Dynamic Time Warping
EM	Expectation Maximization
GLS	Generalized Least Squares
GP	Gaussian Process
GPA	Generalized Procrustes Analysis
ICP	Iterative Closest Point
KKT	Karush–Kuhn–Tucker
LMMSE	Linear Mimimum Mean Squared Error
MAP	Maximum A Posteriori
MDL	Minimum Description Length
ML	Maximum Likelihood
MM	Majorization Minimization
MSE	Mean Squared Error
OLS	Original Least Squares
PCA	Principal Component Analysis
PDM	Point Distribution Model
RGB	Red Green Blue
RMSE	Root Mean Squared Error
SNR	Signal to Noise Ratio
SSM	Statistical Shape Models
SVD	Singular Value Decomposition
UME	Universal Manifold Embedding

Contents

Part I

Introduction and background

Chapter 1
Introduction

"Understand well as I may, my comprehension can only be an infinitesimal fraction of all I want to understand."

— Ada Lovelace

The core of this thesis is the extension of Statistical Shape Models. These are, as the name implies, mathematical models of the variability in the shape of objects. Such an analysis could be understood from different disciplines: Image Processing, Computer Vision, Machine Learning or Artificial Intelligence. In this introductory chapter, these different perspectives are described in Section 1.1 to give the reader a broad understanding of the possibilities of the statistical shape analysis. Then, in Section 1.2, the mathematical description of shape and the typical techniques to model deformation are introduced, as well as a brief overview of some of the most common applications of shape analysis. Finally, in Section 1.3 we motivate the need of our contribution, that is, why the existent shape models should be extended, and how we propose to address these extensions in this work. In Section 1.4 we provide an outlook of the thesis.

1.1 A framework for Statistical Shape Models

In this section we provide a possible framework, as illustrated in Fig. 1.1, to situate the connection between shape analysis and the most related disciplines. Statistical Shape Models (SSM) can be understood from different fields. One is Artificial Intelligence (AI). AI develops computer systems and tasks that perform as if they were human. A very challenging branch is Computer Vision, which aims to emulate or even outperform the human visual perception. This is an ambitious assignment, considering there are studies that show that 80–85% of our learning, cognition and perception-related activities are mediated through our vision [2]. Human vision and its understanding is the result of many years of evolution of our perceptual power to recognize objects and faces [3] [4]. Image Processing deals with the tools to enhance the information within the images. In particular, segmentation methods subdivide an image into meaningful regions [5]. Next, Machine Learning provides the optimiza-

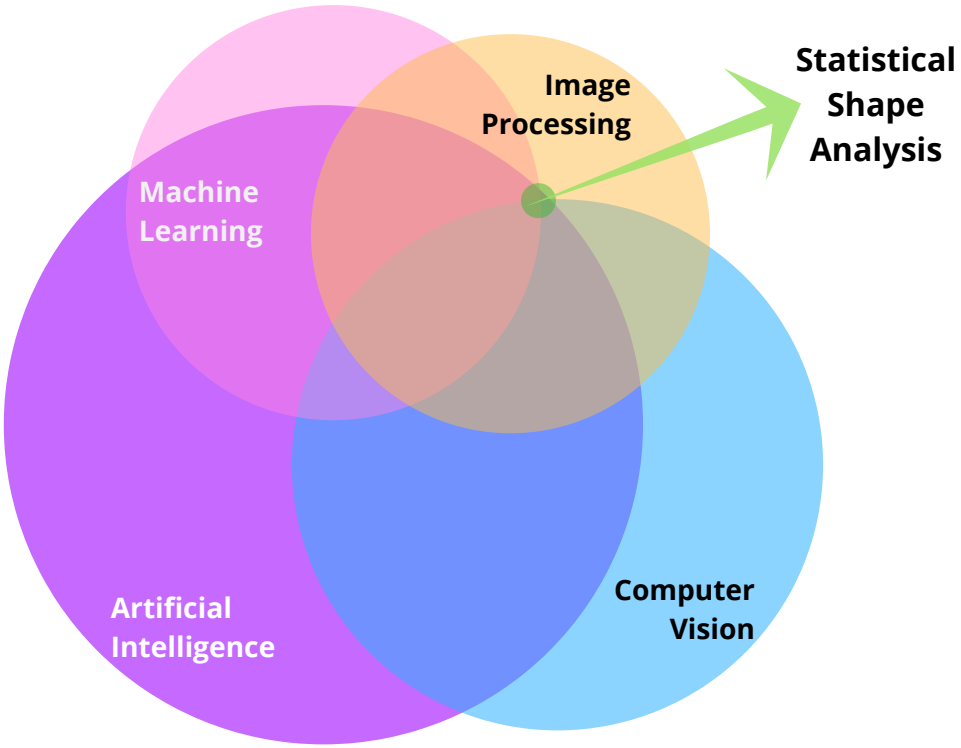


Figure 1.1: Venn's diagram to illustrate the connections between Statistical Shape Analysis and related engineering disciplines [1].

tion techniques and strategies to the computer to recognize data patterns, acquire experience from observations, and perform semantic interpretations of the images.

- **Artificial Intelligence** was first used as a concept at a conference in Dartmouth College in 1956, where it was discussed as a theoretical paradigm. After that, its impact on technology development, and on every engineer discipline is irrefutable. It has been especially remarkable since this last decade, in which the information revolution has established a break point of improvement for AI. Massive amount of data, together with a huge computational power, are transforming machines to a much more competitive level. AI has actually been considered the new electricity [6]. According to Google Dictionary, AI is "*the theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages*". Considering this definition, Computer Vision can be seen as the branch from AI that deals with the visual perception. Computer Vision is also described as a disorganized intellectual frontier [7]. Consequently it is difficult to find the line between Computer Vision and related fields like Image Processing and Machine Learning.

- **Computer Vision** brings visual perception to machines, enhancing the actual human perception. Technology allows humans to perceive more than the visible spectrum thorough the eyes: computers can handle radar, infrared, X-ray or

magnetic resonance, and generate from them many kinds of images, series of images, or in general multidimensional captures of the real world. Computer Vision deals with the understanding of cameras and different sensors, and the generation of every pixel (or voxel) [7]. Also, it handles the interpretation of this information, from the inferences between pixels to the recognition of objects. The statistical shape analysis plays an important role on both parts: a volumetric object has different projected shapes depending on the characteristics and positions of the camera/sensor and the source of illumination/radiation. The variability of these shapes can thus be studied from a Computer Vision perspective. We illustrate this motivation in Fig. 1.2.

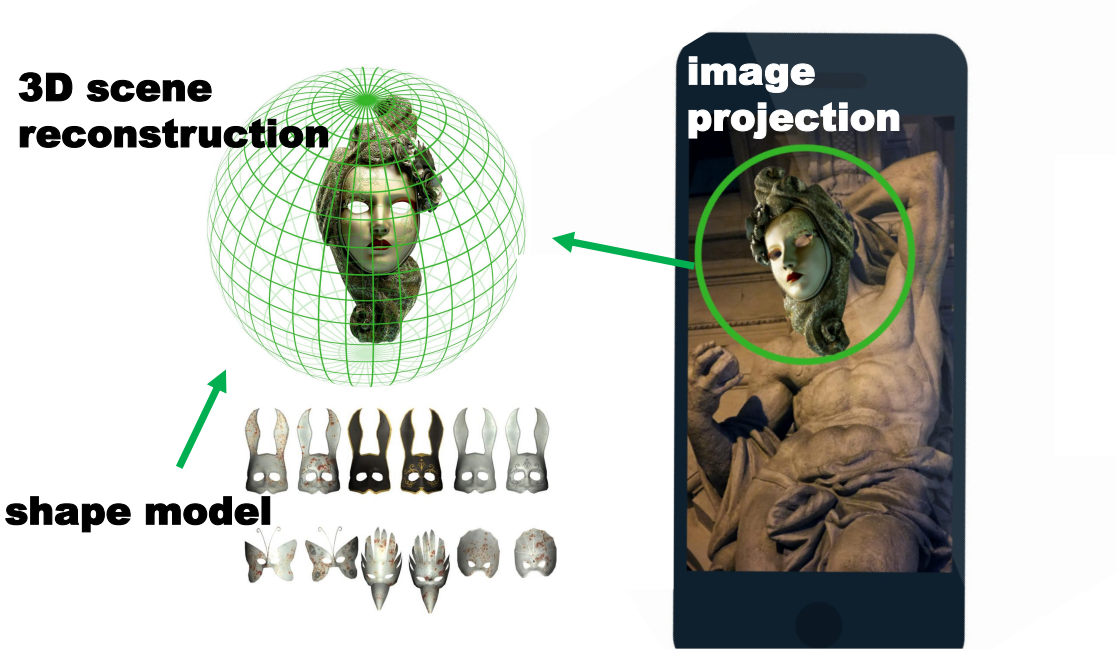


Figure 1.2: An illustration of a Computer Vision perspective of shape analysis. A model of deformation of the object in the scene (a mask) can enhance the understanding of the camera position and the image projection plane.

- **Image Processing** deals with the analysis and transformation of images in order to improve their usability, the performance of the processes depending on them, and the quality of the visual experience they provide. For example, a filter that provides smoothness to enhance the gradients and provide better edge detections [8]. The edges are used to find the boundaries of the object in the images. When the contours of the objects are lines and circles, these are mathematically parametrizable and simple. Hence, their segmentation is relatively efficient with a Hough Transform [5]. When the objects present shapes that are morphologically more challenging the statistical shape analysis enhances the processing. This interpretation is illustrated in Fig. 1.3.

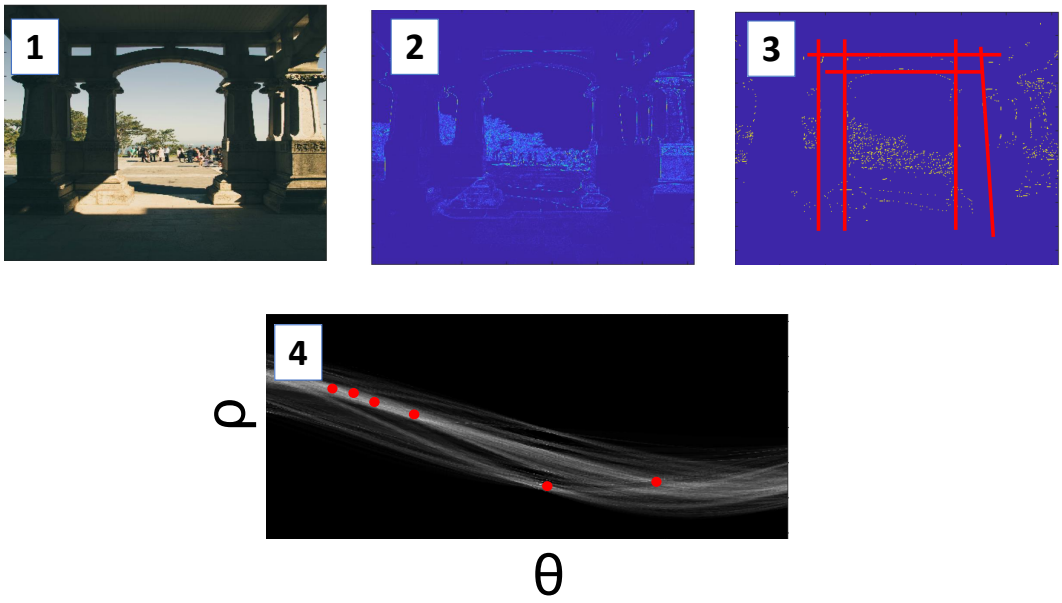


Figure 1.3: An illustration of an Image Processing perspective of shape analysis. From an RGB image (1), and its gradients (2), the edges can be calculated (3). The shape of some objects, such as the arc, can be modelled with basic geometry (red lines in 3). A line is modeled with two parameters (θ and ρ), which can be estimated from the peaks of a Hough Transform (4) of the edge image.

- **Machine Learning** is the combination of data analysis, statistics and optimization techniques to provide computers with the skill of solving complicated tasks that require human intellect, and improve the performance after the experience of solving them. Statistical shape analysis enhances machine learning algorithms providing prior information about the deformation of the objects with witch the learning algorithms can check the feasibility of a solution. For instance, as illustrated in Fig. 1.4, a shape model of human skeleton provides the information to discard output candidates from a machine learning approach that segments the bone in images.

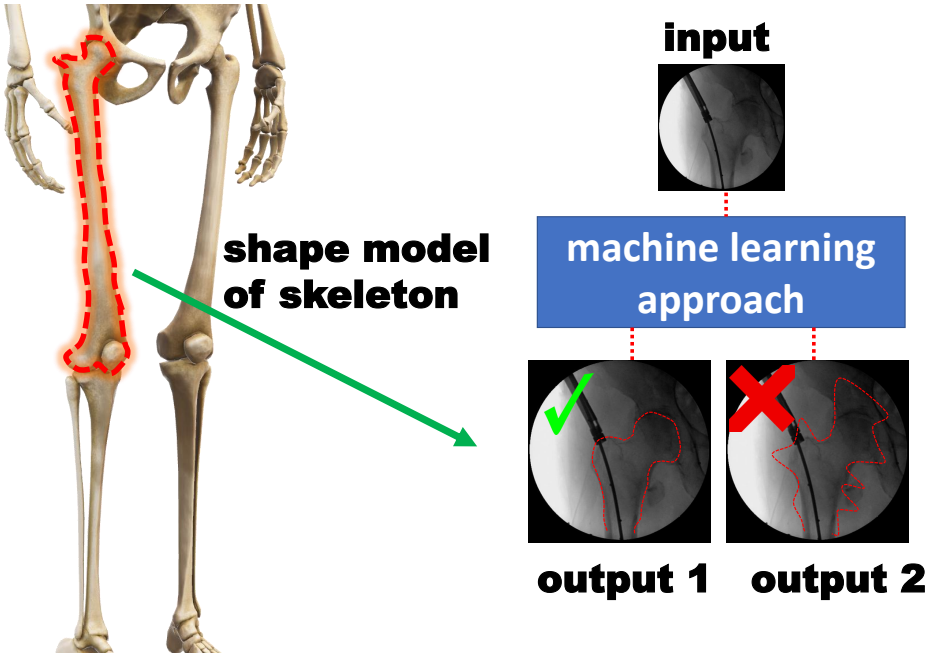


Figure 1.4: An illustration of a Machine Learning understanding of shape analysis. A shape model of the skeleton constraints the possible outputs from the machine learning approach and plausibility. The sources of medical images are described in the Appendix.

1.2 An overview about shape models

As motivated in previous section, Computer Vision, Image Processing and Machine Learning algorithms need mathematical models of the variability of shape in order to handle the shape information. There are different approaches of deformable models in images and volumes. In order to provide an intuition to the reader on how broad these possibilities are, we provide a few examples on how to address deformation with models.

There are many applications where these models are useful. For instance, a deformable model of shape is essential in medical image analysis. The statistical shape theory started motivated by the study of archaeology and anthropology [9]. In general, deformation information about the objects in a scene has big potential in artificial vision. In the second part of this section, we introduce some of these application to motivate the study of shape analysis in this work.

1.2.1 Models for shape and deformation

We give a brief motivation to the mathematical analysis of shape. When the shape of the object is approximated with basic geometry, a hand-crafted model can be considered. To enhance the precision of the shape model, the interpolation of splines can be used. The alternative that we have considered in this work is the landmark-based

models. These are determined by a collection of points that sample the contour of an object. We present an overview of these models used in shape analysis

Hand-crafted combination of basic geometry

Complex shapes can often be described as a combination of simple geometry, such as lines and circles. As an illustrative example we present an X-ray image of the femur in Fig. 1.5. The femoral shaft is approximated by two lines, and the femoral head by a circle. The deformation of these basic geometry is modelled with only a few parameters (two for a line, three for a circle).

A very important inconvenient of these models is that they are hand-crafted, and each new object would need a complete new model. Also, the models do not generally account for the whole contour of the object, as they depend on small geometrical pieces that are very basic. Since they are too simple, they may also not fit accurately the real contour. For instance, the femoral head may sometimes be elliptical or deformed, and a circle may lose precision in the segmentation. Still, a circle model has been successfully used to segment femoral heads in X-ray images [10].

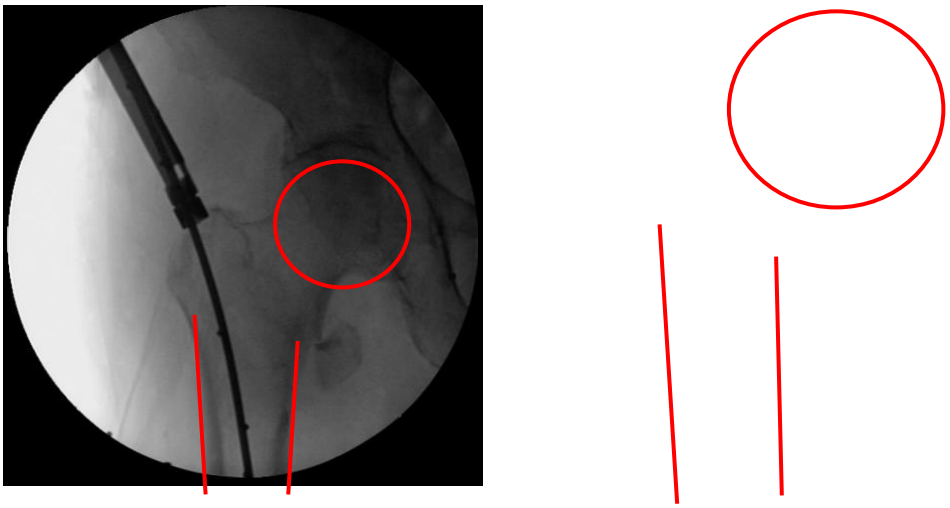


Figure 1.5: A hand-crafted model of the proximal femur shape, based on two lines on the shaft and a circle on the femoral head. (Medical image sources are detailed in the Appendix).

Interpolation of splines

In order to describe the contour with higher precision, it can also be modelled as a collection of spline interpolants [11]. In Fig. 1.6 we present an illustrative example

of a set of splines and their control points that define the femur contour. Cubic B-splines, for instance, have been used in medical image analysis to model anatomical deformation [12].

Another possibility is a Fourier decomposition of the boundaries [13]. A contour is decomposed as a collection of ellipses. Shape constraints can be considered in the parameters that model the Fourier synthesis. These, however, may not be accurate in very sharp corners.

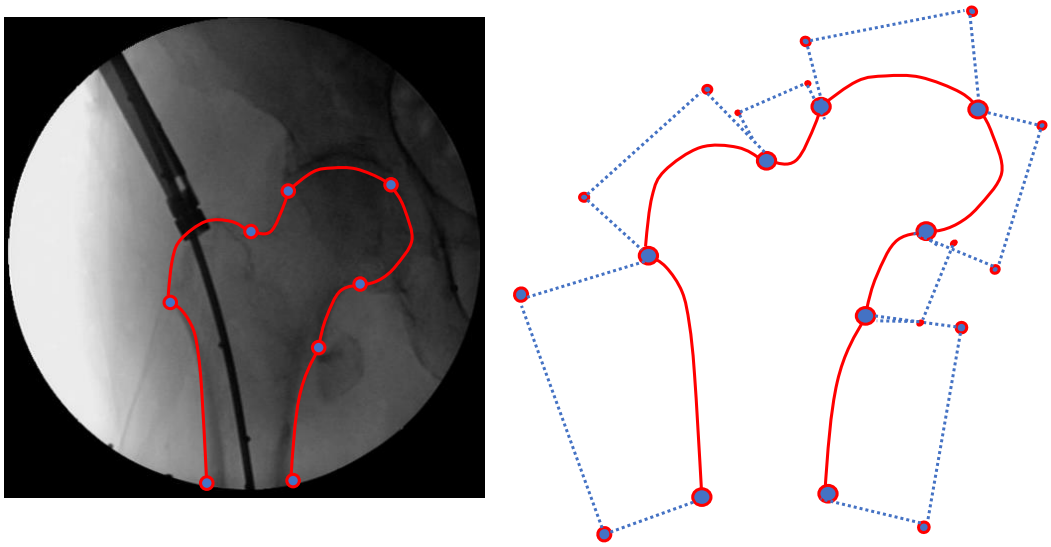


Figure 1.6: Illustrative example of a model of the femur contour as an interpolation of splines. The smaller dots represent the control points of the splines. (Medical image sources are detailed in the Appendix).

Landmark-based shape model

Landmarks are significant points that belong to the object boundary. For instance, the beginning and the ending points in the femoral head contour are anatomically interesting to study the femur shape. Shape boundaries can be described as collections of these landmarks, which may be manually chosen and labeled, or directly sampled from the contour [14]. Shape variability is then modeled as the statistical behaviour of the coordinates of these landmarks. Coming back to our example, we may study the distance between the beginning and ending points of the femoral head, and how much this distance changes among different patients. An illustrative example of a landmark-based shape for the femur is shown in Fig. 1.7.

The coordinates of these landmarks are typically represented in a vector. When considering vectors, linear algebra can be used: shape deformation can be modeled

with linear equations, which provide computational efficiency to the developed algorithms, and at the same time give consistent geometrical representations [15]. Shape statistics are extracted from the covariance matrix of these shape vectors. Typically a Principal Component Analysis, i.e., an eigenvalue decomposition of the covariance matrix, is performed to obtain a collection of eigenvectors describing shape deformation as a linear combination. We illustrate these linear combination of eigenvectors in Fig. 1.8. When the covariance matrix is learnt from training data, these models are known as Point Distribution Models (PDM) [14].

The SSM used in this thesis are landmark-based.

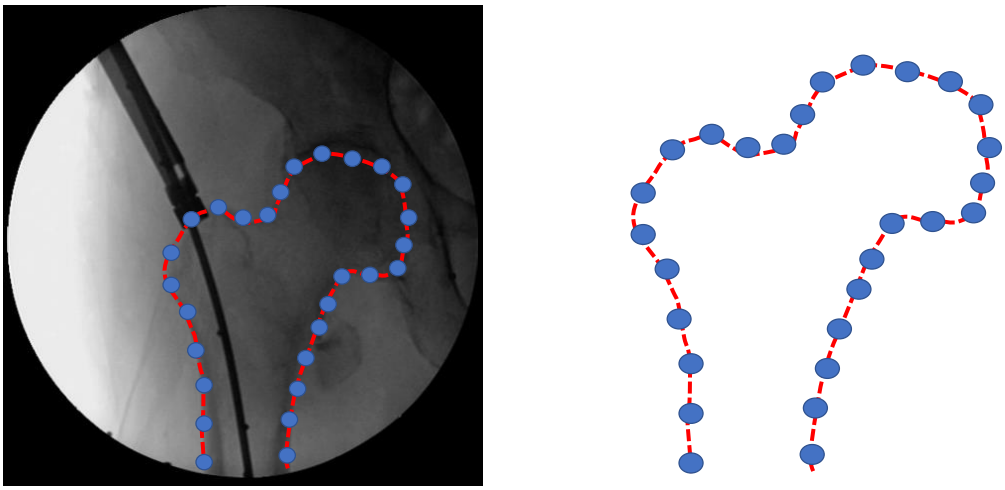


Figure 1.7: A landmark-based shape composed by a collection of points (blue dots) on the boundary (red line) of the object of interest (the femur). (Medical image sources are detailed in the Appendix).

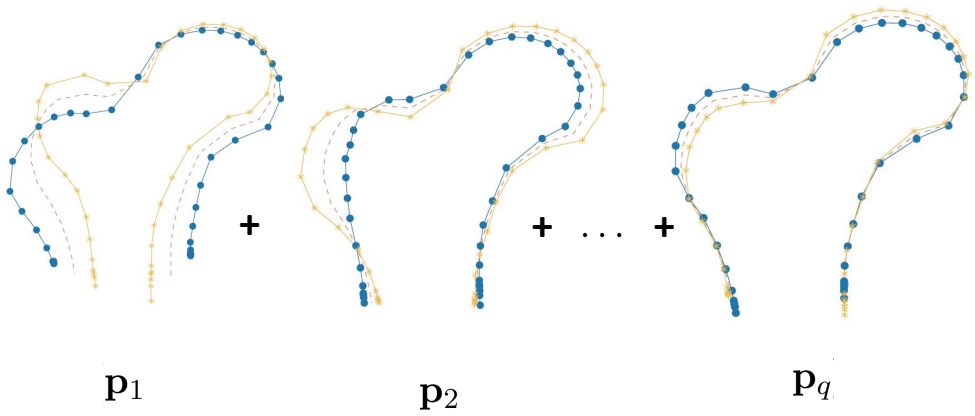


Figure 1.8: The deformation of the landmark-based contour is accounted from a collection of eigenvectors $\mathbf{p}_i$, each representing a mode of deformation that is uncorrelated from the other modes.

1.2.2 Applications of shape analysis

There are multiple and varied applications for shape analysis. Shape models are used as prior knowledge about the deformation of the objects that are segmented or detected. Objects contain deformation for multiple reasons: different views from the same object when a camera changes the focal point, industrial products that present defects, and anatomical parts of the body that are different for each individual.

We overview three of the most common applications of statistical shape analysis. These are the applications that we have observed more frequently during the development of this thesis.

Medical image analysis

A very important application of shape analysis is medical imaging. Radiology is a discipline that can greatly benefit from Image Processing and Machine Learning [16]. The anatomy of the human body changes considerably from individual to individual, and also due to different pathologies, accidents and abnormalities. This collection of deformations are successfully modeled with shape analysis. Radiologists are constantly observing medical images, and they need to accurately define the anatomical boundaries. When this task is performed automatically, the radiologists can save time and be more efficient.

Computer Assisted Surgery (CAS) is also an interesting medical application for SSM [10]. The principal dataset used in this thesis (see Appendix for more details) is formed by CAS images. These images were generated during the collocation of an intramedullary nail in a trauma surgery. In this procedure, the surgeon needs to insert

the nail inside the medullary cavity of the femur to fix a femoral fracture. A CAS navigation system guides the surgery team through the procedure, enhancing the accuracy of the placement, and improving the quality of life of the patients after the surgery. Such CAS navigation system needs an automatic detection of the boundaries of the bones. For this, we can use SSM. An illustration of this CAS application is depicted in Fig. 1.9.

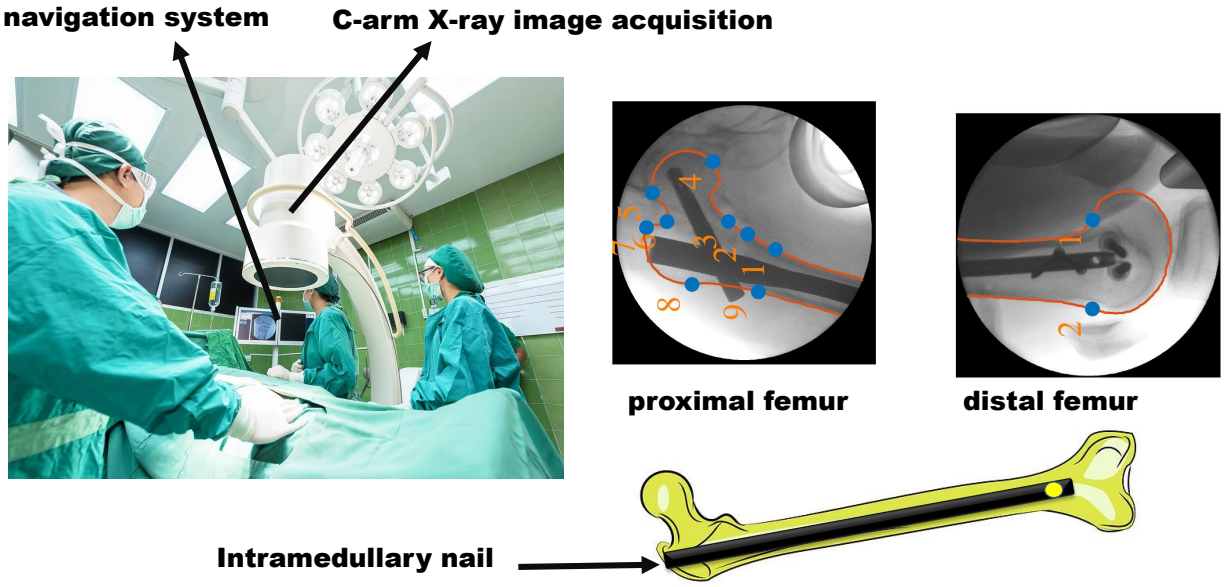


Figure 1.9: An illustration of CAS navigation system inside a surgery room. An automatic process, based on Image Processing and Machine Learning, is able to find the bone contours in the images. This information is then used to guide the surgery team during the procedure. (Medical image sources are detailed in the Appendix. The surgery room and bone drawing were taken from Pixabay and are free licence).

Artificial Vision

Control systems for robots need a good knowledge of the geometry around them to make motion decisions and compute successful grasps [17]. We illustrate an example in Fig. 1.10.

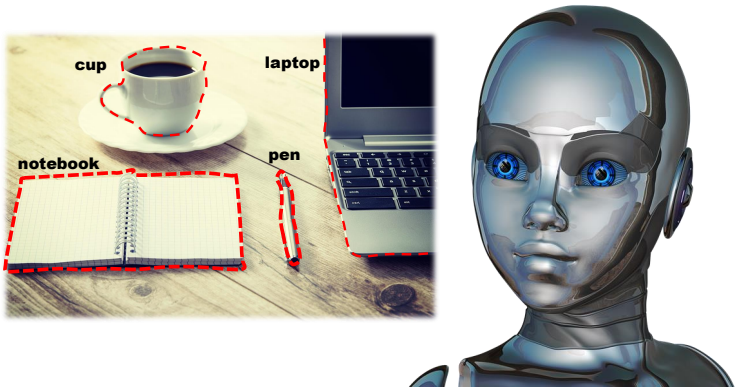


Figure 1.10: An illustration of artificial vision. Prior knowledge of the geometry and shape of the objects can enhance the segmentation performed by the control systems. (Images source: modified from originals in Pixabay with free licence)

Archaeology and Cultural Heritage

Shape analysis has a collection of possible applications in the field of Cultural Heritage. Actually, the mathematical principles of Point Distributions Models, i.e., Kendall's shape theory [9], were first derived motivated by their applications in Archaeology. The study of archaeological pieces, for instance, an anthropological study of the skull as illustrated in Fig. 1.11, is typically based on a collection of landmarks. Shape analysis has also been used to study the evolution of human tools in archaeology sites [18].

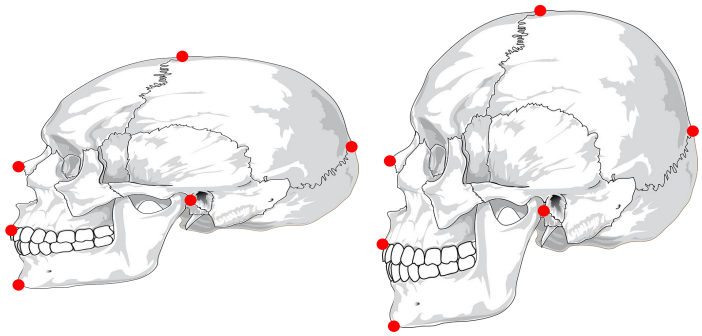


Figure 1.11: Landmarks (red dots) on two different skulls. The morphological changes of the position of these landmarks guide anthropological studies. (Images source: modified from originals in Pixabay with free licence).

1.3 Motivation to extend Statistical Shape Models

As we have already mentioned, this work focuses on landmark-based shape models, such as PDM. The reason to choose these models is that they are a good trade-off between simplicity and flexibility. A collection of landmarks is a very flexible definition for a contour, and can model many kinds of shapes. Also, PDM are linear models, and therefore provide efficiency, make the estimation procedures simple and have an intuitive interpretation [19]. However, these may lead to very simple models that are not robust enough or are not optimized to a specific use. Furthermore, heuristics are frequently present in many SSM-related applications. These Adhoc designs and decisions are also a loss of potential for the SSM.

In this section we overview the problems from SSM that have motivated and inspired the ideas developed in this thesis.

1.3.1 The need of robustness for fluoroscopic X-ray images

Medical imaging is a challenging discipline. The clinical analyses and medical intervention that depend on the processing of the images are critical, and the imprecision of the results can cause very serious consequences to the patients. This is the case, for instance in CAS. The accuracy of the segmentation techniques needs to be very high, as well as anatomically consistent. When the quality of the medical images is high, the classical SSM-based approaches have demonstrated good results [14] [20]. However, if the contrast of the images is not good, and some of the anatomical parts are occluded, SSM are not robust enough to deal with the difficulties. In Fig. 1.12 we show one fluoroscpic vs one diagnostic X-ray to show the consequences of a lower X-ray dose in the image quality.

In this work, one of the medical imaging modalities we have studied is fluoroscopic X-ray in a CAS context. This application demands computation in real time and high accuracy, whereas the contrast and the SNR are low due to the small X-ray dose. Furthermore, there are surgical tools and other artifacts that often occlude part of the anatomy. In such case, after testing the SSM-based Active Shape Models (ASM) by Cootes [14], the results are inaccurate. Therefore, in order to make a successful use of the SSM in more challenging medical image modalities, such as fluorosocpy, there is an important necessity to add more robustness to the shape models, as well as to the segmentation algorithms that make use of them.

We address this problematic in Chapter 3 of this thesis. In this chapter we propose a technique to enhance the segmentation permanence of SSM-based algorithms in fluoroscopic X-ray images.

1.3.2 Heuristic model selection

Selecting the model order of SSM, that is, the number of parameters that determine the model, is a compromise between a small training error and good generalization. We have observed that this order is not optimized in the design of SSM, and is often

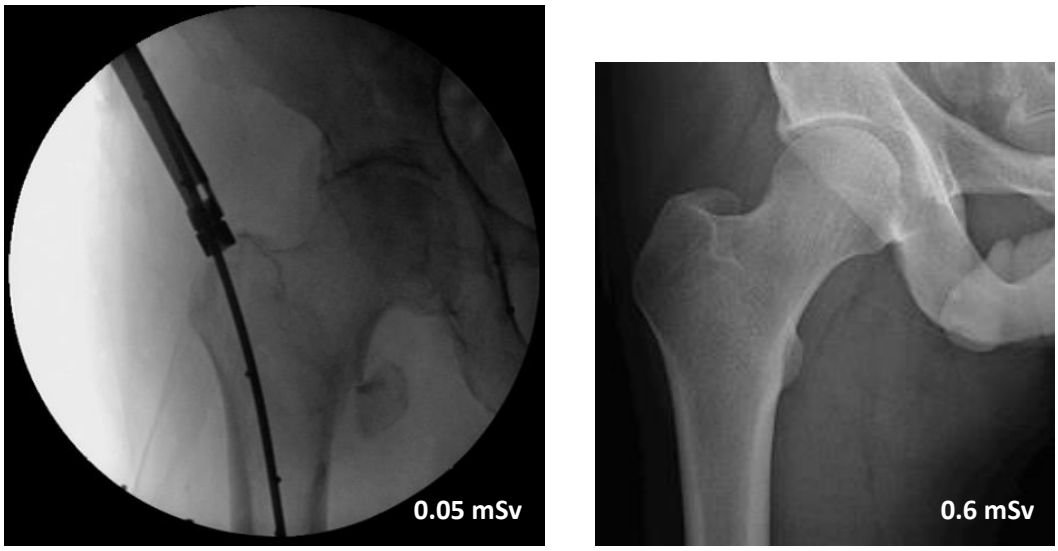


Figure 1.12: The fluoroscopic image (left) has approximately ten times less mSv (mili Sievert, a unit of ionizing radiation dose [21]) than a diagnostic X-ray (right). Therefore, the differences in contrast, resolution and SNR are visually evident. Also, an example of an occluding tool that difficulties segmentation processes is seen (left). The details about the medical image sources are in the Appendix.

chosen heuristically. For instance, in the PDM design, the number of shape eigenvector is typically selected to account for a fixed percentage of training variance [20], which is a "rule-of-thumb".

There are important issues to consider about the over- and underfitting in SSM. The goal of design of a SSM is to learn a statistical parametrization that generalizes well to the unseen data. However, if the order is too large, the model may be too specific to the training data, and therefore not generalize well (overfitting); if the order is too small, the model is too simple and the representation error is high (underfitting). Therefore, this trade-off should be observed. Furthermore, a good choice of the model depends on the number of training samples (typically small in medical applications) and the noise level in the training data set (often unknown and difficult to model).

We address the model selection problem in Chapters 4 and 5 of this thesis. First, in Chapter 4 we propose a novel model-order selection technique, based on information theory, that can be used in general source enumeration problems. Then, in Chapter 5 we show how our proposed technique enhances SSM.

1.3.3 Manual landmark registration

When the SSM are learned from training observations, we require a collection of points (the landmarks) sampled from the training contours. These points are very

often manually selected, and based only on experience or subjective criteria. For instance, a selection of anatomical points that are clinically relevant or relatively easy to identify. Furthermore, these manual labeling is very time consuming and error prone.

In order to calculate, for instance, the sample covariance matrix of a set of training vectors, the landmarks in these vectors need to be in correspondence for every training sample. This means that, the same landmarks are considered for every observation, and in the same order. Therefore, if some landmarks are missing or occluded in some observations, these cannot be included in the training. Also, the number of landmarks has to be sufficiently small, since training sets do not typically contain many observations, and the SSM may become ill-conditioned. In conclusion, the design of SSM is not possible if we only have the object boundaries as training observations instead of the complete collections of landmarks with correspondences. Interesting information may also be missed when the whole boundary is not considered in the training process.

We address these problems in Chapter 6 of this thesis. In this chapter we propose a registration technique that works directly with the object boundary when the correspondence is unknown, and does not need landmarks. We have considered Dynamic Time Warping to solve the correspondence problem.

1.4 Outlines and contributions of this thesis

This doctoral work provides techniques to enhance SSM. The reader has now a motivation of the engineering fields, mathematical models and applications in which shape analysis has an impact, as well as an overview of some of the problems to be addressed. The main application that motivated the use of SSM in this thesis is anatomy segmentation, for navigation systems in CAS. These images, based on fluoroscopy X-ray, are challenging, since the radiation dose is low, which generates low contrasts.

In this section we give an overview of the contributions of this thesis, describing each part of the document. This doctoral work is divided in four parts:

1.4.1 Part I: Introduction and background

The first part of the thesis is meant to provide the motivation and necessary background to the reader to fully understand the contributions and conclusions of the work.

Chapter 2: a background on statistical shape analysis

In this chapter we provide a thorough description of the concepts from Shape Theory that are necessary for a full understanding of the contributions of this thesis. Starting from Kendall's concept of shape [9], and arriving to the segmentation techniques that make use of SSM [14].

1.4.2 Part II: Contributions and developed work

This part contains the contributions of the thesis.

Chapter 3: robustness keeping simplicity in a generalized least-squares

As already motivated in the previous section, SSM-based segmentation techniques lack robustness. In this part we propose to enhance these techniques and deal with the challenging low contrast fluoroscopic images.

The main idea of our contribution is the following. Segmentation algorithms that use SSM often consider a target shape measured from the image. This target is used in a regression to find a valid shape that belongs to a SSM. The fit of a target shape is typically an ordinary least-squares (OLS) minimization. We propose to consider a generalized least-squares (GLS) fit instead to add robustness. In order to find the weighting matrix of the GLS, we calculate an estimate of the covariance matrix of the residuals from training data.

This chapter has produced the following publication:

- A. Eguizabal and P. Schreier, "A weighting strategy for Active Shape Models," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, Beijing, China, Sept. 2017, pp. 1-6.

Chapter 4 and 5: model selection and seeking optimality with information theory

We mentioned in previous section that the model selection of SSM is typically undervalued, while its impact is important in the qualities of the SSM. This part of the contribution deals with the problem of model selection.

A PDM requires choosing a model order q, which determines how much of the variation seen in the training data is accounted for by the PDM. A good choice of the model order is key for a good trade-off between over- and underfitting. Yet the most common approach for choosing it simply keeps a predetermined percentage of the total shape variation. Our contribution consist on choosing an information-theoretic approach that considers the ill-posed problems. The model order is chosen as

$$\text{Model order} = \text{a likelihood term} + \text{a penalty term},$$

where the penalty term depends on a selected information criterion, e.g. Akaike [22]. This approach was proved to be extended to more general array signal processing scenarios.

These chapters have produced the following publications:

- A. Eguizabal, C. Lameiro, D. Ramirez and P. Schreier, "Source enumeration in the presence of colored noise," *IEEE Signal Processing Letters*, vol. 26, no. 3, 2019.

- A. Eguizabal, P. Schreier and D. Ramirez, "Model-order selection in Statistical Shape Models," in *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Aalborg, Denmark, Sept. 2018, pp. 475-479.

Chapter 6: registration and correspondence with Dynamic Time Warping

The landmarks in the training samples for SSM are often manually selected, and need to be in correspondence. The correspondence problematic is also an issue to perform a shape fit, that is, the least-squared problem mentioned before needed in a SSM-based segmentation. Therefore, it is not possible to use the boundaries of the shapes directly neither to design the SSM nor to fit the shape model.

In this chapter we propose a strategy that uses the boundaries directly, instead of a collection of landmarks in correspondence. Our technique is based on Dynamic Time Warping (DTW), a times-series algorithm that uses dynamic programming.

This chapter has produced the following publication:

- A. Eguizabal, P. Schreier, and J. Schmidt, "Procrustes registration of two-dimensional statistical shape models without correspondences," submitted to *IEEE Transactions on Image Processing* 2019, available on ArXiv.

1.4.3 Part III: Conclusions and future lines of research

We finally close the thesis providing the concluding remarks. In this final section we also present the contributions under developments, as well as the future lines of research.

There are many interesting open ideas to explore the potentials of shape analysis. An important step is to continue the line of Chapter 6 with, for example, an extension to three-dimensional surfaces of our proposed registration strategy. In this part we present the preliminary results of our work under development, as well as an introduction to our future lines of research.

1.4.4 Appendix: The data

It is a difficult task to collect medical images for research purposes. During the development of this work, we had to manually label our own database, and had to design a program to perform these tasks consistently and accurately. In the appendix we present the details about our own medical image database. Also, we briefly described other publicly available datasets used in our tests.

Chapter 2
Statistical shape analysis

"Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less."

— Maria Sklodowska-Curie

In order to fully understand this thesis, we need to establish the mathematical fundamentals of the statistical models of shape. Landmark-based models, which are the shape models considered in this work, are based on Kendall's definition of shape [9]. Kendall formulated his shape theory to study the stochastic processes on shapes, and motivated by archaeological studies: the archaeologists usually annotated landmarks on their notebooks. Kendall was willing to present a definition of shape defined by landmarks and forming a polygon. In order to provide a mathematical framework to the statistical shape analysis, these polygons need to be independent of any translation, scale or rotation, that is, those transformation that do not alter the shape. Therefore, a distance that is invariable to these transformations needs to be defined. Section 2.2 provides an introduction to shape theory, to these invariant distances, and to the concept of Procrustes analysis. Section 2.3 focuses on the particular case of shape in two dimensions (i.e., in an image), where certain mathematical advantages can be considered with respect to the general formulation. Having reached that theoretical introduction, we focus particularly in the Point Distribution Models. In Section 2.4 we describe how these shape models are used by the segmentation algorithms, and how they are fit into the images. Finally, in Section 2.5 we introduce the existing solutions to deal with incomplete training samples.

2.1 Landmark-based shape

When we wish to perform statistical analysis of shape we assume that the shape observations belong to real world forms. These may come from a sketch of an archaeologist, an automatic segmentation of a Computer Vision algorithm, or a manually drawn anatomical piece. As a general basis, we understand that a planar shape observation contains the vertices of an outline representing the contour of an object of interest. This object may belong to nature, and thus its contour does not follow

a simple mathematical behaviour, i.e., it cannot be simplified by basic geometrical shapes. These are actually the kind of shapes that generate interest to be statistically analysed.

When we collect shape observations to perform a statistical analysis, we assume that they all contain the same number of N points, and that all points are in one-to-one correspondence for each point $n = 1, \ldots, N$. These points (or vertices) are also known as landmarks, since its position is generally determined by some meaningful information. For instance, in the archaeological study of spears, an evident landmark is the tip. With respect to an anatomical study of a hand outline, also the tips of the fingers are intuitive landmarks. Depending of the nature of the landmark, we classify them into:

- Manual: a human observer decides where the landmark needs to be. These landmarks must be on a point of the curve that is easy to inspect and reproduce in new observations, providing consistency to the identification.
- Automatic (or semi-automatic): typically equidistantly placed along the curve between two manual landmarks.

We show a few examples of landmarks-based shapes in Fig. 2.1.

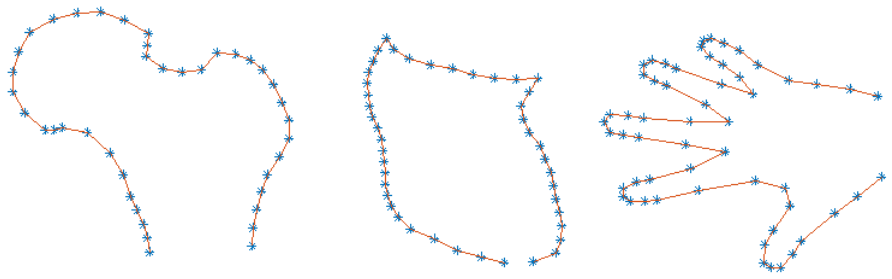


Figure 2.1: Landmark-based shape examples of (from left to right) a femur, a lung and a hand.

2.2 Shape theory

Before defining what the analysis of statistical shape is, we should clarify the terminology of shape. Our starting point is the definition that Kendall published on his shape theory in 1977 [23]:

> *"Shape is all the geometrical information that remains when location, scale, and rotational effects are quotiented (or filtered) out from an object"*

In this dissertation the shapes are considered to be a collection of vertices of a polygon, i.e., landmark-based. Some examples could be, a set of pixels forming a curve in a image, a set of landmarks representing significant points of a curve, or a cloud of points belonging to a volume.

2.2.1 A shape manifold

A manifold is a topological space that is locally Euclidean. Shape theory defines the shape manifold [23]. This manifold attempts to be a framework to shape modeling. A point in a shape manifold represents any collection of N points in a D-dimensional Euclidean space forming the same shape, i.e., a set of N vertices of dimension $D = 2$ (planar shape) or $D = 3$ (volumetric shape). As aforementioned, shape is invariant to scale, rotation and rigid translation. Consequently, the same shape formed by D-dimensional N vertices is mapped to the same point in the manifold, independently of its original position, rotation or size. A distance between two shapes can then be determined as a geodesic between two points of the shape manifold. In Fig. 2.2 we illustrate a shape manifold. Shape theory concludes that, a shape manifold can be

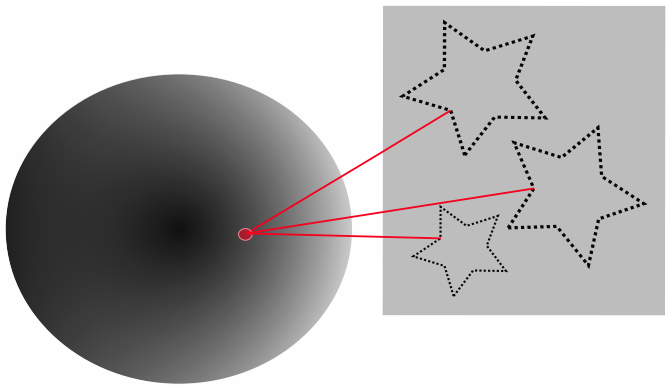


Figure 2.2: Representation of the shape manifold

defined if $D = 2$ but not for $D = 3$ [23]. When $D = 3$ the determined space is not topologically a manifold. Therefore, more generally, shape theory considers the shape space, which fulfils manifold properties only if $D = 2$.

2.2.2 Shape invariant transformations

We assume that the shapes exist naturally in an Euclidean space of dimension D (for planar shapes $D = 2$, for volumetric shapes $D = 3$). Let $\mathbf{x}_n \in \mathbb{R}^{D\times 1}$ be a column vector with the coordinates of the n-th vertex of a D dimensional polygon. Thus, N vertices are contained in a matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$, such that $\mathbf{X} \in \mathbb{R}^{D\times N}$. We also assume that a shape always contains more vertices than dimensions, i.e., $N > D$.

Considering the shapes belong originally to $\mathbb{R}^{D\times N}$, we start by identifying the transformations in this manifold that do not alter the shape of $\mathbf{X}$. These transformations are translation (move a shape in the space), scaling (shrink or dilate), and rotation. The shear transformation deforms the angles between the vertices of $\mathbf{X}$ and thus is not shape invariant. Furthermore, the reflection is not considered shape invariant either. Thus, a mirrored version of a shape is a new shape. For instance, a planar shape of a hand after a reflection changes from a left hand into a right hand

or vice-versa, and consequently, the left and the right hand are considered different shapes.

The following transformations to $\mathbf{X}$ do not alter its shape [24], as illustrated in Fig. 2.3,

$$\text{Translation: } \mathbf{t} \in \mathbb{R}^{D\times 1}, \qquad \mathbf{X} + \mathbf{t}\mathbf{1}_N = [\mathbf{x}_1 + \mathbf{t}, \ldots, \mathbf{x}_N + \mathbf{t}] \tag{2.1}$$

$$\text{Scale: } s \in \mathbb{R}^{+}, \qquad s\mathbf{X} = [s\mathbf{x}_1, \ldots, s\mathbf{x}_N] \tag{2.2}$$

$$\text{Rotation: } \mathbf{R} \in SO(D), \qquad \mathbf{R}\mathbf{X} = [\mathbf{R}\mathbf{x}_1, \ldots, \mathbf{R}\mathbf{x}_N] \tag{2.3}$$

where $\mathbf{1}_N = [1, 1, \ldots, 1] \in \mathbb{R}^{1\times N}$ and $SO(D)$ refers to the special orthogonal group of $\mathbb{R}^D$, i.e., $\mathbf{R}$ is an orthogonal matrix with positive determinant.

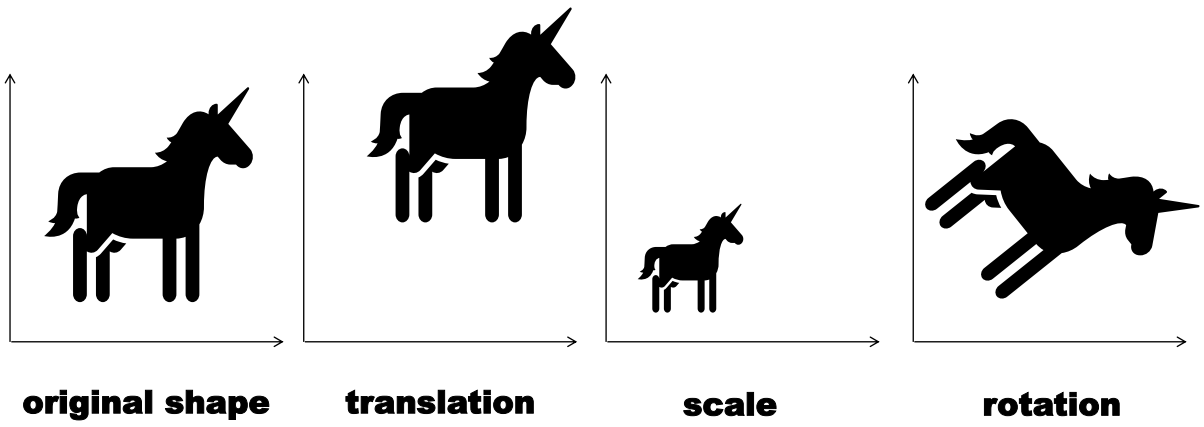


Figure 2.3: An illustrative example of the shape invariant transformations. The translation moves the original sample along the plane. The scale changes its size. The rotation transform the angles without altering the relative angles between vertex. These transformation, as well as a linear combination of them, do not change the shape from the original sample.

2.2.3 A distance between shape observations

Considering the original space $\mathbb{R}^{D\times N}$ as the initial manifold, we define the distance between two shapes in this space as the sum of squared Euclidean distances between

corresponding vertices. Let us define $\mathbf{X}^{(1)}, \mathbf{X}^{(2)} \in \mathbb{R}^{D\times N}$, this distance is

$$d(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}) = \sqrt{\sum_{n=1}^{N} ||\mathbf{x}_n^{(1)} - \mathbf{x}_n^{(2)}||^2} = \sqrt{\mathrm{Tr}\Big((\mathbf{X}^{(1)} - \mathbf{X}^{(2)})(\mathbf{X}^{(1)} - \mathbf{X}^{(2)})^T\Big)}. \quad (2.4)$$

This distance is, however, not invariant to the previously defined transformations. Thus is not valid to measure differences between shape. This is straightforward to demonstrate. Let us assume two matrices with the same shape, $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)} = s\mathbf{R}\mathbf{X}^{(1)} + \mathbf{t}\mathbf{1}_N$. The shape distance $d(\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$ should be 0. From observing (2.4), we see that, when the distance is 0, $||\mathbf{x}_n^{(1)} - \mathbf{x}_n^{(2)}||^2$ must be 0 for all $n = 1, \ldots, N$. We can easily show that the difference between the first elements $x_{11}^{(1)} - x_{11}^{(2)}$ is generally different from 0 and consequently so is $||\mathbf{x}_n^{(1)} - \mathbf{x}_n^{(2)}||^2$. Considering that the first row of the rotation matrix $\mathbf{R}$ is $\mathbf{x}_1 = [r_{11},\ r_{12}]^T$, and the first element of $\mathbf{t}$ is t_1, then $x_{11}^{(2)} = s(r_{11}x_{11}^{(1)} + r_{12}x_{21}^{(1)}) + t_1$, and the difference

$$x_{11}^{(1)} - x_{11}^{(2)} = x_{11}^{(1)}(1 - sr_{11}) - sr_{12}x_{21}^{(1)} - t_1, \quad (2.5)$$

which is different from 0 unless $s = 1$ and $\mathbf{R} = \mathbf{I}$. In order to derive a distance in the shape space that is invariant to the aforementioned transformations, we introduce the concepts of centroid and size.

2.2.4 Centroid and size

Let us define the centroid of $\mathbf{X}$ as

$$C(\mathbf{X}) = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n. \quad (2.6)$$

This centroid operator is linear, and therefore, when $\mathbf{X}$ is rotated, translated or scaled, so is its centroid under the same transformation, that is

$$C(s\mathbf{X}) = \frac{1}{N}\sum_{n=1}^{N} s\mathbf{x}_n = sC(\mathbf{X}), \quad (2.7)$$

$$C(\mathbf{X} + \mathbf{t}\mathbf{1}_N) = \frac{1}{N}\sum_{n=1}^{N} (\mathbf{x}_n + \mathbf{t}) = C(\mathbf{X}) + \mathbf{t}, \quad (2.8)$$

$$C(\mathbf{R}\mathbf{X}) = \frac{1}{N}\sum_{n=1}^{N} \mathbf{R}\mathbf{x}_n = \mathbf{R}C(\mathbf{X}). \quad (2.9)$$

Let us also find a definition to the size of $\mathbf{X}$. Following [9], we determine the size as the distance of each vertex to the shape centroid, that is

$$S(\mathbf{X}) = \sqrt{\sum_{n=1}^{N} ||\mathbf{x}_n - C(\mathbf{X})||^2} = \sqrt{\mathrm{Tr}\{(\mathbf{X} - C(\mathbf{X})\mathbf{1}_N)(\mathbf{X} - C(\mathbf{X})\mathbf{1}_N)^T\}}. \quad (2.10)$$

This size is invariant under translation [24],

$$S(\mathbf{X}+\mathbf{t}\mathbf{1}_N)=\sqrt{\sum_{n=1}^{N}||\mathbf{x}_n+\mathbf{t}-(C(\mathbf{X})+\mathbf{t})||^2}=\sqrt{\sum_{n=1}^{N}||\mathbf{x}_n-C(\mathbf{X})||^2}=S(\mathbf{X}), \tag{2.11}$$

invariant under rotation,

$$S(\mathbf{R}\mathbf{X})=\sqrt{\sum_{n=1}^{N}||\mathbf{R}\mathbf{x}_n-\mathbf{R}C(\mathbf{X})||^2}=$$

$$\sqrt{\sum_{n=1}^{N}||\mathbf{R}(\mathbf{x}_n-C(\mathbf{X}))||^2}=\sqrt{\mathrm{Tr}\{(\mathbf{R}(\mathbf{x}_n-C(\mathbf{X})))(\mathbf{R}(\mathbf{x}_n-C(\mathbf{X})))^T\}}=$$

$$\sqrt{(\mathrm{Tr}\{(\mathbf{X}-C(\mathbf{X})\mathbf{1}_N)^T\mathbf{R}^T\mathbf{R}(\mathbf{X}-C(\mathbf{X})\mathbf{1}_N)\}}=S(\mathbf{X}), \tag{2.12}$$

and homogeneous to the scaling

$$S(s\mathbf{X})=\sqrt{\sum_{n=1}^{N}||s\mathbf{x}_n-sC(\mathbf{X})||^2}=sS(\mathbf{X}). \tag{2.13}$$

2.2.5 A distance invariant to translation, scale and rotation

We define a function that transforms a shape matrix $\mathbf{X}$ so that it is invariant to scale and translation as

$$T(\mathbf{X})=\frac{1}{S(\mathbf{X})}(\mathbf{X}-C(\mathbf{X})\mathbf{1}_N). \tag{2.14}$$

After this transformation, translation and scale are removed. Let us define the matrices $\mathbf{X}$ and its scale and translated transformation, that is, $s\mathbf{X}+\mathbf{t}\mathbf{1}_N$. The distance between $T(\mathbf{X})$ and $T(s\mathbf{X}+\mathbf{t}\mathbf{1}_N)$ is then invariant to translation and scale, that is

$$d(T(\mathbf{X}),T(s\mathbf{X}+\mathbf{t}\mathbf{1}_N))=\sqrt{\sum_{n=1}^{N}||\frac{\mathbf{x}_n-C(\mathbf{X})}{S(\mathbf{X})}-\frac{s\mathbf{x}_n+\mathbf{t}-sC(\mathbf{X})-\mathbf{t}}{sS(\mathbf{X})}||^2}=0. \tag{2.15}$$

However this distance is still not invariant to rotation. There is not a specific rotation reference to perform an equivalent action as in (2.14), where the centroid and size are references for translation and scale. Therefore, we determine the wished invariante Sdistance between two different shape matrices $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ as

$$d_s(\mathbf{X}^{(1)},\mathbf{X}^{(2)})=\min_{\mathbf{R}\in SO(D)}d(T(\mathbf{X}^{(1)}),\mathbf{R}T(\mathbf{X}^{(2)})), \tag{2.16}$$

where the rotation $\mathbf{R}$ is

$$\mathbf{R}^*=\arg\min_{\mathbf{R}\in SO(D)}d(T(\mathbf{X}),\mathbf{R}T(\mathbf{X})). \tag{2.17}$$

2.2.6 Procrustes analysis

According to Greek Mythology, Procrustes (or Damases) [25] was the son of Poseidon, and a monster that lived in the vicinity of Athens. He used to force every visitor of his territory through a trial that always ended with his death: after making them lie on a bed that never had the same proportions, he prolonged with hammer blows the limbs that did not reach the ends. He gives his name to Procrustes Analysis, where a rotation forces two matrices of vertices to fit each other.

Let us consider $\mathbf{X}_1 = T(\mathbf{X}^{(1)})$, $\mathbf{X}_2 = T(\mathbf{X}^{(2)})$, and the optimal rotation defined as

$$\begin{aligned}\mathbf{R}^* &= \underset{\mathbf{R}\in SO(D)}{\arg\min}\, d(\mathbf{X}_1, \mathbf{R}\mathbf{X}_2) = \underset{\mathbf{R}\in SO(D)}{\arg\min}\, \mathrm{Tr}\{(\mathbf{X}_1 - \mathbf{R}\mathbf{X}_2)(\mathbf{X}_1 - \mathbf{R}\mathbf{X}_2)^T\} \\ &= \underset{\mathbf{R}\in SO(D)}{\arg\min}\, \left[\mathrm{Tr}\{\mathbf{X}_1\mathbf{X}_1^T\} - 2\mathrm{Tr}\{\mathbf{R}\mathbf{X}_2\mathbf{X}_1^T\} + \mathrm{Tr}\{\mathbf{X}_2\mathbf{X}_2^T\}\right] = \underset{\mathbf{R}\in SO(D)}{\arg\max}\, \mathrm{Tr}\{\mathbf{R}\mathbf{X}_2\mathbf{X}_1^T\}.\end{aligned} \tag{2.18}$$

This problem is a special case of the Procrustes problem in [26]. In [27], Schönemann proposed a solution to the Orthogonal Procrustes problem defined as

$$\mathbf{R}^* = \underset{\mathbf{R}\mathbf{R}^T=\mathbf{I}_D}{\arg\max}\, \mathrm{Tr}\{\mathbf{R}\mathbf{X}_2\mathbf{X}_1^T\}, \tag{2.19}$$

where $\mathbf{R}$ is constrained to be orthogonal, i.e., $\mathbf{R} \in O(D)$, but not necessarily a rotation matrix. This means that $|\mathbf{R}|$ could be 1 or -1. The solution is the following. Let us consider the Singular Value Decomposition (SVD) of $\mathbf{X}_2\mathbf{X}_1^T = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$, where $\mathbf{U}$ and $\mathbf{V}$ are orthonormal matrices and $\mathbf{\Lambda} = \mathrm{diag}\{\lambda_1, \ldots, \lambda_D\}$ is a diagonal matrix with non-negative elements, $\lambda_i \geq 0 \quad \forall i = 1, \ldots, D$. Therefore, the trace to maximize in (2.19) can be expressed as

$$\mathrm{Tr}\{\mathbf{R}\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T\} = \mathrm{Tr}\{\mathbf{V}^T\mathbf{R}\mathbf{U}\mathbf{\Lambda}\} = \mathrm{Tr}\{\mathbf{S}\mathbf{\Lambda}\} = \sum_{i=1}^{D} \lambda_i s_{ii}, \tag{2.20}$$

where s_{ii} is the i-th element of the diagonal of the matrix $\mathbf{S} = \mathbf{V}^T\mathbf{R}\mathbf{U}$. The matrix $\mathbf{S}$ results from the product of orthonormal matrices and thus it is orthonormal, that is $|s_{ii}| \leq 1$. Therefore, the solution that maximizes the sum in (2.20) is such that $s_{ii} = 1 \quad \forall i = 1, \ldots, D$ [27]. This means that the orthogonal matrix $\mathbf{S}$ is the identity matrix, $\mathbf{S} = \mathbf{I}_D$, and thus $\mathbf{V}^T\mathbf{R}\mathbf{U} = \mathbf{I}_D$. The orthogonal matrix $\mathbf{R}^*$ is then

$$\mathbf{R}^* = \mathbf{V}\mathbf{U}^T. \tag{2.21}$$

However, this solution does not provide necessarily a rotation matrix, since the determinant $|\mathbf{R}^*|$ could still be negative and thus perform reflections, which are not considered shape-invariant transformations. Thus, a modified version with an additional constraint to the determinant, $|\mathbf{R}| = 1$, i.e., $\mathbf{R} \in SO(D)$, is proposed:

$$\max_{|\mathbf{R}|=1} \sum_{i=1}^{D} \lambda_i s_{ii}. \tag{2.22}$$

The expression in (2.21) is a rotation matrix when $|\mathbf{U}||\mathbf{V}| = 1$. In that case, (2.21) is also the solution to (2.22). However, when $|\mathbf{U}||\mathbf{V}| = -1$, (2.21) is not a rotation matrix and hence not a valid solution to (2.22). Since $|\mathbf{R}^*| = 1$ is required, in that case $|\mathbf{S}| = -1$ and thus the solution to $\mathbf{S}$ cannot be the identity matrix. Still, $\mathbf{S}$ would be an orthogonal matrix with negative determinant. As described in [24], a vector $\mathbf{s} = [s_{11}, s_{22}, \ldots, s_{DD}]^T$ is the diagonal of such a matrix if an only if it lies in the convex hull of

$$\boldsymbol{\epsilon} = \{[\pm 1, \ldots, \pm 1] \quad | \quad \text{odd number of minus signs}\}. \tag{2.23}$$

This result can be used to find the optimal solution of (2.22) as follows. First, we notice that the singular values in (2.22) are arranged in decreasing order, i.e., $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_D$. Second, as the objective function of (2.22) is linear, its optimal solution lies on a extreme point of the feasible set, which is the convex hull of $\boldsymbol{\epsilon}$. The extreme points of $\boldsymbol{\epsilon}$ have an odd number of negative signs. More 1s instead of -1s in $\mathbf{S}$ make the sum to maximize in (2.22) bigger. Consequently, the optimal extreme point is only one negative element, corresponding to the minimum singular value, i.e., s_{DD}. Therefore, the new solution is

$$\begin{aligned} \mathbf{S} &= \text{diag}\{1, 1, \ldots, |\mathbf{U}||\mathbf{V}|\}, \\ \mathbf{R}^* &= \mathbf{U}\mathbf{S}\mathbf{V}^T. \end{aligned} \tag{2.24}$$

A more detailed derivation to this solution can be found in [28].

2.2.7 Pre-shape and shape spaces

The action of the function $T(\mathbf{X})$ defined in (2.14) can be seen as a transformation of the shape matrix $\mathbf{X}$ into a new set $\mathbb{T}$, in which the size is fixed to 1, and the centroid is on the origin of $\mathbb{R}^D$. Therefore, the scale and translation are constant for every matrix $\mathbf{X}$ in the set. This scale/translation-normalized set is defined as

$$\mathbb{T} = \left\{\mathbf{X}_T \in \mathbb{R}^{d \times N} : \mathbf{X}_T = \frac{1}{S(\mathbf{X})}(\mathbf{X} - C(\mathbf{X})\mathbf{1}_N)\right\}. \tag{2.25}$$

The set $\mathbb{T}$ is equivalent to the pre-shape sphere $\mathbb{S}_D^N$ defined in shape theory [23]. The term pre-shape means that the space is one step ahead from the shape space, since the rotations are not filtered yet. The pre-shape space is therefore not invariant to rotation. Also, in $\mathbb{T}$ the N vertices can be expressed with respect to their centroid. This means that, a projection of the vertices $\mathbf{Z} = \mathbf{Q}\mathbf{X}$, where $\mathbf{Q}$ is an orthogonal matrix, can be determined such that one of the new projected elements in $\mathbf{Z}$ is the scaled centroid [23]. Since the centroid of the elements in $\mathbb{T}$ is known, a degree of freedom can be neglected from $\mathbf{Z}$, remaining $N-1$ new vertices in $\tilde{\mathbf{Z}} \in \mathbb{R}^{d \times (N-1)}$. Thus, the pre-shape is considered to be the unit hypersphere

$$\mathbb{S}^{D(N-1)-1} = \mathbb{T}/\{0\} = \left\{\tilde{\mathbf{Z}} \in \mathbb{R}^{D \times (N-1)} : C(\tilde{\mathbf{Z}}) = 1\right\}. \tag{2.26}$$

The shape space is defined after the $SO(D)$ group is quotient out of the pre-shape space, that is

$$\Sigma_D^N = \mathbb{S}^{D(N-1)-1}/SO(D). \tag{2.27}$$

The shape space captures what is invariant to scale, translation and rotation. In the shape space, a geodesic distance between two pre-shape $\mathbf{X}_1, \mathbf{X}_2 \in \mathbb{T}$ is defined as [24]

$$g(\mathbf{X}_1, \mathbf{X}_2) = \arccos \max_{\mathbf{R}\in SO(D)} \mathrm{Tr}(\mathbf{R}\mathbf{X}_1\mathbf{X}_2^T), \tag{2.28}$$

where $\mathbf{R}$ is the solution to a Procrustes analysis proposed in section 2.2.6. This geodesic distance $g(\mathbf{X}_1, \mathbf{X}_2)$ is directly related to the shape invariant distance defined in Eq. (2.16) as

$$\begin{aligned} d_s(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}) = &\min_{\mathbf{R}\in SO(D)} d(T(\mathbf{X}^{(1)}), \mathbf{R}T(\mathbf{X}^{(1)})) = \\ &\min_{\mathbf{R}\in SO(D)} \sqrt{\mathrm{Tr}\{\mathbf{X}_1\mathbf{X}_1^T - 2\mathbf{R}\mathbf{X}_1\mathbf{X}_2^T + \mathbf{X}_2\mathbf{X}_2^T\}} = \\ &\max_{\mathbf{R}\in SO(D)} \sqrt{2(\mathrm{Tr}\{\mathbf{R}\mathbf{X}_1\mathbf{X}_2^T\} - 1)} = \\ &\sqrt{2(\cos(g(\mathbf{X}_1, \mathbf{X}_2)) - 1)} \end{aligned} \tag{2.29}$$

The shape space Σ_D^N only fulfils the properties of a Riemannian manifold when $D < 3$ [23]. For higher dimensions the rotations $\mathbf{R} \in SO(D)$ are not free actions, which means that a rotation different from $\mathbf{R} = \mathbf{I}_D$ may preserve the pre-shape, i.e., $\mathbf{R}\mathbf{X}_1 = \mathbf{X}_1$ for some $\mathbf{R} \neq \mathbf{I}_D$. Therefore, we can only talk about a shape manifold when we analyze planar shapes, that is $D = 2$ (we assume $D = 1$ is a trivial an non-interesting case). Σ_2^N is a Riemanninan manifold, and has an inner product defined in the tangent space, allowing to measure the angle at each point and define the lengths of the curves [29]. When $D > 2$ we deal with a shape space that is not a Riemannian manifold, where we can still define a generalized geodesic [30].

2.2.8 Statistical Shape Models

After shape matrices are centered and size-normalized, they live in a shape manifold [23]. The shape space Σ_D^N is a manifold that accounts for all shape structures $\mathbf{X}$ with N vertices in $\mathbb{R}^D$. In order to model shape variability, the distance metric to consider is the geodesic over this manifold.

We initially consider a statistical model of shape as a random process, whose realizations are shape manifold points in a neighborhood around a reference in this manifold. Let us consider the reference shape matrix $\mathbf{X}_{\text{ref}} \in \mathbb{R}^{D\times N}$, and its representation in the shape space Σ_D^N as $\tilde{\mathbf{X}}_{\text{ref}} = \frac{1}{S(\mathbf{X}_{\text{ref}})}\left(\mathbf{X}_{\text{ref}} - C(\mathbf{X}_{\text{ref}})\mathbf{1}_N\right)$. In order to generate shape structures that are deformations of the shape reference, the following function is considered:

$$\tilde{\mathbf{X}} \sim \Phi_{\tilde{\mathbf{X}}_{\text{ref}}}(\boldsymbol{\Theta}), \tag{2.30}$$

where $\tilde{\mathbf{X}} \in \Sigma_D^N$ is a shape structure that distributes as the random process of deformation $\Phi_{\tilde{\mathbf{X}}_{\text{ref}}}(\boldsymbol{\Theta})$, which depends on the reference $\tilde{\mathbf{X}}_{\text{ref}} \in \Sigma_D^N$ and a set of parameters $\boldsymbol{\Theta}$. An ensemble of realizations of $\tilde{\mathbf{X}}$, that is $\{\tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2, \ldots, \tilde{\mathbf{X}}_M\}$, have the same shape property [9].

The random process $\Phi_{\tilde{\mathbf{X}}_{\text{ref}}}$ and its statistical analysis over the shape manifold is non-linear, since the shape variability is determined by geodesic distances. In practice, this is going to be problematic for the performance and understanding of the models [31]. Gaussian Process Morphable Models, formulated in [32], consider a Gaussian Process (GP) for $\Phi_{\tilde{\mathbf{X}}_{\text{ref}}}$. Consequently, a linear model of deformation determines the realizations of $\tilde{\mathbf{X}}$ as

$$\tilde{\mathbf{X}} \sim \text{GP}(\tilde{\mathbf{X}}_{\text{ref}}, \boldsymbol{\Gamma}), \tag{2.31}$$

where $\boldsymbol{\Gamma} \in \mathbb{R}^{DN \times DN}$ is a kernel function that defines the covariance function of the process. The only requirement for $\boldsymbol{\Gamma}$ is that it is a symmetric and positive semi-definite kernel [32]. This kernel function is chosen according to the deformation properties that are expected from the object of interest. GP are used to model shape of anatomical objects [32]. In such cases the variability is assumed to be smooth, and highly correlated between neighbor vertices. Typically more than one reference is available as training data to learn the shape model. In than case, the mean and covariance kernel of the GP can be estimated from this training data. This is also the case for Point Distribution Models [14]. We explain this in detail in the following sections.

The tangent to the shape space

The GP linear model of statistical shape in Eq. (2.31) does not consider that the shape structure must belong to a shape manifold, i.e., the spherical topology of the shape space is disregarded [33]. This means that, the distances between observations are not geodesics of the shape space. Thus, different realizations may have not only different shape but also different scale or rotation from the reference.

An alternative to the shape space is to consider a local tangent space analysis at the reference. In [33] the authors consider a tangent space that preserves distances and angles with an exponential mapping of $\tilde{\mathbf{X}}$. Also, the geodesic distances can be directly considered in the statistical analysis [34]. The concept of the tangent space is also considered for Point Distribution Models in [35] and [31]. The training shape structures are projected onto the so defined tangent space to the reference structure, which is the mean shape in that case. These projections are simply the rescaled observations

$$\tilde{\mathbf{X}}^t = \frac{S(\tilde{\mathbf{X}}_{\text{ref}})}{\tilde{\mathbf{X}}_{\text{ref}}^T \tilde{\mathbf{X}}} \tilde{\mathbf{X}} \tag{2.32}$$

where $\tilde{\mathbf{X}}^t$ is the projection on the tangent space over $\tilde{\mathbf{X}}_{\text{ref}}$.

2.3 Statistical analysis of planar shapes

One of the most common cases in image processing are planar shapes, i.e., shapes with $D = 2$ in $\mathbb{R}^2$. In this case, the shape space Σ_2^N has the structure of a Riemanian manifold [23]: it is differentiable, and the tangent space at each point has an inner product. These properties provide an interesting mathematical framework for the statistical analysis of shape. Also, the rotation $\mathbf{R} \in SO(2)$ can be understood as a complex scalar multiplication. This allows to simplify the formulation of the manifold properties Σ_2^N and consider equivalently a projective complex space $\mathbb{C}P^{(N-2)}$.

Let us consider the complex space $\mathbb{C}$. The vertices (or points) of a shape belong to a complex vector, whose real and imaginary parts correspond to each of the coordinates of $\mathbb{R}^2$. The shape invariant transformations parameters (scale, rotation and translation) are defined by a complex affine transformation. Thus, we model a collection of shape points in a complex vector $\boldsymbol{\kappa} \in \mathbb{C}^{N\times 1}$. The N vertices in $\boldsymbol{\kappa} = [\kappa_1, \ldots, \kappa_N]^T$ are now complex scalars, i.e., $\kappa_n \in \mathbb{C} \quad \forall n = 1, \ldots, N$. A shape invariant transformation on a vector $\boldsymbol{\kappa}^{(1)}$ is defined as:

$$\boldsymbol{\kappa}^{(2)} = r\boldsymbol{\kappa}^{(1)} + t, \tag{2.33}$$

where $\{r, t\} \in \mathbb{C}$ are the parameters of the rigid transformation (r is the scale in magnitude and rotation in phase, and t the translation), and $\boldsymbol{\kappa}^{(2)} \in \mathbb{C}^{N\times 1}$ is the transformed vector. $\boldsymbol{\kappa}^{(1)}$ and $\boldsymbol{\kappa}^{(2)}$ represent the same shape.

Given that the contributions of this thesis are mainly about planar shapes,from this point of the chapter we will consider the complex notation to refer to a shape vector.

2.3.1 A distance in the complex space

We derive the distance defined in Eq. (2.4). Let us define the centroid of $\boldsymbol{\kappa}$ for two vectors $\boldsymbol{\kappa}^{(1)}, \boldsymbol{\kappa}^{(2)} \in \mathbb{C}^{N\times 1}$ as

$$d(\boldsymbol{\kappa}^{(1)}, \boldsymbol{\kappa}^{(2)}) = \sqrt{\sum_{n=1}^{N} |\kappa_n^{(1)} - \kappa_n^{(2)}|^2} = \sqrt{||\boldsymbol{\kappa}^{(1)} - \boldsymbol{\kappa}^{(2)}||^2} = ||\boldsymbol{\kappa}^{(1)} - \boldsymbol{\kappa}^{(2)}||, \tag{2.34}$$

which is the l_2-norm of the difference of the vectors. Since this is a particular case of Eq. (2.4), the same properties are met. The distance is not invariant to translation, scale and rotation. We comparably define the centroid and size of $\boldsymbol{\kappa}$, as well as a distance of shape that is invariant to translation, scale and rotation. The centroid of $\boldsymbol{\kappa}$ is the average of the N vertices in $\boldsymbol{\kappa}$

$$C(\boldsymbol{\kappa}) = \frac{1}{N} \sum_{n=1}^{N} \kappa_n, \tag{2.35}$$

where $C(\boldsymbol{\kappa}) \in \mathbb{C}$. The size of $\boldsymbol{\kappa}$ is defined as

$$S(\boldsymbol{\kappa}) = \sqrt{\sum_{n=1}^{N} |\kappa_n - C(\boldsymbol{\kappa})|^2} = ||\boldsymbol{\kappa} - C(\boldsymbol{\kappa})\mathbf{1}_N||, \tag{2.36}$$

where $S(\boldsymbol{\kappa}) \in \mathbb{R}^+$, and it is the l_2-norm of $\boldsymbol{\kappa} - C(\boldsymbol{\kappa})\mathbf{1}_N$. The transformation function that filters out translation and scale is thus

$$T(\boldsymbol{\kappa}) = \frac{\boldsymbol{\kappa} - C(\boldsymbol{\kappa})\mathbf{1}_N}{||\boldsymbol{\kappa} - C(\boldsymbol{\kappa})\mathbf{1}_N||} = \boldsymbol{\tau}. \tag{2.37}$$

The vector $\boldsymbol{\tau}$ has the geometrical properties of the pre-shape space [23]. We will refer to $\boldsymbol{\tau}$ as the pre-shape vector. We define the shape distance, d_s, that is also invariant to the action of rotations. A rotation matrix $\mathbf{R} \in SO(2)$ is now equivalent to a scalar complex $r \in \mathbb{C}$ with $|r| = 1$. Let us assume two pre-shapes $\boldsymbol{\tau}_1 = T(\boldsymbol{\kappa}^{(1)})$ and $\boldsymbol{\tau}_2 = T(\boldsymbol{\kappa}^{(2)})$, where the shape distance between $\boldsymbol{\kappa}^{(1)}$ and $\boldsymbol{\kappa}^{(2)}$ is

$$d_s(\boldsymbol{\kappa}^{(1)}, \boldsymbol{\kappa}^{(2)}) = \min_{|r|=1} d(\boldsymbol{\tau}_1, r\boldsymbol{\tau}_2) = \min_{|r|=1} ||\boldsymbol{\tau}_1 - r\boldsymbol{\tau}_2||, \tag{2.38}$$

and the rotation we need to compute is

$$r^* = \arg\min_{|r|=1} ||\boldsymbol{\tau}_1 - r\boldsymbol{\tau}_2||. \tag{2.39}$$

This distance d_s in this case is much simpler that the one presented in the previous section, since now there is no need to solve a Procrustes analysis to determine r. We can simply derive it, considering $r = e^{j\theta}$, as

$$\begin{aligned}
\theta^* &= \arg\min_{\theta} ||\boldsymbol{\tau}_1 - e^{j\theta}\boldsymbol{\tau}_2|| = (\boldsymbol{\tau}_1 - e^{j\theta}\boldsymbol{\tau}_2)^H(\boldsymbol{\tau}_1 - e^{j\theta}\boldsymbol{\tau}_2) \\
&= \arg\min_{\theta}(\boldsymbol{\tau}_1^H\boldsymbol{\tau}_1 - e^{-j\theta}\boldsymbol{\tau}_2^H\boldsymbol{\tau}_1 - \boldsymbol{\tau}_1^H\boldsymbol{\tau}_2 e^{j\theta} + e^{-j\theta}\boldsymbol{\tau}_2^H\boldsymbol{\tau}_2 e^{j\theta}) \\
&= \arg\min_{\theta}\left(2 - [(\boldsymbol{\tau}_1^H\boldsymbol{\tau}_2 e^{j\theta}) + (\boldsymbol{\tau}_1^H\boldsymbol{\tau}_2 e^{j\theta})^*]\right) \\
&= \arg\max_{\theta} \quad \mathrm{Re}[\boldsymbol{\tau}_1^H\boldsymbol{\tau}_2 e^{j\theta}]
\end{aligned} \tag{2.40}$$

The real part of a complex number is maximum when this number is real, and thus it is equivalent to the magnitude. Thus, the solution is $\theta^* = -\arg(\boldsymbol{\tau}_1^H\boldsymbol{\tau}_2)$. The shape distance results in

$$d_s(\boldsymbol{\kappa}^{(1)}, \boldsymbol{\kappa}^{(2)}) = ||\boldsymbol{\tau}_1 - e^{-j\arg(\boldsymbol{\tau}_1^H\boldsymbol{\tau}_2)}\boldsymbol{\tau}_2|| = 2\left(1 - |\boldsymbol{\tau}_1^H\boldsymbol{\tau}_2|\right). \tag{2.41}$$

The geodesic between two pre-shapes is then

$$g(\boldsymbol{\tau}_1, \boldsymbol{\tau}_2) = \arccos|\boldsymbol{\tau}_1^H\boldsymbol{\tau}_2|. \tag{2.42}$$

2.3.2 Procrustes registration of two planar shapes

In certain cases, we do not need to project the shape vectors into a shape manifold, but instead, we consider a reference shape vector, $\boldsymbol{\kappa}^{(1)}$. A Procrustes registration is a

linear transformation of the points on a shape vector, i.e., target vector $\boldsymbol{\kappa}^{(2)}$, to best conform the locations of the points of a reference vector shape $\boldsymbol{\kappa}^{(1)}$. We consider this transformation to be shape invariant, that is, to be produced by translation, scale and rotation.

Considering that planar shapes live in a complex space, we model the Procrustes registration (scale, rotation and translation) as a complex affine transformation. We rename the previously define distance between $\boldsymbol{\kappa}^{(1)}$ and $\boldsymbol{\kappa}^{(2)}$ as Procrustes distance d_p as

$$d_p(\boldsymbol{\kappa}^{(1)}, \boldsymbol{\kappa}^{(2)}) = d(\boldsymbol{\kappa}^{(1)}, \boldsymbol{\kappa}^{(2)}) = ||\boldsymbol{\kappa}^{(1)} - \boldsymbol{\kappa}^{(2)}||, \tag{2.43}$$

and formulate this registration as a minimization of the squared Procrustes distance. Let us assume we have the shape vectors $\boldsymbol{\kappa}^{(1)}$ and $\boldsymbol{\kappa}^{(2)}$, and consider the Procrustes registration of $\boldsymbol{\kappa}^{(2)}$ (target) over $\boldsymbol{\kappa}^{(1)}$ (reference), that is, the projection of $\boldsymbol{\kappa}^{(2)}$ into the pose of $\boldsymbol{\kappa}^{(1)}$. This is performed by the pose parameters $r, t \in \mathbb{C}$ that minimize the squared Procrustes distance $d_p(\boldsymbol{\kappa}^{(1)}, \boldsymbol{\kappa}^{(2)})^2$:

$$r, t = \arg\min_{r,t} d_p(\boldsymbol{\kappa}^{(1)}, \boldsymbol{\kappa}^{(2)})^2 = \arg\min_{r,t} ||\boldsymbol{\kappa}^{(1)} - (r\boldsymbol{\kappa}^{(2)} + \mathbf{1}t)||^2. \tag{2.44}$$

Let us express the transformation as $r\boldsymbol{\kappa}^{(2)} + \mathbf{1}t = [\boldsymbol{\kappa}^{(2)} \quad \mathbf{1}]\tilde{\mathbf{r}} = \mathbf{K}^{(2)}\tilde{\mathbf{r}}$, where $\tilde{\mathbf{r}} = [r \quad t]^T$ is a vector containing the pose parameters (scale, rotation and translation). This problem is actually very well-known and straightforward: the vector $\tilde{\mathbf{r}}$ that minimizes the squared Procrustes distance between $\boldsymbol{\kappa}^{(1)}$ and $r\boldsymbol{\kappa}^{(2)} + \mathbf{1}t$ is the solution to the linear least-squares fit

$$\tilde{\mathbf{r}}^* = \arg\min_{\tilde{\mathbf{r}}} ||\boldsymbol{\kappa}^{(1)} - \mathbf{K}^{(2)}\tilde{\mathbf{r}}||^2, \tag{2.45}$$

where the closed-form optimal solution of the pose vector is $\tilde{\mathbf{r}}^* = (\mathbf{K}^{(2)H}\mathbf{K}^{(2)})^{-1}\mathbf{K}^{(2)H}\boldsymbol{\kappa}^{(1)}$.

2.3.3 Group-wise Procrustes registration

When a statistical shape model is learned from a set of training observations of the shape structure, we need to measure variability that is independent to rotation, scale and translation. In order to learn this shape variability, we perform a group-wise registration of this set of training structures. Therefore, these training samples are registered to a reference $\boldsymbol{\kappa}_{\text{ref}}$, which is typically a mean [36]. This is the case, for instance, to design Point Distribution Models [14].

A group-wise registration consists on the following. For every mth shape vector $\boldsymbol{\kappa}_m$ in the training set we find the pose parameters

$$[r_m^*, t_m^*] = \arg\min_{r_m, t_m} \sum_{m=1}^{M} ||\boldsymbol{\kappa}_{\text{ref}} - (r_m\boldsymbol{\kappa}_m + \mathbf{1}t_m)||^2. \tag{2.46}$$

Let us assume a set of M shape vectors, that is $\boldsymbol{\kappa}_1, \ldots, \boldsymbol{\kappa}_M$. According to the proposed definition of shape [9], the shape variability is what remains after these shape vectors

have a common scale, translation and rotation with respect to a pre-shape vector of reference $\boldsymbol{\tau}_{\text{ref}}$, which is assumed to have unit scale and null translation. This process is a group-wise rigid registration or superimposition. The observed shape vectors are transformed, as described in Section 2.3.1, to filter out the translation and the scale to obtain $\boldsymbol{\tau}_m$

$$\boldsymbol{\tau}_m = T(\boldsymbol{\kappa}_m) = \frac{\boldsymbol{\kappa}_m - C(\boldsymbol{\kappa}_m)\mathbf{1}_N}{||\boldsymbol{\kappa}_m - C(\boldsymbol{\kappa}_m)\mathbf{1}_N||}, \tag{2.47}$$

for every observation $m = 1, \ldots, M$. In order to complete the superimposition of the M observation to the shape vector reference $\boldsymbol{\tau}_{\text{ref}}$, we choose rotations r_m, with $|r_m| = 1$ and with the following minimization, as explained in Section 2.3.1,

$$r_m^* = \underset{|r_m|=1}{\arg\min} ||\boldsymbol{\tau}_{\text{ref}} - r_m \boldsymbol{\tau}_m||, \tag{2.48}$$

where the reference vector $\boldsymbol{\tau}_{\text{ref}}$ is typically the mean shape [37].

The Procrustes mean shape

To perform a group-wise registration, we consider the Procrustes mean shape [37] as the reference pre-shape $\boldsymbol{\tau}_{\text{ref}}$. This definition is geometrically convenient in shape theory so that the statistical shape model is defined on a tangent to the hypersphere of the mean shape $\boldsymbol{\mu}$ [36]. This mean shape is determined as follows. Considering we have a set of M training shape vectors, and their preshapes are $\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_M$, the mean shape is defined as

$$\boldsymbol{\mu}^* = \underset{||\boldsymbol{\mu}||^2=1}{\arg\min} \sum_{m=1}^{M} d_s(\boldsymbol{\tau}_m, \boldsymbol{\mu}), \tag{2.49}$$

and assuming we know the rotations r_m to superimpose $\boldsymbol{\tau}_m$ on $\boldsymbol{\mu}$, as explained in Eq. (2.48), then the estimate of $\boldsymbol{\mu}$ is the closed-form solution of a constrained least-squares

$$\boldsymbol{\mu} = \frac{\sum_{m=1}^{M} r_m \boldsymbol{\tau}_m}{||\sum_{m=1}^{M} r_m \boldsymbol{\tau}_m||}. \tag{2.50}$$

However, the rotations r_m are not known and therefore the mean $\boldsymbol{\mu}$ and the rotations are mutually dependent.

The Generalized Procrustes Analysis

The rotations r_m and the mean $\boldsymbol{\mu}$ are mutually dependent. The Generalizes Procrustes Analysis (GPA) is an alternating optimization framework to obtain a solution to both. The GPA iterates between minimizing the distances of the pre-shapes to $\boldsymbol{\mu}$, i.e., calculating the rotations r_m and estimating $\boldsymbol{\mu}$. For each of the m shape vectors, the minimization

$$[r_m^*, \boldsymbol{\mu}^*] = \underset{\substack{|r_m|=1 \\ ||\boldsymbol{\mu}||^2=1}}{\arg\min} \sum_{l=1}^{M} ||r_m \boldsymbol{\tau}_m - \boldsymbol{\mu}||^2 \tag{2.51}$$

is solved. The GPA provides a solution to Eq. (2.51) as described in Algorithm 2.1. We show an example of the results of this analysis over a collection of 50 femur shape vectors in Fig. 2.4.

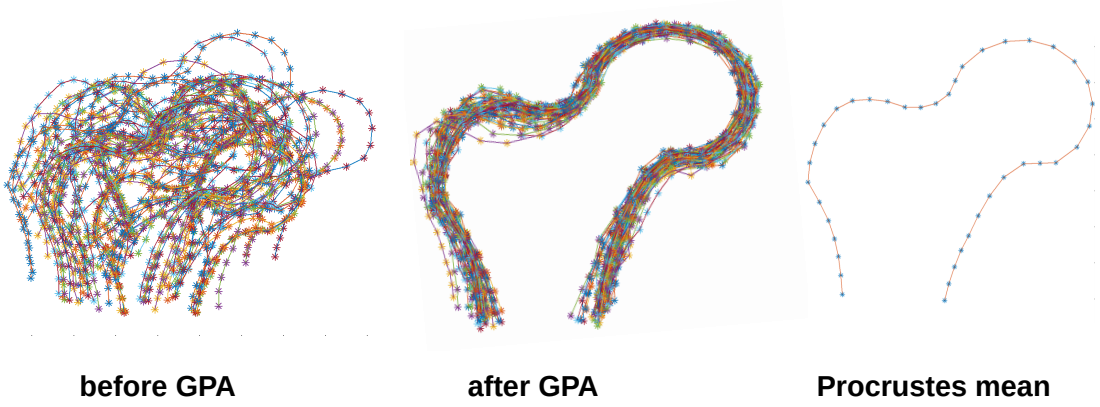


Figure 2.4: An example of a collection of 50 femur shapes before (left) and after (center) a Generalized Procrustes Analysis, as well as the Procrustes mean (right). More information about the source of the data can be found in the Appendix.

Algorithm 2.1 Group-wise rigid registration in a GPA.

Data: $\boldsymbol{\kappa}_1 \ldots, \boldsymbol{\kappa}_M$, with length N and assigned correspondence.
Result: r_m, for $m = 1, \ldots, M$, and $\boldsymbol{\mu}$ (i.e., solution to Eq. (2.51)).
initialization: $i < i_{\max}$, $c > c_{\min}$, $\boldsymbol{\mu}^{(0)} = T(\boldsymbol{\kappa}_1)$, $\boldsymbol{\tau}_m^{(0)} = T(\boldsymbol{\kappa}_m)$ for $m = 1, \ldots, M$
while $c > c_{min}$ *and* $i < i_{max}$ **do**
  for $m = 1, \ldots, M$ **do**
    - Find rotation $\{r_m\}$ to superimpose $\boldsymbol{\tau}_m^{(i)}$ over $\boldsymbol{\mu}^{(i)}$ considering Eq. (2.39).
    - Determine superimposed vector $\boldsymbol{\tau}_m^{(i+1)} = r_m \boldsymbol{\tau}_m^{(i)}$.

  end

  - Re-calculate Procrustes mean as $\boldsymbol{\mu}^{(i+1)} = \frac{\sum_{m=1}^{M} \boldsymbol{\tau}_m^{(i+1)}}{||\sum_{m=1}^{M} \boldsymbol{\tau}_m^{(i+1)}||}$.
  - Compute $c = ||\boldsymbol{\mu}^{(i+1)} - \boldsymbol{\mu}^{(i)}||^2$, and iterate $i = i + 1$

end

As an alternative, the following group-wise registration is typically used in statistical shape analysis,

$$[r_m^*, t_m^*, \boldsymbol{\mu}^*] = \underset{||\boldsymbol{\mu}||^2=1}{\arg\min} \sum_{l=1}^{M} ||(r_m \boldsymbol{\tau}_m + \mathbf{1}_N t_m) - \boldsymbol{\mu}||^2, \tag{2.52}$$

which considers that the shape observation do not need to be unit size neither have a centroid at the origin, and thus the variability remained after the registration to the estimated mean is not necessarily only due to shape as defined by Kendall. An approach like the GPA Algorithm in 2.1, with minor extensions, also provides a solution to Eq. (2.52).

2.3.4 Point Distribution Models

When a set of training samples are available, an average and a sample covariance matrix can be estimated from these samples. The first and second order moments are statistical parameters that can be used to determine a linear model of shape. Point Distribution Models (PDM) are linear models of shape variability [14], in which an observed shape vector is described as

$$\boldsymbol{\kappa}_m = r_m(\boldsymbol{\mu} + \boldsymbol{\delta}) + \mathbf{1} t_m, \tag{2.53}$$

where $\boldsymbol{\delta}$ is considered to be a vector of shape deformation, $\boldsymbol{\mu}$ is the mean shape, and r_m and t_m are the pose parameters. To be consistent with Kendall's shape definition, the vector $\boldsymbol{\delta}$ should accomplish certain constraints so that $\boldsymbol{\mu} + \boldsymbol{\delta}$ is unit size and its centroid is the origin. However, these constraints are not usually considered and

a simplified model of $\boldsymbol{\delta}$ is preferred. This vector of deformation can be considered deterministic or random. When modeled as a random vector, a possible distribution is normal, i.e., $\boldsymbol{\delta} \sim \mathcal{CN}(\mathbf{0}, \boldsymbol{\Sigma})$. Independently of the distribution of $\boldsymbol{\delta}$, its sample covariance matrix can be calculated from a collection of M observations $\boldsymbol{\kappa}_m$. After applying a GPA, as described in Algorithm 2.1 and thus obtaining the mean-registered vectors $\boldsymbol{\tau}_m$ and its mean $\boldsymbol{\mu}$, the sample covariance matrix of the training set remains

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{M-1} \sum_{m=1}^{M} (\boldsymbol{\tau}_m - \boldsymbol{\mu})(\boldsymbol{\tau}_m - \boldsymbol{\mu})^H. \tag{2.54}$$

The principle of PDM is to consider that the N elements within the shape vector $\boldsymbol{\tau}_m$ are highly correlated, and thus most of the information about shape variability is contained in a lower-dimensional subspace. Also, typically a correlation between the two coordinates of the planar space, that is the real and imaginary parts of the shape vector $\boldsymbol{\tau}_m$, is not dismissed. This means that, the complex random vector $\boldsymbol{\delta}$ is supposed to be improper [38]. Instead of considering the complementary matrix with the sample covariance matrix [38], we define a new real vector as a composite of the real and imaginary parts of the initial complex vector. Let us define an invertible function $V : \mathbb{C}^{N\times 1} \rightarrow \mathbb{R}^{2N\times 1}$ as $V(\boldsymbol{\tau}) = [\mathrm{Re}(\boldsymbol{\tau})^T, \mathrm{Im}(\boldsymbol{\tau})^T]^T$, which provides the corresponding real description. Thus, we define the vectors $\mathbf{x}_m = V(\boldsymbol{\tau}_m) \in \mathbb{R}^{2N\times 1}$ and $\boldsymbol{\mu}_{\mathbf{x}} = V(\boldsymbol{\mu}) \in \mathbb{R}^{2N\times 1}$, and the sample covariance

$$\hat{\boldsymbol{\Sigma}}_{\mathbf{x}} = \frac{1}{M-1} \sum_{m=1}^{M} (\mathbf{x}_m - \boldsymbol{\mu}_{\mathbf{x}})(\mathbf{x}_m - \boldsymbol{\mu}_{\mathbf{x}})^T. \tag{2.55}$$

Let us define its eigenvalue decomposition, i.e., $\hat{\boldsymbol{\Sigma}}_{\mathbf{x}} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T$. We perform dimensionality reduction, assuming the linear model

$$\mathbf{x} = \boldsymbol{\mu}_{\mathbf{x}} + \mathbf{P}\mathbf{b} + \boldsymbol{\epsilon}, \tag{2.56}$$

where $\mathbf{b} \in \mathbb{R}^p$ is a vector of $p < 2N$ parameters, and $\mathbf{P} = [\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_p] \in \mathbb{R}^{2N\times p}$ are the p eigenvectors of $\hat{\boldsymbol{\Sigma}}_{\mathbf{x}}$ corresponding to the p largest eigenvalues. In order to preserve shape plausibility, $\mathbf{b}$ is restricted to a set $\mathbb{B}(\boldsymbol{\lambda})$, where $\boldsymbol{\lambda}$ is a p-dimensional vector containing the p largest eigenvalues of $\hat{\boldsymbol{\Sigma}}_{\mathbf{x}}$. The vector $\boldsymbol{\epsilon}$ accounts for the variability that we do not represent with the model. Consequently, any observed shape vector $\boldsymbol{\kappa}_m$ can be linearly approximated as $\mathbf{x}_m \approx \boldsymbol{\mu}_{\mathbf{x}} + \mathbf{P}\mathbf{b}_m$, and therefore any vector of landmarks can be approximated as

$$\boldsymbol{\kappa}_m \approx r_m V^{-1}(\boldsymbol{\mu}_{\mathbf{x}} + \mathbf{P}\mathbf{b}_m) + \mathbf{1}t_m \ , \tag{2.57}$$

characterized by the parameters $\mathbf{b}_m \in \mathbb{R}^p$, which is the shape parameter vector of $\boldsymbol{\kappa}_m$, and $\{r_m, t_m\}$, which are the rotation and translation (pose) parameters.

2.4 Shape model fitting

Now that we have introduced the fundamentals of shape theory, and particularly described the statistical analysis of planar shapes to build shape models, we focus

now on how to use these models in a practical scenario. We describe in this section how segmentation algorithms make use of SSM in order to enhance the accuray of the results.

Segmentation algorithms that use SSM [14] [20] typically consists in two steps. First, a collection of candidate landmarks are found in the image. Second, the shape model is fit to these collection of target points, such that the resulting segmentation is a valid shape observation of the SSM. Very popular algorithm such as Active Shape Models (ASM) [14] and Constraint Local Models (CLM) [20] iterate between these two steps until they converge into a segmentation solution. We illustrate a shape model fit iterative process in Fig. 2.5, and describe the two steps in detail in the following of this section.

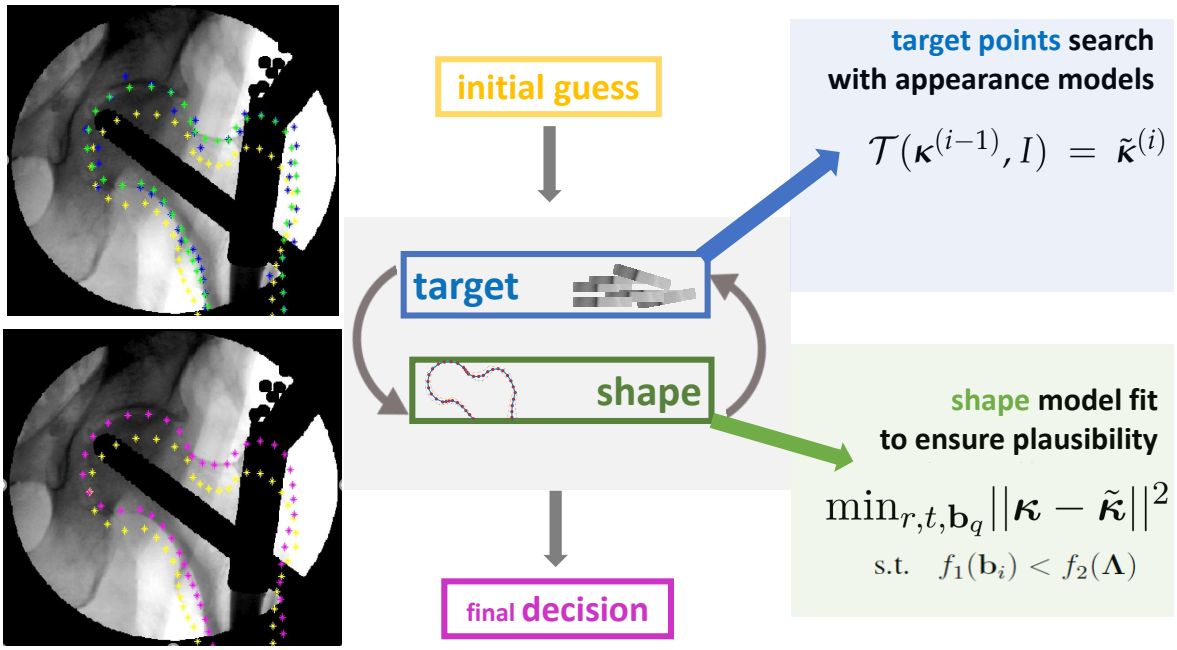


Figure 2.5: Illustration of the process of an iterative shape fitting that was run on the image of example. First, an initial guess (yellow) was given. Then, the algorithms iterates between searching the target landmarks (blue) and fitting the shape model (green). After converge, the resulting landmarks are provided (magenta). More details about the medical images are found in the Appendix.

2.4.1 Target point search

A SSM-based segmentation starts with a search of candidate landmarks. During iteration i we need to find the targets $\tilde{\boldsymbol{\kappa}}^{(i)}$ in the image I. We determine these by means of a detector $\mathcal{T}(\boldsymbol{\kappa}^{(i-1)}, I) = \tilde{\boldsymbol{\kappa}}^{(i)}$, which considers the result of the previous iteration $\boldsymbol{\kappa}^{(i-1)}$. These detectors are named part-based when the search of each nth

landmark in $\tilde{\boldsymbol{\kappa}}^{(i)}$ is independent of the rest of the landmarks. The detectors are known as holistic if the localization of each landmark depends on the other landmarks [39].

The detector can consist on a one-dimensional gray-level path of the image or on a two-dimensional region or patch, as illustrated in Fig. 2.6. In the rest of these section we describe the typical approaches for target search.

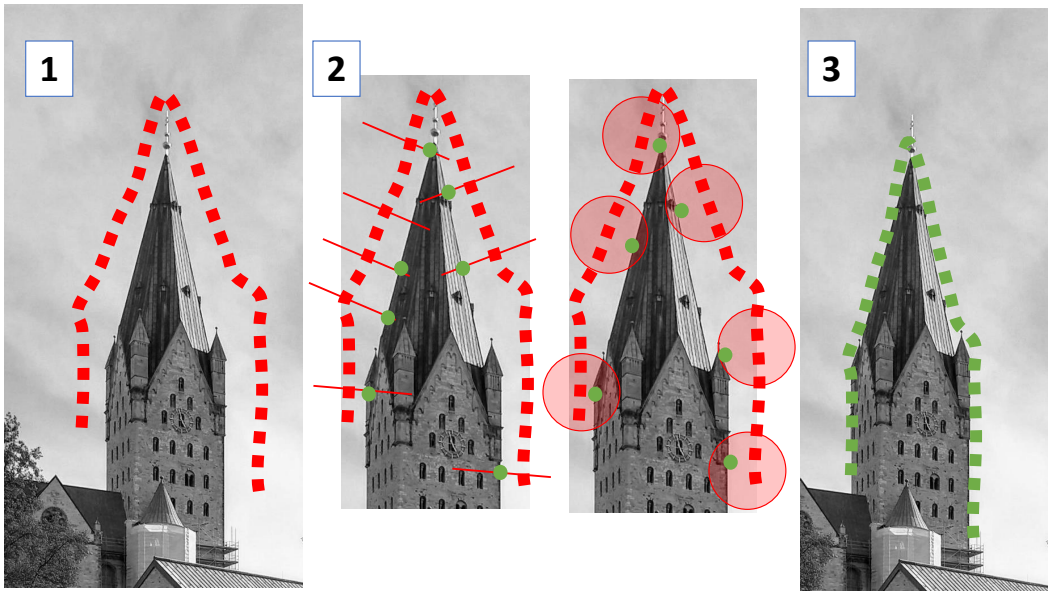


Figure 2.6: An illustration of the target point search. 1. The process starts with the landmarks from previous iteration. 2. The target point search is performed within a one-dimensional path perpendicular to the landmarks (left), or with a patch around the landmarks (right). The found targets are illustrated as green circles. 3. The target points are considered for the shape model fit, which generates the result.

One-dimensional gray-level profile

As described in the first version of the Active Shape Models (ASM) segmentation algorithm [14], a target search can consist in evaluating a grey-level profile for each landmark. ASM considers the current estimate of landmarks $\boldsymbol{\kappa}^{(i)}$, to define a line of search in the image for each landmark. We illustrate this search in Fig. 2.6(2) left.

A vector of one-dimensional intensity values is extracted from for each nth landmark. This line contains the current value of the $\kappa_n^{(i)}$, typically at the middle point, and is approximately perpendicular to the boundary formed by $\boldsymbol{\kappa}^{(i)}$. We illustrate the extraction of this line in Fig. 2.7. The best candidate to be the next target point is selected according to a determined metric. The simplest approach is to consider an intensity change, that is, to assume that the landmark is on an edge. Thus, the derivative of the gray values is approximated by finite differences, and its maximum point is found. A different and more complex approach consists on a model of ap-

pearance of the gray-level profile. Let us denote by $\mathbf{g}_n^{(i)} \in \mathbb{R}^{\ell \times 1}$ the vector containing ℓ intensity (or differences of intensity) values representing the appearance on the image line for landmark n. Assuming $\mathbf{g}_n^{(i)} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{g}_n}, \mathbf{S}_{\mathbf{g}_n})$, an average $\boldsymbol{\mu}_{\mathbf{g}_n}$ and covariance matrix $\mathbf{S}_{\mathbf{g}_n}$ are estimated from the training images. Then, during the target search, the Mahalanobis distance $d_n^{(i)} = (\mathbf{g}_n^{(i)} - \boldsymbol{\mu}_{\mathbf{g}_n})^T \mathbf{S}_{\mathbf{g}_n}^{-1} (\mathbf{g}_n^{(i)} - \boldsymbol{\mu}_{\mathbf{g}_n})$ is measured on vector candidates at different positions, and the smallest distance corresponds to the best target [14]. We illustrate this search on Fig. 2.7. There are several extensions of this approach. For instance, in [40] the authors consider a multi-resolution search, where different models are learnt from coarse to fine resolutions. In [41] the authors build color models instead of gray. Also, robust estimators such as M-estimators and RANSAC have been considered to find the targets [42]. Other strategies consider a regularizated search, as the authors in [43], that define smoothness constraints in a regularized iterative search, and solve it with dynamic programming.

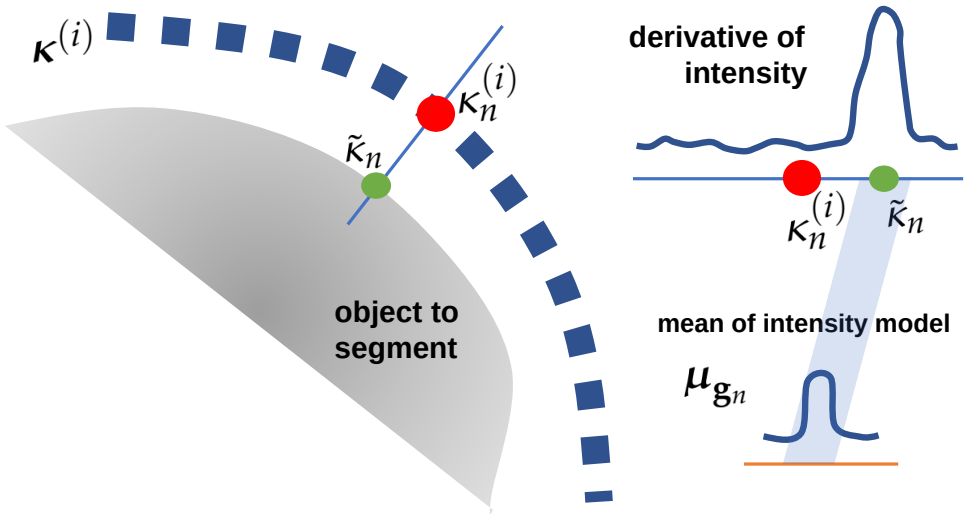


Figure 2.7: An illustration of the one-dimensinal gray-level profile target search, as used in the ASM algorithm [35].

Classification/regression of patch-based features

A grey-level vector is a very simple approach to analyse local appearance. The next level of complexity is to consider a 2-dimensional patch around each landmark. We illustrate this search in Fig. 2.6(2) right. In such a search, different image features are extracted to conform either a regression of distance to a target point, or a classification to decide if a point is the target or not. The Constrained Local Models (CLM) approach [44], with a patch-based search, has been successful in facial landmark detection. Similarly to ASM, in CLM a region is extracted around each nth landmarks at current $\kappa_n^{(i)}$ and at each iteration i, generating a response patch that represents

the likelihood of each of the points in the patch to be the next target. The appearance of each nth patch is modelled as a normalized grey-level vector, from which, similarly to the Point Distribution Model, a linear model of principal components is determined. More recent versions of the CLM consider random forests to vote for the needed displacement from the current point to the next target. In [20] the authors considered Haar features as inputs to the random forests. The best appearance features are, however, uncertain and typically chosen by heuristics. One approach to overcome this is to consider an ensemble of weak classifiers as an AdaBoost [45]. Similar approaches consider cascaded regression models [46], and robust discriminative regressions [47] [48]. Nevertheless, there are many different features and detection strategies that can be considered to determine a target position. To name a few: a generalized Hough Transform [49], Support Vector Machines [50], Scale Invariant Feature transform [51], Histogram Oriented Gradients [52], a multiview approach [53], etc.

Detection with Deep Neural Networks

More recent approaches start considering Deep Learning and Convolutional Neural Networks (CNN) to search for the target landmarks $\tilde{\boldsymbol{\kappa}}^{(i)}$ [54], [55], [56]. Particularly, the U-net architecture has been demonstrated to be very successful in landmark localization, as it is shown in the results of [57], [58], and [59]. At the time of selecting an architecture it is important to consider the computational expense of running a neural net for each target localization, which may compromise real-time requirements.

2.4.2 Shape model fit

After the target search provides a collection of landmarks, these may not form a consistent shape. Fitting a SSM to the target landmarks incorporates the prior knowledge about shape. Let us recall that, as described in previous section, at each iteration new target landmarks $\tilde{\boldsymbol{\kappa}}$ are detected. However, $\tilde{\boldsymbol{\kappa}}$ may not describe a plausible shape nor an accurate pose. In this section we describe how to fit the shape model after searching for the targets, as well as how to solve both problems (find targets and shape fit) simultaneously.

Shape fit with known targets

When these target points $\tilde{\boldsymbol{\kappa}}$ have been calculated independently from each other, the shape fit consist in finding the vector of landmarks $\boldsymbol{\kappa}$ that is closest in pose and shape to the target vector $\tilde{\boldsymbol{\kappa}}$. This model matching is presented as the following least-squares minimization

$$\boldsymbol{\kappa}^* = \underset{r,t,\mathbf{b}\in\mathbb{B}}{\arg\min} ||\boldsymbol{\kappa}(r,t,\mathbf{b}) - \tilde{\boldsymbol{\kappa}}||^2 \,, \tag{2.58}$$

where $\boldsymbol{\kappa}(r,t,\mathbf{b}) = \frac{1}{r}[V^{-1}(\boldsymbol{\mu}_\mathbf{x} + \mathbf{P}\mathbf{b}) - \mathbf{1}t]$ as described by the PDM, r and t are the pose parameters, and $\mathbf{b}$ is the shape parameter vector. To enforce shape plausibility $\mathbf{b}$ must be contained in a subset $\mathbb{B}$. A typical definition for this subset is an element-based constraint with respect to the eigenvalues of the model, $\boldsymbol{\lambda}_p$. This vector is interpreted as the variances of the elements of $\mathbf{b}$ during the training [14]. The feasible set is then expressed as

$$\mathbb{B}_{\text{var}} = \{\mathbf{b} \in \mathbb{R}^{p\times 1} : |b_n| < \alpha\sqrt{\lambda_n}, \forall n = 1,\ldots,p\}, \tag{2.59}$$

where α is typically equal to 3, and λ_i are the elements on the vector of eigenvalues $\boldsymbol{\lambda}_p$. Then, the resulting elements in $\mathbf{b}$ must be contained within α times the standard deviation of the training. Another common subset $\mathbb{B}$, also presented in [14], considers the Mahalanobis distance of $\mathbf{b}$ and a threshold ξ as

$$\mathbb{B}_{\text{Mah}} = \{\mathbf{b} \in \mathbb{R}^{p\times 1} : \mathbf{b}^T\boldsymbol{\Lambda}_p\mathbf{b} \leq \xi\}, \tag{2.60}$$

where $\boldsymbol{\Lambda}_p$ is a diagonal matrix whose diagonal is the vector of eigenvalues of the model $\boldsymbol{\lambda}_p$. In this constraint, the impact of each element of $\mathbf{b}$ is not independent, but it is a weighted part in the total Mahalanobis distance instead. This means that, if there is only one value b_n that is very far from the model, the constraint may not affect the solution.

The proposed fit in (2.58) is a non-linear least-squares problem, since the pose parameters and the shape parameter are non-linearly dependent. Still, an optimal solution can be found in an alternating optimization framework, by solving the following subproblems on each iteration search i:

- Subproblem 1: find the pose parameters $r^{(i)}$ and $t^{(i)}$, assuming $\mathbf{b}$ is fixed (using the previous, $\mathbf{b}^{(i-1)}$), and thus calculating $\mathbf{k} = V^{-1}(\boldsymbol{\mu} + \mathbf{P}\mathbf{b}^{(i-1)})$. We then solve

$$[r^{(i)}, t^{(i)}] = \arg\min_{r,t} ||(r\mathbf{k} + \mathbf{1}t) - \tilde{\boldsymbol{\kappa}}||^2 , \tag{2.61}$$

  which is a linear least-squares problem with closed form solution $[r^{(i)}\ t^{(i)}]^T = (\mathbf{K}^H\mathbf{K})^{-1}\mathbf{K}^H\tilde{\boldsymbol{\kappa}}$, with $\mathbf{K} = [\mathbf{k}\ \mathbf{1}]$.

- Subproblem 2: find the shape parameters $\mathbf{b}^{(i)}$, assuming r and t are fixed, that is, considering the already calculated $r^{(i)}$ and $t^{(i)}$. Computing $\tilde{\mathbf{x}} = V[\frac{1}{r^{(i)}}(\tilde{\boldsymbol{\kappa}} - \mathbf{1}t^{(i)})]$, the problem remains

$$\mathbf{b}^{(i)} = \arg\min_{\mathbf{b}\in\mathbb{B}} ||\boldsymbol{\mu} + \mathbf{P}\mathbf{b} - \tilde{\mathbf{x}}||^2 , \tag{2.62}$$

  which is also a linear least-squares problem. The optimal unconstrained solution is $\mathbf{b}^u = \mathbf{P}^T(\tilde{\mathbf{x}} - \boldsymbol{\mu})$. When the constraints are defined with subset $\mathbb{B}_{\text{var}}$

as in (2.59), the optimal solution to $\mathbf{b}^{(i)}$ consists in rescaling $\mathbf{b}^u$ such that the constraints are fulfilled, that is

$$b_n^{(i)} = \frac{b_n^u}{|b_n^u|} \times \min(|b_n^u|, \alpha\sqrt{\lambda_n}), \quad \forall n = 1, \ldots, p, \tag{2.63}$$

where $b_n^{(i)}$ is the nth element of $\mathbf{b}^{(i)}$, and b_n^u is the nth element of $\mathbf{b}^u$.

If the subset $\mathbb{B}_{\text{Mah}}$ is considered instead, an equivalent rescaling, as suggested in [14], is $b_n^{(i)} = b_n^u \times \left(\sqrt{\frac{\xi}{\mathbf{b}^{u\,T}\mathbf{\Lambda}_p\mathbf{b}^u}}\right), \quad \forall n = 1, \ldots, p$. However, this is not the optimal solution to (2.62). Consequently, there is no guarantee that the alternating optimization converges when $\mathbb{B}_{\text{Mah}}$ is considered. Nevertheless, the problem remains convex, and the optimal solution can be found, although not in closed-form and the computation expense may not be worth it. This suboptimal solution has been often used to solve the ASM problem [14].

This optimization problem is the most common in the ASM fit, where the targets $\tilde{\boldsymbol{\kappa}}$ are obtained with grey-level profiles [14]. Also, it is used in the CLM [60], when the targets are also fixed before the optimization takes place. The described alternating optimization solution to the shape model fit is summarized in Algorithm **??**. In these kind of segmentation strategies, new target points $\tilde{\boldsymbol{\kappa}}^{(i)}$ are calculated every ith search iteration; the iteration finishes when a shape model is fit to the targets. The overall iterative segmentation framework that alternates between searching the targets and fitting the shape model, i.e., based on the ASM, is shown in Algorithm 2.2. Notice that, there are two iterative process: one is the target search, and the second one (inside this) is the alternating optimization for the shape model fit.

Data: $\tilde{\boldsymbol{\kappa}}$ (target points), and $\{\mathbf{P}, \boldsymbol{\mu}, \boldsymbol{\lambda}\}$ (SSM).
Result: $\boldsymbol{\kappa}(r, t, \mathbf{b})$ (shape model fit)
initialization: $c > c_{\min}$, $i = 0$, $\boldsymbol{\kappa}^{(0)} = V^{-1}(\boldsymbol{\mu})$
while $c > c_{min}$ *and* $i < i_{max}$ **do**

- Obtain $r^{(i+1)}$ and $t^{(i+1)}$ considering Subproblem 1 (2.61).
- Obtain $\mathbf{b}^{(i+1)}$ considering Subproblem 2 (2.62)
- $\boldsymbol{\kappa}^{(i+1)} = r^{(i+1)}V^{-1}(\boldsymbol{\mu} + \mathbf{P}\mathbf{b}^{(i+1)}) + \mathbf{1}t^{(i+1)}$.
- $c = ||\boldsymbol{\kappa}^{(i+1)} - \boldsymbol{\kappa}^{(i)}||^2$, and iterate $i = i + 1$

end

Algorithm 2.2 Segmentation framework that alternates between two independent steps: searching for the targets $\tilde{\kappa}$, and fitting a shape model.

Data: I (image), $\kappa^{(0)}$ (initial guess), and $\{\mathbf{P}, \mu, \lambda\}$ (SSM).
Result: κ (segmentation result)
initialization: $c > c_{\min}$, $i = 0$, $\kappa^{(0)} = V^{-1}(\mu)$
while $c > c_{min}$ *and* $i < i_{max}$ **do**

- Search for the new target points $\tilde{\kappa}^{(i+1)} = \mathcal{T}(\kappa^{(i)}, I)$, as in Section 2.4.1.
- Fit a plausible shape $\kappa^{(i+1)}$ considering Algorithm **??** with input $\tilde{\kappa}^{(i+1)}$.
- $c = ||\kappa^{(i+1)} - \kappa^{(i)}||^2$, and iterate $i = i + 1$

end

Find targets and fit shape simultaneously

In the least-squares formulation in (2.58) the detection of the targets $\tilde{\kappa}$ is independent from the shape fit optimization. However, in practice the shape fit optimization has an influence on the target search $\mathcal{T}(\kappa^{(i)}, I)$ and vice versa. Therefore, they should be considered dependent parameters, and be determined jointly in a global optimization that entails the targets search and the shape model fit simultaneously. Nevertheless, it is not trivial to establish a relation between both process using regularization terms and constraints. Also, the resulting problems may become more complex, computationally demanding, and only feasible with heuristic solutions.

Nevertheless, there are different approaches that consider to solve both, target search and shape fit, simultaneously. An alternative to a global optimization is a Maximum a Posteriori (MAP) approach as in [43], [39], [44] and [61]. The MAP comprises a target search (either based on grey-level profile models or patch-based classifications) and the assumption of the shape model parameters as a prior. The output of the target detection, $\mathcal{T}(\kappa^{(i)}, I)$, is interpreted as a maximum likelihood of the targets $\tilde{\kappa}$, given the shape fit parameters $[r, t, \mathbf{b}]$. Then, the shape plausibility is endorsed considering a prior probability of these parameters. The posterior probability after a Bayes setting remains

$$[r, t, \mathbf{b}] = \arg\max(\underbrace{\log p(\tilde{\kappa}|r, t, \mathbf{b})}_{\text{likelihood of targets}} + \underbrace{\log p(r, t) + \log p(\mathbf{b})}_{\text{prior pobab.}}), \tag{2.64}$$

where $p(.)$ stands for probability density function. The likelihood term may be a function of a set of examined grey-level profiles, such as $p(\tilde{\kappa}|r, t, \mathbf{b}) = \prod_{n=1}^{N} p(\mathbf{g}_n|\mathbf{g}_n, \mathbf{S}_{\mathbf{g}_n})$ [43], or a probabilistic interpretation of the regression result of a patch-based search [62] [61]. With respect to the priors, the pose parameters (r and t) are typically assigned a non-informative prior [61]. On the contrary, the prior of the shape vector $\mathbf{b}$, statistically independent from r and t, is defined considering the shape model parameters ($\mathbf{P}$, μ, and Λ_t) as $\log p(\mathbf{b}) \propto -\mathbf{b}^T \Lambda_t \mathbf{b}$ [44] [61] [39]. The complex solution of these objective functions typically leads to approximated solutions. One

option is a Nelder-Meade simplex method [43] [44], or an approximated Expectation Maximization iterative approach [61].

The shape matching can also be performed optimizing a quality-of-fit with no explicit probability interpretation, considering an independent sum of costs for each landmark [63] as

$$[r, t, \mathbf{b}] = \underset{r^{(i)}, t^{(i)}, \mathbf{b}^{(i)} \in \mathbb{B}}{\arg\min} \sum_{n=1}^{N} \mathcal{T}(\boldsymbol{\kappa}^{(n)}) \,, \tag{2.65}$$

which in [63] is heuristically optimized in an iterative process after a patch-based search result. Notice that the optimization procedures to solve either (2.64) or and (2.65) need to be computed several times during the segmentation, since the search and shape match are inside an iterative segmentation framework similar to the one described in Algorithm 2.2. Therefore, these approaches may become too computationally complex.

More recently, after the strong incorporation of Deep Neural Networks in Computer Vision, the research focus has moved to integrate shape information into the back-propagation processes of the network training [64] [65]. Therefore, novel approaches such as deep structured Convolutional Neural Network [65] and deep Regressors [64] consist in an end-to-end segmentation system that comprehends the shape constraints in a SSM-based layer, together with the search of best segmentation landmarks, all integrated in a single architecture.

2.5 Statistical shape analysis with incomplete data

SSM may also be used to estimate a collection of landmarks on a shape vector that are occluded or missing. These incomplete observations are calculated considering the observed points as a starting point, together with the statistical information of the shape model. Also, another topic concerning incomplete data is the shape model design when the training samples are incomplete. This is a very common pre-processing step required for real data in data mining algorithms [66].

Hence, shape vectors present missing points and outliers in both training and to-be-segmented images. We present in this section how we have dealt with both problematics: finding missing landmarks in a shape vector, and dealing with incomplete data to train a shape model.

2.5.1 Linear Minimum Mean Square Error for missing landmarks

There are many applications where the object to segment may be partially occluded and the goal is to estimate this hidden section of the boundary of the object or to reconstruct a damage section of an object. A naive estimate of these occlusions is the mean shape $\boldsymbol{\mu}$. An assumed distribution of shape deformation may also be used to estimate a posterior shape [67]. Finding a collection of missing landmarks is an

estimation problem. In order to find a trade-off between the error variance and the bias of the estimate, we consider to minimize the mean squared-error of the estimation.

When we consider PDM we can a linear estimator. Let us recall the real descriptor vector of shape, $\mathbf{x} = V(\boldsymbol{\tau}) \in \mathbb{R}^N$, where $\boldsymbol{\tau} \in \mathbb{C}^N$ is a shape vector registered to the Procrustes shape mean. Let $\mathbf{x}^l \in \mathbb{R}^{N_l}$ denote the elements in $\mathbf{x}$ that correspond to the lost landmarks (or missing), and $\mathbf{x}^o \in \mathbb{R}^{N_o}$ the elements corresponding to the observed landmarks, and $N = N_l + N_o$. The PDM parameters, that is, $\mathbf{P}$, $\boldsymbol{\Lambda}$ and $\boldsymbol{\mu}_\mathbf{x}$ are re-ordered to determine $\tilde{\mathbf{P}}$, $\tilde{\boldsymbol{\Lambda}}$ and $\tilde{\boldsymbol{\mu}}_\mathbf{x}$ so that the first N_l elements in $\tilde{\boldsymbol{\mu}}_\mathbf{x}$, and the first N_l rows in $\tilde{\mathbf{P}}$ and $\tilde{\boldsymbol{\Lambda}}$ correspond to the missing elements $\mathbf{x}^m$, that is

$$\begin{aligned} \mathbf{x}^l &= \boldsymbol{\mu}_\mathbf{x}^l + \mathbf{P}^l \mathbf{b}^l, \\ \mathbf{x}^o &= \boldsymbol{\mu}_\mathbf{x}^o + \mathbf{P}' o \mathbf{b}^o, \\ \tilde{\boldsymbol{\mu}}_\mathbf{x} &= [(\boldsymbol{\mu}_{,\mathbf{x}}^{\,l})^T,\ (\boldsymbol{\mu}_\mathbf{x}^o)^T]^T, \\ \tilde{\mathbf{P}} &= [(\mathbf{P}^l)^T,\ (\mathbf{P}^o)^T]^T, \end{aligned} \tag{2.66}$$

where $\mathbf{P}^l$ and $\mathbf{P}^o$ are the rows of $\mathbf{P}$ that correspond to N_l and N_o respectively, and $\boldsymbol{\mu}_\mathbf{x}^l$ and $\boldsymbol{\mu}_\mathbf{x}^o$ are the N_l and N_o elements in $\boldsymbol{\mu}_\mathbf{x}$. Let us assume that $\mathbf{x}^m$ is a random vector of shape, and that we define a linear estimate from the observed vector $\mathbf{x}^o$ calculating

$$\tilde{\mathbf{P}}\tilde{\boldsymbol{\Lambda}}\tilde{\mathbf{P}}^T = \tilde{\boldsymbol{\Sigma}} = \begin{bmatrix} \boldsymbol{\Sigma}^{ll} & \boldsymbol{\Sigma}^{lo} \\ (\boldsymbol{\Sigma}^{lo})^T & \boldsymbol{\Sigma}^{oo} \end{bmatrix}, \tag{2.67}$$

where $\boldsymbol{\Sigma}^{lo}$ denotes the sample cross-covariance matrix between the missing landmarks $\mathbf{x}^l$ and the remaining available landmarks $\mathbf{x}^o$, and $\boldsymbol{\Sigma}^{oo}$ is the covariance matrix of the available landmarks. The Linear Minimum Mean Squared Error (LMMSE) [68] estimator of the missing landmark from the remaining landmarks is then

$$\hat{\mathbf{x}}^l = \boldsymbol{\Sigma}^{lo}(\boldsymbol{\Sigma}^{oo})^{-1}(\mathbf{x}^o - \boldsymbol{\mu}_\mathbf{x}^o) + \boldsymbol{\mu}_\mathbf{x}^l. \tag{2.68}$$

2.5.2 Building shape models with incomplete training samples

For building PDM we need to estimate the mean and the covariance matrix of the training shape vectors, from which we extract the corresponding eigenvectors and eigenvalues. Training vectors are collected from real-world data, which contains very frequently imperfections and artifacts . Also, there may be occlusions in these training contours. All these effects translate into missing landmarks, randomly located in different positions of the training vectors. In order to train the PDM we need to estimate these missing landmarks first [66].

This missing data estimation can be done in an Expectation-Maximization (EM) framework. One of the applications of the EM algorithm is to find maximum likelihood solutions for models having latent variables [69], which are the missing landmarks in this case. In spite of these incomplete samples, we need to find the maximum likelihood estimate of the covariance matrix from the training samples. In each iteration, the Expectation (E-step) consists in calculating the missing values for each

training sample $\mathbf{x}_m$, that is $\mathbf{x}_m^l$, considering an LMMSE estimate as described in the previous section [70]. When the observations are considered jointly Gaussian, i.e., $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{x}, \boldsymbol{\Sigma})$, the LMSSE-estimate $\hat{\mathbf{x}}_{\text{LMMSE}}^l$, coincides with the expected value of the posterior likelihood of the missing values , and

$$\mathbf{x}^l|_{\mathbf{x}^o, \boldsymbol{\mu}_\mathbf{x}, \boldsymbol{\Sigma}} \sim \mathcal{N}(\hat{\mathbf{x}}_{\text{LMMSE}}^l, \mathbf{Q}), \tag{2.69}$$

with $\mathbf{Q} = \boldsymbol{\Sigma}^{ll} - \boldsymbol{\Sigma}^{lo}(\boldsymbol{\Sigma}^{oo})^{-1}(\boldsymbol{\Sigma}^{lo})^T$. Then, when this is done for all the M samples, the Maximization (M-step) consist in finding the maximum likelihood estimates of the mean and covariance matrix of the training data, $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}_\mathbf{x}$, given the LMSSE estimates of the missing values. The strategy is described in Algorithm 2.3.

Algorithm 2.3 EM-based approach to estimate the mean and covariance matrix of training data when there are missing values.

Data: training data $\mathbf{x}_1, \ldots, \mathbf{x}_M$.
Result: ML estimates $\boldsymbol{\Sigma}$, $\boldsymbol{\mu}_\mathbf{x}$
initialization: $\boldsymbol{\mu}^{(0)}$,$\boldsymbol{\Sigma}^{(0)}$, $c > c_{\min}$, $i = 0$
while $c > c_{min}$ *and* $i < i_{max}$ **do**
E-step:
for $m = 1, \ldots, M$ **do**

- LMMSE estimate of the missing variables

$$\hat{\mathbf{x}}_{\text{LMMSE}}^l = \boldsymbol{\Sigma}^{lo(i)}(\boldsymbol{\Sigma}^{oo(i)})^{-1}(\mathbf{x}_m^o - \boldsymbol{\mu}^{o(i)}) + \boldsymbol{\mu}^{o(i)} \tag{2.70}$$

- Impute observation in correct variable order:

$$\tilde{\mathbf{x}}_m = \text{reorder}[\mathbf{x}_m^o,\ \hat{\mathbf{x}}_{\text{LMMSE}}^l] \tag{2.71}$$

end
M-step:

- ML estimates:

$$\boldsymbol{\mu}^{(i+1)} = \frac{1}{M}\sum_{m=1}^{M} \tilde{\mathbf{x}}_m \tag{2.72}$$

$$\boldsymbol{\Sigma}^{(i+1)} = \frac{1}{M-1}\sum_{m=1}^{M} (\tilde{\mathbf{x}}_m - \boldsymbol{\mu}^{(i+1)})(\tilde{\mathbf{x}}_m - \boldsymbol{\mu}^{(i+1)})^T \tag{2.73}$$

- Check convergence: $c = ||\boldsymbol{\Sigma}^{(i)} - \boldsymbol{\Sigma}^{(i+1)}||_F$.
- Increase iteration: $i = i + 1$.

end

The resulting sample covariance matrix, $\boldsymbol{\Sigma}$ is positive-semi-definite. However, it may be ill-conditioned if too many samples are missing. In such cases, the submatrix

$\mathbf{\Sigma}^{oo}$ may already be poorly conditioned, providing unstable and erroneous results to the LMMSE and maximum likelihood estimates. To overcome this, a regularized version of the EM algorithm is proposed in [70]. This EM-based solutions does not guarantee optimality, nor convergence, given the unknown nature of the training data [71], which may not be normal, and the missing values may be distributed under certain patters, providing correlation between the observations. Still, this approach has been empirically demonstrated to provide successful results [66].

Part II

Developed work and contributions

Chapter 3
Robust Active Shape Models

"Ideas do not last long. We must do something with them."

— Santiago Ramón y Cajal

As described in Chapter 2, Active Shape Models are an iterative segmentation technique to find a landmark-based contour of an object in an image. In each iteration, a least-squares fit of a plausible shape to some detected target landmarks is determined. However, as motivated in Chapter 1, ASM-based segmentation algorithms fail to work in fluoroscopic images, since these have low resolution. Finding the target landmarks is a critical step in ASM: some landmarks are more reliably detected than others, and some landmarks may not be within the field of view of their detectors. To add robustness while preserving simplicity at the same time, a generalized least-squares approach can be used instead, where a weighting matrix can incorporate reliability information about the landmarks. In this chapter we propose a strategy to choose this matrix. We evaluate our strategy on fluoroscopic X-ray images to segment the femur. We show that our technique outperforms the standard ASM as well as other more heuristic weighted least-squares strategies.

In Section 3.1 we motivate the contribution with is a state-of-the-art, as well as with an overview of the preliminary concepts. In Section 3.2 we discuss our approach in detail. In Section 3.3 we evaluate the performance of our technique, which is based on a leave-one-out test.

The results of this chapter haven been published in:

- A. Eguizabal and P. Schreier, "A weighting strategy for Active Shape Models," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, Beijing, China, Sept. 2017, pp. 1-6.

3.1 Motivation and preliminaries

ASM adapt to many kinds of shapes and imaging modalities. However, they do require images with reasonably good contrast in order to find the object contours. An example of a challenging scenario is intraoperative fluoroscopic X-ray imaging, which

produces low-quality images due to the low X-ray dose. More robust alternatives to ASM exist (e.g., [20]), but they are typically more computationally expensive. Motivated by the real-time computation requirements in CAS, our goal is to keep the simplicity of ASM and still improve robustness. For this purpose we propose replacing the least-squares procedure in the ASM with a generalized least-squares (GLS) approach. The idea of GLS is to weight individual landmarks according to their reliability: reliably identified landmarks should be trusted more than less reliable landmarks. This raises the question of how to identify the right weighting strategy. This question has already been addressed before: in [72], where a directional regularization is proposed; in [42], which suggests to use robust parameter estimation; in [73] and [74], where the weights change in every iteration depending on a score of the target detectors; in [75], which proposes a pose-invariant metric; and in [76], which measures the reliability of the landmarks based on models of local appearance. Many of these weighting strategies are heuristic and also prone to over-fitting. Some of them also add computational complexity to the ASM algorithm.

Our contribution in this chapter is a strategy for choosing the weights, which has a theoretical justification and still keeps the simplicity of the ASM algorithm. Our proposed strategy measures the reliability of the target landmarks based on the covariance of the residuals of the fit obtained from training data. Additionally, we perform a test to determine whether a landmark is valid, i.e., whether it is within the field of view of its target detector, and we incorporate this test into the weighting matrix.

3.1.1 The least-squares fit in ASM

As described in Chapter 2, the ASM places a set of N landmarks on the contour of the object of interest. We use M training images with manually placed landmarks to learn the shape variability. We model the landmarks in a complex space $\mathbb{C}$, where the pose parameters are defined by a complex affine transformation: the landmarks in the mth training image are complex vectors $\boldsymbol{\kappa}_m \in \mathbb{C}^{N\times 1}$, whose real and imaginary parts correspond to the coordinates of the two-dimensional Euclidean space.

Recalling the description of PDM in Chapter 2, a vector of landmarks can be approximated as

$$\boldsymbol{\kappa}_m \approx \frac{1}{r_m}\left(V^{-1}(\hat{\boldsymbol{\mu}}_{\mathbf{x}} + \mathbf{P}\mathbf{b}_m) - \mathbf{1}t_m\right) , \tag{3.1}$$

where $\mathbf{b}_m \in \mathbb{R}^{p\times 1}$ is the shape parameter vector of $\boldsymbol{\kappa}_m$, and $\{r_m, t_m\}$ its pose parameters. Let us assume that $\boldsymbol{\kappa}^{(i)}$ is the resulting vector of landmarks after iteration i of the ASM. As described in Chapter 2, each iteration first examines local regions around each landmark in $\boldsymbol{\kappa}^{(i-1)}$ in order to detect new target landmarks $\tilde{\boldsymbol{\kappa}}^{(i)}$. However, $\tilde{\boldsymbol{\kappa}}^{(i)}$ may not describe a plausible shape or an accurate pose. In order to find the closest plausible shape and pose, that is, the landmarks in vector $\boldsymbol{\kappa}^{(i)}$, the following least-squares problem is solved:

$$\min_{r^{(i)},t^{(i)},\mathbf{b}^{(i)}\in\mathbb{B}} ||\boldsymbol{\kappa}^{(i)} - \tilde{\boldsymbol{\kappa}}^{(i)}||^2 . \tag{3.2}$$

Ordinary least-squares as a maximum likelihood problem

The typical ASM is based on an OLS fit that obtains the pose (r, t) and shape ($\mathbf{b}$) parameters. The residuals $\boldsymbol{\epsilon}$ of this regression are the difference between the parametrized vector following the model, $\boldsymbol{\kappa}(r,t,\mathbf{b}) \in \mathbb{C}^N$, and the observed vector $\tilde{\boldsymbol{\kappa}} \in \mathbb{C}^N$, that is, $\boldsymbol{\epsilon} = \boldsymbol{\kappa} - \tilde{\boldsymbol{\kappa}}$. The OLS fit consists in minimizing the squared sum of the elements on the residual vector $\boldsymbol{\epsilon}$

$$\min_{r,t,\mathbf{b}} ||\boldsymbol{\epsilon}||^2. \tag{3.3}$$

Let us assume $\boldsymbol{\epsilon} \sim \mathcal{CN}(0, \sigma^2\mathbf{I})$, where each of the components in $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_n, \ldots, \epsilon_N]^T$ correspond to the residual of one landmark. The likelihood of the parameters $r, t, \mathbf{b}$ given a single landmark ϵ_n follows

$$\mathcal{L}(r,t,\mathbf{b}|\epsilon_n) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-\epsilon_n^2}{2\sigma^2}\right). \tag{3.4}$$

Since we have assumed white noise, each residual ϵ_n is independent of each other, and the expression of the likelihood considering the N landmarks remains

$$\mathcal{L}(r,t,\mathbf{b}) = \prod_{n=1}^{N} \mathcal{L}(r,t,\mathbf{b}|\epsilon_n). \tag{3.5}$$

Therefore, under these conditions, the least-squares problem in (3.3) is equivalent to a maximum likelihood estimation of the parameters, that is

$$[r^\star, t^\star, \mathbf{b}^\star] = \arg\max_{r,t,\mathbf{b}} \mathcal{L}(r,t,\mathbf{b}) = \arg\min_{r,t,\mathbf{b}} [-\log \mathcal{L}(r,t,\mathbf{b})] = \\ \arg\min_{r,t,\mathbf{b}} \sum_{n=1}^{N} \left(\frac{\epsilon_n^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)\right) = \arg\min_{r,t,\mathbf{b}} ||\boldsymbol{\epsilon}||^2. \tag{3.6}$$

Our proposed approach in this chapter is based on interpreting the least-squares in ASM as an ML in order to give a statistical meaning to the residuals.

3.1.2 Robustness: why and how

Outliers and robustness

Real-life measurements often contain outliers. These are samples that do not behave as expected, and differ significantly from the rest of the samples. These may cause error-prone model parameters when these are trained from measurements with outliers. Also, when models are fit to data with outliers, the resulting fit may differ significantly from a good fit.

In regression models it is often assumed that the measurements contain additive Gaussian noise. In fact, as described in the previous section, the OLS fit in ASM is equivalent to an ML estimate if the noise is white and Gaussian. This means that the

difference between an observed vector $\tilde{\mathbf{x}}$ and its regressed (or estimated) value $\mathbf{x}$, i.e., the residual $\boldsymbol{\epsilon} = \mathbf{x} - \tilde{\mathbf{x}}$, is modeled as $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$. However, this Gaussian model would represent the residuals very poorly if there were outliers in the fitting data, for instance, when some landmarks are not detected by the target search or their detection is very poor. Robust statistics deal with this problematic [77]. They consist in preventing the estimation accuracy to break-down in the presence of outliers, and also adapting to possible deviations from the typical estimation assumptions (such as the Gaussian white noise model). Therefore, as illustrated in Fig. 3.1, making a technique more robust consists in a trade-off between optimally obtaining model-based parameters and fully depending on observed data.

The typical ASM problem entails solving an OLS to obtain the pose and shape parameters. If the residuals in an OLS minimization are not Gaussian and white, the parameter calculation may be biased, as illustrated on Fig. 3.2, or even break down. Consequently, in this chapter we consider the following points to make the ASM more robust:

1. Contemplate the residuals of the least-squares fit as different from Gaussian and white.

2. Detect the outliers to discard them before the least-squares fit, and therefore avoid the bias (as in the illustrative example of Fig. 3.2).

A generalized least-squares in ASM

An approach to add robustness while keeping simplicity is to substitute an OLS by a generalized least-squares (GLS). The expression in (3.2) is an OLS, where every component of the difference between the vectors $\boldsymbol{\kappa}^{(i)}$ and $\tilde{\boldsymbol{\kappa}}^{(i)}$ has the same impact on the minimization. These components, i.e. the landmark positions, can be weighted according to their reliability in a GLS:

$$\min_{r^{(i)},t^{(i)},\mathbf{b}^{(i)}\in\mathbb{B}} (\boldsymbol{\kappa}^{(i)} - \tilde{\boldsymbol{\kappa}}^{(i)})^H \mathbf{W} (\boldsymbol{\kappa}^{(i)} - \tilde{\boldsymbol{\kappa}}^{(i)}) \ , \tag{3.7}$$

where $\mathbf{W}$ is the matrix of weights that controls the importance of individual landmarks in the optimization. We will consider both diagonal and nondiagonal weight matrices. Our goal is to find a good definition of the matrix of weights $\mathbf{W}$ that has a theoretical explanation in the context of the problem.

The chi-square test

Another incorporation that we propose to the robust ASM is a chi-square test. This test is based on the analysis of a chi-square distribution.

A chi-square distribution with ℓ degrees of freedom is the probability density function of the sum of ℓ squared independent normal random variables. A chi-square test is based on determining whether a random variable d is distributed as chi-square or not. This test has many possible applications in statistical analysis, such as deciding

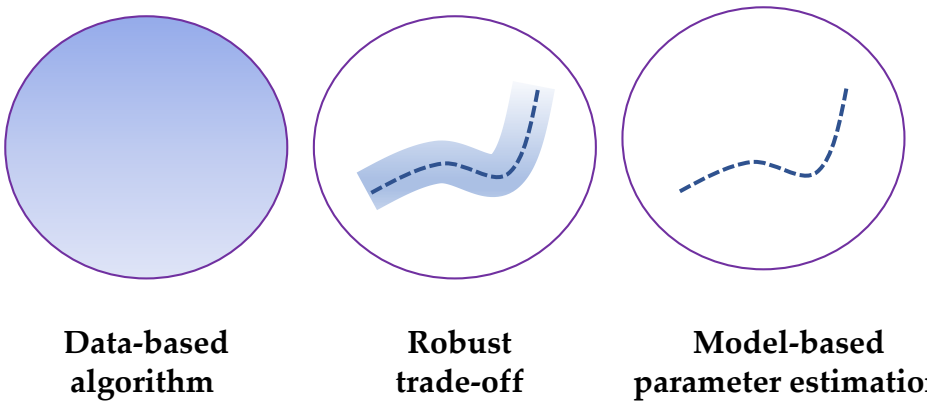


Figure 3.1: Visual intuition of the trade-off given by a robust approach in comparison to a fully data-based algorithm (such as a neural net or a decision tree), and to a model-based parameter estimation (such as a linear regression based on an ordinary least-squares, which is optimized for Gaussian and white residuals). Data-based algorithms are fully determined by the observed data and a priori more solutions are possible. However, model-based approaches are typically more constrained since they need to follow the model. A robust solution allows deviations from a model in order to account unexpected changes on the data. This chart is inspired in [77] and [78].

if a set of random variables are independent and normal, or, as in our case, determining whether the observed data fits a model of statistical behaviour. Assuming that a vector $\mathbf{g} \in \mathbb{R}^{\ell}$ behaves as normal, i.e., $\mathbf{g} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{g}}, \mathbf{S}_{\mathbf{g}})$, and that we know the parameters of the distribution (mean vector $\boldsymbol{\mu}_{\mathbf{g}}$ and covariance matrix $\mathbf{S}_{\mathbf{g}}$), we may detect an outlier observation of such vector by considering a chi-square test. This test can be applied during the target search of the ASM algorithm, where the Mahalanobis distance to an appearance model of a landmark is calculated as

$$d = (\mathbf{g} - \boldsymbol{\mu}_{\mathbf{g}})^T \mathbf{S}_{\mathbf{g}}^{-1} (\mathbf{g} - \boldsymbol{\mu}_{\mathbf{g}}). \tag{3.8}$$

The null hypothesis of the test assumes that d is a sample from a chi-square distribution of order ℓ, and the alternative hypothesis means that it does not, that is

$$\begin{aligned} \mathcal{H}_0 &: d \sim \chi^2_{\ell} \\ \mathcal{H}_1 &: d \not\sim \chi^2_{\ell}. \end{aligned} \tag{3.9}$$

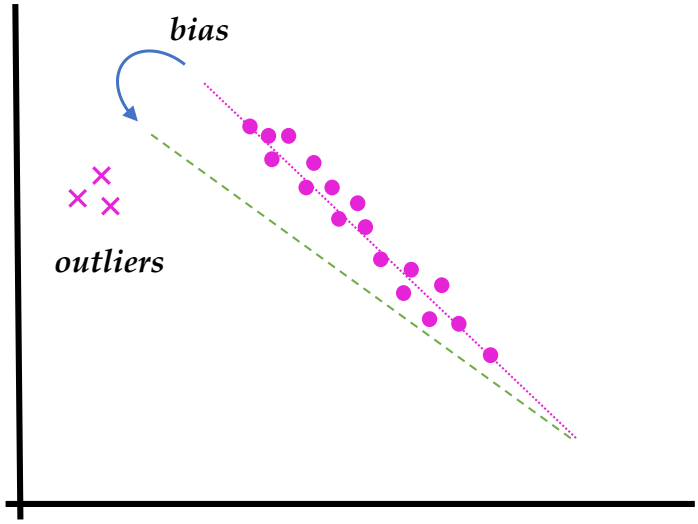


Figure 3.2: Illustrative example on how a set of outliers in the observed data (crosses), which differ from the typical observed data (dots), may bias the model parameters. In this case, the parameters model the slope of a line in a linear regression.

We consider a false alarm probability p for this outlier test. Therefore, p is the probability that the null hypothesis is erroneously dismissed, meaning that a sample that was not an outlier is considered so. Consequently, $\mathcal{H}_0$ is assumed to be correct if $d \leq f$, where f is the value at which χ^2_ℓ fulfils that the area under its probability density function from f to ∞ is p, i.e., the probability $P(d > f) = p$. In other words, when the cumulative distribution function of χ^2_ℓ at f is $1 - p$. We illustrate this explanation in Fig. 3.3.

3.2 A weighting strategy for ASM

Our main contribution consists in designing the matrix $\mathbf{W}$ in Eq. (3.7). In each ASM iteration we need to find the new targets $\tilde{\boldsymbol{\kappa}}^{(i)}$. We determine these by means of a detector $\mathcal{T}(\boldsymbol{\kappa}^{(i-1)}) = \tilde{\boldsymbol{\kappa}}^{(i)}$, which explores the local regions around the previous landmarks $\boldsymbol{\kappa}^{(i-1)}$. We assume that this detector searches for the best match of the gray-level profile around each landmark as measured by the Mahalanobis distance. However, our technique can be generalized to other strategies, such as the ones described in Chapter 2. A complicating factor is that the true landmarks $\boldsymbol{\kappa}^\star$ are not necessarily within the local regions explored by the detector $\mathcal{T}(\boldsymbol{\kappa}^{(i-1)})$, i.e. they are not within its field of view. This is the case when the true landmarks are occluded

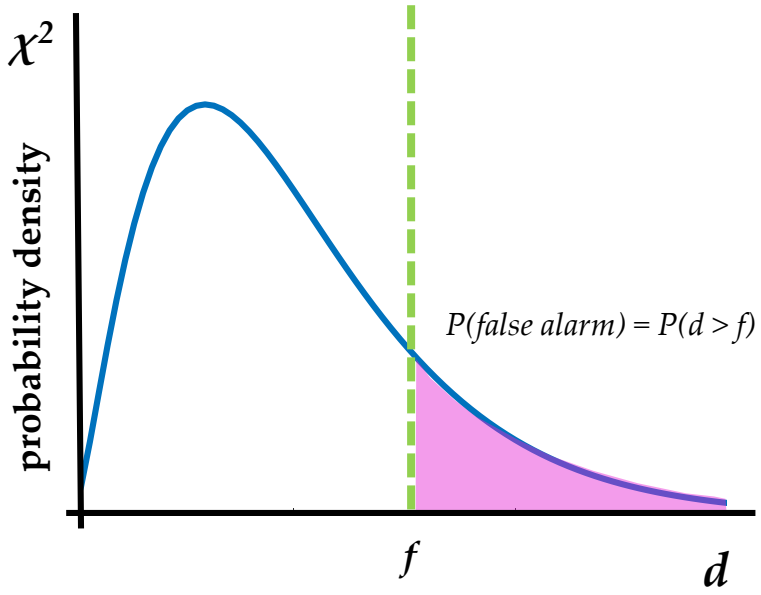


Figure 3.3: Illustration of a chi-square test to detect outliers. We depict the probability density function of χ^2_ℓ with $\ell = 5$. The magenta area represents the value of the probability of false alarm.

or out of alignment (especially during the first iterations of the ASM algorithm). We therefore incorporate into our strategy a chi-square test to determine whether the true landmark is visible to the detector. If the test results that a landmark is not visible, then this landmark is excluded by setting the corresponding entry in the weighting matrix $\mathbf{W}$ to zero.

3.2.1 GLS as a maximum likelihood problem

Let us first assume the true landmarks $\boldsymbol{\kappa}^\star$ are within the field of view of the detector $\mathcal{T}$ in iteration i, i.e., when determining $\tilde{\boldsymbol{\kappa}}^{(i)}$. We define the residual error vector as $\boldsymbol{\epsilon}^{(i)} = \boldsymbol{\kappa}^{(i)} - \tilde{\boldsymbol{\kappa}}^{(i)}$. We also assume that $\boldsymbol{\epsilon}^{(i)}$ is complex normal with zero mean and covariance matrix $\mathbf{R}$. Thus, the likelihood of the pose and shape parameters $\{r^{(i)}, t^{(i)}, \mathbf{b}^{(i)}\}$ given the error $\boldsymbol{\epsilon}^{(i)}$ is

$$\mathcal{L}(r^{(i)}, t^{(i)}, \mathbf{b}^{(i)}) = \frac{1}{\pi^N |\mathbf{R}|} \exp\left(- \boldsymbol{\epsilon}^{(i)H} \mathbf{R}^{-1} \boldsymbol{\epsilon}^{(i)} \right) . \tag{3.10}$$

Given these conditions, and equivalently to Section 3.1.1, the ML estimation of the parameters $\{r^{(i)}, t^{(i)}, \mathbf{b}^{(i)}\}$ is equivalent to a GLS problem [79], that is

$$\max_{r^{(i)}, t^{(i)}, \mathbf{b}^{(i)}} \mathcal{L}(r^{(i)}, t^{(i)}, \mathbf{b}^{(i)}) = \min_{r^{(i)}, t^{(i)}, \mathbf{b}^{(i)}} \boldsymbol{\epsilon}^{(i)H} \mathbf{R}^{-1} \boldsymbol{\epsilon}^{(i)} , \tag{3.11}$$

where the constraints on $\mathbf{b}^{(i)}$ imposed in (3.7) affect the solution, but do not alter the equality in (3.11). Therefore, if we assume that the residual vector is normally distributed, finding the weighting matrix $\mathbf{W}$ in (3.7) is equivalent to estimating the covariance matrix of the residuals, i.e., we set $\mathbf{W} = \mathbf{R}^{-1}$. If the training dataset for estimating $\mathbf{R}$ is small, it is typically preferable to constrain $\mathbf{W}$ to be diagonal in order to avoid over-fitting. This is equivalent to the assumption of statistical independence between the residuals of each landmark.

3.2.2 Empirical determination of the residual errors

We need to estimate $\mathbf{R}$ based on a set of residual errors that are empirically determined from training images. We first simulate the search of the target landmarks using a detector $\mathcal{T}'$. We split our available training data set containing $2S$ images into two subsets of equal size. With the first subset we train $\mathcal{T}'$; with the second subset we measure the residual errors when employing $\mathcal{T}'$. For this, we simulate the detection of target landmarks, $\mathcal{T}'(\boldsymbol{\kappa}^{(i-1)})$, for each of the S images in the second subset, assuming that the true landmarks of each image are within the field of view of the detector $\mathcal{T}'$. This is achieved by setting $\boldsymbol{\kappa}^{(i-1)} = \boldsymbol{\kappa}_s^\star + \boldsymbol{\delta}$, where $\boldsymbol{\delta} \in \mathbb{C}^N$ is a vector of translations small enough so that the true landmarks of the sth image, $\boldsymbol{\kappa}_s^\star$, remain within the field of view of $\mathcal{T}'(\boldsymbol{\kappa}^{(i-1)})$. We employ the detector that searches for the best match of a gray-level profile as in [14]; hence we determine $\boldsymbol{\delta}$ to place $\boldsymbol{\kappa}^{(i-1)}$ on the line perpendicular to the true object contour, with translations small enough so that $\boldsymbol{\kappa}_s^\star$ is sampled by $\mathcal{T}'(\boldsymbol{\kappa}^{(i-1)})$.

For each of the S images, we perform $\mathcal{T}'(\boldsymbol{\kappa}_s^\star + \boldsymbol{\delta}) = \hat{\boldsymbol{\kappa}}_s$ and measure $\hat{\boldsymbol{e}}_s = \boldsymbol{\kappa}_s^\star - \hat{\boldsymbol{\kappa}}_s$, which is the residual error obtained from training sample s. To determine the sample covariance matrix of the residuals we assemble the error matrix $\hat{\mathbf{E}} = [\hat{\boldsymbol{e}}_1, \ldots, \hat{\boldsymbol{e}}_S] \in \mathbb{C}^{N \times S}$, which contains the residuals from all training samples. The sample covariance matrix is then

$$\hat{\mathbf{R}} = \frac{1}{S} \hat{\mathbf{E}} \hat{\mathbf{E}}^H. \tag{3.12}$$

We note that the so determined $\hat{\mathbf{R}}$ does not depend on the iteration i of the ASM algorithm.

3.2.3 Testing whether a target landmark is valid

The sample covariance matrix of the residual errors $\hat{\mathbf{R}}$ is determined under the assumption that the true landmarks $\boldsymbol{\kappa}^\star$ are within the field of view of the detectors $\mathcal{T}(\boldsymbol{\kappa}^{(i-1)})$. If the true landmarks are not visible to the detector, the determined target landmarks are certainly incorrect. In such a case, the target landmarks are not valid. In order to test whether a target landmark in $\tilde{\boldsymbol{\kappa}}^{(i)}$ returned by the ASM algorithm is valid, we run a hypothesis test based on the metric that is used to determine the target landmark. In our case, this metric is a Mahalanobis distance $d_n^{(i)}$ that measures the distance between the observed and the modeled local appearance. Let us denote by $\mathbf{g}_n^{(i)} \in \mathbb{R}^{\ell \times 1}$ the vector containing ℓ intensity values representing

the appearance observed by detector $\mathcal{T}(\boldsymbol{\kappa}^{(i-1)})$ to determine landmark n. The null hypothesis $\mathcal{H}_0$ is that the target landmark, which corresponds to the point that minimizes $d_n^{(i)}$, is valid. We assume $\mathbf{g}_n^{(i)} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{g}_n}, \mathbf{S}_{\mathbf{g}_n})$. Thus, the Mahalanobis distance $d_n^{(i)} = (\mathbf{g}_n^{(i)} - \boldsymbol{\mu}_{\mathbf{g}_n})^T \mathbf{S}_{\mathbf{g}_n}^{-1} (\mathbf{g}_n^{(i)} - \boldsymbol{\mu}_{\mathbf{g}_n})$ follows a chi-squared distribution with ℓ degrees of freedom under the null hypothesis

$$\mathcal{H}_0 : d_n^{(i)} \sim \chi_\ell^2 \tag{3.13}$$

Therefore, for each iteration i and each landmark n, we perform a chi-square test of $d_n^{(i)}$: if the null hypothesis is not rejected, then we set an indicator variable $f_n^{(i)} = 1$; otherwise $f_n^{(i)} = 0$. We can fix a false alarm rate for this test based on the χ_ℓ^2 distribution. The idea of this approach is to work only with target landmarks whose observed gray-level profiles are close enough to the model profiles. This test can be incorporated into the weighting matrix $\mathbf{W}$. For each iteration i, we define a diagonal matrix

$$\mathbf{F}^{(i)} = \begin{pmatrix} f_1^{(i)} & 0 & \cdots & 0 \\ 0 & f_2^{(i)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & f_N^{(i)} \end{pmatrix}. \tag{3.14}$$

The weighting matrix $\mathbf{W}^{(i)}$ is then obtained as

$$\mathbf{W}^{(i)} = \mathbf{F}^{(i)} \hat{\mathbf{R}}^{-1} \mathbf{F}^{(i)}. \tag{3.15}$$

3.3 Results and discussion

To evaluate the performance of our proposed method we choose a challenging segmentation task: the contour of the femur in fluoroscopic (low-dose) X-ray images. The quality of this image modality is low and thus the performance of the standard ASM algorithm suffers. Our database contains 350 gray-scale images that have been acquired during surgeries treating hip fractures in an approximate anterior-posterior orientation. These images show the upper part of the femur and part of the hip. They belong to different surgical interventions and C-arm devices. The image sizes range between 450x450 to 510x510 pixels. We have a ground truth consisting of a manually segmented femur contour and landmarks for every image. We show one of these images in Fig. 3.4. As these images come from surgeries treating hip fractures, they contain an intermedullary nail and a screw, which further complicate the segmentation task with occlusions. More details about these images can be found in the Appendix.

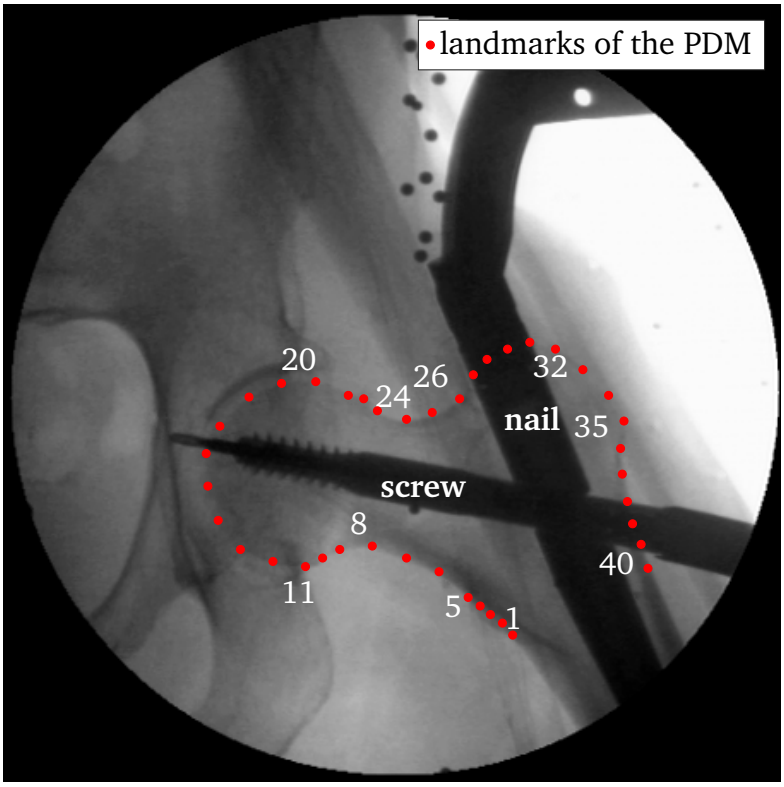


Figure 3.4: Landmarks of a PDM (1 to 40) of the femur in a fluroscopic X-ray image in anterior-posterior orientation. More information about the medical images source can be found in the Appendix.

3.3.1 Leave-one-out test

We evaluate the performance of our method using a leave-one-out test. For all the images in the database we “leave one image out" to test and keep the remaining as training images. The process is repeated for every available image and the results are averaged over all test images. For each leave-one-out test iteration t, we obtain $\hat{\mathbf{P}}_{(t)}$, $\hat{\mathbf{\Lambda}}_{(t)}$ and detector $\hat{\mathcal{T}}_{(t)}$ from the training images. To perform the chi-square test, for every landmark n we also obtain $\hat{\mu}_{\mathbf{g}_n(t)}$ and $\hat{\mathbf{S}}_{\mathbf{g}_n(t)}$. We determine the sample covariance of the residuals $\hat{\mathbf{R}}_{(t)}$ as described in Section 3.2.2.

We performed this leave-one-out test for all considered strategies. The same conditions (initial guess, shape parameters, number of ASM iterations...) were used for all, only changing the definition of the weighting matrix $\mathbf{W}$. For each strategy, we ran 100 ASM iterations on each image. We then measured the squared distance of every resulting landmark to the ground truth contour. We show the root mean square error (RMSE) for each PDM landmark in Fig. 3.5.

3.3.2 Evaluation of the proposed method

We implemented two different versions of our approach: one where we use the full matrix $\mathbf{W}$, and one where we restrict it to be diagonal. We compared these to the following strategies:

1. the standard ASM [14] without weighting, i.e., $\mathbf{W} = \mathbf{I}$.
2. $\mathbf{W}$ as a diagonal matrix with diagonal elements $1/\hat{d}_n$, as proposed by [73].
3. $\mathbf{W}$ as a diagonal matrix with diagonal elements $1/(1+\text{trace}(\mathbf{S}_{\mathbf{g}_n}))$, as proposed by [76].

As shown in Fig. 3.5, our strategy outperforms all the other strategies on average, in particular when $\mathbf{W}$ is diagonal, i.e., assuming independent residuals (this could be due to the small sample size available). The improvement our strategy provides over other strategies is especially significant in the area of the femoral head (between landmark 11 and 20, with $\sim$30% improvement achieved). This is the most challenging area of segmentation since there is overlap from other hip bones, and contour edges are particularly weak. Also, for the surgical procedure from which we obtained these images, this is an important region of interest [80].

As an example, we show in Fig. 3.6 the landmarks found after 100 ASM iterations for our proposed weighting strategy with diagonal $\mathbf{W}$ compared to the standard ASM, for a particular image in the database. Our strategy follows the real contour much more closely.

We also further investigated the hypothesis test described in Section 3.2.3: We measured the norm d_n after the simulation of several ASM iterations. Then we collected the observed $\hat{d}_n$ values either as $\hat{d}_n|f_n = 1$, if the true point was valid (null hypothesis), or $\hat{d}_n|f_n = 0$ otherwise. We computed normalized histograms of $\hat{d}_n$, as shown in Fig. 3.7, to see how well a chi-squared distribution fits $\hat{d}_n|f_n = 1$. The approximation is fairly good, although it does not account for a strong tail in the histogram of $\hat{d}_n|f_n = 1$. We set a probability of false alarm, i.e., the probability of classifying a valid landmark as invalid, of 10% based on the approximating χ^2_ℓ distribution. This covers $\sim$75% of the values from $\hat{d}_n|f_n = 1$, but only $\sim$30% from $\hat{d}_n|f_n = 0$. We suggest to set a relatively low false alarm threshold, which prevents the least-squares solution to be implausible. If too many landmarks are regarded as invalid and excluded, the problem may become underdetermined, and solutions may be unreliable.

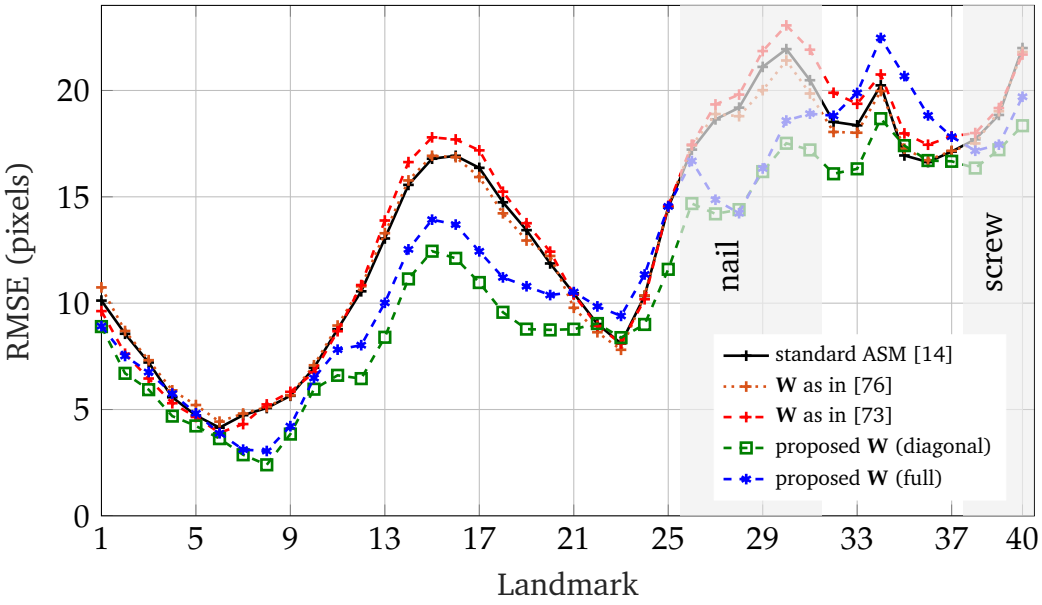


Figure 3.5: RMSE of our proposed strategy compared with [76], [73] and [14]. We mask in gray the areas where the surgical implants (nail and screw) most likely occlude the bone contour.

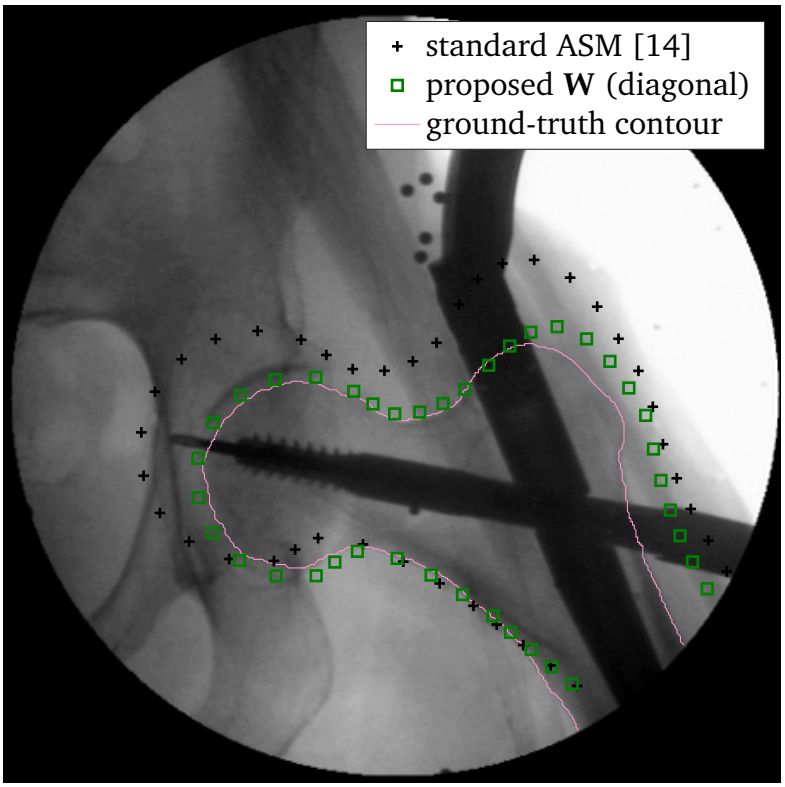


Figure 3.6: One example of a segmented femur, comparing our technique with the standard ASM algorithm (without weighting). More information about the medical images source can be found in the Appendix.

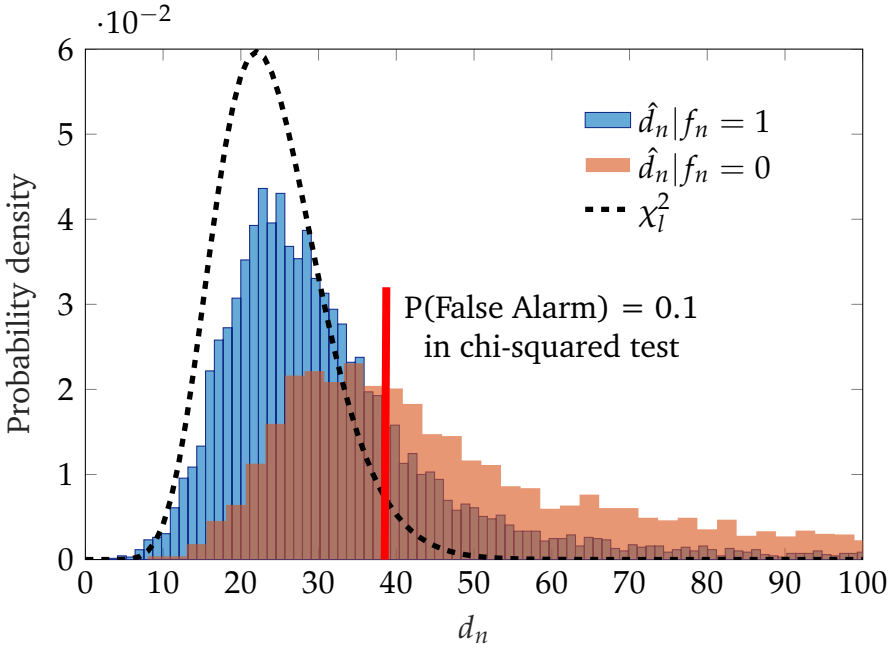


Figure 3.7: Normalized histograms of the distances $\hat{d}_n|f_n = 1$ and $\hat{d}_n|f_n = 0$ and the approximating chi-squared distribution for comparison.

Chapter 4
A model-order selection technique

"We consider it a good principle to explain the phenomena by the simplest hypothesis possible."

— William of Occam

As explained in Chapter 2, PDM contain statistical information of a collection of shape landmarks, represented by an affine space of eigenvectors obtained by an eigenvalue decomposition of the sample covariance matrix estimated from a collection of training data. We refer to model order as the number of eigenvectors that compose a PDM. Deciding the number of eigenvenctors is an important problem in Blind Source Separation (BSS) to select the number of underlying sources. The selection establishes a trade-off: if the order is too large, the model may not generalize well (overfitting, high variance); if the order is too small, the model may be too simplified and miss important variability (underfitting, high bias).

During the development of this thesis, we found potential in researching about the model order in PDM, since the most common way of choosing it is to keep the eigenvectors that account for a given percentage of variance [20]. This selection is only heuristic and does ignore the importance of the trade-off between overfitting and underfitting. This motivation from PDM inspired the development of a new model selection technique, which is applicable to more general BSS problems to enumerate the number of sources, for instance in array signal processing. In this proposed technique we have considered information theory, on the basis that *"the simplest explains the best*", as Occam's razor motivates.

In this chapter we present in detail the proposed new model-order selection strategy. In Section 4.1 we introduce the motivation and basics concepts about information theory to fully comprehend the technique. Section 4.2 contains the problem formulation, and the proposed solution. The results are discussed in Section 4.3. In Chapter 5 we will formulate the proposed technique for the particular case of shape models, providing a study within typical shape data.

The results of this chapter haven been published in:

- A. Eguizabal, C. Lameiro, D. Ramirez and P. Schreier, "Source enumeration in the presence of colored noise," *IEEE Signal Processing Letters*, vol. 26, no. 3, 2019.

4.1 Motivation and preliminaries

A collection of training shape vectors may be considered, for instance, as a sensor array processing problem: when building a PDM we obtain a set of observations such as $\mathbf{x} \in \mathbb{R}^N$ described as $\mathbf{x} = \mathbf{P}\mathbf{b}_q$ (without loss of generality let us assume $\mathbf{x}$ is zero-mean). $\mathbf{P}$ are the model eigenvectors, and the shape parameters are a set of underlying sources with lower dimension $\mathbf{b}_q \in \mathbb{R}^q$ with $q < N$. Actually, the vector of observations is modeled as such a superimposition of signals in many signal processing problems [81]. Array signal processing comprises many and varied applications, such as radar, sonar, multiple-input multiple-output (MIMO) wireless communications, and electroencephalography (EEG) [82]. In many cases, they share this mathematical framework: the observed data is modeled as a superposition of a finite number of independent sources that are embedded in additive noise. In this section we start motivating the need of a new strategy to select the model order in these problems. In the foregoing subsections we introduce the theoretical background of the proposed technique.

4.1.1 Source enumeration in array signal processing

One of the key processing steps for many of these array signal processing applications is to estimate the number of source signals [81]. These signals and the noise are often assumed to be random processes with certain statistical properties. As we have already motivated in the previous section, it is common to assume that the noise is spatially white, meaning that its covariance matrix is $\sigma^2\mathbf{I}$, and the noise eigenvalues are identical and equal to σ^2. Yet white noise is unrealistic in many applications. Noise can be filtered, and there are certain cases where it shows structure. For example, in undersea sonar, shipping noise is directional [82]. Also, often different sensors experience different noise power [83], and there typically exists stronger correlation between sensors that are close [84]. In PDM the noise may be due to imprecise training landmarks, presence of non-representative variability in the training set, insufficient samples, or quantization errors. Thus, the noise at different landmarks may be correlated and may have different variances. Furthermore, the prior Procrustes alignment typically introduces color in the noise.

When the noise is arbitrarily spatially correlated, the problem may become untractable [84]. However, if the noise is sufficiently weaker than the signal, detection is still possible. There are few approaches in the literature for the case of colored noise. Furthermore, in such approaches, certain structure in the noise covariance matrix is typically assumed. In [83] the authors considered non-uniform noise, which is still spatially uncorrelated. In [85] the noise model is colored, but two arrays of sensors that have uncorrelated noise between each other are assumed. Similarly in [84], the authors modeled the noise covariance matrix as block-diagonal and assumed to know the block size. The authors in [86] considered colored noise in their model, but required a different data set with noise-only observations.

Therefore, we design a new strategy to detect the number of sources, which does not require an independent noise-only observation set or uncorrelated arrays, and

which allows an unknown noise covariance matrix. The only requirement is that the noise is sufficiently weaker than the signal, which is also an implicit requirement of existing techniques.

4.1.2 Information theory: entropy, differential entropy and mutual information

Information theory, as it name suggests, considers a theoretical definition for the ambiguous term *information*. From a communication point of view, *information* could be the content of the messages sent through a channel. In the 1940s Shannon formulates the channel capacity [87] as a tight upper-bound of the reliable transmission rate, which is determined considering the *entropies* of the input and output of the channel in their *mutual information*. This concept of *entropy* to account for *information* is related to the one used in Physics for thermodynamics, as well as the one in Computer Science for the *entropy* of Kolmogorov complexity, and more generally in Probability Theory [88].

The entropy measures the average uncertainty of a random variable [88]. We model the observations as complex vectors $\mathbf{x} \in \mathbb{C}^N$. Therefore, we need to consider the entropy of a continuous random vector, i.e., differential entropy of a random vector. Let us define the entropy of the random vector $\mathbf{x}$, with probability density function $f(\mathbf{x})$, as

$$h(\mathbf{x}) = -\int_{\mathbb{X}} f(\mathbf{x}) \log f(\mathbf{x}) d\mathbf{x}, \tag{4.1}$$

where $\mathbb{X}$ is the support set of the random vector $\mathbf{x}$. Let us consider a second random vector $\mathbf{y} \in \mathbb{C}^N$ with probability density $g(\mathbf{y})$. The relative entropy (also known and Kullback-Leibler distance) between two density functions is defined as

$$D(f(\mathbf{x})||g(\mathbf{y})) = \int f(\mathbf{x}) \log \left(\frac{f(\mathbf{x})}{g(\mathbf{y})}\right) d\mathbf{x} d\mathbf{y}, \tag{4.2}$$

where the region of integration is the entire space. The mutual information $I(\mathbf{x}, \mathbf{y})$ between two random variables, in this case vectors, is the relative entropy between the joint distribution of $\mathbf{x}$ and $\mathbf{y}$, i.e., $f(\mathbf{x}, \mathbf{y})$, and the product $f(\mathbf{x})g(\mathbf{y})$

$$I(\mathbf{x}, \mathbf{y}) = D(f(\mathbf{x}, \mathbf{y})||f(\mathbf{x})g(\mathbf{y})) = \int f(\mathbf{x}, \mathbf{y}) \log \left(\frac{f(\mathbf{x}, \mathbf{y})}{f(\mathbf{x})g(\mathbf{y})}\right) d\mathbf{x} d\mathbf{y}. \tag{4.3}$$

4.1.3 Model selection

Model selection consists in choosing the best model from a set of different model candidates [19]. The goal of design of an SSM is not to learn an exact representation of the training data, but a statistical parametrization that generalizes to the unseen data as well [89]. To select the model order, that is, the number of parameters that determine the model, there is a compromise between a small training error and a good generalization. Typically, and especially in linear models, the higher

the complexity of the model, the smaller the training error, but also the worse the generalization. If the order is too large, the model may be too specific to the training data, and therefore not generalize well to unseen data (overfitting); if the order is too small, the model is too simple and the representation error is generally high (underfitting). We illustrate this effect in Fig. 4.1. This behavior reveals when the training error is decomposed in bias and variance.

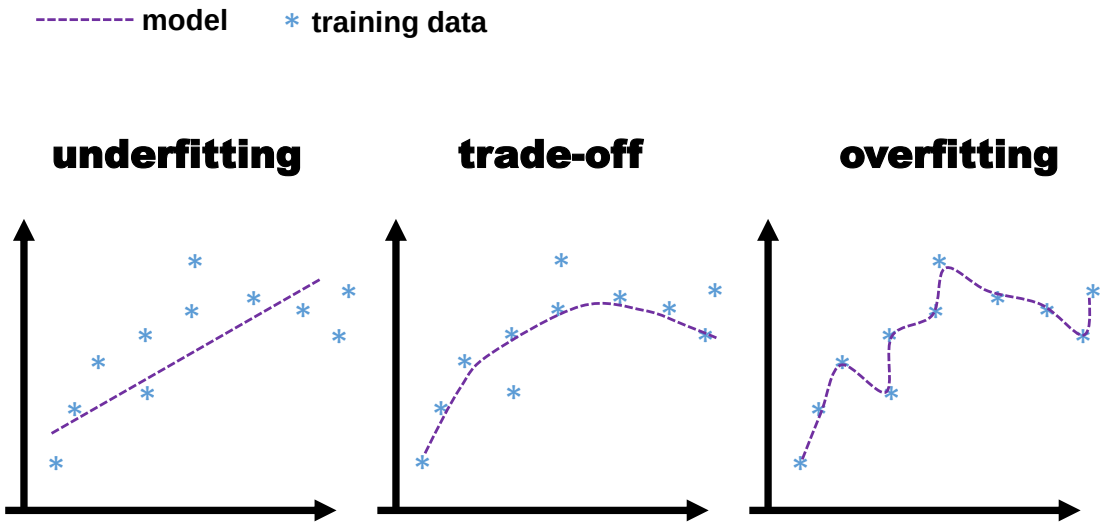


Figure 4.1: Illustrative example to the underfitting and overfitting effect in model order selection.

The bias-variance trade-off

As an optimization objective function to train the PDM we consider the mean squared error (MSE), which is a common optimization criterion in machine learning and estimation [68] [89]. The MSE measures the mean squared deviation from an estimator to the value to estimate. However, the MSE estimate tends to be biased to the particular training samples available for the model learning [89], and cannot be written as a function of only the training data [68]. To understand this problem, we rewrite the MSE as a combination of bias and variance. Let us consider $\hat{\mathbf{x}}$ as the estimate generated by a PDM for the unknown true shape $\mathbf{x}$. The MSE of the model, i.e., the estimator $\hat{\mathbf{x}}$, is then

$$\mathrm{MSE}(\hat{\mathbf{x}}) = \mathrm{E}\left[||\hat{\mathbf{x}} - \mathbf{x}||^2\right] = \underbrace{||\mathrm{E}\left[\hat{\mathbf{x}}\right] - \mathbf{x}||^2}_{\text{bias}^2} + \underbrace{\mathrm{E}\left[||\hat{\mathbf{x}} - \mathrm{E}[\hat{\mathbf{x}}]||^2\right]}_{\text{variance}}. \tag{4.4}$$

From Eq. (4.4), we observe that the MSE is a trade-off between bias and variance [89]. The bias quantifies the expectation of how much the model $\hat{\mathbf{x}}$ varies from the

true unobserved $\mathbf{x}$. The variance, on the other hand, is a metric of how much the model $\hat{\mathbf{x}}$ varies from its expected value $\mathrm{E}[\hat{\mathbf{x}}]$. If a model $\hat{\mathbf{x}}$ is too simple and does not fit accurately the training data, the bias will tend to be high, decreasing as the complexity increases. On the contrary, if the model is too complex and fits very closely the available training data, it may also be fitting the noise in the training set, as well as over-representing the limited samples available. Hence, the variance will be very high, since $\hat{\mathbf{x}}$ will deviate from its yet unknown expected value $\mathrm{E}[\hat{\mathbf{x}}]$, which depends on all possible observable data. This intuition is illustrated in Fig. 4.2. Nevertheless, in practice the expected values in the MSE can only be estimated from the training data, or assumed to follow certain model. We only have a training set containing a limited number of samples of the true shape, and the exact model of representation for $\mathbf{x}$ is unknown [90]. Consequently, calculating the value for the theoretical MSE is unrealizable [68], and this MSE trade-off is compromised by the training data.

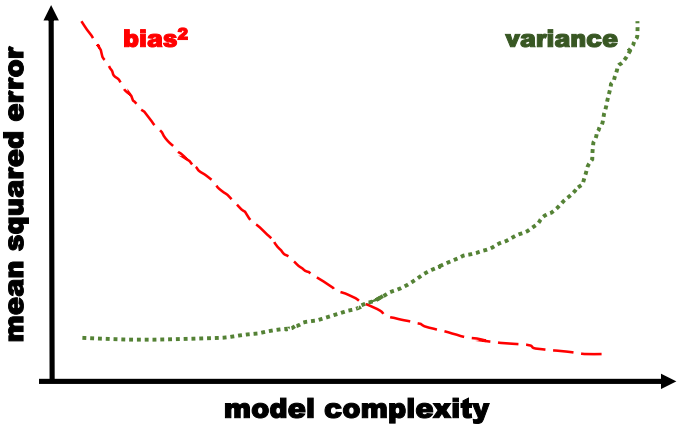


Figure 4.2: Illustrative example of the variability of the bias and the variance of the model with respect to its complexity.

Quality of fit and Maximum Likelihood parameter estimation

A PDM is a parametric model of the shape. Considering that we already know the matrix of eigenvectors $\mathbf{P}_q \in \mathbb{R}^{N\times q}$, let us recall the linear model

$$\mathbf{x} = \mathbf{P}_q \mathbf{b}_q + \boldsymbol{\epsilon}, \tag{4.5}$$

where $\mathbf{x} \in \mathbb{R}^N$, and, without loss of generality, the mean shape $\boldsymbol{\mu}_\mathbf{x}$ is a vector of zeros. The noise is Gaussian and white, i.e., $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. The parameters to estimate in the model are $\mathbf{b}_q \in \mathbb{R}^q$ (shape descriptor), σ^2 (noise variance), and the model order q. Let us define the parameter vector $\boldsymbol{\theta}_q = [\mathbf{b}_q \ \sigma^2]^T \in \mathbb{R}^{q+1}$. In order to select a model order q we need to evaluate the best model fit among all possible orders, i.e., consider the possible parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$. We can infer the likelihood of model q,

$L(\theta_q)$, from the probability distribution $p(\mathbf{x}) = \mathcal{N}(\mathbf{P}_q\mathbf{b}_q, \sigma^2\mathbf{I})$, as

$$L(\theta_q) = p(\mathbf{x}|\mathbf{b}_q, \sigma^2) = \prod_{n=1}^{N} p(x_n|\mathbf{b}_q,\ \sigma^2), \tag{4.6}$$

where, under the white Gaussian assumption, the likelihood reduces to the expression

$$L(\theta_q) = p(\mathbf{x}|\mathbf{b}_q, \sigma^2) = \frac{1}{(2\pi)^{N/2}(\sigma^2)^{N/2}} \exp\big(-\frac{||\mathbf{x} - \mathbf{P}_q\mathbf{b}_q||^2}{\sigma^2}\big). \tag{4.7}$$

As the logarithm (log) is a monotonically increasing function, we deduce from Eq. (4.7) that the parameter values that maximize the likelihood also minimize the expression

$$-2\log p(\mathbf{x}|\mathbf{b}_q, \sigma^2) = N\log 2\pi + N\log\sigma^2 + \frac{||\mathbf{x} - \mathbf{P}_q\mathbf{b}_q||^2}{\sigma^2}. \tag{4.8}$$

Since $\mathbf{b}_q$ is an unknown deterministic parameter, and the variance of the noise σ^2 is not given, we need to calculate their Maximum Likelihood (ML) estimates, $\hat{\mathbf{b}}_q$ and $\hat{\sigma}^2$, and evaluate

$$\hat{\mathbf{b}}_q = \arg\min_{\mathbf{b}_q\in\mathbb{B}} ||\mathbf{x} - \mathbf{P}_q\mathbf{b}_q||^2, \tag{4.9}$$

$$\hat{\sigma}^2 = \frac{1}{N}||\mathbf{x} - \mathbf{P}_q\hat{\mathbf{b}}_q||^2. \tag{4.10}$$

Then, the evaluated likelihood for θ_q, given the ML estimate, is

$$-2\log p(\mathbf{x}|\hat{\mathbf{b}}_q, \hat{\sigma}^2) = N\log\hat{\sigma}^2 + \text{constant}. \tag{4.11}$$

When we evaluate this likelihood for all possible model orders, with $q = 1, \ldots, N$, its quality of fit is

$$Q(q) = [\log p(\mathbf{x}|\hat{\mathbf{b}}_1, \hat{\sigma}^2), \ldots, \log p(\mathbf{x}|\hat{\mathbf{b}}_N, \hat{\sigma}^2)]^T, \tag{4.12}$$

and a selection of q that maximizes this quality, also maximizes the likelihood of the respecting parameters, as well as minimizes the least-squares error of fitS. This q selection is

$$q^\star = \arg\max Q(q). \tag{4.13}$$

However, to calculate the quality of fit $Q(q)$ we can only consider a limited training data, and thus $\log p(\mathbf{x}|\hat{\mathbf{b}}_q, \hat{\sigma}^2)$ is not good enough as indicator because it tends to overfit with increasing complexity (bigger q) [69]. We illustrate this effect in Fig. 4.3, where an example from our femoral shape dataset has been used to estimate the quality of fit $Q(q)$ with respect to the complexity (order q). We see that the error of fit, i.e., the negation of the quality of fit $-Q(q)$, is monotonically decreasing. Thus, the model that best represents the observed data (with higher likelihood as measured by $Q(q)$) is simply the more complex, which generally produces over-fitting.

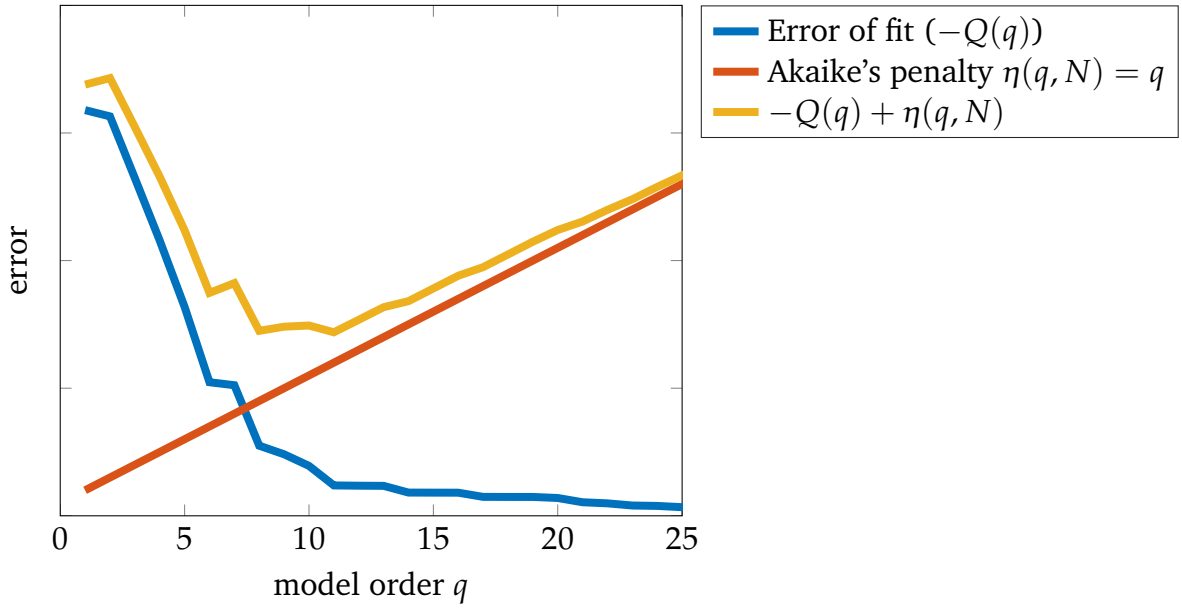


Figure 4.3: A practical example of the trade-off between likelihood of the parameters of the model given the training data, and the information-theoretic regularization. The results were obtained after computing the ML parameters estimates, as well as the likelihood expressions proposed in section 4.1.3, from samples from our femur data.

Information-theoretic criteria

Regularization is a possible technique to control over-fitting. It consists in adding a penalty term to the training error function to discourage very complex models [69]. Therefore, we need to choose a function of regularization to penalize high orders. The ML example just described is equivalent to a least-squares problem, since we estimate a frequentist parameter $\mathbf{b}_q$ in white Gaussian noise. Therefore,

$$\hat{\mathbf{b}}_q = \arg\max\left[\log p(\mathbf{x}|\mathbf{b}_q, \hat{\sigma}^2)\right] = \arg\min\left[||\mathbf{x} - \mathbf{P}_q\mathbf{b}_q||^2\right]. \tag{4.14}$$

Let us consider the regularized least-squares error

$$E_{\text{RLS}} = \underbrace{||\mathbf{x} - \mathbf{P}_q\mathbf{b}_q||^2}_{\text{training error}} + \underbrace{\gamma E(\mathbf{b}_q)}_{\text{regularization term}}, \tag{4.15}$$

where a typical selection for the penalty is $E(\mathbf{b}_q) = ||\hat{\mathbf{b}}_q||^2$ [69]. This regularization could imply heuristics to choose the penalty term and its weigh γ.

Unlike regularization, information theory provides a trade-off that does not require such subjective settings [81]. The analysis of the entropies and the statistical properties of the likelihoods provides a metric on how informative a model is. In 1970s Akaike [22] proposed a bias correction based on the mean differential entropy (the Kullback-Liebler distance, as defined in Eq. (4.2)) between the likelihood of the model, $\log p(\mathbf{x}|\mathbf{b}_q, \sigma^2)$, and its available estimate $\log p(\mathbf{x}|\hat{\mathbf{b}}_q, \hat{\sigma}^2)$ [22] [81]. Inspired

by Akaike's work, others studied information-theory based penalties, like Rissanen [91], who considered the minimum description length (MDL) principle to select the model that yields the minimum code length. Both derivations (Akaike's and Rissanen's) can be expressed as a compromise between the log-likelihood of the observed data and a penalty depending on the order q [92]. This can be seen as

$$q^* = \arg\min_q (\underbrace{-\log p(\mathbf{x}|\hat{\mathbf{b}}_q, \hat{\sigma}^2)}_{\text{likelihood term}} + \underbrace{\eta(q,N)}_{\text{penalty}}), \tag{4.16}$$

where q is the model order, and N is the number of observations used to compute the likelihood term. We illustrate in Fig. 4.3 the trade-off provided when the information-theoretic criteria is considered to correct the quality of fit obtained from the training data. The penalty term $\eta(q,N)$ is different depending on the contemplated information-theoretic criterion. For the example of this section, and considering the previously mentioned derivations, we obtain the penalties

$$\text{AIC: } \eta(q,N) = q \tag{4.17}$$

$$\text{BIC: } \eta(q,N) = \frac{q}{2}\log N, \tag{4.18}$$

where AIC stands for Akaike's Information Criteria, and BIC for Bayesian Information Criteria, which is equivalent to the MDL criteria described in by Rissanen [91]. Other information-theory-based model selection formulations have the same resulting derivation as in Eq. (4.16) [92]. Also, the same scheme is applicable to more complex models apart from this ordinary least-squares example.

Outlook of the information-theoretical model selection

Consequently, in order to perform a model order selection based on information theory, the following steps apply:

1. Choose a model to describe the training data. The order of such model is the open question.

2. Find an expression for the log-likelihood of the training data and the ML estimates of the model parameters.

3. Choose an order selection rule (AIC, BIC...) and verify that the assumptions made to derive its corresponding penalty term hold for the chosen model.

There is, however, not a closed rule to select the optimal penalty criterion. The best penalty choice will depend on the specific application [92]. The general assumptions made to derive the information-theoretic penalty terms are weak and usually hold [92]. AIC tends to asymptotically overestimate the model order, whereas BIC yields a consistent estimate [81].

4.1.4 The majorization-minimization optimization

The model under white Gaussian noise considered in the previous section is, in most cases, unrealistic. When more general noise models are assumed, the problem formulation becomes more complicated. Also, the fact that the parameters to estimate are constrained to a set $\mathbb{B}$ adds difficulty to calculate its ML estimates. Therefore, these estimates will not remain simply as a closed-form solution to an ordinary least-squares optimization. To deal with more complicated ML estimates we have considered the majorization-minimization (MM) algorithmic framework [93]. Closely related to the EM algorithm described in Chapter 2, this technique can be widely used to find minima in objective functions, and has applications in statistics, image processing and machine learning [93].

The MM framework consists in the following. Let us consider the optimization problem

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & f(\mathbf{x}), \\ \text{s.t.} \quad & \mathbf{x} \in \mathcal{X} \end{aligned} \tag{4.19}$$

where $\mathcal{X} \in \mathbb{R}^N$ in the feasible set, and $f(\mathbf{x})$ a continuous objective function $f: \mathcal{X} \rightarrow \mathbb{R}$. The process initializes at the starting point $\mathbf{x}_o \in \mathcal{X}$. The algorithm generates a sequence of feasible points $\mathbf{x}_t$, with $t \in \mathbb{N}$, iterating between the following steps:

1. **Majorization**: find a surrogate function $g(\mathbf{x})$ that locally around $\mathbf{x}_t$ approximates the objective function $f(\mathbf{x})$, and is also an upperbound, that is

$$g(\mathbf{x}|\mathbf{x}_t) \geq f(\mathbf{x}) + \mathbf{c}_t, \forall \mathbf{x} \in \mathcal{X}, \tag{4.20}$$

   with $\mathbf{c}_t = g(\mathbf{x}_t|\mathbf{x}_t) - f(\mathbf{x}_t)$ and the difference between $g(\mathbf{x}|\mathbf{x}_t)$, and $f(\mathbf{x})$ is minimum at $\mathbf{x}_t$.

2. **Minimization**: find the value $\mathbf{x}_{t+1}$ that minimizes the surrogate function $g(\mathbf{x}|\mathbf{x}_t)$, that is,

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x} \in \mathcal{X}} g(\mathbf{x}|\mathbf{x}_t). \tag{4.21}$$

In Fig. 4.4 we show two iteration steps of this process, considering the problem of minimizing the objective function $f(x)$.

The MM algorithm converges towards a stationary point of the original problem. If $f(\mathbf{x})$ is convex all stationary points of the problem are global optima. Therefore, MM converges to global optima for convex problems.

The surrogate function

There is no closed rule to choose the surrogate function in the MM process. Still, there are some very commonly used functions, such as the first and second order Taylor expansions of the objective function. A summary of the most frequent surrogate functions and their applications can be found in [93].

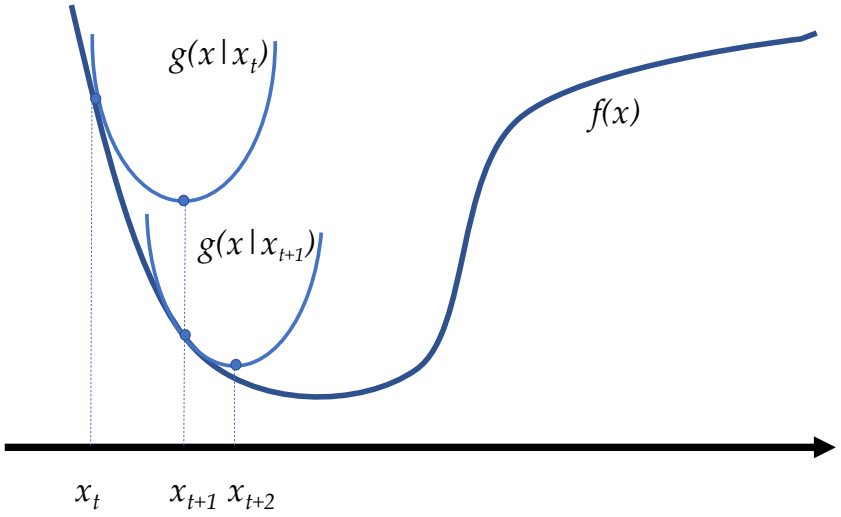


Figure 4.4: Illustrative example of the MM algorithm to minimize the scalar function $f(x)$ considering the surrogate $g(x|x_t)$, inspired by [93].

Let us we consider the second order Taylor expansion at $\mathbf{x}_t \in \mathcal{X}$ of the objective function $f(\mathbf{x})$, and that there exists a matrix $\mathbf{M} \succeq \nabla^2 f(\mathbf{x})$, $\forall \mathbf{x} \in \mathcal{X}$. Then, the following inequality holds

$$f(\mathbf{x}) \leq f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^T(\mathbf{x} - \mathbf{x}_t) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_t)^T\mathbf{M}(\mathbf{x} - \mathbf{x}_t). \tag{4.22}$$

In the previously described ML estimation, that is, minimizing the cost $f(\mathbf{x}) = ||\mathbf{x}||^2$ when the noise $\boldsymbol{\epsilon}$ is white, the problem is an ordinary least-squares fit. This becomes a generalized least-squares problem when the noise $\boldsymbol{\epsilon}$ is not white, i.e., $f(\mathbf{x}) = \mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}$ when $\boldsymbol{\epsilon} \in \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. When we consider a more general complex case, with $\mathbf{x} \in \mathbb{C}^N$, the cost function becomes $f(\mathbf{x}) = \mathbf{x}^H\boldsymbol{\Sigma}^{-1}\mathbf{x}$, where $\boldsymbol{\Sigma} = \mathrm{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^H]$ is a Hermitian matrix. Following Eq. (4.22), and by applying Wirtinger calculus for complex-valued matrix differentials [93], this complex valued function can be upperbounded as

$$\mathbf{x}^H\boldsymbol{\Sigma}^{-1}\mathbf{x} \leq \mathbf{x}^H\mathbf{M}\mathbf{x} + 2\mathrm{Re}[\mathbf{x}^H(\boldsymbol{\Sigma}^{-1} - \mathbf{M})\mathbf{x}_t] + \mathbf{x}_t^H(\mathbf{M} - \boldsymbol{\Sigma}^{-1})\mathbf{x}_t, \tag{4.23}$$

where the matrix $\mathbf{M}$ can have a desired structure to simplify the optimization process, for instance diagonal. Notice that the quadratic term of the bound only depends on $\mathbf{M}$ and not on $\boldsymbol{\Sigma}$. For further details about the bound derivations we refer to [93].

4.2 An information-theoretical approach

Let us now formulate the problem to deal with. Giving that the PDM can be consider as a particular case of an array processing problem, instead of N landmarks let us consider the more general case of an array of N sensors, and the observed N-dimensional vector $\mathbf{x}(m) \in \mathbb{C}^N$, which corresponds to the mth observation. We collect M observations according to model

$$\mathbf{x}(m) = \mathbf{A}\mathbf{s}(m) + \mathbf{n}(m), \tag{4.24}$$

where $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_q]$ is the array mixture matrix with full column rank, and $\mathbf{s}(m) = [s_1(m), \ldots, s_q(m)]^T$ is the vector of sources. The number of sources, q, is unknown, and we wish to determine it. These sources are modeled as unknown Gaussian random variables [81]. The q source signals are assumed to be zero-mean and independent. Typically, the noise vector $\mathbf{n}(m)$ is assumed to be independent of the sources, Gaussian and white, i.e., with covariance matrix $\sigma^2\mathbf{I}$. In that case the $N-q$ smallest eigenvalues of the covariance matrix $\mathbf{R}_{xx} = \mathrm{E}[\mathbf{x}\mathbf{x}^H]$ correspond to the noise subspace and are equal to σ^2. The separation of signal from noise is typically based on the assumption that the noise eigenvalues are identical [81], [94]. This does not work if the noise is not white.

Let us now assume the noise $\mathbf{n}(m)$ is colored, i.e., a zero-mean random vector with an arbitrary covariance matrix $\mathbf{\Sigma}$. If the source vector $\mathbf{s}(m)$ has covariance matrix $\mathbf{R}_{ss}$, the covariance matrix of the observed data $\mathbf{x}(m)$ is

$$\mathbf{R}_{xx} = \mathbf{A}\mathbf{R}_{ss}\mathbf{A}^H + \mathbf{\Sigma}. \tag{4.25}$$

If the signal-to-noise ratio (SNR) is sufficiently high, the signal subspace is at least approximately spanned by the q eigenvectors corresponding to the q largest eigenvalues of $\mathbf{R}_{xx}$. However, the problem is that, in order to separate signal from noise eigenvalues, the approaches in [81] and [94] need to assume that the $N-q$ smallest eigenvalues (due to noise) are equal. If this is not the case, they will fail. We now propose a strategy that can estimate q accurately under these conditions.

Inspired by statistical shape models, we propose an interpretation of the problem as a multivariate regression. In PDM, the eigenvectors of the sample covariance matrix obtained from the observed (training) data determine the regression parameters. Furthermore, in PDM there are constraints on the regression in order to ensure shape plausability, and that the values of the explained variables are sufficiently likely considering the examples seen in the training data. Therefore, we separate signal from noise by regarding everything that looks non-representative as noise.

4.2.1 Regression interpretation

We separate the observed data in two subsets of sizes M_1 and M_2, with $M_1 + M_2 = M$, in order to obtain M_1 training and M_2 test observations. We show that our method is fairly robust with respect to the precise split into M_1 and M_2, and choosing M_1 of similar size as M_2 works best. Thus, these subsets are $\mathbf{X}_1 =$

$\left[\mathbf{x}(1),\ldots,\mathbf{x}(M_1)\right] \in \mathbb{C}^{N\times M_1}$, and $\mathbf{X}_2 = \left[\mathbf{x}(M_1+1),\ldots,\mathbf{x}(M_1+M_2)\right] \in \mathbb{C}^{N\times M_2}$. Within the first set we obtain the matrix of regressors, for which we estimate the sample covariance matrix of the observed data, $\hat{\mathbf{R}}_{xx} = \frac{1}{M_1}\mathbf{X}_1\mathbf{X}_1^H$, and compute its eigenvalue decomposition as $\hat{\mathbf{R}}_{xx} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^H$. We also define $\boldsymbol{\lambda} = \left[\lambda_1,\ldots,\lambda_q,\ldots,\lambda_N\right]$ as the diagonal of $\boldsymbol{\Lambda}$, with eigenvalues in decreasing order. These will determine constraints on the regression. We assume the first q values in $\boldsymbol{\lambda}$ correspond to signal and the remaining $N-q$ to noise.

Let us assume the following interpretation of (4.24) for a fixed q:

$$\mathbf{x}(m) = \mathbf{P}_q\mathbf{b}_q(m) + \mathbf{e}(m), \tag{4.26}$$

where $\mathbf{P}_q$ is the matrix of regressors containing the eigenvectors corresponding to the q largest eigenvalues of $\hat{\mathbf{R}}_{xx}$, $\mathbf{b}_q(m) \in \mathbb{C}^q$ is a deterministic vector of regression parameters, and $\mathbf{e}(m) \sim \mathcal{CN}(\mathbf{0},\, \boldsymbol{\Sigma}_q)$ is the residual error. Let $\boldsymbol{\epsilon} = \left[\epsilon_1,\ldots,\epsilon_N\right]$ denote the eigenvalues of $\boldsymbol{\Sigma}_q$, arranged in descending order. In order to make sure that $\mathbf{x}(m)$ generated by (4.26) have statistics sufficiently similar to the observations from the first split $\mathbf{X}_1$, the following structure is assumed in the regression [95]:

1. An energy constraint is placed on the vector of unknown parameters $\mathbf{b}_q(m) = [b_1,\ldots,b_q]^T$, such that $|b_n|^2 \leq \lambda_n, \forall n = 1\ldots q$ (equivalently to shape plausibility constraint).

2. $\boldsymbol{\Sigma}_q$ is assumed to be full rank. In order to enforce this, its smallest eigenvalue ϵ_N is constrained to be $\epsilon_N \geq \lambda_N$.

Considering the samples in $\mathbf{X}_2$, we define the following multivariate linear regression:

$$\mathbf{X}_2 = \mathbf{P}_q\mathbf{B}_q + \mathbf{E}, \tag{4.27}$$

where $\mathbf{B}_q = \left[\mathbf{b}_q(1),\ldots,\mathbf{b}_q(M_2)\right] \in \mathbb{C}^{q\times M_2}$ is the matrix of unknown regression parameters, and the columns of the error matrix $\mathbf{E} = \left[\mathbf{e}(1),\ldots,\mathbf{e}(M_2)\right] \in \mathbb{C}^{N\times M_2}$ are samples of the residual error $\mathbf{e}(m)$.

The problem is therefore to determine the model order q of the regression in (4.27), which is equal to the number of source signals. In this interpretation, $\mathbf{P}_q$ is assumed to be a known matrix of regressors and $\mathbf{B}_q$ is modeled as a deterministic matrix of unknown parameters. In order to determine the best trade-off between over- and underfitting, we propose an information-theoretic criterion for multivariate regression [96], which we modify in order to include the constraints and the proposed colored noise model. An information-theoretic criterion is composed of a likelihood and a penalty term, and the best model order q^* is obtained as

$$q^* = \arg\min_q \Big(\underbrace{-\log p(\mathbf{X}_2|\hat{\mathbf{B}}_q, \hat{\boldsymbol{\Sigma}}_q)}_{\text{likelihood term}} + \underbrace{\eta(q)}_{\text{penalty}}\Big), \tag{4.28}$$

where $\hat{\mathbf{B}}_q$ and $\hat{\boldsymbol{\Sigma}}_q$ are the constrained Maximum Likelihood (ML) estimates of the model parameters for model order q, and $p(\cdot)$ stands for the likelihood function.

The penalty $\eta(q)$ depends on the selected criterion (Akaike, Bayesian, etc. [92]). We assume a frequentist interpretation of the regression. Thus, the only random variable in the regression is the residual error matrix $\mathbf{E} = \mathbf{X}_2 - \mathbf{P}_q\mathbf{B}_q$, which we have defined as samples of a complex Gaussian distribution with covariance matrix $\mathbf{\Sigma}_q$. Consequently, the log-likelihood expression in (4.28), ignoring the constant terms, becomes

$$\log p(\mathbf{X}_2|\hat{\mathbf{B}}_q, \hat{\mathbf{\Sigma}}_q) = -M_2 \log|\hat{\mathbf{\Sigma}}_q| - \mathrm{Tr}\{(\mathbf{X}_2 - \mathbf{P}_q\hat{\mathbf{B}}_q)^H \hat{\mathbf{\Sigma}}_q^{-1}(\mathbf{X}_2 - \mathbf{P}_q\hat{\mathbf{B}}_q)\}. \quad (4.29)$$

4.2.2 ML estimates of the model parameters

The ML estimates of parameters $\mathbf{\Sigma}_q$ and $\mathbf{B}_q$ are mutually dependent. Therefore we propose an alternating optimization, in which the optimal solution for each subproblem is obtained to ensure the convergence.

Estimate of $\mathbf{\Sigma}_q$

This is obtained from the expression

$$\hat{\mathbf{\Sigma}}_q = \underset{\mathbf{\Sigma}_q \succeq \lambda_N \mathbf{I}}{\arg\min} \Big[M_2 \log|\mathbf{\Sigma}_q| + \mathrm{Tr}\{(\mathbf{X}_2 - \mathbf{P}_q\hat{\mathbf{B}}_q)^H \mathbf{\Sigma}_q^{-1}(\mathbf{X}_2 - \mathbf{P}_q\hat{\mathbf{B}}_q)\}\Big], \quad (4.30)$$

where $\mathbf{\Sigma}_q \succeq \lambda_N\mathbf{I}$ denotes that $\mathbf{\Sigma}_q - \lambda_N\mathbf{I}$ is positive semi-definite. The optimal solution to (4.30) is derived from the unconstrained ML estimate of $\mathbf{\Sigma}_q$ [97], which is the sample covariance matrix, that is $\hat{\mathbf{\Sigma}}_q^u = \frac{1}{M_2}(\mathbf{X}_2 - \mathbf{P}_q\hat{\mathbf{B}}_q)(\mathbf{X}_2 - \mathbf{P}_q\hat{\mathbf{B}}_q)^H$. Let its eigenvalue decomposition be $\hat{\mathbf{\Sigma}}_q^u = \mathbf{V}\mathbf{\Psi}\mathbf{V}^H$. The eigenvalues in $\mathbf{\Psi}$ that are smaller than λ_N are replaced with λ_N. Based on results in [97], this leads to the solution

$$\hat{\mathbf{\Sigma}}_q = \mathbf{V}\tilde{\mathbf{\Psi}}\mathbf{V}^H, \quad [\tilde{\mathbf{\Psi}}]_{nn} = \max([\mathbf{\Psi}]_{nn}, \lambda_N), \quad \forall n = 1 \ldots N, \quad (4.31)$$

where $[\mathbf{\Psi}]_{nn}$ denotes the element on the diagonal of $\mathbf{\Psi}$. This $\hat{\mathbf{\Sigma}}_q$ is the optimal solution to (4.30).

Notice that $\mathbf{\Sigma}_q \succeq \lambda_N\mathbf{I} \Leftrightarrow \mathbf{\Sigma}_q^{-1} \preceq \lambda_N^{-1}\mathbf{I}$. Thus, the optimization problem in (4.30) is equivalent to

$$\underset{\mathbf{0} \preceq \tilde{\mathbf{\Sigma}}_q \preceq \lambda_N^{-1}\mathbf{I}}{\text{minimize}} \; -\log\left|\tilde{\mathbf{\Sigma}}_q\right| + \mathrm{Tr}\left(S\tilde{\mathbf{\Sigma}}_q\right), \quad (4.32)$$

where $\tilde{\mathbf{\Sigma}}_q = \mathbf{\Sigma}_q^{-1}$ and $S = M_2^{-1}(\mathbf{X}_2 - \mathbf{P}_q\hat{\mathbf{B}}_q)(\mathbf{X}_2 - \mathbf{P}_q\hat{\mathbf{B}}_q)^H$. The above problem is similar to the one considered in Lemma 3.2 of [97]. Following the proof provided in [97], we readily obtain

$$\tilde{\mathbf{\Sigma}}_q = \mathbf{V}\tilde{\mathbf{\Psi}}'\mathbf{V}^H, \; [\tilde{\mathbf{\Psi}}']_{nn} = \min\left\{([\mathbf{\Psi}]_{nn})^{-1}, \lambda_N^{-1}\right\}, \; \forall n, \quad (4.33)$$

where $\mathbf{V}$ and $\mathbf{\Psi}$ contain the eigenvectors and eigenvalues, respectively, of S. Therefore, the optimal solution of (4.30) is $\hat{\mathbf{\Sigma}}_q = \tilde{\mathbf{\Sigma}}_q^{-1}$, i.e., Eq. (4.31).

Estimate of $\mathbf{B}_q$

The constrained ML estimation of $\mathbf{B}_q$ is obtained by maximizing the likelihood given in (4.29), or, equivalently, solving the generalized least-squares problem

$$\hat{\mathbf{B}}_q = \underset{\mathbf{B}_q \in \mathbb{B}(\boldsymbol{\lambda})}{\arg\min} \mathrm{Tr}\{(\mathbf{X}_2 - \mathbf{P}_q\mathbf{B}_q)^H \hat{\boldsymbol{\Sigma}}_q^{-1}(\mathbf{X}_2 - \mathbf{P}_q\mathbf{B}_q)\}, \tag{4.34}$$

where $\mathbb{B}(\boldsymbol{\lambda}) = \{[\mathbf{y}(1),\ldots,\mathbf{y}(M_2)] \in \mathbb{C}^{q\times M_2} : |y_n(m)|^2 \leq \lambda_n, \forall n = 1\ldots q, \forall m = 1\ldots M_2\}$. Even though the foregoing problem is convex and hence its optimal solution can be found by standard numerical methods [98], the computational cost is expensive since it does not admit a closed-form solution. As an alternative, we propose a more focused algorithm based on majorization-minimization (MM) [93] that is faster and converges to the optimal solution. Considering the least-squares form in (4.34), and $\mathbf{E} = \mathbf{X}_2 - \mathbf{P}_q\mathbf{B}_q$, we construct the surrogate function $g(\mathbf{E}|\mathbf{E}_{(t)})$ that is derived from the second-order Taylor expansion of (4.34). Thus, the cost function in (4.34) is upper-bounded by $g(\mathbf{E}|\mathbf{E}_{(t)})$, as introduced in Section 4.1 and derived in [93],

$$\begin{aligned} g(\mathbf{E}|\mathbf{E}_{(t)}) = \mathrm{Tr}\{\mathbf{E}^H\boldsymbol{\Sigma}_{\min}^{-1}\mathbf{E} + 2\mathrm{Re}[\mathbf{E}(\hat{\boldsymbol{\Sigma}}_q^{-1} - \boldsymbol{\Sigma}_{\min}^{-1})\mathbf{E}_{(t)}] \\ + \mathbf{E}_{(t)}^H(\boldsymbol{\Sigma}_{\min}^{-1} - \hat{\boldsymbol{\Sigma}}_q^{-1})\mathbf{E}_{(t)}\}, \end{aligned} \tag{4.35}$$

where $\boldsymbol{\Sigma}_{\min}^{-1} = \frac{1}{\epsilon_N}\mathbf{I}$. The original cost function in (4.34) is equal to (4.35) only when $\mathbf{E} = \mathbf{E}_{(t)}$. At each iteration of the MM algorithm, the feasible points $\mathbf{E}_{(t)}$ are updated as

$$\mathbf{E}_{(t+1)} = \underset{\mathbf{B}_q \in \mathbb{B}(\boldsymbol{\lambda})}{\arg\min} \quad g(\mathbf{E}|\mathbf{E}_{(t)}). \tag{4.36}$$

This surrogate objective function, $g(\mathbf{E}|\mathbf{E}_{(t)})$, is convex since $\boldsymbol{\Sigma}_{\min}$ is positive definite. Additionally, the feasible set $\mathbb{B}(\boldsymbol{\lambda})$ is also convex, which makes (4.36) a convex optimization problem. Furthermore, it satisfies Slater's condition, i.e., the feasible set has a non-empty interior, which implies that the Karush-Kuhn-Tucker (KKT) conditions are necessary and sufficient for optimality [98]. Hence, we start with the Lagrangian function, which is given by

$$\mathcal{L} = g(\mathbf{E}|\mathbf{E}_{(t)}) + \sum_{m=1}^{M_2} \mathrm{Tr}\left[\boldsymbol{\Phi}_m\left(\mathbf{b}_q(m)\mathbf{b}_q(m)^H - \tilde{\boldsymbol{\lambda}}\tilde{\boldsymbol{\lambda}}^H\right)\right], \tag{4.37}$$

where $\tilde{\boldsymbol{\lambda}} = [\sqrt{\lambda_1},\ldots,\sqrt{\lambda_q}]^T$, and $\boldsymbol{\Phi}_m$ is a diagonal matrix with diagonal $\boldsymbol{\phi}(m)$, whose nth element $[\boldsymbol{\phi}(m)]_n$ is the Lagrange multiplier associated with the constraint $|[\mathbf{b}_q(m)]_n|^2 \leq \lambda_n$. The derivative of the Lagrangian with respect to $\mathbf{b}_q^*(m)$ can be shown to be given by

$$\frac{\partial\mathcal{L}}{\partial\mathbf{b}_q^*(m)} = -\mathbf{P}_q^H\boldsymbol{\Sigma}_{\min}^{-1}\left(\mathbf{x}_2(m) - \mathbf{P}_q\mathbf{b}_q(m)\right) - \mathbf{P}_q^H\mathbf{v}(m) + \boldsymbol{\Phi}_m\mathbf{b}_q(m), \tag{4.38}$$

where $\mathbf{v}(m)$ and $\mathbf{x}_2(m)$ are the mth columns of $(\hat{\boldsymbol{\Sigma}}_q^{-1} - \boldsymbol{\Sigma}_{\min}^{-1})\mathbf{E}_{(t)}$ and $\mathbf{X}_2$, respectively. Equating (4.38) to zero yields

$$\mathbf{b}_q(m) = \left(\mathbf{P}_q^H \boldsymbol{\Sigma}_{\min}^{-1} \mathbf{P}_q + \boldsymbol{\Phi}_m\right)^{-1} \mathbf{P}_q^H \left(\boldsymbol{\Sigma}_{\min}^{-1} \mathbf{x}_2(m) + \mathbf{v}(m)\right). \tag{4.39}$$

Furthermore, since $\boldsymbol{\Sigma}_{\min}^{-1} = \frac{1}{\epsilon_N}\mathbf{I}$ and $\mathbf{P}_q^H \mathbf{P}_q = \mathbf{I}$, (4.39) can be further simplified to

$$\mathbf{b}_q(m) = \left(\frac{1}{\epsilon_N}\mathbf{I} + \boldsymbol{\Phi}_m\right)^{-1} \mathbf{P}_q^H \left(\frac{1}{\epsilon_N}\mathbf{x}_2(m) + \mathbf{v}(m)\right). \tag{4.40}$$

The foregoing expression is the optimal value of the mth column of $\mathbf{B}_q$ in terms of the Lagrange multipliers. Now we observe the following. First, since $\boldsymbol{\Phi}_m$ is a diagonal matrix, we have that $(\frac{1}{\epsilon_N}\mathbf{I} + \boldsymbol{\Phi}_m)^{-1}$ is also a diagonal matrix with elements

$$\left[\left(\frac{1}{\epsilon_N}\mathbf{I} + \boldsymbol{\Phi}_m\right)^{-1}\right]_{nn} = \frac{\epsilon_N}{[\boldsymbol{\phi}(m)]_n \epsilon_N + 1}, \tag{4.41}$$

where $[\cdot]_{nn}$ indicates the nth element along the diagonal and $[\boldsymbol{\phi}(m)]_n$ is the nth element of $\boldsymbol{\phi}(m)$. Thus, the Lagrange multiplier $[\boldsymbol{\phi}(m)]_n$ only affects the nth element of $\mathbf{b}_q(m)$. Second, by the KKT conditions $[\boldsymbol{\phi}(m)]_n \neq 0$ when $|[\mathbf{b}_q(m)]_n|^2 = \lambda_n$, and $[\boldsymbol{\phi}(m)]_n = 0$ otherwise. Therefore, if the nth element of the unconstrained solution (obtained by taking $\boldsymbol{\Phi}_m = \mathbf{0}$ in (4.40)) fulfills $|[\mathbf{b}_q(m)]_n|^2 \leq \lambda_n$, this is also the optimal solution of the original problem. Otherwise, the nth element of the unconstrained solution has to be scaled such that $|[\mathbf{b}_q(m)]_n|^2 = \lambda_n$.

Through these conditions it is shown, that the optimal solution to (4.36) can be obtained by scaling the unconstrained solution. This unconstrained solution can be obtained straightforwardly as

$$\begin{aligned}\mathbf{B}_{q(t+1)}^u &= \epsilon_N \mathbf{P}_q^H \big(\boldsymbol{\Sigma}_{\min}^{-1}\mathbf{X}_2 + (\hat{\boldsymbol{\Sigma}}_q^{-1} - \boldsymbol{\Sigma}_{\min}^{-1})\mathbf{E}_{(t)}\big) \\ &= \epsilon_N \big(\mathbf{P}_q^H \hat{\boldsymbol{\Sigma}}_q^{-1}\mathbf{X}_2 - \mathbf{P}_q^H (\boldsymbol{\Sigma}_{\min}^{-1} - \hat{\boldsymbol{\Sigma}}_q^{-1})\mathbf{P}_q \mathbf{B}_{q(t)}\big).\end{aligned} \tag{4.42}$$

The optimal solution $\mathbf{B}_{q(t+1)}$ is therefore obtained by scaling each element of $\mathbf{B}_{q(t+1)}^u$ such that the constraints are fulfilled. That is, each column $\mathbf{b}_{q(t+1)}(m)$ of $\mathbf{B}_{q(t+1)}$ is

$$\mathbf{b}_{q(t+1)}(m) = \tilde{\mathbf{b}}, \quad \tilde{b}_n = \frac{b_n^u(m)}{|b_n^u(m)|} \times \min(|b_n^u(m)|, \sqrt{\lambda_n}) \qquad \forall n = 1, \ldots, q \quad \forall m = 1, \ldots, M_2, \tag{4.43}$$

where $\tilde{b}_n$ is the nth element of $\tilde{\mathbf{b}}$, and $b_n^u(m)$ is the nth element of the mth column of $\mathbf{B}_{q(t+1)}^u$. The resulting ML estimate $\hat{\mathbf{B}}_q$ is then given by the value of $\mathbf{B}_{q(t+1)}$ upon convergence of the MM algorithm.

Algorithm 4.1 Alternating optimization to obtained constrained ML estimates of $\mathbf{B}_q$ and $\mathbf{\Sigma}_q$.

Data: Inputs are $\mathbf{P}_q$, $\boldsymbol{\lambda}$, $\mathbf{X}_2$, $i_{\max}$ and $c_{\min}$
Result: $\hat{\mathbf{B}}_q$ and $\hat{\mathbf{\Sigma}}_q$
initialization $i = 1$, $\hat{\mathbf{\Sigma}}_q^{(0)} = \mathbf{I}$, $c > c_{\min}$
while $c > c_{min}$ **do**

1. Obtain $\hat{\mathbf{B}}_q^{(i)}$ as in (4.34) with MM.
2. Obtain $\hat{\mathbf{\Sigma}}_q^{(i)}$ as in (4.30) and (4.31).
3. $c = ||\hat{\mathbf{\Sigma}}_q^{(i)} - \hat{\mathbf{\Sigma}}_q^{(i-1)}||_F$, $i = i + 1$.

end

4.2.3 Determining the model order

As shown in (4.30) and (4.34), the ML estimates $\hat{\mathbf{B}}_q$ and $\hat{\mathbf{\Sigma}}_q$ are mutually dependent, which prohibits finding a closed-form solution. In this section we have proposed an alternating optimization algorithm (see Algorithm 4.1) to maximize the log-likelihood in (4.28). Since $\hat{\mathbf{B}}_q$ and $\hat{\mathbf{\Sigma}}_q$ are obtained optimally, the algorithm is guaranteed to converge.

The procedure to estimate q is as follows. For every possible model order q (with $q = 1, \ldots, \min(N, M_1) - 1$), we compute the ML estimates $\hat{\mathbf{B}}_q$ and $\hat{\mathbf{\Sigma}}_q$ by means of Algorithm 4.1. Then, following (4.28), we estimate the model order as

$$q^* = \arg\min_q \left[M_2 \log |\hat{\mathbf{\Sigma}}_q| + \mathrm{Tr}(\hat{\mathbf{E}}^H \hat{\mathbf{\Sigma}}_q^{-1} \hat{\mathbf{E}}) + \eta(q) \right], \tag{4.44}$$

where $\hat{\mathbf{E}} = \mathbf{X}_2 - \mathbf{P}_q \hat{\mathbf{B}}_q$.

Sample-poor case

Sometimes we have access to only a small number M of observed training vectors relative to the number N of landmarks (or sensors). Under these circumstances, the covariance matrix of residuals in (4.30) may be ill-conditioned. In order to address this situation and reduce the number of parameters to estimate in (4.30), we propose the following solution. We model $\mathbf{\Sigma}_q$ as block-diagonal (even though it has an arbitrary structure) with blocks of size $s \times s$. This results in $d = \lceil \frac{N}{s} \rceil$ blocks, that is $\mathbf{\Sigma}_q = \mathrm{blkdiag}\{\mathbf{\Sigma}_q^1, \cdots, \mathbf{\Sigma}_q^d\}$, where the size of the last block $\mathbf{\Sigma}_q^d$ differs from the other blocks if $\frac{N}{s}$ is not an integer. The size s of the blocks gives a trade-off between a well-conditioned and a biased estimate of the true $\mathbf{\Sigma}_q$. Considering this, the ML estimate $\hat{\mathbf{\Sigma}}_q$ is simply the block-diagonal version of the matrix defined in (4.30). We

reformulate (4.44) such that the model order q^* is obtained as

$$q^*_{\text{sp}} = \arg\min_q \Big[\sum_{n=1}^{d} \left(M_2 \log(|\hat{\mathbf{\Sigma}}^n_q|) + \text{Tr}(\hat{\mathbf{E}}^H (\hat{\mathbf{\Sigma}}^n_q)^{-1} \hat{\mathbf{E}}) \right) + \eta_{\text{sp}}(q) \Big], \tag{4.45}$$

where the subscript of q^*_{sp} stands for sample poor.

4.3 Results and discussion

In this section we validate our strategy with numerical results. We have considered simulated data, where the model order is controlled and hence there is a ground-truth to validate the results.

4.3.1 Simulation settings

We consider the linear model in (4.24) to generate simulated observations. We assume the matrix $\mathbf{A}$ to be unitary (in order to control the SNR) and choose it as uniformly distributed on the unitary group [99]. We fix the number of sources q at a known value. We generate random values uniformly distributed between 1 and 10 as the signal eigenvalues in $\mathbf{R}_{ss}$. First, we evaluate the technique when the number of samples M is hight. Then, we propose relatively sample-poor scenarios, so our estimate in the presented simulations is based on q^*_{sp} in (4.45) with block size $s = 2$. Consequently, the number of parameters to estimate in $\mathbf{\Sigma}_q$ is reduced, and its estimate is less biased than a diagonal estimate. We choose the Akaike information criterion [22], as in [94]. Therefore $\eta(q) = 2M_2 q + \frac{N(N-1)}{2}$, and for the sample-poor case $\eta_{\text{sp}}(q) = 2M_2 q + \frac{3}{2}N$. $\eta(q)$ and $\eta_{\text{sp}}(q)$ are the number of degrees of freedom in the respective likelihood expressions. To simulate the noise, we randomly choose $\mathbf{U}$ and $\mathbf{\Omega}$, from which we compute $\mathbf{R}_{nn} = N\frac{\mathbf{U}\mathbf{\Omega}\mathbf{U}^H}{\text{Tr}\{\mathbf{\Omega}\}}$. The trace of $\mathbf{R}_{nn}$ is N, $\mathbf{U}$ is uniformly distributed on the unitary group, and $\mathbf{\Omega}$ is a diagonal matrix of N eigenvalues uniformly distributed between 0.1 and 1. We consider different structures of covariance matrices $\mathbf{\Sigma}$ to address four models of noise: white noise, $\mathbf{\Sigma}_{\text{s1}} = \mathbf{I}$; a mixture of white and colored noise, $\mathbf{\Sigma}_{\text{s2}} = \mathbf{I} + \mathbf{R}_{nn}$; colored noise, $\mathbf{\Sigma}_{\text{s3}} = \mathbf{R}_{nn}$; and colored noise with block-diagonal structure, $\mathbf{\Sigma}_{\text{s4}}$, whose blocks are equal to those along the block-diagonal of $\mathbf{R}_{nn}$. Finally, we scale the noise observations in order to accomplish the SNR settings. We define the SNR by a parameter β that is the ratio of the smallest signal eigenvalue in $\mathbf{R}_{ss}$ and the largest noise eigenvalue in $\mathbf{\Sigma}$.

Comparative strategies

We compare our proposed strategy with four others:

1. [81], the classical approach that considers additive white Gaussian noise.
2. [94], which assumes additive white Gaussian noise and a sample-poor scenarios

3. [83], which considers non-uniform, i.e., spatially uncorrelated noise with diagonal (but not identity) covariance matrix; and

4. [84], which allows colored noise, but assumes that the noise is block-diagonal with known block size.

4.3.2 Splitting data into training and test data

We present an empirical study of the impact of sizes M_1 and M_2 to show that the precise split of data into training and test data does not strongly affect the results of the estimate in the observed scenarios. As we can see in Fig. 4.5, the precise split was not critical, thus there is no good reason for deviating much from an even split with $M_1 = M_2$. This choice can be further justified by looking at how many parameters we need to estimate. Within $\mathbf{X}_1$, the calculation of $\mathbf{P}_q$ and $\boldsymbol{\lambda}$ involves $q(2N-q)+1$ parameters [81], and the expression $\log p(\mathbf{X}|\hat{\mathbf{B}}_q, \hat{\boldsymbol{\Sigma}}_q)$, in (4.34), has $2(Mq) + \frac{(s+1)}{2}N$ degrees of freedom. In many scenarios, M and N are values of similar order, and q is not very high. Thus, the degrees of freedom are similar for both training and test, so a similar size for both sets seems reasonable. Therefore, in further simulations we chose $M_1 = M_2$.

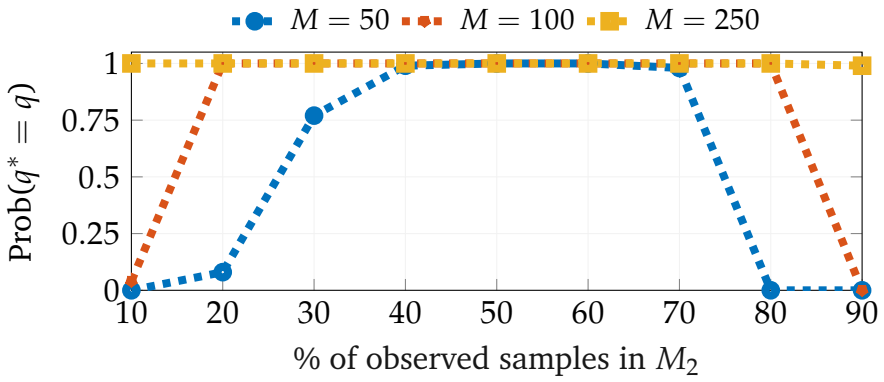


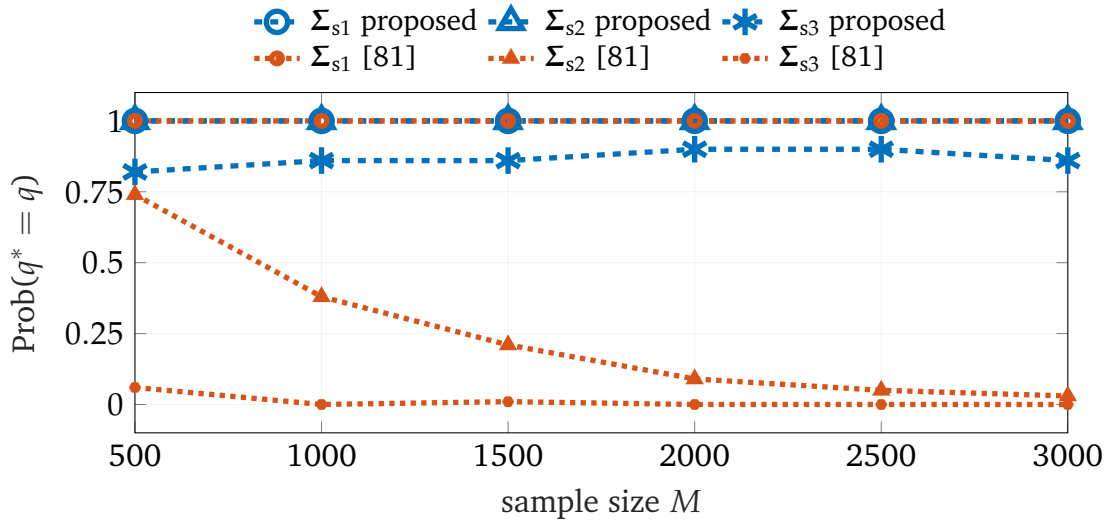
Figure 4.5: Probability of correctly choosing the model order q. The settings are $\beta = 10$ dB, $N = 50$ sensors, $q = 10$ sources, and simulated noise follows the covariance matrix $\boldsymbol{\Sigma}_{s2} = \mathbf{I} + \mathbf{R}_{nn}$. We averaged 100 simulations.

4.3.3 Comparison with a classical white-noise based approach

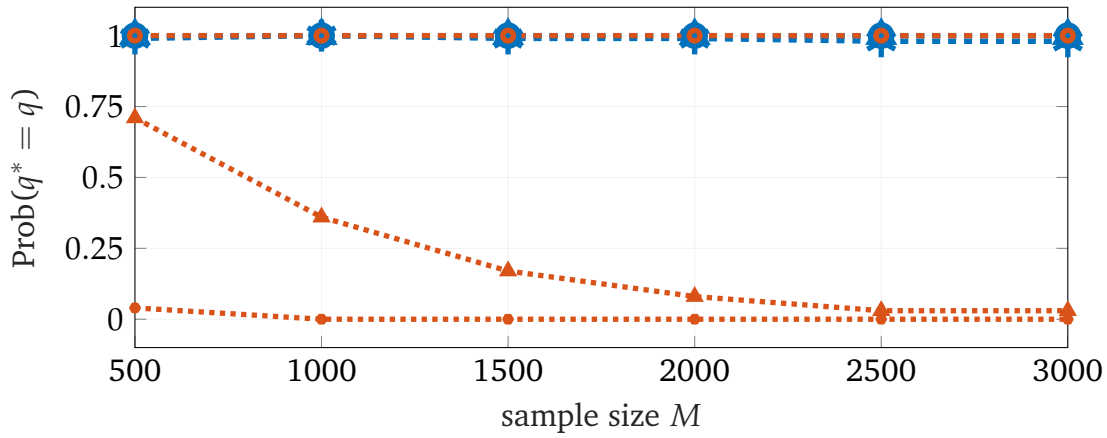
The classical approach in [81] considers white noise in the model. Therefore, its performance is very poor in our simulations when the noise differs white. In Fig. 4.1, we see that, in a colored-noise scenario our strategy maintains a much better performance when the noise is relatively high ($\beta = 5$ dB), and it almost does not alter with respect to a white-noise scenario when the SNR improves ($\beta = 10$ dB). The chassical strategy in [81] deteriorates much more as the number of observations M increases. The problematic is not experiences by our proposed strategy.

4.3.4 Sample-poor case

In Fig. 4.7 we illustrate the performance with respect to the number of samples for a fixed SNR parameter β, as well as with respect to β for a fixed number of samples. In Figs. 4.7(a) and 4.7(b) the noise structures are $\mathbf{\Sigma}_{s1}$ (white noise, optimal conditions for the strategy in [94]) and $\mathbf{\Sigma}_{s2}$. In Figs. 4.7(c) and 4.7(d) we consider $\mathbf{\Sigma}_{s3}$ (noise covariance without structure) and $\mathbf{\Sigma}_{s4}$ (block-diagonal, which constitutes the optimal conditions for the strategy in [84]; we choose block size $\frac{N}{2}$). In Figs. 4.7(a) and 4.7(b) we see that the strategy based on white noise [94] fails when the noise is colored. While having similar performance to our strategy when the noise is white, in the case of colored noise its performance is poor even when the noise is considerably weaker than the signal. Similarly, in Figs. 4.7(c) and 4.7(d), the block-diagonal strategy [84] fails when the covariance matrix of the noise is not block-diagonal. Even when the noise is block-diagonal, unlike our strategy, [84] needs to know the actual block-size. It also requires more samples and larger β for a performance similar to our proposed technique. The strategy [83], which assumes non-uniform noise, delivers good results for large β and large number of samples. However, our strategy requires fewer samples and smaller β to achieve the same probability of correct model-order choice.

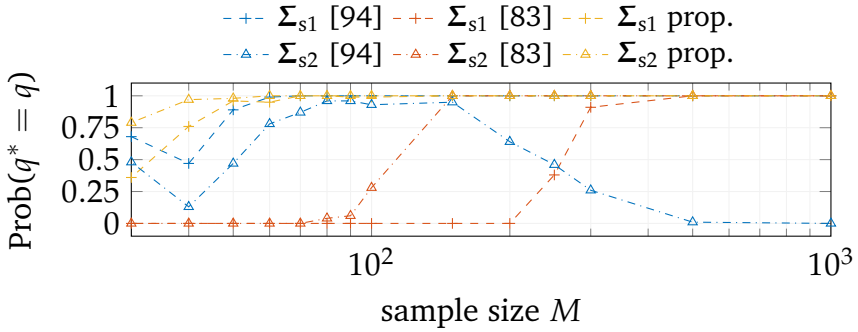


(a) $\beta = 5$ dB

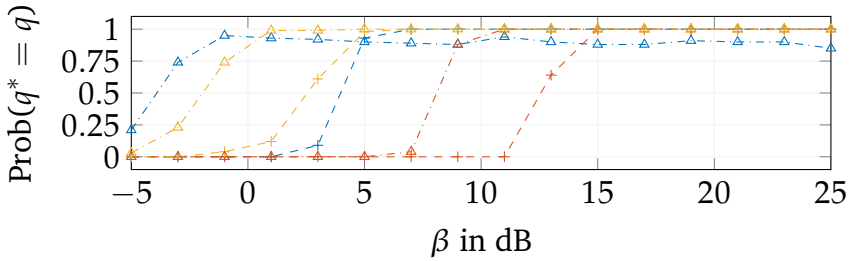


(b) $\beta = 10$ dB

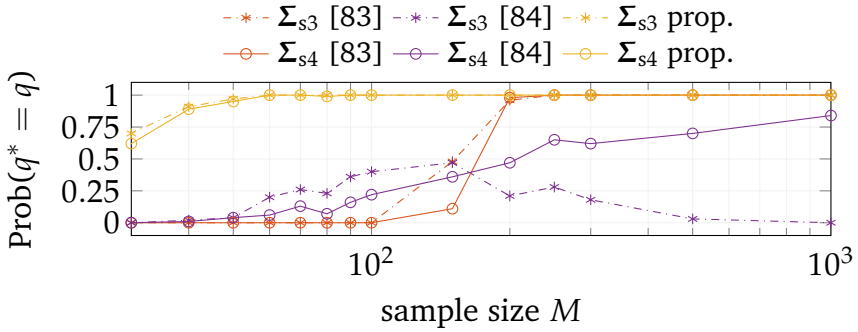
Figure 4.6: Comparison of the performance with respect to the number of samples M of the proposed (prop.) estimator with the classical one in [81], depicted in different colors (prop. in blue, and [81] in red]. We show results averaged over 100 simulations with parameter setting $N = 10$ and $q = 3$. In (a) $\beta = 5$ dB. In (b) $\beta = 10$ dB. We have generated three different noise structures.



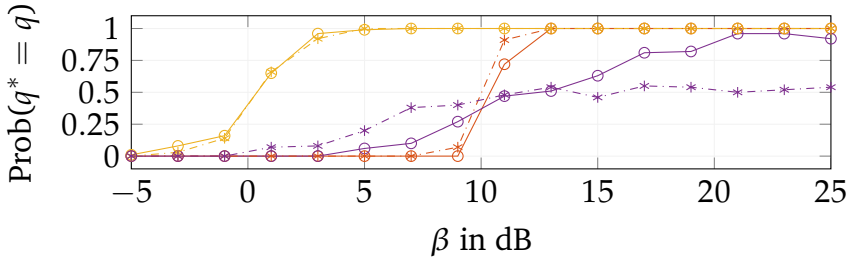
(a) Sample size variation with noise covariance matrices $\boldsymbol{\Sigma}_{s1}$ and $\boldsymbol{\Sigma}_{s2}$.



(b) β variation with noise covariance matrices $\boldsymbol{\Sigma}_{s1}$ and $\boldsymbol{\Sigma}_{s2}$.



(c) Sample size variation with noise covariance matrices $\boldsymbol{\Sigma}_{s3}$ and $\boldsymbol{\Sigma}_{s4}$.



(d) β variation with noise covariance matrices $\boldsymbol{\Sigma}_{s3}$ and $\boldsymbol{\Sigma}_{s4}$.

Figure 4.7: Comparison of the performance of the proposed (prop.) estimator with the ones in [94], [83], and [84], depicted in different colors. We show results averaged over 100 simulations with parameter setting $N = 50$ and $q = 10$, emulating the typical shape model scenario. In (a) and (c): $\beta = 10$ dB and we vary M. In (b) and (d): $M = 150$ and we vary β. We generate four different noise structures.

Chapter 5
Model-order selection in statistical shape models

"One accurate measurement is worth a thousand expert opinions."
— Grace Hopper

As we have already motivated in Chapter 4, SSM require choosing a model order, which determines how much of the variation seen in the training data is accounted for by the SSM. A good choice of this order also depends on the number of training samples and the noise level in the training data set. Yet the most common approach for choosing the model order in SSM simply keeps a predetermined percentage of the total shape variation of the training set. This is a heuristic design that undervalues of the potential of the models.

In this chapter, we particularize the technique proposed in Chapter 4 for SSM. We empirically show that the model order chosen by our proposed technique provides a better trade-off between over- and underfitting than the existing techniques. We start describing the state-of-the-art of model selection in PDM in Section 5.1, and motivate the need of model selection in shape analysis. Then, in Section 5.2 we consider the model order strategy proposed in Chapter 4 applied to PDM. In Section 5.3, we discuss results over artificial and real shape data, and the importance of model selection in shape analysis.

The results of this chapter haven been published in:

- A. Eguizabal, P. Schreier and D. Ramirez, "Model-order selection in Statistical Shape Models," in *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Aalborg, Denmark, Sept. 2018, pp. 475-479.

5.1 Motivation and preliminaries

The most common way of choosing the model order of a PDM is to keep the eigenvectors that account for a given percentage of variance (typically 90-98% [20]). This

consists simply in calculating the accumulative sum of the eigenvalues of the PDM of N landmarks, i.e., $\boldsymbol{\lambda}_N = [\lambda_1, \ldots, \lambda_N]$, an choose the order q such that

$$\frac{\sum_{n=1}^{q} \lambda_n}{\sum_{n=1}^{N} \lambda_n} < 90\%, \tag{5.1}$$

where in this case the percentage is 90%. We illustrate this heuristic selection in Fig. 5.1, where we have calculated the expression in (5.1) considering the PDM of the proximal femur.

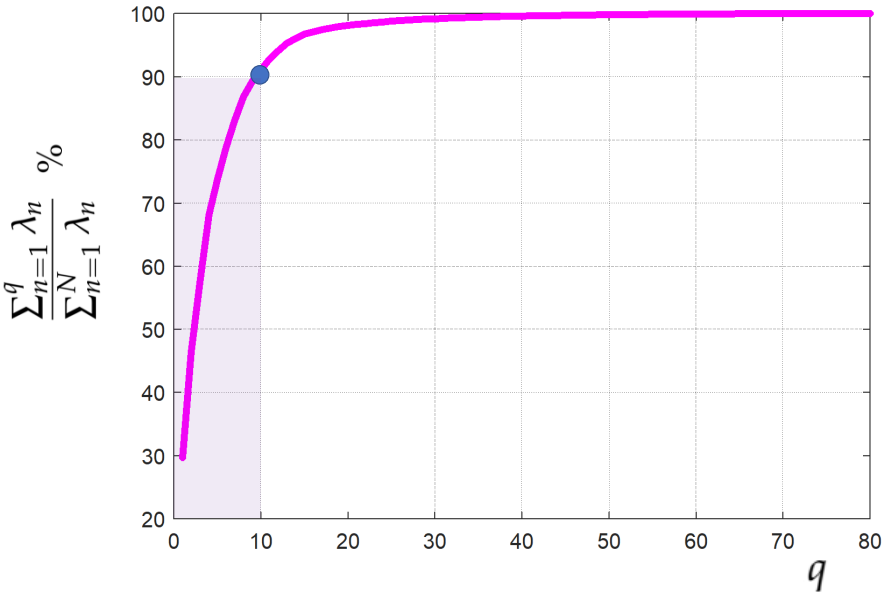


Figure 5.1: Calculation of the cumulative variance of the PDM from the proximal femur. According to this heuristic criterion, based on keeping 90% of the training variance, the selected order is $q = 10$.

Many landmark-based shapes, such as anatomical shapes in medical image analysis, are high-dimensional, and often only few observations are available. These observations may also contain noise artifacts. The heuristic approach of choosing the model order based on a kept fraction of total variance may therefore be suboptimal. The best model order varies significantly depending on the number of samples and noise level of the training data set. Our motivation is to design a model-order selection rule that has a theoretical justification and leads to a statistical shape model with good representation ability. We consider therefore information theory, which has successfully been used before to enhance registration and detection algorithms [100][101], as well as to place landmarks automatically in statistical shape models [102].

This model-order selection problem has been addressed before in statistical shape model design. In [103] the authors suggested a t-test of bootstrap stability of the PCA modes of the PDM, and they validated the strategy on simulated anatomical shapes

with white noise. The authors in [94] proposed a strategy based on an information-theoretic criterion for small sample support in a more generic array-processing context, also assuming white noise. These techniques may fail if the noise is not white. In [83], also in an array-processing context, the authors considered nonuniform noise.

In PDM design, there is no obvious model for the noise. Therefore, techniques that consider a specific noise structure may not work well. In order to address this, we propose a new strategy, based on information-theoretic criteria, that assumes a more generic colored-noise model. Our strategy is specifically designed to determine the model order in a PDM, although it may also be applied to other model-order selection problems with colored noise. We interpret the PDM as a multivariate regression, where the model order is determined considering the statistical properties of the regression residuals. We also perform a comparative study, with simulated and real shapes, where we prove the good performance of our strategy, as well as the importance of an accurate model order in PDMs.

5.2 Determining the model order in PDM

Let us assume we obtain M observations of the shape vector $\mathbf{x}. \in \mathbb{R}^N$ that contains the two concatanated coordinades in the Euclidean space of $\frac{N}{2}$ landmarks. We consider the M shape training samples $\mathbf{x}^{(m)}$, $m = 1, ..., M$, after a Procrustes alignment. The proposed solution in Chapter 4 divides the problem into two steps. First, we obtain estimates for the model parameters $\mathbf{P}$, $\mathbf{\Lambda}$, and $\boldsymbol{\mu}_{\mathbf{x}}$. Then, we estimate the model order q. For this, we split the observed data into two subsets of sizes $\mathbf{X}_1$ and $\mathbf{X}_2$. Within the first set $\mathbf{X}_1$, we compute the sample mean $\hat{\boldsymbol{\mu}}_{\mathbf{x}} = \frac{1}{M_1}\sum_{m=1}^{M_1}\mathbf{x}^{(m)}$ as well as the sample covariance matrix $\hat{\mathbf{R}}_{yy} = \frac{1}{M_1}(\mathbf{X}_1\mathbf{X}_1^T - \hat{\boldsymbol{\mu}}_{\mathbf{x}}\hat{\boldsymbol{\mu}}_{\mathbf{x}}^T)$ and its eigenvalue decomposition $\hat{\mathbf{R}}_{yy} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^T$. Let us define the matrix $\mathbf{Y}$ that contains the entries of $\mathbf{X}_2$ with mean removed, and consider the following multivariate linear regression:

$$\mathbf{Y} = \mathbf{P}_q\mathbf{B}_q + \mathbf{E}, \tag{5.2}$$

We propose the information-theoretic formulation for model selection in multivariate linear regressions, similarly as described in Chapter 4, where the model order q^* is chosen as

$$q^* = \arg\min_q \Big(\underbrace{-\log p(\mathbf{Y}|\hat{\mathbf{B}}_q, \hat{\mathbf{\Sigma}}_q)}_{\text{likelihood term}} + \underbrace{\eta(q)}_{\text{penalty}}\Big), \tag{5.3}$$

and the log-likelihood expression in (5.3) can be written as

$$\log p(\mathbf{Y}|\hat{\mathbf{B}}_q, \hat{\mathbf{\Sigma}}_q) = -\frac{M_2}{2}\log|\hat{\mathbf{\Sigma}}_q| - \frac{1}{2}\mathrm{Tr}\{(\mathbf{Y} - \mathbf{P}_q\hat{\mathbf{B}}_q)^T\hat{\mathbf{\Sigma}}_q^{-1}(\mathbf{Y} - \mathbf{P}_q\hat{\mathbf{B}}_q)\} + \text{constant}. \tag{5.4}$$

5.2.1 ML estimation of the regression parameters

According to the PDM definition, the columns of $\mathbf{B}_q$, denoted by $\mathbf{b}_q^{(m)}$ for $m = 1, \ldots, M_2$, are contained in a set $\mathbb{B}(\boldsymbol{\lambda}_q)$. Thus, the ML estimate of $\mathbf{B}_q$ is also con-

strained. We define $\mathbb{B}(\boldsymbol{\lambda}_q) = \{\mathbf{b}_q \in \mathbb{R}^{q\times 1} : |b_i| \leq \sqrt{\hat{\lambda}_i}, \forall i = 1 \dots q\}$, where $\mathbf{b}_q = [b_1, \dots, b_q]^T$, and $\boldsymbol{\lambda}_q = [\hat{\lambda}_1, \dots, \hat{\lambda}_q]^T$ contains the q largest eigenvalues of $\hat{\mathbf{R}}_{yy}$. Then, as demonstrated in Chapter 4, the ML estimation of $\mathbf{B}_q$ is equivalent to the regularized least-squares minimization

$$\hat{\mathbf{B}}_q = \underset{\mathbf{B}_q \in \mathbb{B}(\boldsymbol{\lambda}_q)}{\arg\min} \ \mathrm{Tr}\{(\mathbf{Y} - \mathbf{P}_q\mathbf{B}_q)^T \boldsymbol{\Sigma}_q^{-1} (\mathbf{Y} - \mathbf{P}_q\mathbf{B}_q)\}, \tag{5.5}$$

where $\mathbf{B}_q \in \mathbb{B}(\boldsymbol{\lambda}_q)$ is applied column-wise. We calculate the solution to (5.5) following the lines of Chapter 4. That is, we first obtain the unconstrained solution to (5.5), i.e., $\hat{\mathbf{B}}_q^u = (\mathbf{P}_q^T \boldsymbol{\Sigma}_q^{-1} \mathbf{P}_q)^{-1} (\mathbf{P}_q^T \boldsymbol{\Sigma}_q^{-1}) \mathbf{Y}$. Then, we scale it such that the constraints are fulfilled.

The ML estimate of the covariance matrix of the residual noise $\boldsymbol{\Sigma}_q$ is, as long as $N < M_2$, the sample covariance matrix:

$$\hat{\boldsymbol{\Sigma}}_q = \frac{1}{M_2} (\mathbf{Y} - \mathbf{P}_q\mathbf{B}_q)(\mathbf{Y} - \mathbf{P}_q\mathbf{B}_q)^T. \tag{5.6}$$

The ML estimates $\hat{\mathbf{B}}_t$ and $\hat{\boldsymbol{\Sigma}}_q$ are mutually dependent, which prohibits finding a closed-form solution. We use the proposed alternating optimization algorithm in Chapter 4.

5.2.2 Choosing the model order

The shape data is often high-dimensional but with small number of samples. Under these circumstances, the estimate of the matrix $\boldsymbol{\Sigma}_q$ may be ill-conditioned. In order to deal with this, we reduce the number of parameters to be estimated by assuming $\boldsymbol{\Sigma}_q$ to be a diagonal matrix, with $\boldsymbol{\sigma}^2 = [\sigma_1^2, \sigma_2^2, \dots, \sigma_N^2]$ on the diagonal. Consequently, its ML estimate is $\hat{\boldsymbol{\Sigma}}_q = \mathrm{diag}(\frac{1}{M_2}(\mathbf{Y} - \mathbf{P}_q\mathbf{B}_q)(\mathbf{Y} - \mathbf{P}_q\mathbf{B}_q)^T)$. Following the lines of [96] and Chapter 4, we choose the Akaike information criterion (AIC), so the penalty term in (5.3) is $\eta(q) = M_2 q + N$, which corresponds to the degrees of freedom in (5.4). Notice that AIC has been used as well for similar problems (for example in [94]), and that the assumptions made to derive it are quite weak [92]. Finally, the model-order estimate q^* is obtained by minimizing the terms in (5.3) that depend on q:

$$q^* = \arg\min_q \Big[M_2 \Big(\sum_{i=1}^{N} \log(\hat{\sigma}_i^2) + 2q \Big) + \sum_{i=1}^{N} \sum_{m=1}^{M_2} \frac{\hat{\epsilon}_i^{2(m)}}{\hat{\sigma}_i^2} \Big], \tag{5.7}$$

where $\hat{\epsilon}_i^{(m)}$ is the ith element of vector $\hat{\boldsymbol{\epsilon}}^{(m)} = \mathbf{y}^{(m)} - \mathbf{P}_q \hat{\mathbf{b}}_q^{(m)}$; and $\hat{\mathbf{b}}_q$ and $\hat{\sigma}_i^2$ are the estimated values after convergence of the alternating optimization in Algorithm 4.1 in Chapter 4.

5.3 Results and discussion

In this section we validate with numerical results the used of our proposed model selection strategy in shape analysis. In a PDM trained from real data sets, the true

model order of the shape model is not known. Hence, we have also used simulated shape data (following the model in Appendix), where the model order q is set beforehand and thus known.

5.3.1 Data description

We simulate realistic synthetic shapes following the model in the Appendix and similarly to the simulated data in [103]. We use the eigenvectors $\mathbf{P}_t$ obtained from available real shape data sets and choose values for $\mathbf{b}_q$ that are consistent with the sample data. Then we add white Gaussian noise $\boldsymbol{\epsilon}$ with different noise levels β, which we define as the ratio between the smallest kept signal eigenvalue and the noise variance. We choose these noise levels β such that the produced shapes still look realistic. Then we randomly rotate, scale, and translate these synthetic shapes and use Procrustes to re-align them. Procrustes alignment typically colors the noise, so it may not longer be white when the model order has to be selected.

We have used different shape databases of real contours, in order to show that the proposed technique generalizes to different structures of shape:

- Proximal femur bone outlines (168 samples with 40 landmarks) and the distal femur bone outline (111 samples with 40 landmarks) as seen in fluoroscopic X-rays.
- Lung outlines (246 samples with 44 landmarks) and heart outlines (246 samples with 26 landmarks) from chest X-rays [104].
- Hand outlines from photographs [31] (38 samples with 20 landmarks).

We show the average shape of each data set in Fig. 5.2. More information about the data can be found in the Appendix.

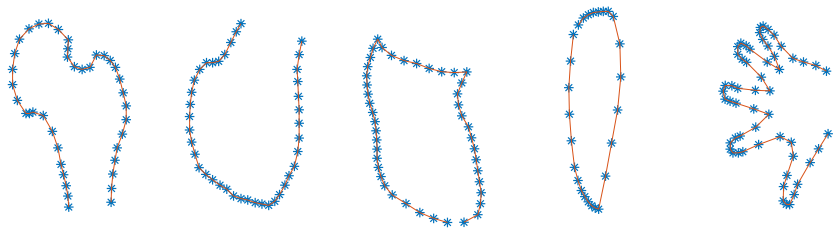
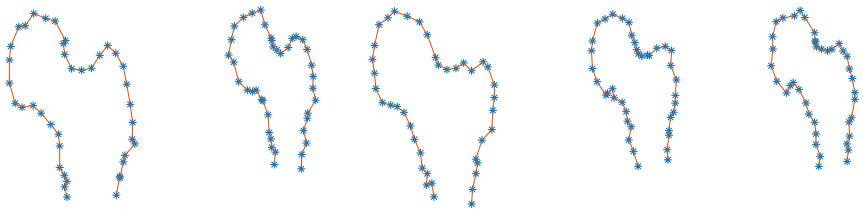
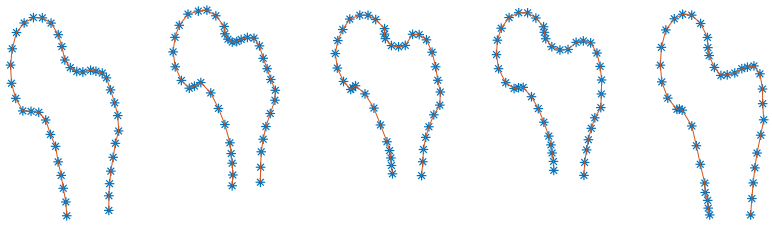


Figure 5.2: Examples from the datasets of proximal femur, distal femur, lung, heart and hand. The landmarks are shown as blue stars.

These data sets are similar: they belong to human anatomy shapes, the number of samples is limited, each sample contains a fixed number of landmarks in correspondence, and the noise is unknown. In each data set, the shapes are composed by a few landmarks that are anatomical and manually labeled, and the rest of the landmarks are equally distributed between these.



(a) $\beta = 5$ dB



(b) $\beta = 20$ dB

Figure 5.3: Examples from the artificial femur shapes.

5.3.2 Comparative strategies

We compare our proposed strategy with four others:

1. A variance threshold of 95%, as described in [14] and validated in [20].
2. An information-theoretic technique that considers white noise and a sample poor scenario [94].
3. An information-theoretic criterion considering non-uniform noise [83].
4. A boostrap t-test designed for PDMs [103] that considers white noise.

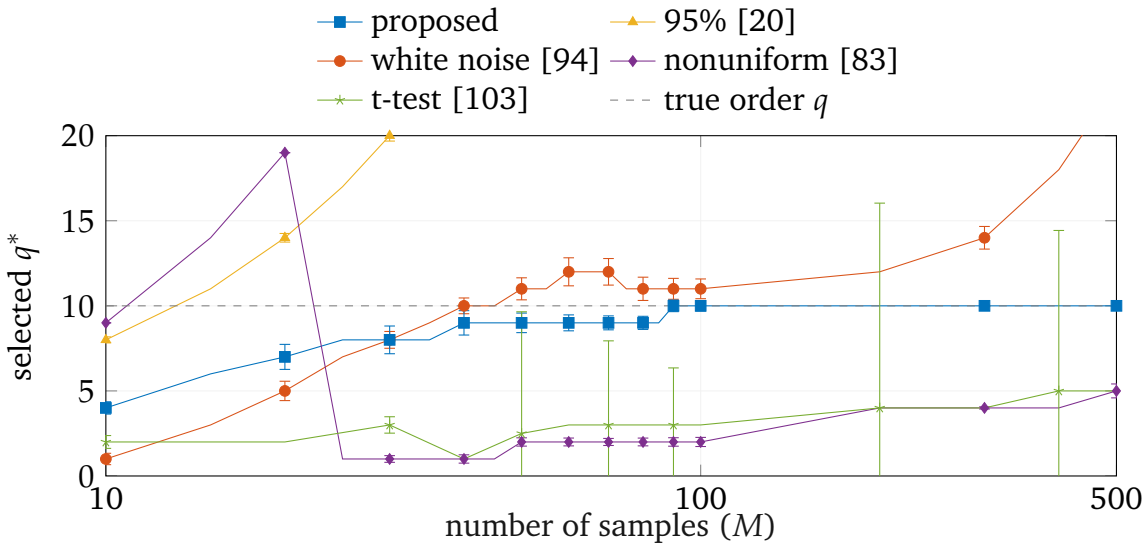
5.3.3 Results on simulated shape data

The evaluation of simulated data is straightforward since there is a known ground truth for the model order q.

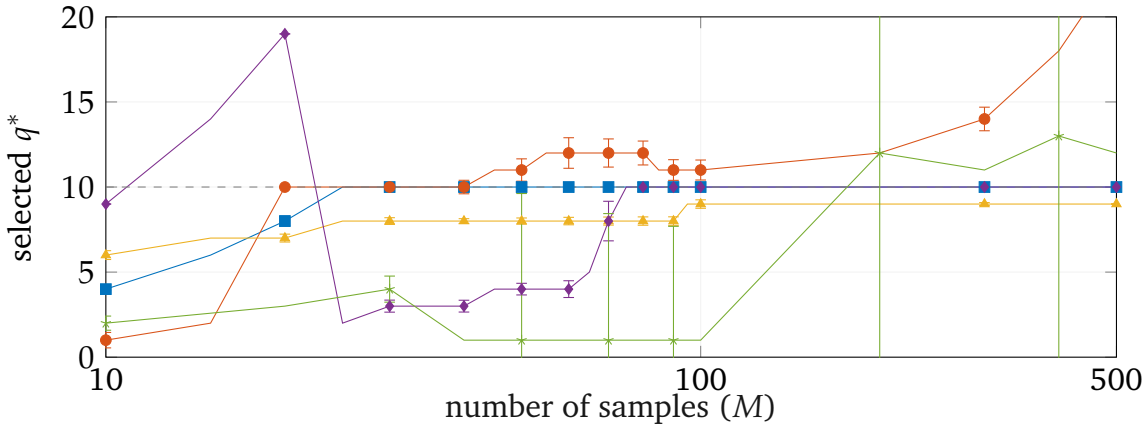
In Fig. 5.4 we show the model order obtained from 1000 Monte Carlo simulations. We see that the performance of the 95%-approach [20] (triangle, yellow lines) depends considerably on the level of noise: there is a tendency to overestimate if β is moderate (5 dB) and to underestimate if β is high (20 dB). The white-noise strategy in [94] (circles, red lines) tends to overestimate, especially when the number of samples increases. We believe this is due to color in the noise, introduced in the simulation by the Procrustes alignment. The t-test strategy [103] (stars, green lines) does not perform well and leads to results with high variance. The reason for this may be that this approach assumes white noise to evaluate stability. The nonuniform

noise strategy [83] (diamond, purple lines) tends to underestimate if β is moderate (5 dB), and it provides an incorrect estimation if the number of samples is very small. Our proposed strategy (square, blue lines) outperforms the competitors: it provides the best model-order estimate in general, it needs fewer samples to find the correct estimate, it is not highly dependent on the noise level, and it is consistent with increasing number of samples.

The results in Fig. 5.4 empirically show that AIC was suitable for the problem. Nevertheless, the proposed solution considers a general penalty term. Our approach can still account for other information theoretical penalties.



(a) $\beta = 5$ dB



(b) $\beta = 20$ dB

Figure 5.4: Average model order over 1000 Monte Carlo simulations, using simulated femur shapes of 40

5.3.4 Results on real shape data

There is no known ground truth for model order q in a PDM that is trained with real shape data sets. Nevertheless, we may still evaluate how plausible the model-order estimate is when the number of available samples changes and compare this with the behavior in artificial data. Additionally, we illustrate the importance of the model order in PDMs with a numerical experiment that shows the impact of the selection of t when a PDM of order q is used to deal with partial occlusions in shapes.

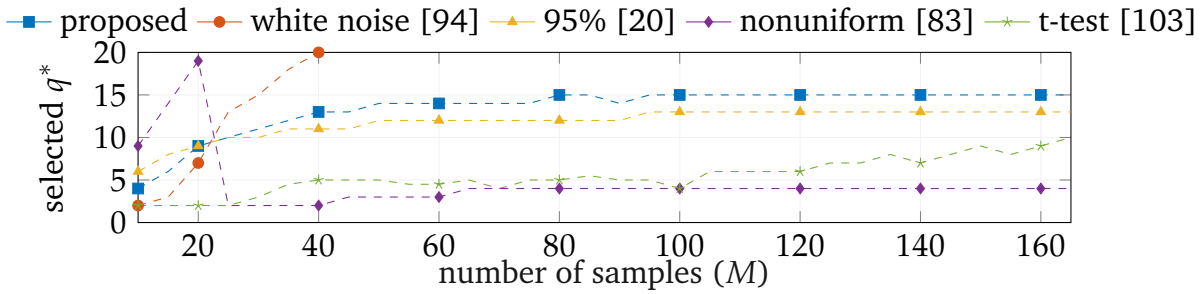
In Fig. 5.5 we show the estimated model order for different number of samples on the three data sets. There are similarities with Fig. 5.4: “95%" [20], t-test [103], and “nonuniform" [83] provide small model orders, while “white-noise" [94] provides large (probably too large) model orders. Our strategy seems to provide a consistent model order that starts converging with fewer samples.

As an illustrative example of the importance of the model order in statistical shape models, we performed an experiment that shows how well the PDM with the selected model order can deal with partial occlusions. Considering an obtained PDM as the prior information about shape deformation, we perform an estimate of an occluded (or missing) landmark in a new observed shape. The test consists in the following: inside a leave-one-out test, this is, for all $m = 1, \ldots, M$ available samples in a data set, within the mth “left-out" sample we delete one landmark from that shape. Then, we estimate it from the remaining landmarks using a linear minimum mean-squared error (LMMSE) estimator. The $M-1$ “not left-out" samples are used to design the PDM, in which we evaluate all possible orders t. Let $\mathbf{y}_i$ denote the missing landmark (which consists of its x- and y-coordinates) and $\mathbf{y}_a$ the remaining available landmarks. Let $\hat{\mathbf{R}}_{ia}$ denote the sample cross-covariance matrix between the missing landmark and the remaining available landmarks, and $\hat{\mathbf{R}}_{aa}$ the covariance matrix of the available landmarks. These matrices are calculated from the available PDM of order q. The LMMSE estimator of the missing landmark from the remaining landmarks is then $\hat{\mathbf{y}}_i = \hat{\mathbf{R}}_{ia}\hat{\mathbf{R}}_{aa}^{-1}\mathbf{y}_a$. In our experiment, we successively estimate one landmark i from the others, repeating this for all $i = 1, ..., N/2$ landmarks. We average the error over all landmarks and over all available samples and obtain

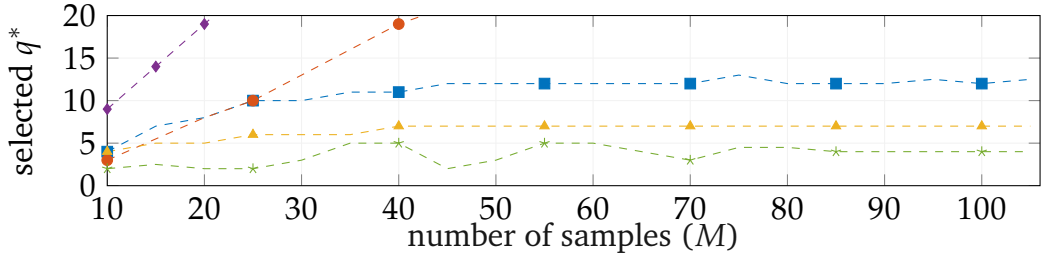
$$e_{\text{LMMSE}}(q) = \frac{1}{M}\frac{1}{N/2}\sum_{m=1}^{M}\sum_{i=1}^{N/2}||\hat{\mathbf{y}}_i^{(m)} - \mathbf{y}_i^{(m)}||^2, \tag{5.8}$$

which is evaluated for all possible model orders, i.e., $q = 1, \ldots, \min(N, M)$. Figure 5.6 shows this metric for the three data sets as a function of considered model order q. In Fig. 5.6, we observe that the evaluated empirical error decreases until it reaches a minimum, which is different for each data set. A shape model with too small an order may suffer from underfitting (thus, the error decreases if we add more complexity), and too large an order may lead to overfitting (and therefore, the error decreases if we reduce the complexity). We conclude that choosing the right model order is critical in order to minimize the LMMSE. We see that the model order determined by our technique (blue squares) leads to the smallest LMMSE among all competing techniques. We also notice that there is a relatively large interval of model orders

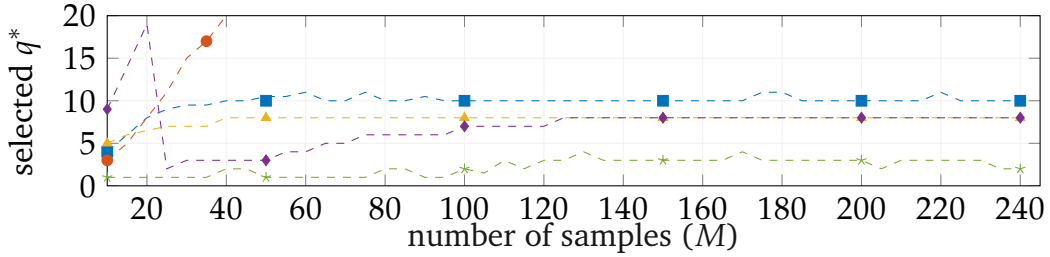
that lead to similar LMMSEs. The principle of parsimony dictates that in such a case a smaller order is to be preferred. Our technique observes this principle.



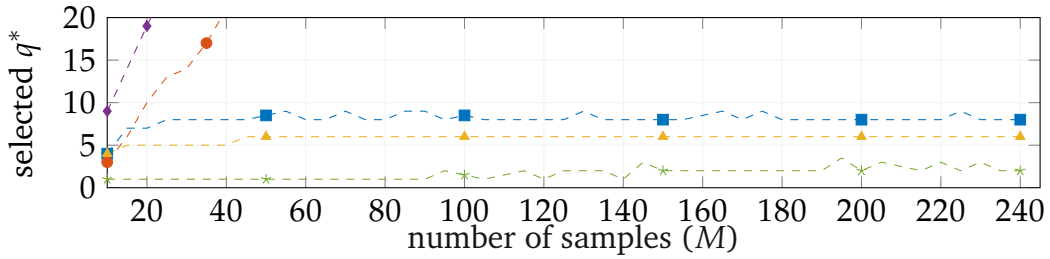
(a) Proximal Femur



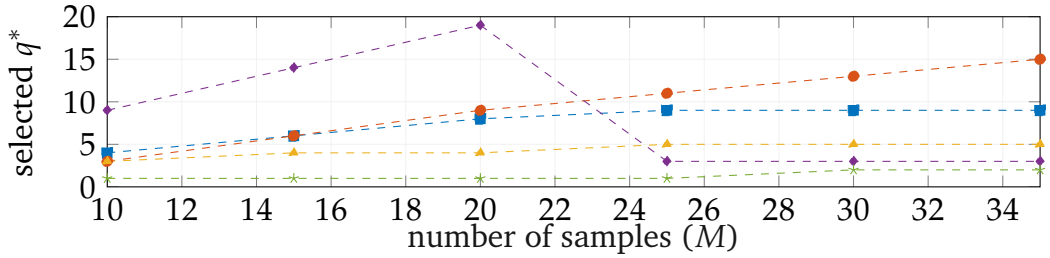
(b) Distal Femur



(c) Lung



(d) Heart



(e) Hand

Figure 5.5: Selected model order for five different real shape data sets. Notice that the different shapes have different number of total samples M.

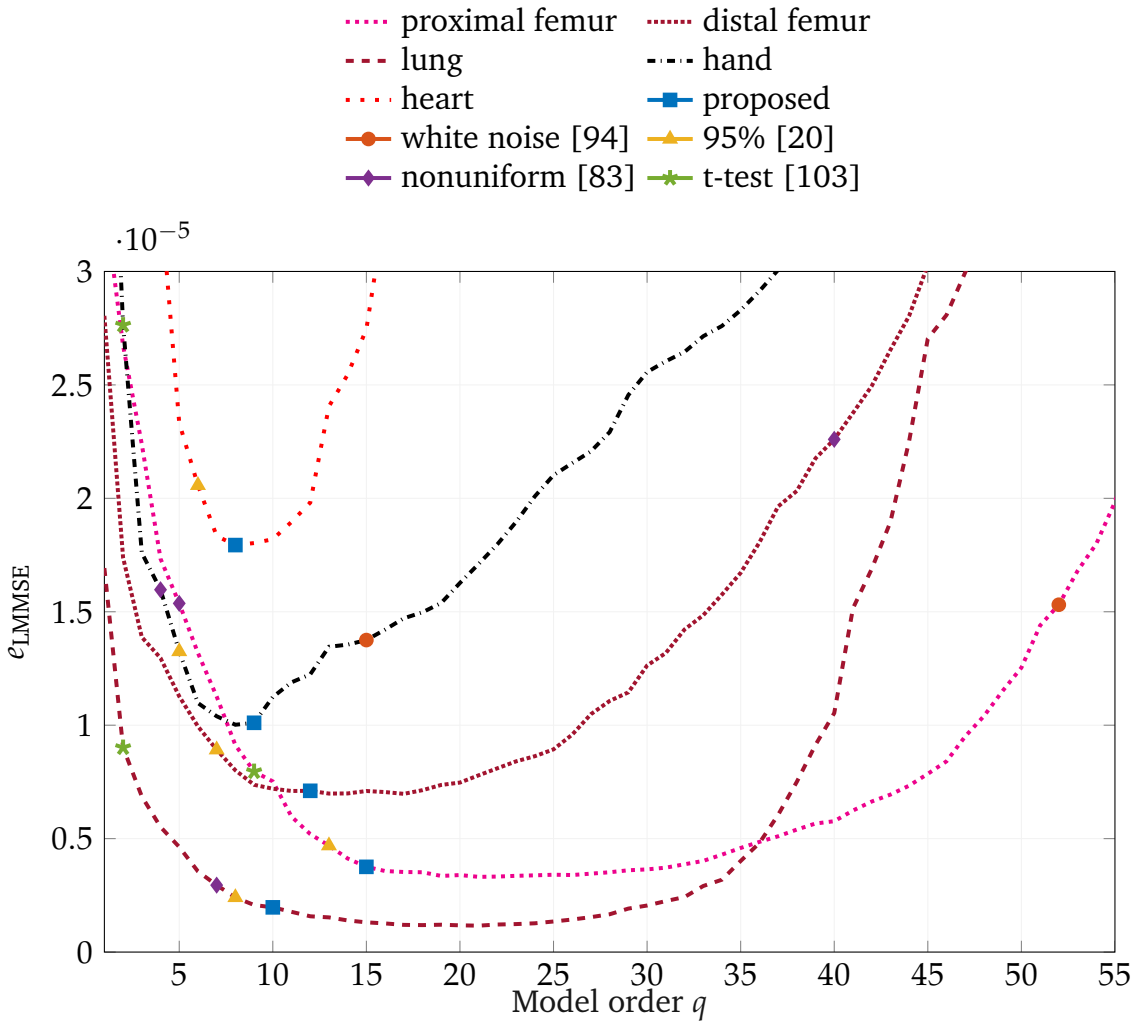


Figure 5.6: e_{LMMSE} for all possible order q. The evaluated model orders are obtained using all available samples to train the PDMs. The order selected by “white noise" (orange circle) for the lung data set is not depicted because is out of plot limits.

Chapter 6

Procrustes registration of contours without correspondences

"All creative people want to do the unexpected."

— Hedy Lamarr

Shape registration is key in image processing and computer vision. Two contours of the same shape must be registered in many different applications, such as the design of statistical shape models, shape retrieval, or segmentation. With anatomical shapes, the variability among individuals adds difficulty. As it is shown in Chapter 2, a Procrustes registration between two contours with known correspondences is straightforward. However, these correspondences are not generally available. Manually placed landmarks are often used for correspondence in the design of SSM. However, determining landmarks on contours is time-consuming and often error-prone. The conventional solution to simultaneously find correspondence and registration is the Iterative Closest Point (ICP) algorithm. However, ICP requires an initial position of the contours that is close to registration, and it is not robust against outliers. In this Chapter we propose a new strategy, based on Dynamic Time Warping (DTW), that efficiently solves the registration problem without correspondences. We study the registration performance in a collection of different shape data sets and show that our technique outperforms competing techniques based on the ICP approach. Furthermore, we extend our strategy so that it can be applied to an ensemble of contours of the same shape, which is an extension of the generalized Procrustes analysis accounting for a lack of correspondence. The chapter is organized as follows. Section 6.1 presents a motivation, an introduction to the ICP algorithm, and an overview about DTW and dynamic programming. Then, in Section 6.2 we present the problem formulation and the DTW-based solution. Finally, we present the experiments and the corresponding results in Section 6.3.

The results of this chapter will be submitted to be published as:

- A. Eguizabal, J. Schmidt, and P. Schreier, "Registration of contours without correspondences using dynamic time warping," to be submitted to *IEEE Transactions on Image Processing*, 2019.

6.1 Motivation and preliminaries

Iterative Closes Point (ICP) is the conventional solution to find correspondence and registration simultaneously. This iterative technique, however, is not an optimal solution and depends considerably on the initial guess. Also, the algorithm does not account for any parametrization of the contours, and hence, needs to assume that every point in one contour may correspond to any other point on the other contour. This may become computationally very expensive and error-prone. Dynamic Time Warping (DTW), on the other hand, is a technique to optimize the temporal assignment between two time series. A set of constraints are considered in the path of correspondence, thus a more efficient and accurate solutions can be found using dynamic programming.

In this section we discuss the state-of-the-art and motivate the use of our approaches. Then, we recall some concepts from Chapter 2 and provide an introduction to the techniques that are necessary to understand our proposal.

6.1.1 Why to consider registration without correspondences

Image registration is a big challenge in computer vision and medical image analysis [105]. The registration of shapes is necessary to enhance segmentation and retrieval algorithms, and more specifically, to build statistical shape models [14], [106]. The training contours that are required to learn a shape model are generally manually or semi-automatically obtained from training images, which is very time-consuming [107]. The training contours can also be automatically extracted from a different image domain where segmentation is less challenging. For instance, in order to learn a bone shape for segmentation in fluoroscopic X-ray images, the training contours may be obtained from diagnostic CT scans. In all these circumstances, registration of the training contours is required to build the shape model.

A Procrustes registration (rotation, scale, and translation, as defined in [108]) between two contours with known correspondences is a linear least-squares problem: when the correspondences are available, a Procrustes analysis can be employed [9]. However, determining correspondences is challenging. Coming back to statistical shape models, corresponding points are often determined manually with a few landmarks [35]. The most common choice of landmarks in medical images are anatomical references, which allow to maintain consistency between different samples in the annotation process. However, it is tedious to determine anatomical landmarks [63]. Also, the definition of an anatomical landmark typically refers to a point on the surface of the anatomy, whose projection in the image plane may not belong to the two-dimensional boundary. This fact is especially problematic in applications where the focal point of the image may vary, such as in fluoroscopic images that are taken intraoperatively with a C-arm [109]. Another difficulty arises when the contour is open and occasionally occluded on the extremes, since the correspondences at the start and end points are not trivial. Also, establishing manual correspondences is done based on experience, which is typically not optimal. Furthermore, manual la-

beling is generally not very dense (the number of landmarks is much smaller than the image resolution).

When correspondences are unknown, they may also be assigned automatically. Registration is then typically performed with an ICP approach [110]. This algorithm iterates between solving the registration and finding the corresponding points, which are chosen as the closest in terms of a defined distance (e.g. Euclidean). However, the algorithm only converges if the two contours to be registered are initially close, and it is not robust against outliers. Assigning correspondences and simultaneously solving a registration problem has aroused interest before, such as in [111] and [112]. In [111] the problem is presented as an extension of Procrustes alignment, and it uses a gradient descent approach to determine correspondences. In [112], the authors incorporate probability and cross-entropy in the correspondence and determine a cost function to minimize. Another approach is given in [113], which optimizes a similar objective function in a more rigorous way and presents a probabilistic approach called Coherent Point Drift (CPD). Finally, [114] shows a technique that considers M-estimators in ICP to assign the correspondences. These approaches, like ICP, typically assume that any point on the reference contour may correspond to any other point on the target contour. Then, the correspondence is modeled in a matching matrix whose dimensions are reference length times target length.The proposed technique in [115] introduces an order-preserving constraint in the correspondences, which allows to define an optimal and computationally efficient solution. It models the correpondence with graphical models and solves it with dynamic programming. However, it considers only closed contours. Similarly, the authors in [116] fix the order of the points of the contours and estimate the correspondence, optimizing a geodesic distance between them. However, they assume that scale and location have been filtered out and only consider closed shape contours. In this chapter we propose DTW [117] to establish an automatic correspondence between two boundaries. DTW is a common technique in time-series analysis to align signals, and it has also been successfully used in shape retrieval. The authors in [118] combine DTW with Fourier analysis. Since they used closed-form contours of similar arc-length, they correct the scale and the translation normalizing the contours and find a rotation-invariant representation with the Fourier coefficients. However, their technique does not apply to open contours. In DTW the input signals contain thousands of points and the solution is still computationally efficient [119]. Unlike ICP and CPD, DTW imposes constraints on the order of corresponding points between two contours, and this leads to more efficient and accurate solutions. Our approach based on DTW is able to deal with both open and closed contours. Also, since DTW is robust to scaling and translation of the inputs, there is no need for a very good initialization [117]. Furthermore, to simultaneously solve the registration and the correspondence problems, we combine DTW with a weighted Procrustes registration to add robustness against outliers. Additionally, we propose an extension of the generalized Procrustes analysis considering DTW to assign correspondences.

Group-wise registration

In order to design a statistical shape model, registration needs to be performed group-wise for a collection of training contours, that is, an ensemble of contours of the same shape needs to be registered to the same reference. When correspondences are known, the typical solution is the generalized Procrustes analysis proposed in Chapter 2 [120], [121]. When correspondences are missing, there are approaches where the group-wise registration is performed plane-to-plane considering image warping. In [122] the authors use thin-plate B-splines to model the deformation of the shapes, together with a generalized Procrustes analysis to perform the registration. However, the contours they use are not very dense (around 15 landmarks) and the thin-plate deformation may become very computationally demanding with denser contours of higher resolution. The authors in [123] define a plane-to-plane warping based on free deformation and mutual information. They assume, however, a good initial guess in terms of rotation, scale, and translation. Similarly, in [124] they also consider a plane-to-plane warping and the image appearance in the registration. They require, however, a good initialization of the samples since they perform an exhaustive search to determine the affine transformations. Other solutions are based on a contour-to-contour warping. The approach in [17] performs a generalized Procrustes analysis determining a geodesic distance between the contours. Nevertheless this distance depends on ad-hoc parameters and requires a good initialization of the registration. In [102] the authors describe a method to build shape models when correspondences are not available based on information theory. However, in order to establish the correspondences, they need to select a reference shape whose parameterization is fixed. Also, the contours are not longer than 30 points, which is orders of magnitude smaller than the typical image size.

In our proposed approach we also consider DTW to establish a high resolution correspondence in a group-wise manner, as an extension of the generalized Procrustes analysis. We register a set of shape vectors and estimate its Procrustes mean when there is no correspondence information. We assume that the contours to be registered have the same shape according to Kendall's definition [9], i.e., they share the same geometrical information after scale, rotation, and translation are removed.

6.1.2 Procrustes registration with correspondences

Recalling Chapter 2, let us assume we have the shape vectors $\mathbf{x}_1$ and $\mathbf{x}_2$, both containing N points, and that all points are in one-to-one correspondence for $n = 1, \ldots, N$. Let us recall the Procrustes registration, described in Chapter two, to register $\mathbf{x}_2$ onto $\mathbf{x}_1$. This is performed by the pose parameters $[r, t] \in \mathbb{C}$ that minimize the squared distance between $\mathbf{x}_1$ and $\mathbf{x}_2$. We define this distance as

$$d^2 = \sum_{n=1}^{N} |x_1[n] - x_2[n]|^2 = ||\mathbf{x}_1 - \mathbf{x}_2||^2. \tag{6.1}$$

As discussed in Chapter 2, we can rewrite the transformation as $r\mathbf{x}_2 + \mathbf{1}t = [\mathbf{x}_2 \quad \mathbf{1}]\mathbf{p}$, where $\mathbf{p} = [r \quad t]^T$ is a vector containing the pose parameters. The vector $\mathbf{p}$ that

minimizes the squared distance between $\mathbf{x}_1$ and $r\mathbf{x}_2 + \mathbf{1}t$ is the solution to a linear least-squares fit. Let us define the matrix $\mathbf{X}_2 = [\mathbf{x}_2 \quad \mathbf{1}]$. The pose $\mathbf{p}^\star$ that minimizes the distance after the Procrustes registration is

$$\mathbf{p}^\star = \arg\min_{r,t} ||\mathbf{x}_1 - (r\mathbf{x}_2 + \mathbf{1}t)||^2, \tag{6.2}$$

where $\mathbf{p}^\star = (\mathbf{X}_2^H \mathbf{X}_2)^{-1} \mathbf{X}_2^H \mathbf{x}_1$.

6.1.3 Point set registration and the Iterative Closest Point algorithm

The ICP algorithm is a well known approach to infer simultaneously correspondence and rigid-registration parameters between two sets of points [105]. We assume a set of N points of dimension D in the matrix $\mathbf{X} = \big[\mathbf{x}[1], \ldots, \mathbf{x}[N]\big]$, where $\mathbf{x}[n] \in \mathbb{R}^D \ \forall \quad n = 1 \ldots, N$. Let us consider two point sets $\mathbf{X}_1 \in \mathbb{R}^{D \times N_1}$ and $\mathbf{X}_2 \in \mathbb{R}^{D \times N_2}$. In order to find the registration parameters $\mathbf{R} \in SO(D)$ (rotation) and $\mathbf{t} \in \mathbb{R}^D$ (translation), the ICP algorithm solves [125]

$$[\mathbf{R}^\star, \mathbf{t}^\star, \mathbf{c}] = \underset{\mathbf{R} \in SO(D),\ \mathbf{t},\ \mathbf{c}}{\arg\min} \sum_{k=1}^{N_2} ||\mathbf{x}_1[c_k] - (\mathbf{R}\mathbf{x}_2[k] + \mathbf{t})||^2, \tag{6.3}$$

where the vector $\mathbf{c} = [c_1, \ldots, c_k, \ldots, c_{N_2}]^T$ contains that indexes of correspondence, that is, c_k is the index of the point in $\mathbf{X}_1$ that corresponds to the $k$the point in $\mathbf{X}_2$.

The ICP solution is an iterative process, in which during each iteration i the following steps are performed:

- Step 1: find the correspondence index in $\mathbf{X}_1$ for the kth point in $\mathbf{X}_2$, that is, the c_k-th point in $\mathbf{X}_1$, considering the closest point criterion, and assuming fixed registration parameter from previous iteration,

$$c_k^{(i)} = \arg\min_{c_k} ||\mathbf{x}_1[c_k] - (\mathbf{R}^{(i-1)}\mathbf{x}_2[k] + \mathbf{t}^{(i-1)})||^2, \tag{6.4}$$

  and repeat for every $k = 1, \ldots, N_2$.

- Step 2: given the correspondence index $c_k^{(i)}$ for every value of k, we can establish correspondence between $\mathbf{X}_1$ and $\mathbf{X}_2$ and the problem becomes a Procrustes analysis, whose solution is described in Chapter 2,

$$[\mathbf{R}^{(i)}, \mathbf{t}^{(i)}] = \underset{\mathbf{R} \in SO(D),\ \mathbf{t}}{\arg\min} \sum_{k=1}^{N_2} ||\mathbf{x}_1[c_k^{(i)}] - (\mathbf{R}\mathbf{x}_2[k] + \mathbf{t})||^2. \tag{6.5}$$

Notice that in the two dimensional case with planar shapes $D = 2$. Notice also that, when $D = 2$, this algorithm could also be expressed in complex number notation as already discussed in Chapter 2. In that case, the shape vectors are $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{C}^N$, and the rotation and translation are expressed with the complex scalars $[r, t] \in \mathbb{C}$.

Originally ICP was implemented to perform a rigid registration between two point sets, that is, without changing the scale of the sets [125]. Considering changes of scale does not add complexity to the formulation [126], but may compromise the performance of the registration. Nevertheless, since the correspondence is based only on the squared Euclidean distance between points, and all points in one set could potentially correspond to any of the other points in the other set, the ICP algorithm needs that the initial position of the sets is already close to registration, and it is also very sensitive to outliers in the sets. There are many different methods that overcome the ICP limitations and make it more robust. The authors in [113] consider soft-assignments of correspondence, that is, they establish the correspondence according to some probability. In addition, the strategy proposed in [114] applies M-estimation and therefore replaces the least squares minimization in (6.3) with robust criteria that handles possible outliers. In any case, the ICP minimization may get stuck on a local minimum if the initial position is very far from registration. It is certainly inefficient to assume that the correspondence is allowed between any possible pair of points, since these points typically have a structure and belong to a particular surface or contour, which can be parametrized.

6.1.4 Dynamic Programming and Dynamic Time Warping

If the correspondence path has a known structure, that is, we take into consideration that the points belong to certain surface or contour and, therefore, not every point of one set can correspond to any other point of the other, then the correspondence problem simplifies considerably. A solution to this problem was already proposed for time domain in the 70s [127], in which they solve word recognition in speech processing. The correspondence between two time series is constrained, since time is continuous and monotonic. The authors in [127] modeled the time fluctuations with a nonlinear warping with very specific properties. This made the minimization process more efficient, and optimally solvable with dynamic programming. We show an example of two times series and their correspondence path with DTW in Fig. 6.1.

In DTW we assume a set of N "time instants" in the vector $\mathbf{x} = \left[x[1], \ldots, x[N]\right]^T$, where $x[n] \in \mathbb{C} \quad \forall \quad n = 1 \ldots, N$. Then, let us define $\mathbf{x}_1$ and $\mathbf{x}_2$, containing N_1 and N_2 points respectively, as two shape vectors with unknown correspondences. We need to determine the optimal warping-path matrix $\mathbf{C} \in \mathbb{R}^{2\times L}$ in terms of Euclidean distance. This matrix $\mathbf{C}$ is composed of L correspondence vectors $\mathbf{c}_l = (n_1^{(l)}, n_2^{(l)})^T$, with $l = 1, \ldots, L$. The vectors $\mathbf{c}_l$ establish correspondences between points $x_1[n_1^{(l)}]$ and $x_2[n_2^{(l)}]$. The optimal warping path $\mathbf{C}^\star$ between $\mathbf{x}_1$ and $\mathbf{x}_2$ is the one that minimizes the sum of distances

$$\mathbf{C}^\star = \arg\min_{\mathbf{C}\in\mathbb{P}} \sum_{l=1}^{L} |x_1[n_1^{(l)}] - x_2[n_2^{(l)}]|^2, \tag{6.6}$$

where $\mathbb{P}$ is the set of allowed warping paths. A warping path in $\mathbb{P}$ must satisfy the following contraints [117]:

1. Boundary condition: The beginning and end points of the shape vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ are in correspondence, i.e., $\mathbf{c}_1 = (1,1)$ and $\mathbf{c}_L = (N_1, N_2)$.

2. Monotonicity condition: The topology of the curve is respected in the correspondence assignment, meaning that $n_i^{(1)} \leq n_i^{(2)} \leq \cdots \leq n_i^{(L)}$ for $i = 1, 2$.

3. Step size condition: Each element of $\mathbf{x}_1$ corresponds to at least one element in $\mathbf{x}_2$ and vice versa. Therefore, the elements in the matrix $\mathbf{C}$ satisfy $\mathbf{c}_{l+1} - \mathbf{c}_l \in \{(1,0),(0,1),(1,1)\}$ for all $l \in \{1,\ldots,L-1\}$.

4. Adjustment window condition [127] [128]: the time-axis fluctuation is constrained, for instance to be always smaller than a threshold α [127]. Two different windows are depicted in Fig. 6.2.

These constraints on the warping path allow an efficient computation of $\mathbf{C}^\star$ with dynamic programming [117].

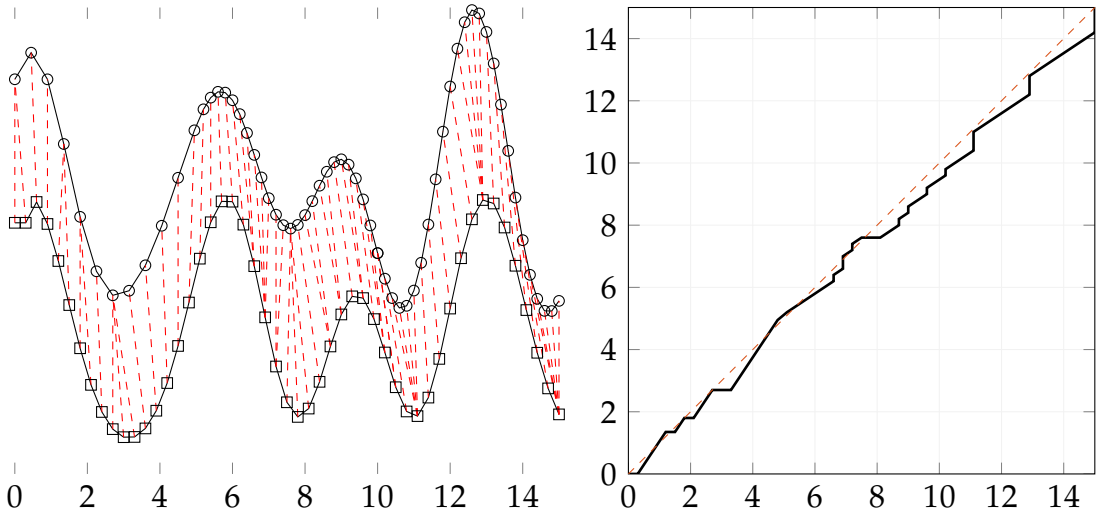


Figure 6.1: An example of Dynamic Time Warping. Left: two time series and its optimal alignment. Right: corresponding warping path (black) and the path that is assigned with direct correspondence (red, dotted).

A dynamic programming solution

Dynamic programming is a mathematical tool for making a sequence of decisions that are interrelated [130]. In opposition to linear programming, there is no standard formulation for a dynamic programming problem, and the particular equations depend on each situation. In general, these problems have the property of being explained in smaller subproblems, and accomplish an optimal substructure [117].

Let us overview an example of a solution process through the concerned DTW. Let us consider the vectors $\mathbf{x}_1 \in \mathbb{R}^{N_1}$ and $\mathbf{x} \in \mathbb{R}^{N_2}$. We need to evaluate a grid of possible correspondence pairs. This grid is a matrix $\mathbf{D} \in \mathbb{R}^{N_1 \times N_2}$ with elements $D(n_1, n_2)$. A

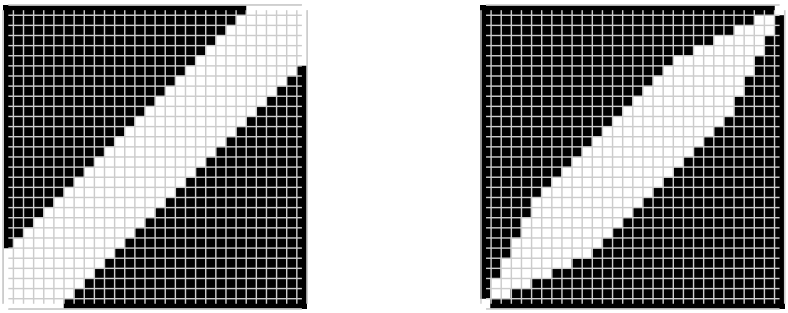


Figure 6.2: Illustration of two corresponding path (or cost) grids with window constraints. The allowed correspondences are in white. We depict the Sakoe-Chiba [127] (left) and the Itakura-parallelogram (right) [129].

dynamic programming stage is considered for each possible pair of correspondences (n_1, n_2), where the cost $D(n_1, n_2)$ is then evaluated for each possible stage. The algorithm calculates each stage cost as

$$D(n_1, n_2) = d(x_1[n_1], x_2[n_2]) + \min\{D(n_1 - 1, n_2 - 1), D(n_1, n_2 - 1), D(n_1 - 1, n_2)\}, \quad (6.7)$$

in which d can be, for instance, the Euclidean distance, i.e., $d(x_1[n_1], x_2[n_2]) = |x_1[n_1] - x_2[n_2]|^2$. In the calculation of this cost the monotonic and continuity of the correspondence path are taken into account, since it depends only on three previous stages $D(n_1 - 1, n_2 - 1), D(n_1, n_2 - 1), D(n_1 - 1, n_2)$. Due to the boundary conditions, the costs $D(1, 1)$ and $D(N_1, N_2)$ are always considered in the final path. We illustrate in detail the process of calculating the stages more visually in Fig. 6.3 with a toy example.

The simplest solution of implementation of this algorithm is an iterative approach to evaluate all possible combinations, and hence every possible stage cost $D(n_1, n_2)$. This solution may have a high computational cost, with complexity $\mathcal{O}(n^2)$ with respect to the number of points the vectors. However, there are strategies to reduce the computation complexity. For instance, the use of an adjustment window α decreases the complexity to $\mathcal{O}(n\alpha)$. Also, several other hardware accelerations could improve the approach even to $\mathcal{O}(n)$ [131].

6.1.5 Dynamic Time Warping to establish correspondences

Once we obtain the warping-path matrix $\mathbf{C}$ with DTW, we define the resulting shape vectors in point-to-point correspondence as $\tilde{\mathbf{x}}_1 = \left[x_1[n_1^{(1)}], \ldots, x_1[n_1^{(L)}]\right]^T$ and $\tilde{\mathbf{x}}_2 = \left[x_2[n_2^{(1)}], \ldots, x_2[n_2^{(L)}]\right]^T$. Each vector contains L points, and $\max(N_1, N_2) \leq L \leq N_1 + N_2 - 1$.

A Procrustes registration as defined in (6.1) requires a one-to-one correspondence between vector elements. In most applications this is not available. In order to

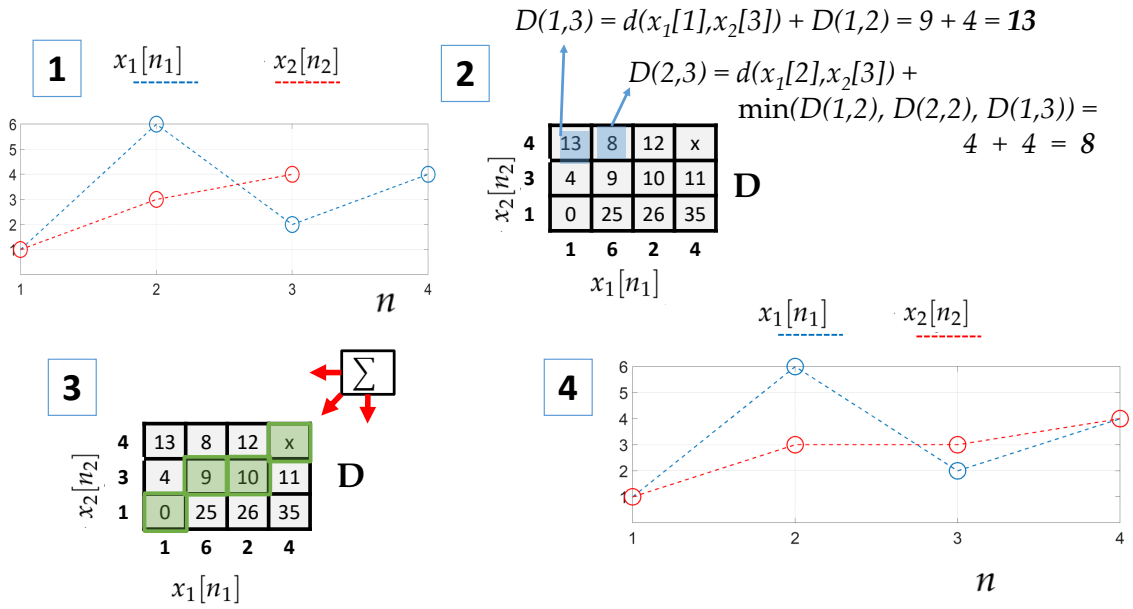


Figure 6.3: A toy example of the use of dynamic programming to find the correspondence path of minimum cost in DTW. In 1. we show how the points in the two series (in blue and red) look initially. In 2. we illustrate the calculation of the cost grid **D** and show the particular cases of $D(1,3)$ and $D(2,3)$. In 3. we see how the correspondence path (in green) is constructed, starting from the end point and going backwards (following the constrained directions depicted with red arrows), and summing the neighbour position with minimum cost in the grid; notice that the top-right corner and bottom-left corner are always inside the path due to the constraints. We plot the result of the alignment in 4., where we see an example of a many-to-one correspondence, since $x_2[2]$ corresponds both with $x_1[2]$ and $x_1[3]$.

assign correspondences between two shape vectors, we consider DTW. A non-linear warping of the signals is considered in order to determine the corresponding points. The objective of DTW is to align these signals so that the sum of the distances (e.g. Euclidean) between the corresponding points is smallest. Consequently, DTW can determine a warping path between two shape vectors that is optimal in terms of the sum of Euclidean distances between corresponding points. Considering the path constraints, there is at least one corresponding point in $x_1[n]$ and $x_2[n]$ for every n, but there may also be more than one point in $x_1[n]$ corresponding to one point in $x_2[n]$ and vice versa. We call this a many-to-one correspondence. This case has occurred in the toy example in Fig. 6.3.

Simultaneously determining correspondences and registration

When we employ DTW to find a path of correspondence between two boundaries, the result may change when a linear transformation is applied to the boundaries. The correspondence path that best explains the shape deformation is the one considered

after a Procrustes registration between the boundaries. At the same time, this Procrustes registration depends on the determined correspondences. Hence, we need to obtain the parameters that solve the overall minimization problem

$$[r^\star, t^\star, \mathbf{C}^\star] = \arg \min_{r,t,\mathbf{C}\in\mathbb{P}} \sum_{l=1}^{L} |x_1[n_1^{(l)}] - (rx_2[n_2^{(l)}] + t)|^2, \tag{6.8}$$

where the transformation parameters r, t and warping path $\mathbf{C}$ are mutually dependent.

6.1.6 Group-wise correspondence and registration

In order to learn statistical shape models we need to perform the group-wise registration of a set of training shape vectors [14], [108]. Let us consider a set of M shape vectors of the same length N and with one-to-one correspondence, i.e., each nth element of each of the M vectors corresponds to the same landmark. Recalling the definition of shape in [9], shape variability is what remains after accounting for scale, translation and rotation (i.e. after a group-wise Procrustes registration). A typical way to remove the effects of size and translation is to normalize the shape vectors to unit size and translate their centroid to the origin of coordinates [23]. That is, for a given shape vector $\mathbf{x}$, we obtain

$$\mathbf{x}_o = \mathbf{x} - \frac{1}{N}\sum_{n=1}^{N} x[n], \text{ and } \boldsymbol{\tau} = \frac{\mathbf{x}_o}{||\mathbf{x}_o||}, \tag{6.9}$$

where $\mathbf{x}_o$ is moved to the origin, and dividing by $||\mathbf{x}_o||$ is a size normalization. We call the vector $\boldsymbol{\tau}$ a preshape, equivalently to the geometrical definition in[23]. The term preshape refers to the fact that the vector is one step away from registration since rotation still needs to be removed [108]. Thus, the group-wise registration problem is reduced to find rotations.

As described in Chapter 2, the preshapes belong to a hypersphere, where the distance between two preshapes is a geodesic [108] defined as $d_s(\boldsymbol{\tau}_1, \boldsymbol{\tau}_2) = \cos^{-1}|\boldsymbol{\tau}_2^H \boldsymbol{\tau}_1|$. Considering we have a set of M training shape vectors, and their preshapes are $\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_M$, the Procrustes mean is defined as

$$\boldsymbol{\mu}^\star = \underset{\boldsymbol{\mu}}{\operatorname{arginf}} \sum_{m=1}^{M} d_s(\boldsymbol{\tau}_m, \boldsymbol{\mu}), \tag{6.10}$$

where $\boldsymbol{\mu}$ is also considered a preshape so that $||\boldsymbol{\mu}||^2 = 1$ and $\frac{1}{N}\sum_{n=1}^{N} \mu[n] = 0$. The solution to (6.10) is typically obtained through a generalized Procrustes analysis [121], [108]. This analysis iterates between minimizing the distances of the preshapes to $\boldsymbol{\mu}$ and estimating $\boldsymbol{\mu}$. However, when we assume not to have correspondences and, furthermore, the observed vectors may be occluded at the extremes, it is not possible to determine the preshape space and thus, compute the mean shape as in (6.10).

Statistical Shape Models

As discussed in Chapter 2, in a statistical linear model of shape, an observed shape vector $\mathbf{x}_m$ follows

$$\mathbf{x}_m = r_m(\boldsymbol{\mu} + \boldsymbol{\delta}_m) + \mathbf{1}t_m, \tag{6.11}$$

where r_m and t_m are the Procrustes registration parameters to minimize the squared distance between $\mathbf{x}_m$ and the mean shape $\boldsymbol{\mu}$, and $\boldsymbol{\delta}_m$ is a realization of the random vector that models shape variability, typically Gaussian $\boldsymbol{\delta}_m \sim \mathcal{CN}(\mathbf{0}, \boldsymbol{\Sigma})$ [37].

6.2 DTW-based solution

Our goal is to determine a Procrustes registration of shape vectors when the correspondences are not available. Our approach proceeds along the following lines. We first use DTW to compute a dense correspondence between two shape vectors and thus perform a Procrustes registration. However, the distance in (6.1) is not robust against outliers, that is, spurious points of the vector that do not belong to the actual contour. These may occur, for example, when an edge detector selects parts of a neighboring object. Hence, we propose a probabilistic interpretation of the Procrustes registration in (6.1) to add robustness against the outliers in the assigned correspondences.

We then solve an overall minimization problem to determine simultaneously the registration and the correspondence allocation. In order to overcome the dependencies between registration and correspondence parameters, while aiming for a tractable solution, we propose an alternating optimization, where we find independently a solution for $\mathbf{p}$ (registration) and $\mathbf{C}$ (warping path of correspondence).

Furthermore, we formulate a group-wise registration approach, in which we calculate the mean shape and hence extend the generalized Procrustes analysis to deal with the lack of correspondence.

6.2.1 A probabilistic Procrustes registration

Let us recall the vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ with no assigned correspondences. After DTW we obtain the vectors $\tilde{\mathbf{x}}_1$ and $\tilde{\mathbf{x}}_2$, which are in one-to-one correspondence and have the same length L. Let us define the distance vector $\tilde{\mathbf{d}} = \tilde{\mathbf{x}}_1 - (r\tilde{\mathbf{x}}_2 + \mathbf{1}t)$, where r and t are the pose parameters of the registration we need to obtain. Let $\tilde{d}_l$ denote the elements in the vector $\tilde{\mathbf{d}} = [\tilde{d}_1, \ldots, \tilde{d}_L]^T$. Thus, $\tilde{d}_l = \tilde{x}_1[l] - (r\tilde{x}_2[l] + t)$ is the distance between the lth corresponding point, whose contribution should be considered in the registration calculation only if there is true correspondence between $\tilde{x}_1[l]$ and $\tilde{x}_2[l]$. Therefore, we give a probabilistic interpretation to $\tilde{d}_l$.

We model the probability of correspondence as a Bernoulli random variable γ_l that takes value 1 when there is correspondence, so

$$\gamma_l = \begin{cases} 1 & \text{with probability } w_l \text{ (correspondence)} \\ 0 & \text{with probability } 1 - w_l \text{ (no correspondence).} \end{cases}$$

Therefore, the registration distance to consider is $\tilde{d}_l\gamma_l$, which is random. Instead of minimizing the sum of squared distances, as in (6.2), the proposed problem is to minimize the expected value of the sum of these squared random distances, that is

$$\mathbb{E}\left[\sum_{l=1}^{L}\tilde{d}_l{}^2\gamma_l\right] = \sum_{l=1}^{L}\tilde{d}_l{}^2\mathbb{E}[\gamma_l] = \sum_{i=1}^{L}\tilde{d}_l{}^2 w_l = \tilde{\mathbf{d}}^H\mathbf{W}\tilde{\mathbf{d}}, \tag{6.12}$$

which results in a weighted least-squares formulation, where $\mathbf{W}$ is a diagonal matrix whose diagonal elements are the probabilities $w_1, w_2, \ldots, w_L$, which will be determined below.

The pose vector $\mathbf{p}^\star = [r^\star \quad t^\star]^T$ that minimizes the sum of probabilistic distances defined in (6.12) is the solution to the weighted least-squares problem

$$\mathbf{p}^\star = \arg\min_{\mathbf{p}} \tilde{\mathbf{d}}^H\mathbf{W}\tilde{\mathbf{d}}\,, \tag{6.13}$$

which is $\mathbf{p}^\star = (\tilde{\mathbf{X}}_2^H\mathbf{W}\tilde{\mathbf{X}}_2)^{-1}\tilde{\mathbf{X}}_2^H\mathbf{W}\tilde{\mathbf{x}}_1$ with $\tilde{\mathbf{X}}_2 = [\tilde{\mathbf{x}}_2 \quad \mathbf{1}]$.

6.2.2 Determining the weights

In order to obtain the correspondence probabilities, i.e., the weights w_l, we first transform $\tilde{\mathbf{x}}_1$ and $\tilde{\mathbf{x}}_2$ into their respective preshapes $\tilde{\boldsymbol{\tau}}_1$ and $\tilde{\boldsymbol{\tau}}_2$, as defined in (6.9). To this end, we assume that the shape vectors $\tilde{\mathbf{x}}_1$ and $\tilde{\mathbf{x}}_2$ belong to the same statistical shape model as defined in (6.11). This model is determined by the mean $\boldsymbol{\mu}$ and a deformation random vector $\boldsymbol{\delta}$. Assuming there is no prior information about the shape model, we fix the preshape of $\tilde{\mathbf{x}}_1$, i.e., $\tilde{\boldsymbol{\tau}}_1$, to be the mean. Then, the preshape of $\tilde{\mathbf{x}}_2$, i.e., $\tilde{\boldsymbol{\tau}}_2$, can be expressed in terms of $\tilde{\boldsymbol{\tau}}_1$ as

$$\tilde{r}\tilde{\boldsymbol{\tau}}_2 = \tilde{\boldsymbol{\tau}}_1 + \boldsymbol{\delta}', \tag{6.14}$$

where $\tilde{r} = \exp(-j\arg(\tilde{\boldsymbol{\tau}}_1^H\tilde{\boldsymbol{\tau}}_2))$, and $\boldsymbol{\delta}' = [\delta'_1, \ldots, \delta'_L]^T$ is an obervation of the random vector of shape deformation. Gaussian deformation models are typically used in shape analysis to model non-rigid deformations [32]. Hence, we choose $\boldsymbol{\delta}' \sim \mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I})$, where $\mathbf{I}$ is the identity matrix. Therefore, each δ'_l is assumed to be independent and identically distributed (i.i.d.) as complex Gaussian with mean zero and variance σ^2. Real and imaginary parts are independent, each with variance $\frac{\sigma^2}{2}$. We may thus employ a chi-squared test since the normalized distances Δ_l, for $l = 1, \ldots, L$, are chi-squared distributed with two degrees of freedom, i.e.,

$$\Delta_l = \frac{2|\delta'_l|^2}{\sigma^2} \sim \chi^2_2. \tag{6.15}$$

Therefore, the weights w_l may be determined from the cumulative chi-squared distribution function as

$$w_l = 1 - \int_0^{\Delta_l} \frac{e^{-u/2}}{2\Gamma(1)}du. \tag{6.16}$$

Since the elements of $\boldsymbol{\delta}'$ are i.i.d., we estimate σ^2 as the sample variance, that is, $\hat{\sigma}^2 = \frac{1}{L}(\tilde{r}\tilde{\boldsymbol{\tau}}_2 - \tilde{\boldsymbol{\tau}}_1)^H(\tilde{r}\tilde{\boldsymbol{\tau}}_2 - \tilde{\boldsymbol{\tau}}_1)$.

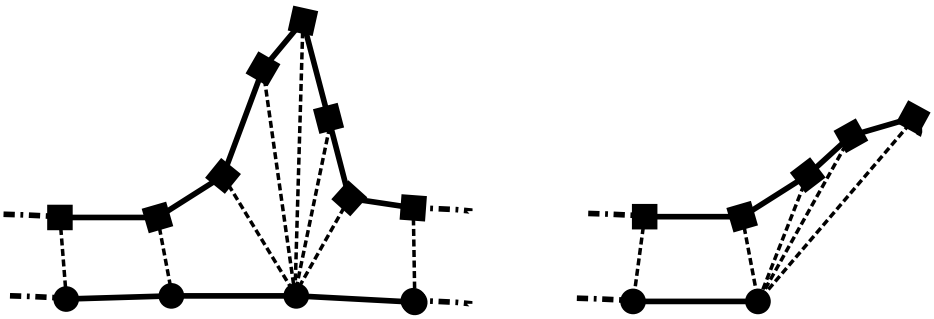


Figure 6.4: Two examples of arc-length fluctuations. Left: There is a many-to-one correspondence due to local differences in the arc-length. Right: The squares-line is longer than the circles-line at the end, and thus many points from the squares-line correspond to a single point from the circles-line.

6.2.3 Soft boundary condition

One of the constraints in DTW is the boundary condition, which requires that the beginning and end points of the two contours are always in correspondence. Therefore, when dealing with contours of different lengths, the DTW algorithm assigns the same correspondences to the additional points at the extremes, as illustrated in Fig. 6.4 (right), generating many-to-one correspondences at the start or end of the warping path. These many-to-one correspondences represent a missing part of the contour and should not influence the weighted least-squares minimization in (6.13). We detect these situations by analyzing the corresponding values in the warping path matrix $\mathbf{C}$ at the start ($l = 1$) and end ($l = L$). Then, we assign zero weight to any repeated corresponding points. This can be achieved with Algorithm 6.1.

Algorithm 6.1 Adjusting the weights to account for many-to-one correspondences in $\mathbf{W}$ that are due to the soft boundary condition.

Inputs: $\mathbf{W}$ with diagonal elements $w_l = 1 - \int_0^{\Delta_l} \frac{e^{-u/2}}{2\Gamma(1)} du \quad \forall l = 1, \ldots, L$ (as in (6.16)).
Path vectors $\mathbf{c}_l = (n_1^{(l)}, n_2^{(l)}) \quad \forall l = 1, \ldots, L$.
for $i = \{1, 2\}$ **do**

1. $l' = \arg\max_l (n_i^{(1)} = n_i^{(l)})$, $w_1, \ldots, w_{l'-1} = 0$

2. $l'' = \arg\min_l (n_i^{(L)} = n_i^{(l)})$, $w_{l''+1}, \ldots, w_L = 0$.

end

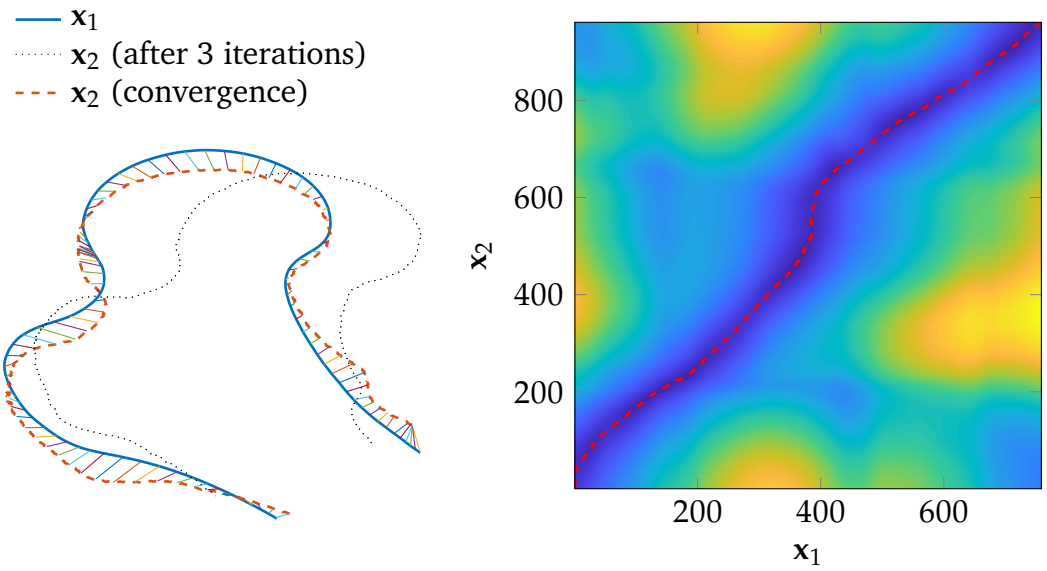


Figure 6.5: Left: Result from Algorithm 2 after 3 iterations (black, dotted) and converged result (red, dashed). Right: DTW cost function for every pair of points, and corresponding warping path of correspondence between the two boundaries after convergence.

6.2.4 Simultaneous pose and correspondences estimation

The whole procedure of jointly estimating the Procrustes registration and correspondences works as follows. We solve the minimization problem

$$[r^\star, t^\star, \mathbf{C}^\star] = \arg\min_{r,t,\mathbf{C}\in\mathbb{P}} (\tilde{\mathbf{x}}_1 - \tilde{\mathbf{x}}_2)^H \mathbf{W} (\tilde{\mathbf{x}}_1 - \tilde{\mathbf{x}}_2) \tag{6.17}$$

iteratively with an alternating optimization. The solution is described in detail in Algorithm 6.2. In Fig. 6.5 we show one example of the warping path $\mathbf{C}$ as well as the resulting cost function and registration obtained by the proposed algorithm.

6.2.5 Group-wise solution

Within a set of M shape vectors with no assigned correspondences, we aim at minimizing the DTW distance between their mean $\boldsymbol{\mu}$ and their pose- and correspondence-corrected versions and hence extending the generalized Procrustes analysis in [108]. Starting from the proposed pair-wise solution, let us assume that $\mathbf{x}_1$ is a reference vector and $\mathbf{x}_2$ is a target vector and determine the corresponding points with respect to the reference vector. We define $\hat{\mathbf{x}}_2(\mathbf{x}_1)$, of length N_1, to contain those points of the target $\mathbf{x}_2$ that correspond to the points in the reference $\mathbf{x}_1$. Since $\mathbf{x}_1$ and $\mathbf{x}_2$ do not have to have the same lengths and due to possible many-to-one correspondences, we agree on the following:

Algorithm 6.2 Proposed algorithm to determine simultaneously the registration and correspondences based on dynamic time warping.

Input: $\mathbf{x}_1 \in \mathbb{C}^{N_1}$ (reference) and $\mathbf{x}_2 \in \mathbb{C}^{N_2}$ (target).
Result: Correspondence $\mathbf{C}$ and pose $\{r, t\}$.
initialization: $i = 0$, $c > c_{\min}$, $\mathbf{y}^{(0)} = \mathbf{x}_2$
while $c > c_{\min}$ *and* $i < i_{\max}$ **do**

1. Find correspondence $\mathbf{C}$ between $\mathbf{y}^{(i)}$ and $\mathbf{x}_1$, as defined in (6.6).
2. Use $\mathbf{C}$ to determine $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{x}}_1$.
3. Calculate $\mathbf{W}$ as in Algorithm 6.1.
4. Find pose parameters $\mathbf{p} = [r \quad t]^T$ as in (6.13).
5. Determine $\mathbf{y}^{(i+1)} = r\mathbf{y}^{(i)} + \mathbf{1}t$.
6. $c = ||\mathbf{y}^{(i)} - \mathbf{y}^{(i+1)}||^2$, $i = i + 1$

end

- If there is a many-to-one corresponding set of points such as in Fig. 6.4 (left), and there are more points in the target than in the reference, we keep only the point that is closest to the mid-point of the fluctuation.
- If the many-to-one correspondences occur the opposite way (more points in the reference than in the target) we assume the points are repeated, unless they appear at the extremes (as in Fig. 6.4, right). In such a case, in order to force the length of $\hat{\mathbf{x}}_2(\mathbf{x}_1)$ to be N_1, these points are labeled as "empty".
- If the target is longer at the extremes than the reference, we delete the extra points in the target.

We need to register a set of shape vectors to its estimated mean $\boldsymbol{\mu}$. Following the above agreements, for each of the M training vectors $\mathbf{x}_m$, $m = 1, \ldots, M$, we determine the vector $\hat{\mathbf{y}}_m(\boldsymbol{\mu}) = r_m \hat{\mathbf{x}}_m(\boldsymbol{\mu}) + \mathbf{1}t_m$, which is registered to the mean $\boldsymbol{\mu}$ with one-to-one correspondence. Then, the mean shape is estimated as

$$\boldsymbol{\mu}^\star = \arg\min_{\boldsymbol{\mu}} \sum_{l=1}^{M} ||\boldsymbol{\mu} - \hat{\mathbf{y}}_m(\boldsymbol{\mu})||^2. \tag{6.18}$$

Notice that $\boldsymbol{\mu}$ is a preshape. In order to register each $\tilde{\mathbf{x}}_m$ to the mean $\boldsymbol{\mu}$ we need to determine

$$[r_m^\star, t_m^\star, \mathbf{C}_m^\star] = \arg\min_{r,t,\mathbf{C}\in\mathbb{P}} (\boldsymbol{\mu} - \hat{\mathbf{y}}_m(\boldsymbol{\mu}))^H \mathbf{W}_m (\boldsymbol{\mu} - \hat{\mathbf{y}}_m(\boldsymbol{\mu})). \tag{6.19}$$

The expression in (6.19) is solved for all $m = 1, \ldots, M$. As expected, the estimation of the mean in (6.18) depends on the extraction of the registration parameters in

(6.19). This is handled in an iterative process. We describe this method in detail in Algorithm 6.3. Notice that the points that are labeled as “empty" are not used when computing the mean.

Algorithm 6.3 Group-wise correspondence and registration as an extended generalized Procrustes analysis.

Input: $\mathbf{x}_m$, for $m = 1, \ldots, M$, with lengths N_m and no assigned correspondences.
Result: Correspondence $\mathbf{C}_m$, poses $\{r_m, t_m\}$ and mean $\boldsymbol{\mu}$.

initialization: $i = 0$, $c > c_{\min}$
$\boldsymbol{\mu}^{(0)} = \boldsymbol{\tau}_m$ (preshape of any $\mathbf{x}_m$, preferably the longest)
while $c > c_{\min}$ *and* $i < i_{\max}$ **do**
  for $m = 1, \ldots, M$ **do**
    1. Use Algorithm 6.2 to find $\mathbf{C}_m$, r_m and t_m (inputs: $\boldsymbol{\mu}^{(i)}$ and $\mathbf{x}_m$).
    2. Determine $\mathbf{y}_m = r_m \mathbf{x}_m + \mathbf{1} t_m$.
    3. Calculate $\hat{\mathbf{y}}_m(\boldsymbol{\mu})$ using $\mathbf{C}_m$.
    4. $\mathbf{x}_m = \mathbf{y}_m$.
  end
  1. Compute $\boldsymbol{\mu}^{(i+1)}$ with elements $\mu^{(i+1)}[n] = \frac{1}{M_n} \sum_{m \in \mathcal{M}_n} \hat{y}_m(\boldsymbol{\mu})[n]$ where $\mathcal{M}_n$ is the set of indices such that $\hat{\mathbf{y}}_m(\boldsymbol{\mu})$ is not empty, and M_n the number of such indices.
  2. $c = ||\boldsymbol{\mu}^{(i+1)} - \boldsymbol{\mu}^{(i)}||^2$, $i = i + 1$
end

6.3 Results and discussion

We validate our proposed strategy with contours of the femur extracted from fluoroscopic (low dose, and thus low quality) X-ray images. In this context, a registration technique may be used to design a statistical shape model of the femur for automatic segmentation [132]. Our database contains manually traced contours of both proximal and distal sides of the femur. We also apply our technique to an open source database of hand images [133], where the contours are automatically extracted with conventional edge detectors.

In this section, we study the accuracy of the registration of a pair of shape vectors in comparison with the competing strategies described in Section 6.3.2. We also compare our group-wise approach with a generalized Procrustes analysis based on

manual correspondences, since this is the typical solution for designing statistical shape models.

6.3.1 Data description

Proximal femur

This set contains contours from the proximal femur extracted from fluoroscopic X-rays images. These images are in anterior-posterior orientation and acquired with a C-arm during surgery implanting a cephalomedullary nail for osteosynthesis, which treats fractures of the femurOur X-ray images therefore also show implants and surgical tools (nail, blade, or k-wire), as seen in Fig. A.1. We collected 350 manually drawn boundaries that contain between 500 and 1000 points, are one pixel wide, and 8-connected. On those contours, 9 landmarks are manually annotated. More details about these images can be found in the Appendix.

Distal femur

This set contains contours from the distal part of the femur, also extracted from fluoroscopic images from the same surgical interventions as with the proximal femur [109]. The images show the femur contour and the medial condyle. We collected 116 manually drawn contours and on those manually placed two landmarks that delimit the condyle. The contours contain between 450 and 950 points and are one pixel wide and 8-connected. We show some examples in Fig. A.1. More details can be found in the Appendix.

Hands

The set contains 1000 closed-form contours that are automatically extracted with conventional edge detectors from natural images of hands. The starting point of the contour is automatically obtained as approximately located around the wrist [133]. The contours contain between 1200 and 2200 points. This data set does not contain manually annotated landmarks.

6.3.2 Competing techniques

We present an overview of the competing strategies that we consider in our experiments.

Manual correspondences

We consider a Procrustes registration based on manually determined correspondences. This is the typical registration to train statistical shape models [20]. These correspondences are determined using the available manually annotated landmarks on the femur. In order to account for more corresponding points, a fixed number of

equidistant landmarks is additionally extracted between the manual landmarks [35]. This process is time-consuming and error-prone because the manual landmarks were tedious to annotate and difficult to define. Due to the C-arm movements, ensuring consistency of the annotated landmarks is very challenging. Also, anatomically meaningful points, such as “the most proximal point to greater trochanter", are difficult to determine in the image or may not even lie on the bone boundary. With manual correspondences the registration parameters are obtained by a least-squares minimization as defined in (6.2).

Coherent Point Drift [113]

This iterative algorithm, based on ICP, models the points on the contours as Gaussian mixtures in order to add robustness to the registration. In each iteration, a correspondence matrix that accounts for every possible pair of correspondences is computed and used to determine the registration parameters.

Robust Iterative Closest Point [114]

This is an iterative algorithm based on ICP. It enhances the robustness of ICP through M-estimators. The correspondences are calculated in a way similar to the CPD strategy.

6.3.3 Test of performance

We do not have available a ground truth of the registration and correspondence parameters. Therefore, to evaluate the accuracy of registration, we consider two different metrics, each with their own particular limitations. To test the robustness against the presence of outliers, we design a test in which we simulate typical errors that an automatic edge detector might introduce into the segmentation. Additionally, since DTW requires ordered input signals, we study the impact of an unknown topology with unordered inputs. Finally, we consider the total variance of the group-wise registration as a quality measure.

Accuracy of pairwise registration

Since there is no ground truth of the registration and correspondence parameters, we consider the following performance metrics. First, the sum of minimum distances between a reference vector $\mathbf{x}_{\text{ref}} \in \mathbb{C}^{N_r}$ and a target vector $\mathbf{x}_{\text{target}} \in \mathbb{C}^{N_t}$, that is

$$d_{\text{test}} = \frac{1}{N_t} \sum_{n_t=1}^{N_t} \min_{n_r=1,\dots,N_r} |x_{\text{target}}[n_t] - x_{\text{ref}}[n_r]|. \tag{6.20}$$

This distance can be interpreted as a modified Hausdorff distance [133]. The interpretation of d_{test} may be misleading, for instance, in the registration of the hand contours, where a short distance may be achieved if a finger in the target is aligned

to an incorrect finger in the reference. We therefore also evaluate a second metric: the intersection-over-union (IoU) of the areas that the contours cover after the registration [134]. However, this metric may also be misleading for the evaluation of incomplete contours, as is the case in our femur database, since a perfect registration may represent a small intersection area with respect to the union area. Unfortunately, there does not seem to be one single quantity that perfectly measures the accuracy of registration in this problem.

Outliers

Outliers are points that do not belong to the contour or have been distorted and disrupt the shape. The femur contours in our database are almost free of outliers because they were manually obtained. The competing strategies [113] and [114] are designed to deal with outliers and noisy contours. In order to compare the robustness of our strategy to the competing techniques, we perform a test where we add outliers artificially to our database by emulating the distortion possibly induced by automatic segmentation. First, we add a noise component to the shape vector $\mathbf{x}_m$, i.e., $\mathbf{z}_m = \mathbf{x}_m + \boldsymbol{\gamma}$, with $\boldsymbol{\gamma} \sim \mathcal{CN}(0, \sigma_n^2 \mathbf{I})$, to consider a small noisy deformation. Then, for each noisy shape vector, $\mathbf{z}_m$, we randomly contort 10 segments of random length l_s, in different sections of the contour, each starting at a random index n_s. Let us denote the points in these segments as $\mathbf{z}_{\text{outliers}}^{(s)}$, for $s = 1, \ldots, 10$. We displace the points from the contour emulating the typical errors of an edge detector occurring when an edge from a neighboring structure is detected instead of the true contour. The resulting displaced and distorted segment is

$$\mathbf{z}_{\text{outliers}}^{(s)} = \mathbf{z}^{(s)} + \mathbf{1}\beta_s, \tag{6.21}$$

where $\mathbf{z}^{(s)} = \left[z_m[n_s], \ldots, z_m[n_s + l_s]\right]^T$ and $\beta_s \sim \mathcal{CN}(0, \sigma_t^2)$. We choose σ_t^2, σ_n^2 and the range of l_s by visual inspection, such that the resulting contours look realistic. We show an example of a contour from the proximal femur database affected by such displacements and distortions in Fig. 6.7, where $\sigma_t = 12$ pixels, $\sigma_n = 1$ pixel, and l_s is a uniform random value between 1 and 10 percent of the total length of the contour.

Unknown order

DTW requires that the order of the entries of each vector follow the contour, with the first entry corresponding to the start point and the last entry corresponding to the end point. In our database this is the case because the contours were manually determined. However, when the contours are automatically extracted, e.g. with an edge detector, this order may be unknown. The inputs to the competing techniques are point clouds, with no given topology. In order to determine to what extent the performance of our technique depends on an a priori known order, we propose the following test: We randomly shuffle the points in the shape vectors so that the order is unknown and add the following preprocessing step to estimate the order before

applying our registration technique. Start and end points are determined heuristically. Since we assume that the target and reference follow the same shape model, we have not contemplated mirroring. We may assume w.l.o.g. to follow the contour in an anti-clockwise direction, and the start and end points are assigned based on their proximity to the image border. Then, we use alpha shapes [135] to calculate an approximate order and reorder the points accordingly.

Performance of the group-wise registration

Following the properties of a good shape model described in [102] and [11], we evaluate the total variance of the group-wise result as a quality metric. We compare our proposed extended generalized Procrustes alignment to a regular generalized Procrustes alignment, which is based on landmarks whose correspondence was manually determined.

6.3.4 Procrustes registration of two shape vectors

We first discuss the accuracy of the pair-wise registration. We evaluate the distance of registering each shape vector to the longest vector in each data set. We measured the described distance d_{test} and the IoU. We show the resulting boxplots in Fig. 6.8. We considered the femoral shapes from our database (proximal and distal) as well as the hand shapes. Our strategy performs best median results among all competing techniques, for all three data sets, in terms of both metrics d_{test} and IoU. It also has the smallest variability in performance as measured by these metrics. Our technique even outperforms the manually determined correspondences (where available). We show one example of the registration results in Fig. 6.9, where our strategy performed a much more accurate registration.

6.3.5 Outliers and unknown order

To determine how well the techniques handle outliers and possibly unknown order of points, we perform two further registration tests for the proximal femur contours.

Test with outliers

In Fig. 6.10 (labeled as “Outlier") we show a boxplot of the registration results with the outlier model discussed in Section 6.2.2 in terms of d_{test}. We observe that our strategy, even though it experiences some loss of accuracy with respect to the outlier-free result, still outperforms the competing strategies. Moreover, if the proposed weighted least-squares minimization in (6.13) is substituted with an ordinary least-squares approach, i.e., $\mathbf{W} = \mathbf{I}$, there is some loss in performance. Hence, the weighted least-squares registration adds robustness.

Test with unknown order

In Fig. 6.10 (labeled as “Unsorted") we present the results of a test in which the points in the vector were shuffled. As expected, this has some effect on our technique but not the competitors since those do not use a prior point ordering. However, our strategy still outperforms the competition, and the weights in (6.13) enhance the accuracy.

Test with outliers and unknown order

In Fig. 6.10 (labeled as “Outliers + Unsorted") we show the results of the combined effects of outliers and unsorted vectors. Our proposed registration still outperforms all competing techniques.

6.3.6 Group-wise registration

We performed a group-wise registration of each of our three shape data sets. We show the results of the obtained mean shape in Fig. 6.11, which look as expected for each dataset. The estimate is worst at the extremes of the femur contours due to a smaller number of samples in these regions. For the hands, the wrist experiences more outliers and variability [133].

We compare the result of our group-wise registration to a registration based on manual correspondences. We performed this study with the femoral data sets, where manual landmarks are available. We show the qualitative results in Fig. 6.12. In order to improve the visualization, we show only 40 points, equidistantly chosen on the registered contours. We see that the proposed registration provides a more compact representation. As a quantitative metric we also evaluate the total variance of the points, which is the trace of the sample covariance matrices of the registered vectors. This metric is used to quantify the quality of the correspondences in [11]. The ratio between the total variances of the proposed and the manual models is 0.05 for the femoral head, and 0.1 for the condyle. This means that the total variance of the proposed registration is one order of magnitude smaller, and hence more compact and better registered [11].

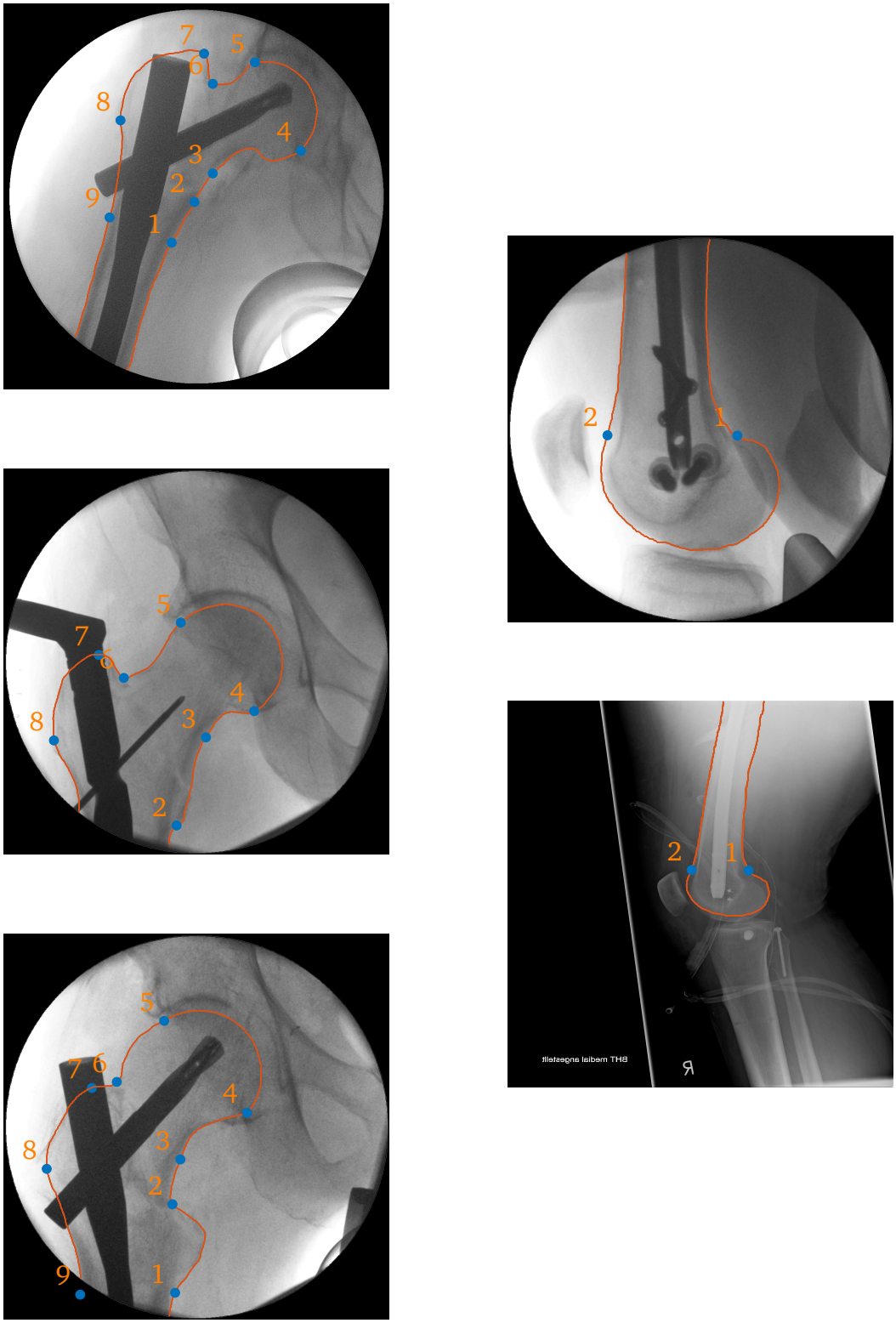


Figure 6.6: Fluoroscopic images of the femur in our database. The proximal femur (first three images) contains 9 manually annotated landmarks: the first and second images have different lengths due to movements of the C-arm; the third image differs due to a more medio-lateral position. The distal femur (last two images) contains two manual landmarks, and the lengths of the visible shaft may be different. More information about the medical images source can be found in the Appendix.

Figure 6.7: Comparison of an original contour (left) and a contour with artificial distortion for the outlier test (right).

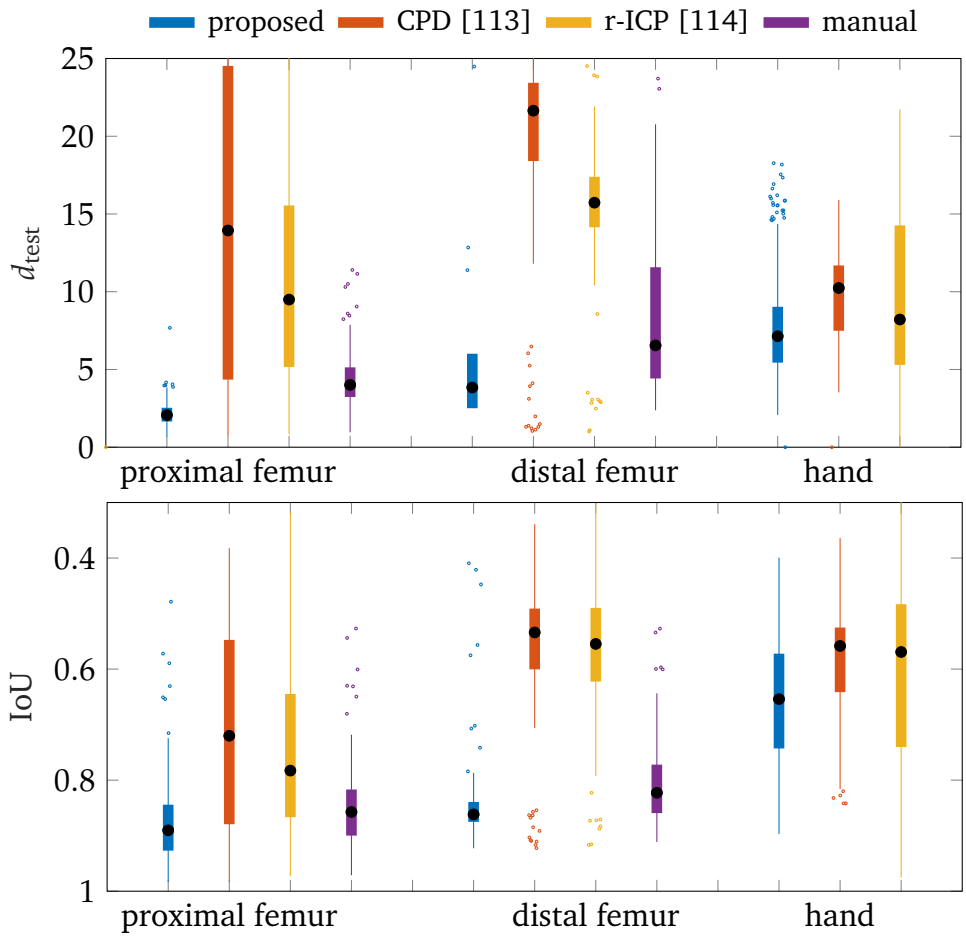


Figure 6.8: Boxplots comparing the registration results using our proposed strategy (blue), CPD [113] (red), r- (robust) ICP [114] (yellow), and manually determined correspondences (purple). There are no manual correspondences available for the hand shapes. The left plot shows d_{test}, which is measured in pixels (in the fluoroscopic images the pixel size is ≈ 0.45 mm, and in the hand images ≈ 0.5 mm [133]). The right plot shows IoU. In Algorithms 6.2 and 6.3 we have employed $i_{\max} = 100$ (maximum number of iterations) and $c_{\min} = 10^{-4} \min_{m=1,\dots,M} ||\mathbf{x}_m||$ (tolerance stopping criterion).

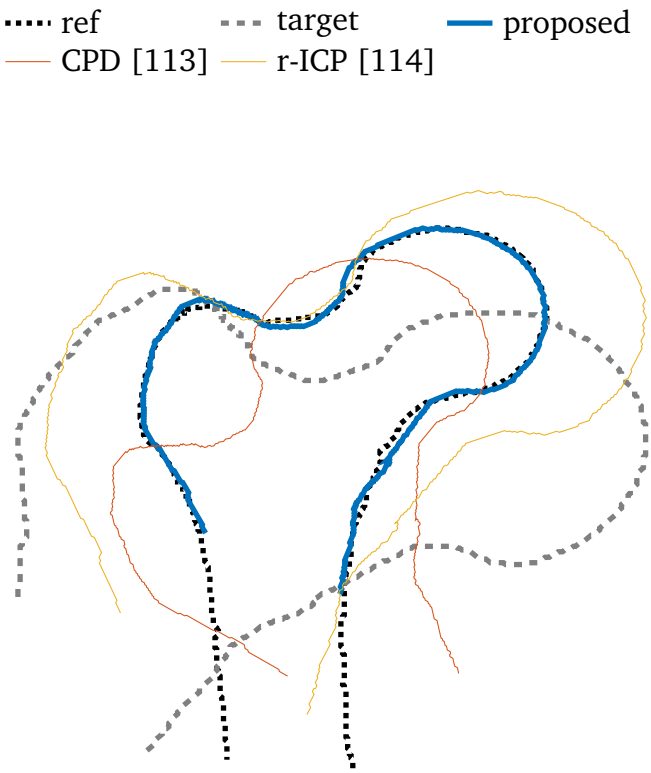


Figure 6.9: Comparative example of the proposed registration of the contour of the proximal femur. The reference (ref) and target have different lengths, sizes, and original positions.

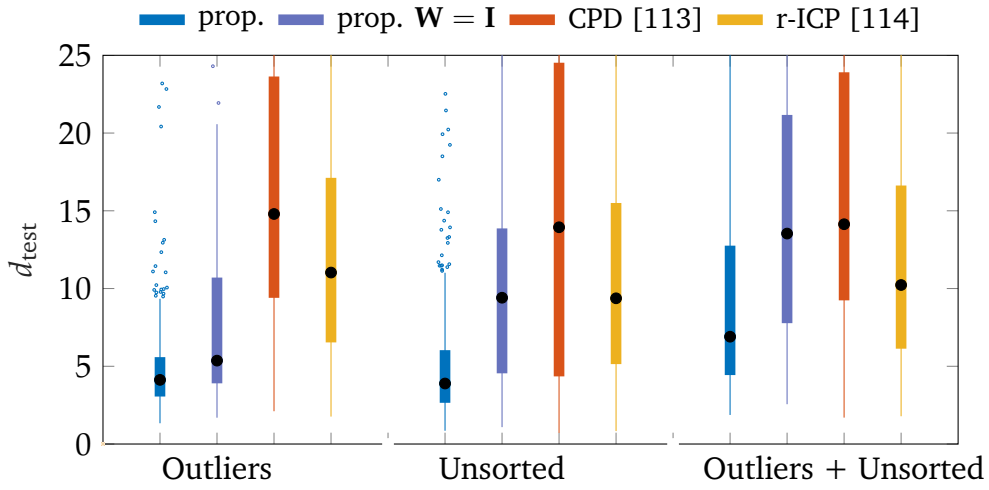


Figure 6.10: Boxplot of error metric d_{test} in pixels (pixel size ≈ 0.45 mm) for registration with outliers and/or without prior ordering, for the proximal femur database. We considered our proposed strategy (prop., blue) as well as an unweighted version of our proposed strategy (prop. $\mathbf{W} = \mathbf{I}$, violet).

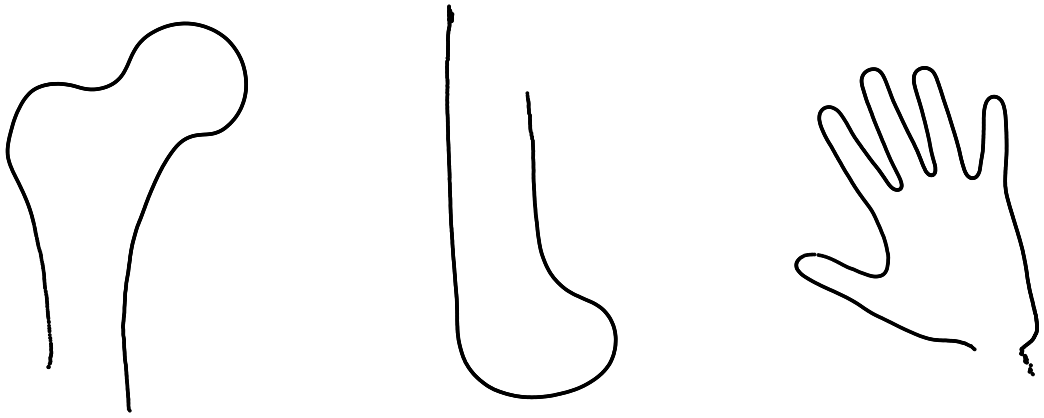


Figure 6.11: Estimated mean shape (proximal femur, distal femur, hand) obtained with proposed group-wise solution as described in Algorithm 6.3.

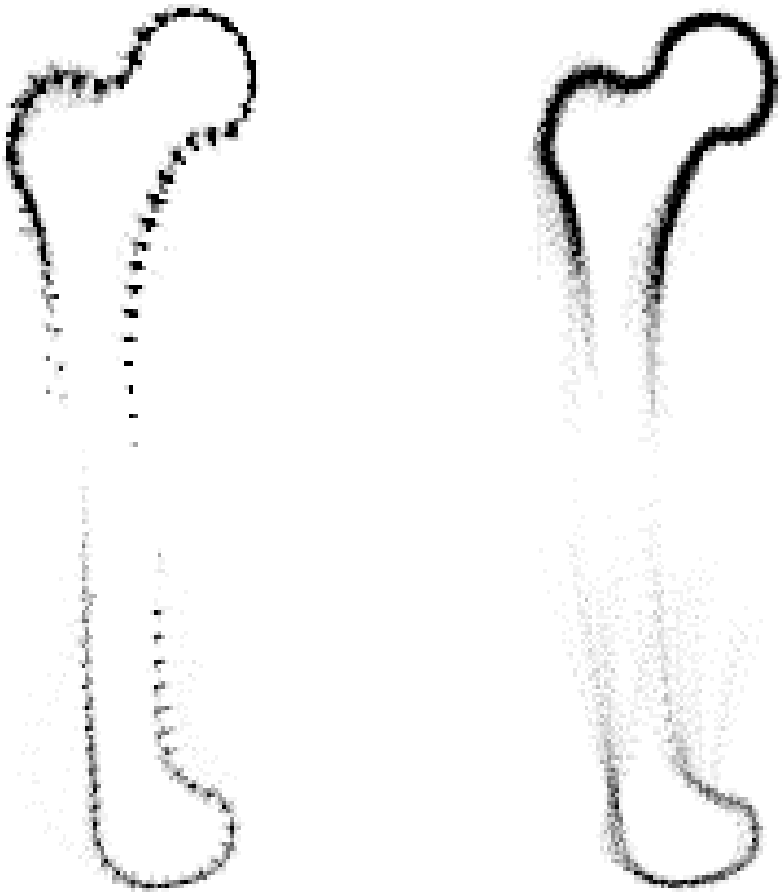


Figure 6.12: 40 equidistant points from proximal and distal femoral contours after applying the proposed Algorithm 6.3 (left images), and a typical generalized Procrustes analysis that is based on manual landmarks (right images).

Part III

Conclusions and future lines of research

Chapter 7
Conclusions and closing remarks

"No amount of experimentation can ever prove me right; a single experiment can prove me wrong."

— Albert Einstein

In this thesis we have studied how to improve SSM in order to enhance their segmentation performance and applicability. The goal of this enhancement is that the SSM will become more useful and more frequently applied as prior information in Image Processing, Machine Learning and Computer Vision.

We have particularly focused on medical imaging. Our first motivation was CAS and fluoroscopic images from the femur, where the femoral contour needs to be segmented. In this scenario, SSM-based segmentation algorithms did not provide enough accuracy for the medical application. Therefore, our starting point was to find the required modifications to make SSM-based segmentation algorithms work in these challenging image modalities like fluoroscopy. We investigated how to add robustness while keeping the simplicity of the algorithms, since real-time computation is critical in CAS applications. A good generalization to other shapes was also a feature to achieve: our proposed enhancement is not only applicable to the femur.

Additionally, we noticed the heuristic design of SSM parameters, such as as the model order and the position of the manual landmarks, and studied whether these heuristic selections were preventing SSM to provide their full benefit and potential. To this end, we designed a novel model-order selection based on information theory. Also, we introduced DTW in the generalized Procrustes analysis to avoid the manual correspondence from the landmarks.

Consequently, our contributions focused on these three points: add robustness to SSM-based segmentation techniques, select the model order more optimally, and avoid the need of manual landmarks during the training process. These proposed improvements and techniques are extensible to different anatomical parts, image modalities, or even other applications not related to medicine (Computer Vision in general).

As a closing remark, we provide the following outlook with the conclusions from each contributions of our work.

7.1 Add robustness, keep simplicity

Firstly, we observed that SSM-based segmentation algorithms did not perform accurately in fluoroscopic images, which need to be processed real-time to be used in CAS applications. These algorithms, such as ASM and CLM, are iterative. In each iteration, a least-squares fit of a plausible shape to some detected target landmarks is determined. Finding these targets is a critical step: some landmarks are more reliably detected than others, and some landmarks may not be within the field of view of their detectors

We proposed a GLS instead of an OLS fit as a straightforward and simple strategy to add robustness to the ASM algorithm. The weights in the GLS fit should not be selected only assuming heuristic confidence metrics of the landmarks, since these do not generalize and are prone to over-fitting. We chose the weights based on empirically determined errors of the least-squares fit, where the weighting matrix is determined as the inverse of the sample covariance of the residuals. This can be interpreted as a maximum likelihood solution of the least-squares problem. Additionally we used a chi-squared test to identify target landmarks that are likely to be incorrect and thus exclude them from the fit. Therefore, we combine prior knowledge about the performance of the target detectors in the weighting matrix designed with training residuals, with additional information about the unreachable landmarks (either occluded or out of alignment) in the chi-squared test.

We tested the strategy in fluoroscopic X-ray images of the femur taken in actual surgeries. We showed that our strategy outperforms the standard ASM as well as other weighting strategies. While our approach was based on a particular metric for identifying target landmarks (the Mahalanobis distance), our idea should generalize to other metrics as well.

7.2 The importance of the model order

Secondly, we studied the importance of the model order in statistical shape models, which determines how much of the variation seen in the training data is accounted for by PDM. A good choice of the model order depends on the number of training samples and the noise level in the training data set. Yet the most common approach for choosing the model order simply keeps a predetermined percentage of the total shape variation. We proposed a model-order selection strategy that is based on information-theoretic criteria and thus has a theoretical justification. Furthermore, this technique can be used in a more general source enumeration problem with observations in additive colored noise. We considered colored noise because it is more realistic than white noise, although it is not typically considered in source detection problems.

In this approach we interpreted the model selection in a different way. That is, in the typical PCA model-order selection, the way to determine the number of sources is to consider that the sources are random and the order is inferred from the covariance matrix of the observations. In our scenario, however, we assume that the sources are

deterministic, so the order only modulates the mean of the observations. Considering a generic model of colored noise, we provided a technique for determining the number of sources when the noise observations have an unknown covariance matrix.

We validated the selection performance of our technique on simulated data, and it outperformed other model-order selection strategies under different conditions of sample support and noise level. We have also evaluated the technique on shapes from real data sets, showing results similar to the evaluation with artificial data. Additionally, we performed an empirical test to illustrate the impact of the model order of shape models, and how the choice of order provided by our technique results on a model with better performance.

7.3 Registration without manual correspondences

Finally, in order to avoid the manual selection of landmarks, we investigated the problem of registration when there are not correspondence references. Solving the problem of the correspondence makes the registration of training contours possible when the landmarks are not available. Furthermore, an accurate registration is fundamental in segmentation and retrieval applications, as well as for learning shape models and atlases, which may be used not only in CAS but in other Computer Vision applications.

We solved simultaneously the correspondence and the Procrustes registration problem. The proposed correspondences are based on DTW, and a probabilistic alignment that adds robustness. DTW preserves the order of points on a contour, providing higher accuracy and preserving efficiency, which is the main difference with respect to the strategies based on the ICP algorithm. Our strategy outperformed competing approaches with similar computation time when tested on three different anatomical contours, even in the presence of outliers and with unordered points.

We also proposed a group-wise solution that can be used to train SSM. Our technique saves annotation time and does not require manually placed landmarks. It also provides a training set that is accurately registered, and hence more compact and meaningful shape models can be learned.

Chapter 8
Contributions under development

"The alchemists in their search for gold discovered many other things of greater value."

— Arthur Schopenhauer

During the development of this work we have discussed many interesting ideas that were under development at the time of writing this thesis. Chronologically, the last contribution of the thesis has been Chapter 6. This DTW-based registration of contours has provided the most promising results in terms of future work, and has motivated many interesting extensions. The first and most obvious one is a three-dimensional extension of this registration approach. The registration of three-dimensional point sets is very important in computer vision and machine learning, and therefore our DTW-based registration would be incomplete if this extension was not feasible. Therefore, we propose in Section 8.1 of this chapter a working solution for the extension, which we have been designing during the last few months prior to the submission of this thesis.

Another important issue when registering a collection of contours with DTW to train a shape model is that the number of points in the contour is typically high, since it is initially as dense as the pixel resolution of the training images. The SSM-based segmentation algorithms like ASM and CLM need to find the landmarks in new images in their iterative process, as described in Chapter 2. If the SSM contain too many points this search becomes very slow and may compromise the real-time requirements of, for instance, CAS applications. Therefore, we have been working on a landmark selection technique, based on sparse optimization, that chooses a small collection of landmarks from the initially dense contours. The criterion of selection is that the non-selected points are estimated from the selected ones using the SSM, and therefore they do not need to be searched as targets by the segmentation algorithms. Consequently, this technique can also be considered as a sampling approach. We present our preliminary results of this promising sampling technique in Section 8.2 of this chapter.

8.1 Procrustes registration of surfaces without correspondences

The ICP-based algorithms discussed in Chapter 6, such as CPD, are not specific only for two-dimensional contours. These techniques also work with three-dimensional point sets. These type of data is very common, since anatomy, archaeological pieces, and in general everyday objects are volumes. Therefore, their three-dimensional shape information has great potential in Image Processing and Machine Learning. For instance, a three-dimensional shape model of a bone can enhance prosthesis design for trauma surgery. Also, a CAS surgical scene can be controlled by a navigation system much better with knowledge about the volumes. Consequently, we have considered important to extend our proposed registration strategy in Chapter 6. In this chapter we present our three-dimensional approach to register surfaces with DTW, which is currently under development.

8.1.1 A three-dimensional extension of the DTW-based solution

The classical DTW algorithm deals with time series analysis. This means that, the signals are in the time domain, and therefore can be parametrized by a one variable: time. However, if we contemplate surfaces, these would need equivalently two parameters. The warping in this case needs to happen over a two dimensional space. There are solutions to high dimensional extensions of DTW. For instance, the authors in [136] consider a two-dimensional warping solved by dynamic programming. However, such an approach would increase the computational requirements of the registration, and the resulting technique may not be feasible in real time.

If the three-dimensional points inside a set are ordered under certain pattern, and this order only depends of one parameter, such as time-dependent series are ordered, the extension of the DTW becomes much simpler. The complex number cost function, based on the Euclidean distance, is substituted by a sum of Euclidean distances of points in $\mathbb{R}^3$. Then, an equivalent dynamic programming solution to the one in Chapter 6 can be applied to obtain the path of correspondence. Still, establishing a point ordering is a challenging task. We have implemented a heuristic approach to determine an ordering, being aware that better ordering criteria could be defined. Nevertheless, our proposed three-dimensional registration has outperformed the competing strategy in [113] in our preliminary results.

Our strategy consists in the following. Let us consider the Procrustes registration of the set $\mathbf{X}_2 \in \mathbb{R}^{3\times N_2}$ onto the set $\mathbf{X}_1 \in \mathbb{R}^{3\times N_1}$, with points $\mathbf{X}_1 = \left[\mathbf{x}_1[1], \ldots, \mathbf{x}_1[N_1]\right]$ and $\mathbf{X}_2 = \left[\mathbf{x}_2[1], \ldots, \mathbf{x}_2[N_2]\right]$. This is performed by the pose parameters $s \in \mathbb{R}^+$ (scale), $\mathbf{R} \in SO(3)$ (rotation), and $\mathbf{t} \in \mathbb{R}^3$ (translation) that minimize the trace of the difference between $\mathbf{X}_1$ and $\mathbf{X}_2$.

As a pre-processing step to arrange the points of each the point sets we consider the distance of these points to a reference point inside the volume, similarly as in Chapter 6. Once the points in the volumetric sets are ordered, the following registration problem

$$[s^\star, \mathbf{R}^\star, \mathbf{t}^\star, \mathbf{C}^\star] = \arg\min_{s,\mathbf{R},\mathbf{t},\mathbf{C}} \left(\sum_{l=1}^{L} ||\mathbf{x}_1[n_1^{(l)}] - (s\mathbf{R}\mathbf{x}_2[n_2^{(l)}] + \mathbf{t})||^2 \right), \tag{8.1}$$

wich can be solved with the DTW-based registration in Chapter 6. The main difference in the DTW is that each cost stage $D(n_1, n_2)$ is now computed considering the distance

$$d(n_1, n_2) = ||\mathbf{x}_1[n_1] - \mathbf{x}_2[n_2]||^2. \tag{8.2}$$

We show an example of our DTW-based registration in Fig. 8.1.

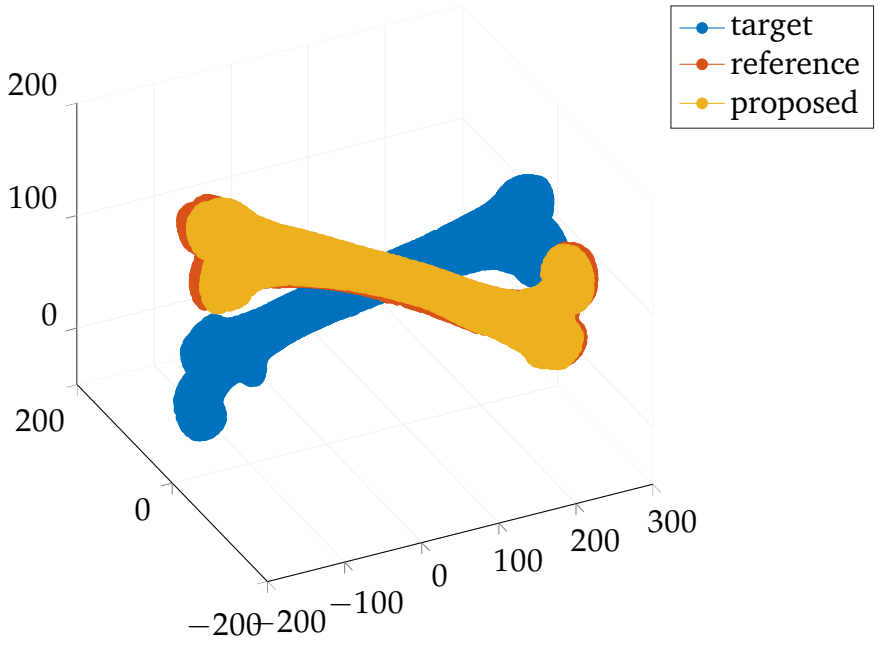


Figure 8.1: An example of a result of our proposed extension for the DTW-based registration strategy.

8.1.2 Preliminary results

We present these experimental results to show that the extension of the approach in Chapter 6 to a three-dimensional case is feasible. Therefore, we have considered a dataset of femoral volumes and applied our DTW-based registration strategy. We then compare the registration performance, in terms of accuracy and computation time, with the CPD algorithm in [113].

We use volume data of the complete femur bone. This set contains 50 surfaces composed of triangle meshes that belong to FutureLearn course on Statistical Shape Modelling database [137]. The volumes may correspond to either left of right leg. More details about this dataset can be found in the Appendix. We validate the DTW-based registration on these 50 meshes of the femur bone surface. We transform the

triangle meshes into points in $\mathbb{R}^3$ and evaluate the performance of our approach in terms of registration accuracy (with the so defined in Chapter 6 d_{test}), and compare against the CPD algorithm. We study the tendency of these two metrics (accuracy and computation time) with respect to the number of points considered in the volumes. For that purpose we uniformly downsampled the volumes.

In this test our approach outperforms CPD on average over the 50 samples in terms of accuracy, as shown in Fig. 8.2, even though the proposed order parametrization of the surface is very basic. A better parametrization could be based on an estimate of the gradients of the surface. Furthermore, our implementation (which is based on the MATLAB function *dtw.m*) performs the registration five times faster that the CPD MATLAB approach provided by the authors in [113].

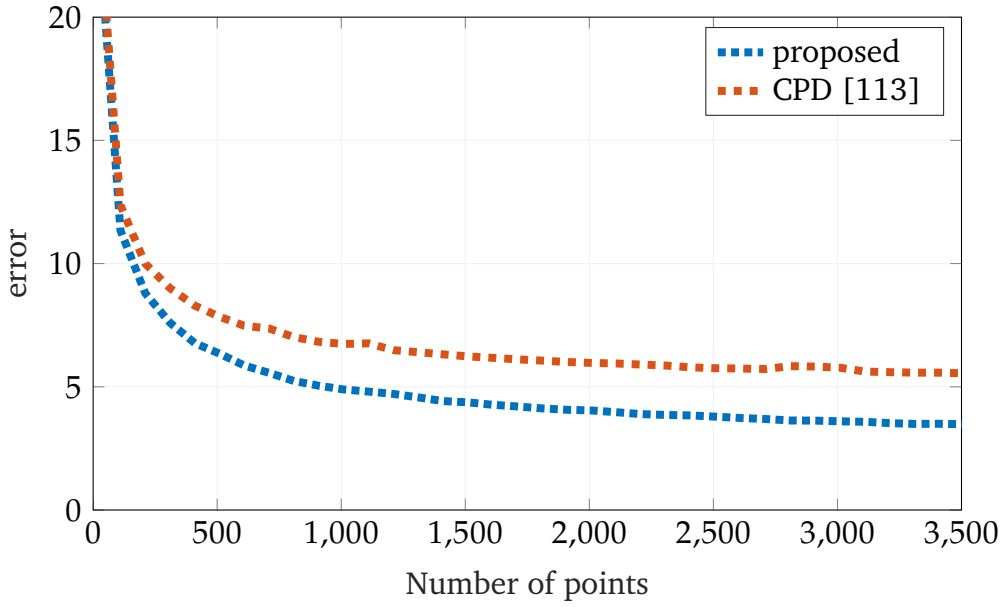


Figure 8.2: Accuracy error (d_{test}) with respect to the number of points in the point sets.

8.2 Target landmark selection via sparse optimization

The contour vectors to train SSM, which are group-wise registered with the DTW-based strategy proposed in Chapter 6, contain as many points as the resolution of the training images. This quantity of points may generate SSM fits that are too dense: if too many landmarks need to be searched to fit a shape model, the resulting segmentation algorithms may become computationally expensive and not feasible in real time. Therefore, a landmark selection needs to be performed. Still, SSM with many points provide good resolution. Consequently we need to find a good trade-off between neglecting and keeping points during the SSM fit.

We propose to consider a LMMSE to minimize the error of calculating the neglected values in the contour vector $\mathbf{x} \in \mathbb{C}^N$. If some landmarks can be efficiently estimated from a few points in $\mathbf{x}$, fewer landmarks need to be searched by segmentation algorithms, and a dense SSM becomes more efficient. We use a sparse optimization to select n landmarks from the initially N points in $\mathbf{x}$ so that only n targets need to be searched for the SSM fit.

8.2.1 A sampling method based on sparse optimization

As explained before, the motivation of this technique is to reduce the computational complexity of each iteration of a segmentation technique, such as ASM, considering the search of only a few target landmark during the SSM fit. With an LMMSE estimation of the non-searched points the high resolution of the DTW-registered SSM is maintained. Let us assume we have M observations of the shape vector $\mathbf{x}$, that is $\mathbf{X} = \left[\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)}\right] \in \mathbb{C}^{N \times M}$. Using an LMMSE estimator, the problem can be cast as

$$\underset{\mathbf{A} \in \mathbb{R}^{N \times N}}{\text{minimize}} ||\mathbf{X} - \mathbf{A}\mathbf{X}||_{\mathrm{F}}^2 + \rho ||\mathbf{A}||_{2,0}, \tag{8.3}$$

where $\rho \geq 0$ is the sparsity parameter [138], which forces the sampling matrix $\mathbf{A} \in \mathbb{C}^{N \times N}$ to have a particular sparsity structure , i.e., some of the rows are zero valued as

$$\mathbf{A} = \begin{bmatrix} \cdots & \cdots & \cdots \\ \text{——} & 0 & \text{——} \\ \cdots & \cdots & \cdots \\ \text{——} & 0 & \text{——} \\ \cdots & \cdots & \cdots \end{bmatrix}. \tag{8.4}$$

This enforces the selection of n landmarks such that the linear estimation of the discarded $N - n$ landmarks is minimum. This problem, however, is highly non-convex, which makes it intractable. Following the lines in [138], we replace the ℓ_0-norm by the tractable surrogate function and consider MM framework, as the one described in Chapter 4. Therefore, (8.3) is replaced by

$$\underset{\mathbf{A} \in \mathbb{R}^{N \times N}}{\text{minimize}} \quad ||\mathbf{X} - \mathbf{A}\mathbf{X}||_{\mathrm{F}}^2 + \rho \sum_{j=1}^{N} g_p^{\epsilon}\Big(\sqrt{\sum_{i=1}^{N} |a_{ij}|^2}\Big), \tag{8.5}$$

where a_{ij} are the elements in the ith row and the jth columnd of matrix $\mathbf{A}$, and $g_p^{\epsilon}(\cdot)$ is any of the surrogate functions in [138].

8.2.2 Preliminary results

We consider the estimation error of the neglected points in the contour as a metric of comparative performance. We present a study on the estimation error of recovery of our proposed method. As competing strategies we use a uniform sampling and a landmark selection technique presented in [139].

As shape data we use two different contours in the analysis:

1. The femoral head contours that have been group-wise registered with the approach in Chapter 6.

2. The hand contours that have been group-wise registered with the approach in Chapter 6.

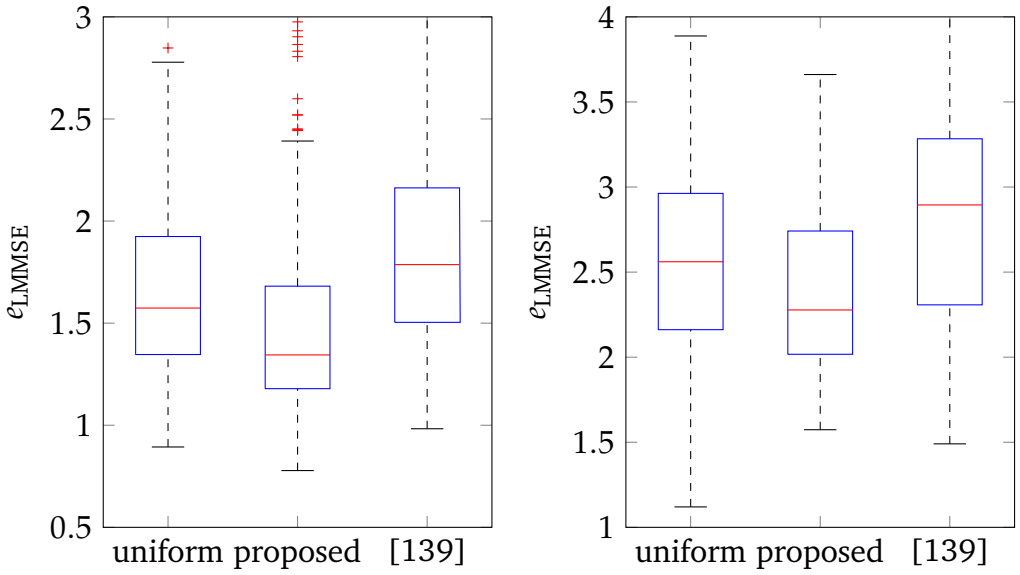
We have downsampled the contours before applying the selection techniques. Therefore, the input contours contain 100 points. The goal of the selection is to obtain 10 out of these 100 contour points. The metric considered is the estimation error of remaining 90 points. The sparsity parameter, that is ρ, is manually adjusted to select 10 points of the contour.

We perform a leave-one-out test to estimate the error, similarly as in Chapter 5. Let $\mathbf{x}_n$ denote the neglected landmarks during the sparse selection, and $\mathbf{x}_s$ the selected landmarks. Let $\hat{\mathbf{R}}_{ns}$ denote the sample cross-covariance matrix between the neglected landmark and the selected landmarks, and $\hat{\mathbf{R}}_{ss}$ the covariance matrix of the selected landmarks. The LMMSE estimator of the neglected landmarks from the selected landmarks is then $\hat{\mathbf{x}}_n = \hat{\mathbf{R}}_{ns}\hat{\mathbf{R}}_{ss}^{-1}\mathbf{x}_s$. We average the error over all neglected landmarks and obtain

$$e_{\text{LMMSE}}(m) = \sqrt{\frac{1}{N-n}||\hat{\mathbf{x}}_n^{(m)} - \mathbf{x}_n^{(m)}||^2}, \tag{8.6}$$

which is evaluated for all samples, i.e., $m = 1, \ldots, M$.

We present the boxplots of e_{LMMSE}, representing the errors for each of the techniques under analysis in Fig. 8.3. Our sparse optimization selection (proposed) outperforms the other techniques by approximately 0.5 pixels on average over the complete contour, as measured error in Fig. 8.3.



(a) proximal femur (b) hand

Figure 8.3: Boxplot of the performance metric e_{LMMSE} from all the leave-one-out iterations over all samples. Values are in pixels (pixel size is ≈ 0.45 mm, and in the hand images ≈ 0.5 mm).

Chapter 9
Future Work

"Science will always be a search, and never a true discovery. Science is the journey, never the destination."

— Karl Popper

During the development of this doctoral research, many promising ideas have been left open. Most of the contributions presented in Part II are related to the analysis of shape models: model order, group-wise registrations or optimal sampling. However, we did not invest as much time to develop strategies to fit of the models. For instance, in an ASM segmentation, finding better target landmarks is critical. It is a promising path of research to design new target landmark search approaches, as an alternative to the Mahalanobis distance of a one-dimensional profiles, to provide more accurate SSM fits and at the same time preserve real-time computation.

Deep Learning is also a very promising tool in image analysis. Although Deep Learning has been mentioned in Chapter 2, and the presented results in Part II also have applications in DNNs, its potential in shape analysis has not been studied in detail in this work, and hence an interesting area of research is still open.

In order to motivate and inspire the readers of this thesis, we describe in this chapter two promising techniques for future research.

1. A Universal Manifold Embedding [140] of local patches of the image to search and register target landmarks.

2. A Deep structured learning solution [141], which considers the statistical shape model as prior knowledge, and its influence has an impact during the the back-propagation of the error in the learning process.

9.1 A Universal Manifold Embedding to search landmarks

The target point search in segmentation algorithms an ASM or a CLM is very important. As described in Chapter 2, ASM target search is based on a one-dimensional line search and a Mahalanobis distance to a model of appearance of this line. This

strategy is very simplified, as well as dependent on the direction of the line of search. CLM target dectectors are based on features from local image patches around the landmarks. Still, these are dependent of the affine transformations happening between observations. A promising techniques to find the landmarks is DNN. However, we need an adequate number of training samples, which is frequently very difficult to obtain for some applications like medicine.

We propose a different alternative: the Universal Manifold Embedding (UME). UME provides a mathematical framework to deal with affine transformations in images: two images (or image patches) are projected to a subspace that is independent of affine deformations. Thus, target landmarks could be identified by means of these subspace projections. We illustrate this idea in Fig. 9.1.

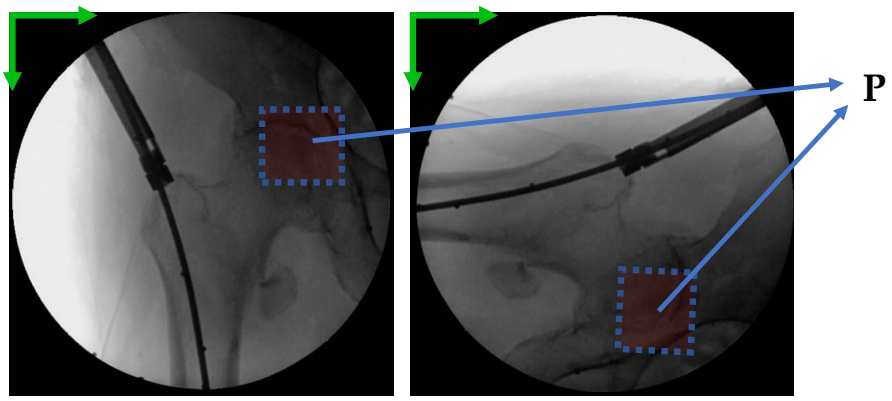


Figure 9.1: Considering the UME strategy in [142], we would project an image patch and an affined tranformed version of it into the same subspace **P**.

9.1.1 The Universal Manifold Embedding

As an encouragement to the reader about this research line, we provide an introduction about the UME. Let us consider two images f and g as functions $\mathbb{R}^2$ (pixel coordinates) to $\mathbb{R}$ (pixel gray value). Image g is the same image as f but after an affine transformation, that is,

$$h(\mathbf{x}) = g(\mathbf{A}^{-1}\mathbf{x} + \mathbf{c}), \tag{9.1}$$

where $\mathbf{x} \in \mathbb{R}^2$ are the pixel coordinates in the Euclidean space of the image, $\mathbf{A} \in \mathbb{R}^{2\times 2}$ is an invertible matrix that defines the affine transformation, and $\mathbf{c} \in \mathbb{R}^2$ a translation vector. Consequently, g is an observation of h undergoing an affine transformation [140]. Therefore, the coordinates of image $g(\mathbf{y})$ relate to the coordinates in $h(\mathbf{x})$ as

$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{d}$, and

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{y} + \mathbf{d} = \mathbf{D}\tilde{\mathbf{y}} \tag{9.2}$$

with $\mathbf{D} = [\mathbf{d}\ \mathbf{A}^{-1}] \in \mathbb{R}^{2\times 3}$ and $\tilde{\mathbf{y}} = [\mathbf{y}^T\ 1]^T$ and $\mathbf{y} = [y_1\ y_2]^T$. Thus, the concept of Universal Manifold Embedding consists on the following. Let us consider a set of $M \in \mathbb{N}$ Lesbegue measurable functions [140], that is, the set of function $w_m : \mathbb{R} \to \mathbb{R}$ with $m = 1, \ldots, M$. Thus, there is a mapping for image f such that

$$\mathbf{T}(f) = \begin{bmatrix} \sum\sum w_1 f(\mathbf{y}) & \sum\sum y_1 w_1 f(\mathbf{y}) & \sum\sum y_2 w_1 f(\mathbf{y}) \\ \cdots & \cdots & \cdots \\ \sum\sum w_M f(\mathbf{y}) & \sum\sum y_1 w_M f(\mathbf{y}) & \sum\sum y_2 w_M f(\mathbf{y}) \end{bmatrix}_{M\times 3}, \tag{9.3}$$

where matrix $\mathbf{T}(f)$ is rank 3, and $\sum\sum$ represents the sum of all the pixels in an image. One possible solution to w_l could be [143]

$$w_m f(\mathbf{y}) = \begin{cases} 1 & \text{at } f(\mathbf{y}) = m \\ 0 & \text{elsewhere} \end{cases} \tag{9.4}$$

and M the number of gray-level values in the images, e.g. 255. If we consider the matrix $\tilde{\mathbf{D}} = [\mathbf{e}_1\ \mathbf{D}^T]$ with $\mathbf{e}_1 = [1, 0, 0]^T$, then the image h, which is an observation of g undergoing an affine deformation, then the following equality applies

$$\mathbf{T}(g)|\mathbf{A}^{-1}|\tilde{\mathbf{D}} = \mathbf{T}(h). \tag{9.5}$$

Consequently $\mathbf{T}(g)$ and $\mathbf{T}(h)$ have the same column space. Projection matrices have one-to-one correspondence to subspaces [142]. Therefore, we can evaluate the projection matrices of $\mathbf{T}(g)$ and $\mathbf{T}(h)$ to determine if they correspond to the same subspace, and therefore represent the same image but with different affine deformations. The projection matrix $\mathbf{P}$ of $\mathbf{T}$ is determined as

$$\mathbf{P} = \mathbf{T}(\mathbf{T}^H\mathbf{T})^{-1}. \tag{9.6}$$

In the absence of noise, the distance between the projection matrices of $\mathbf{T}(g)$ and $\mathbf{T}(h)$, i.e., $\mathbf{P}_{\mathbf{T}(g)}$ and $\mathbf{P}_{\mathbf{T}(h)}$, is

$$d(h, g) = ||\mathbf{P}_{\mathbf{T}(g)} - \mathbf{P}_{\mathbf{T}(h)}||_F = 0 \tag{9.7}$$

However, the distance $d(h, g)$ is exactly zero if the only deformation between the images is affine, and if the images are noise free. Therefore, if order to determine if two images are affine related we may consider the distance to an average projection matrix of the appearance of that image under different deformation and noise observations. An approach to determine the mean projection matrix $\bar{\mathbf{P}}$ is proposed in [142].

9.1.2 Applications of the UME in the ASM-CLM algorithm

We could use the UME to find landmarks in the images. For instance, if we calculate the projection to the UME of a collection of patches in the image, we could find a target landmark, as illustrated in Fig. 9.2. This could be applied globally in the image, and find a small collection of landmarks to compute an initial guess for the shape fit. Also, it could be applied locally around the area of one landmark, and fine-tune the fit. We illustrate the global and local search in Fig. 9.3

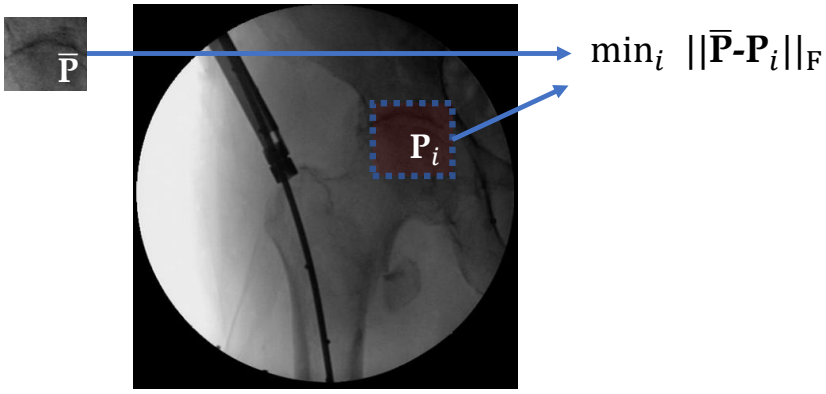


Figure 9.2: Illustration of the use of UME [142] to find a target landmark.

9.2 Deep Learning and shape models

In the era of Big Data, making use of deep learning together with shape models seems promising [144]. Deep learning has demonstrated to be a very powerful tool to solve problems involving image analysis. Neural networks approximate synthesis problems with a very non-linear parametric solution. This solution can be, indeed, constrained by models, as well as prior knowledge about the nature of the problem. SSM can be this prior knowledge [54] [55] and be used to regularize the solution provided by a deep neural net. Therefore, a good trade-off between model and data-based can be studied. There are promising research paths about this idea. Two examples are: how to incorporate SSM in the deep nets training and back-propagation process, and how to enhance SSM to provide benefits to the results. To give the reader an intuition of the answer to these questions, in this section we describe our future research line in this direction.

9.2.1 Deep Structured Active Shape Models

As mentioned in the previous section, a very challenging step in the segmentation problems based on ASM and CLM is to find the target landmarks. An interesting solution is to generate mappings of likelihood of the image to determine which pixel in is more likely to be a landmark. This problem has traditionally been solved with image gradients, wavelet analysis and machine learning with image features. Therefore, finding the target landmarks is very dependent on the image data, and probably composed by non-linear variability that is difficult to model accurately. However, the position of those targets has a particular structure from which we can learn a prior: they belong to a family of shapes. This prior can be provided by SSM.

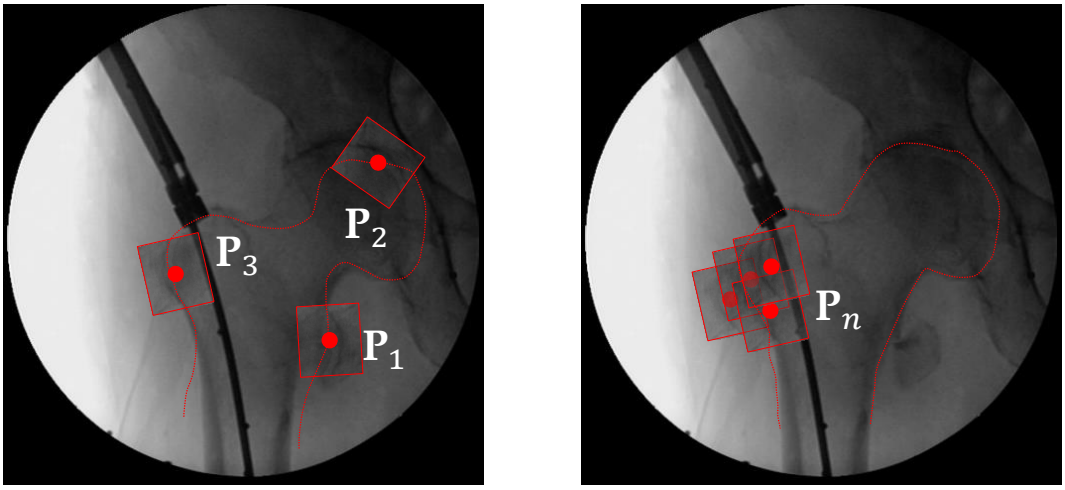


Figure 9.3: Illustration of the use of UME [142] globally (left) and locally (right).

Nevertheless, how could we consider then an approach that accounts for the "data power" and at the same time is conditioned or regularized to a given model/prior? The authors in [65] already considered this problem in an Active Contour Model context. Their goal was to segment the contours of houses as seen from satellite images. They realized that, if they train a DNN to find these contours, the resulting segmentation was sometimes inaccurate in ways that would have been easy to avoid with prior knowledge: the contours were not rectangular, did not have straight lines, etc. Therefore, instead of feeding the DNN to train a resulting contour, the authors train the net to learn the input of an active contour fit. Consequently, the output of the overall deep learning strategy is a contour that fulfils the constraints of the model-based active contour fit, i.e, it considers prior contour properties like smoothness.

The same scheme as in [65] can be extended to SSM and ASM, as illustrated in Fig. 9.4. A DNN, for instance a U-net, could learn an output image D that maps the likelihood of a pixel being a landmark of the SSM. This could be interpreted as a data-based edge detector that searches for target landmarks. Then, a Point Distribution Model can be fit to the resulting mapping. The output of the overall deep learning strategy is thus a collection of PDM parameters r, t and $\mathbf{b}$. These parameters are the result of

$$[r, t, \mathbf{b}] = \arg \max_{\mathbf{b} \in \mathbb{B}(\boldsymbol{\lambda})} D\big(r(\boldsymbol{\mu} + \mathbf{P}\mathbf{b} + t\mathbf{1})\big). \tag{9.8}$$

Following similar optimization techniques as in [65], the influence of the last PDM-step could be back-propagated and hence condition the parameters of the deep net. One disadvantage of such an approach would be that, in order to consider the impact

on net gradients, approximations and relaxations are needed to obtain expressions of the cost functions. Still, there are other possible prior structures in the cost that allow to include SSM [145].

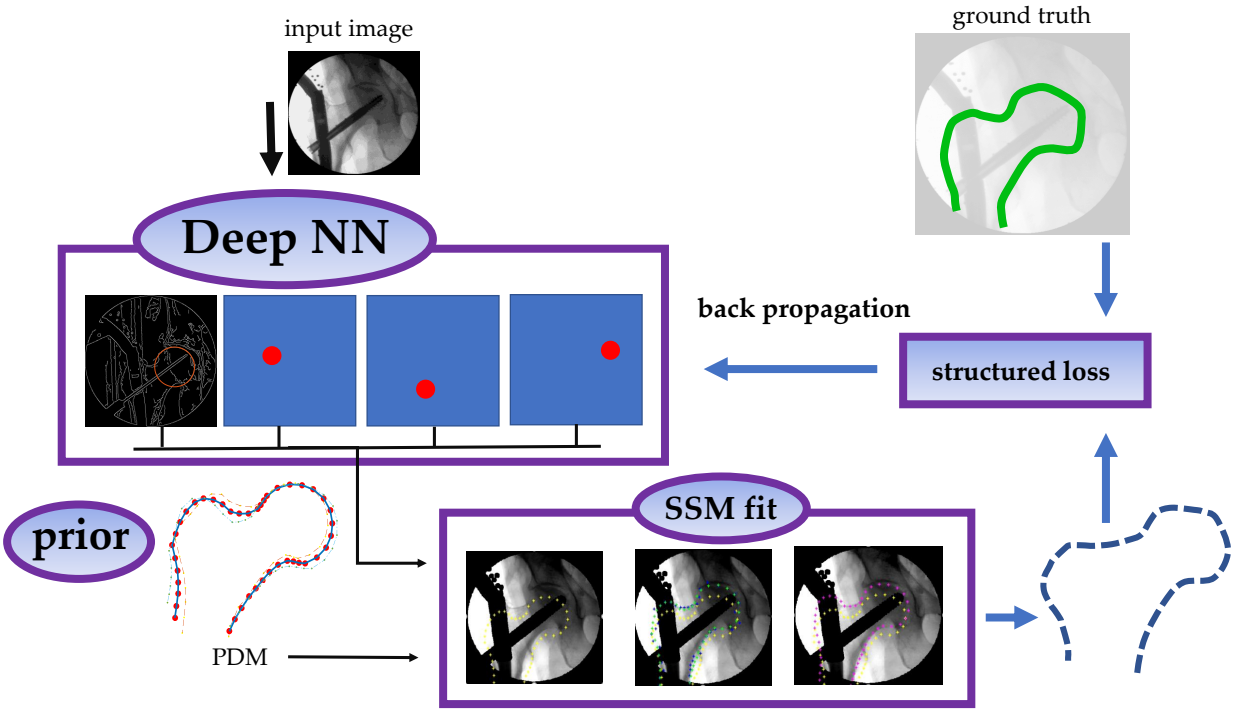


Figure 9.4: Inspired by [65], we propose a strategy in which a DNN predicts a mapping of landmark likelihood. This map is the input of an SSM-fit layer. After this, a structured loss can be computed and back-propagated.

Part IV

Appendix

Appendix A
Let there be Data

"It is a capital mistake to theorize before one has data."

— Sherlock Holmes

Many engineers and scientists agree on the fact that the 21st century is the era of information and data. Statistical models are used to make decisions and predictions, and are drastically becoming data-driven [146]. However, there are certain challenges with respect to the data that is needed in SSM. These training annotations typically consist on boundaries or surfaces around the objects of interest that are manually labeled [35]. This means that, the ground-truth generation is very time-consuming. For these reasons, they are not frequently shared with open access.

SSM are very frequently used in medical imaging to segment anatomy, since the inter-patient morphological variability is contained in the shape information. Medical data is especially hard to access. Most researchers in the medical area are hesitant to open their databases and there is a strong concern about the privacy of the patients. Therefore, there is an important lack of data-sharing incentives in the community [147]. For certain medical image disciplines, such as intraoperative fluoroscopy, there are very few images available . The main reason is that, typically there are no protocols of image storage during surgery. In conclusion, medical image data sets are relatively small, and expensive to acquire. Also, they are difficult to label, since an expert observer needs to supervise the annotation.

In this chapter we present the design of ground-truth annotation of a collection of medical images, more particularly intraoperative fluoroscopic X-rays, where the femur bone needs to be segmented. These database is unfortunately not available for the scientific community. However, we share the segmentation design of the anatomy under study, as well as the description of the application that we developed to acquire high quality annotations.

A.1 Generation of a database of femoral shape

A.1.1 Fluoroscopic images for computer-assisted surgery

Fluoroscopic X-ray images are frequently used in trauma surgery during intervention such as fracture reduction, joint replacement and screw insertions. Typically a C-arm mobile fluoroscopic machine belongs to the surgery room and is used to determine bone positions, implants alignment, and localize surgical tools [148]. We show one C-arm example in Fig. A.1. X-ray fluoroscopy contains lower radiation dose, which makes it safer medical staff and patients. The rotation possibility of the mobile C-arm also allows to take images from many different views during the intervention. However, fluoroscopy also involves important limitations, such as narrow field of view, low contrast and image distortion [148].

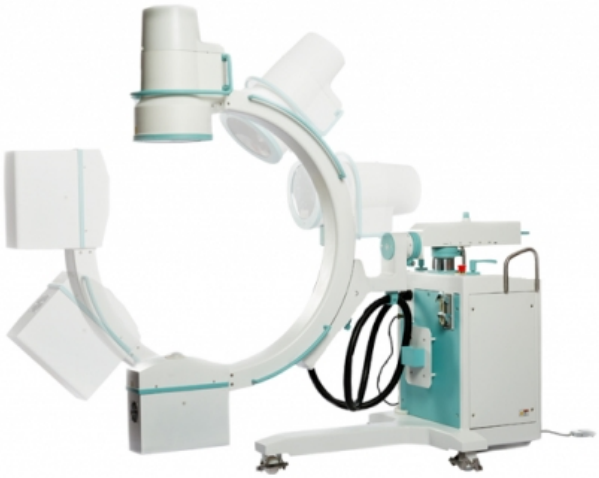


Figure A.1: A C-arm and its rotations. Image source: https://www.kiranxray.com

The fixation of nails and screws during orthopaedic surgery frequently requires an interpretation from the surgeon to estimate the relative position between the tools and the anatomy. The limitations of fluoroscopy may lead to positioning errors. CAS systems are therefore developed to guide the surgeon. A software that detects and segments the anatomy in the images can provide navigation guidance during the surgery without altering the surgical procedure, and it is safer and less expensive than a more powerful X-ray machine [149]. The precise localization of the anatomy in a number of fluoroscopic images can also be used to reconstruct a three dimensional scene, providing very valuable information to the surgeon [10]. Under such low-contrast and noisy conditions, extracting automatically the anatomy contours, or more specifically, bone contours, is a challenging task. SSM with landmark-based segmentation approaches solve this problem.

During the development of this doctoral work, we had access to a set of fluoroscopic images, that were extracted from a femoral trochanteric and shaft fractures [150]. To heal these fractures, typically an intramedullary nail is intersected inside

the femur bone, as depicted in Fig. A.2. The images were routinely acquired during surgery at the Clinic for Trauma Surgery and Orthopaedics at Augsburg University Hospital. All applicable data privacy regulations were observed, and we only worked with anonymized data.

Figure A.2: An example of an intralmedullary nail, inserted inside the a femur with multiple fractures, a fixed with a scre in the femoral head. Image source: [150].

A.1.2 A Graphical User Interface to collect data

We have developed a graphical user interface in MATLAB [151] in order to extract the ground-truth contours for the SSM analysis. This is a user-friendly platform, as shown in Fig. A.3, intended for the manual labelling of:

- The contour of the structure, i.e. the hip bones.
- Landmarks on this contour, i.e. available biological/geometrical points on the boundaries that can be found consistently in every different training image. These landmarks will lead to a straight-forward correspondence between images.

The developed application incorporates the following features to guide the user through the boundary and landmark selection:

- Edges: the user can show the Canny edges [8] automatically detected, with different threshold parameter. An example of the shown edges is depicted in Fig. A.4
- Pixel grid: shows the pixel-size separation to perform annotations with pixel precision.
- Colormap: changes the fake-color of the input gray image to enhance visualization.

The developed platform guarantees that the contours are continuous in a eight-connectivity, and the width of the contour is one pixel.

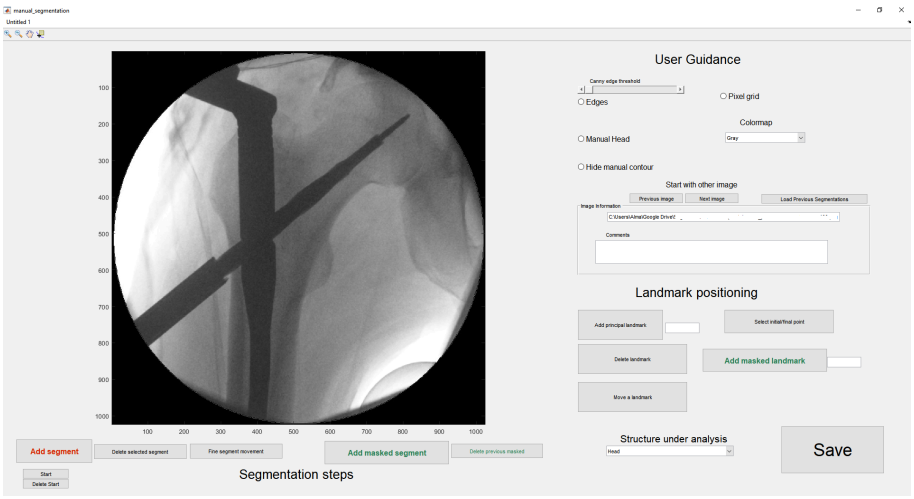


Figure A.3: User interface of the developed application

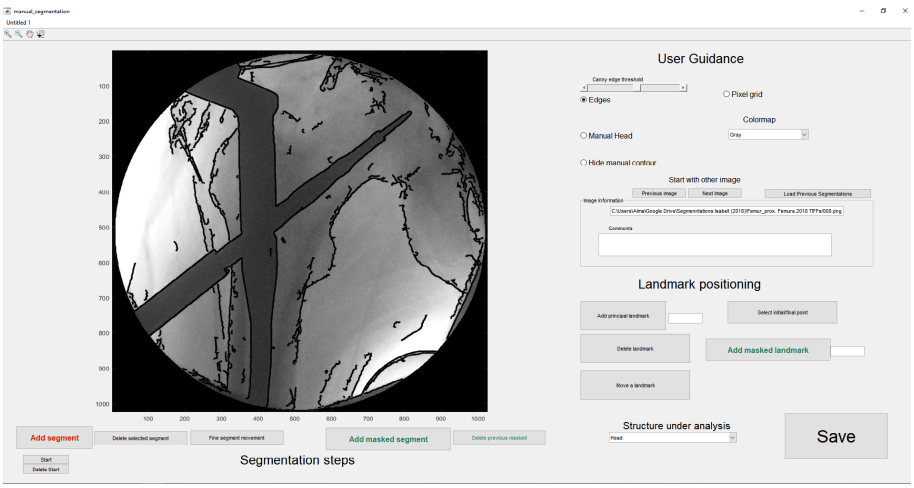


Figure A.4: Input image, in gray color-map and with depicted Canny edges.

Additionally, more than one bone or anatomical structures can be segmented in every image. In this particular analysis, also the acetabulum has a special interest, due to the risk to be confused with the femoral head. Therefore, the platform is ready to sequentially segment first the proximal femur and then the acetabulum in the same image.

Segmentation protocol

In order to be avoid human errors during the manual segmentation process as much as possible, we required three different users on each image: two independent users performed both the boundary and the landmarks selection; and the third user inspected the results of the other two to generate the final result.

The goal is to obtain an accurate manual contour of the bone, as well as a collection of shape landmarks that are easy to locate consistently from image to image, and from user to user. An example of a femur being segmented is shown in Fig. A.5.

We took into account the following remarks:

- The left and right legs are labeled beforehand in order to avoid mirroring.
- The annotation of the start point and the direction of the drawing are agreed, so that the topology of the contour is always the same.
- The contour is segmented in small individual segments, which can be deleted and repeated as many times as necessary.
- When there are different pieces of the boundary and the user is not certain which belongs to the contour, they may choose the one with strongest Canny edge (the one that stays with a highest threshold).
- The areas that are occluded by operation tools or other artifacts, as well as the ones that are outside the view of the C-arm, can be guessed by the user and depicted as masked.
- After segmenting the contour, the user places the landmarks always on the bone contour. Masked landmarks can also be added.

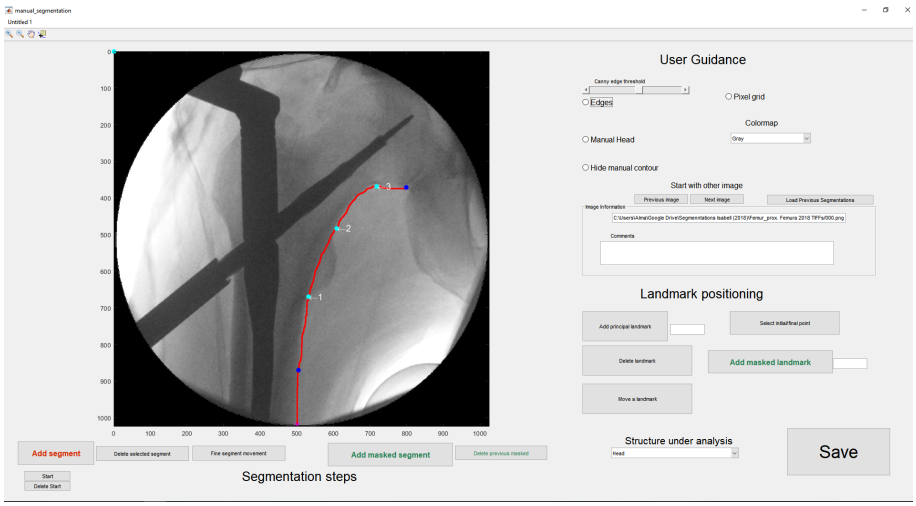


Figure A.5: An example of a contour being segmented (in red), and a few landmarks (light blue) being placed.

Revision and quality check

When the results from the segmentation of the first two users are available, then the results are reviewed. During the revision phase a new independent user (the reviewer) who did not observe the image yet is shown the conflict parts (in which

the othre two users did not agree) in orange color. Then, the reviewer can modify an editable contour accordingly. The reviewer can inspect both segmentation results from the other users, as well as the highlighted areas in witch the two previous users disagreed the most. The following metric is used to find the regions to highlight as conflictive to the reviewer:

$$d_{\text{revision}} = \frac{1}{N_1} \sum_{n_t=1}^{N_1} \min_{n_r} |x_1[n_t] - x_2[n_r]|, n_r = 1, \ldots, N_2. \tag{A.1}$$

When the distance d_{revision} is bigger than an agreed threshold, it is highlighted in orange as shown in Fig. A.6.

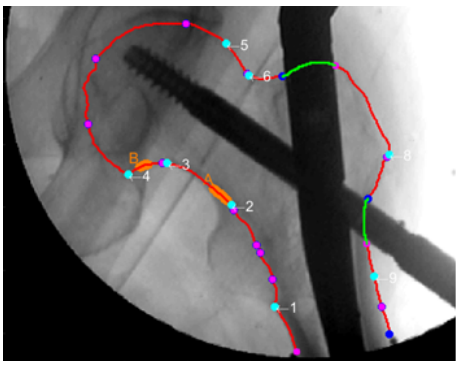


Figure A.6: Highlighted regions (A and B, in orange) indicate during a revision that the users who segmented this image differed considerably in these parts.

A.1.3 Anatomical landmarks in the proximal femur

The proximal femur is the extreme of the femur that is closer to the center of the body, that is, the extreme with the femoral head. The images used are in front-view. This means that, the X-rays are crossing the anatomy in an anterior-posterior way (from the front to the back of the body). Therefore, we design a front-view model of the observed proximal femur. Following the model in [20], we exclude the lesser trochanter unless the shaft contour is not visible in that region. We apply the same criterion with the greater trochanter, which is not included unless the neck is not visible.

The position of these landmarks has been agreed as the following:

1. Beginning of lesser trochanter
2. End of lesser trochanter
3. Change of curvature neck-head
4. Beginning of head circle
5. End of head circle

6. Change curvature head-neck
7. Connection neck-greater trochanter
8. Inferior end of greater trochanter
9. Perpendicular point in shaft direction to point 1

An example of a segmented proximal femur is shown in Fig. A.7.

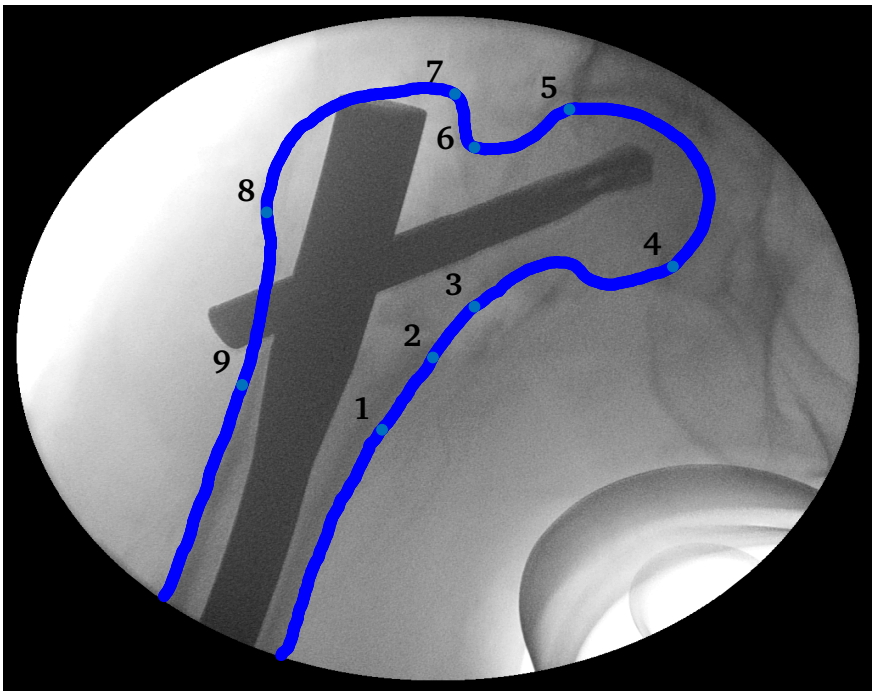


Figure A.7: The result of segmenting the proximal femur contour and the described landmarks.

A.1.4 Anatomical landmarks distal femur

The distal femur is the extreme of the femur that is closer to the feet and contains the condyles. We define a contour that contains the femur shaft and the medial condyle. The anatomical manual landmarks are two:

1. Back side limit of medial condyle
2. From side limit of medial condyle

An example of a segmented distal femur is shown in Fig. A.8.

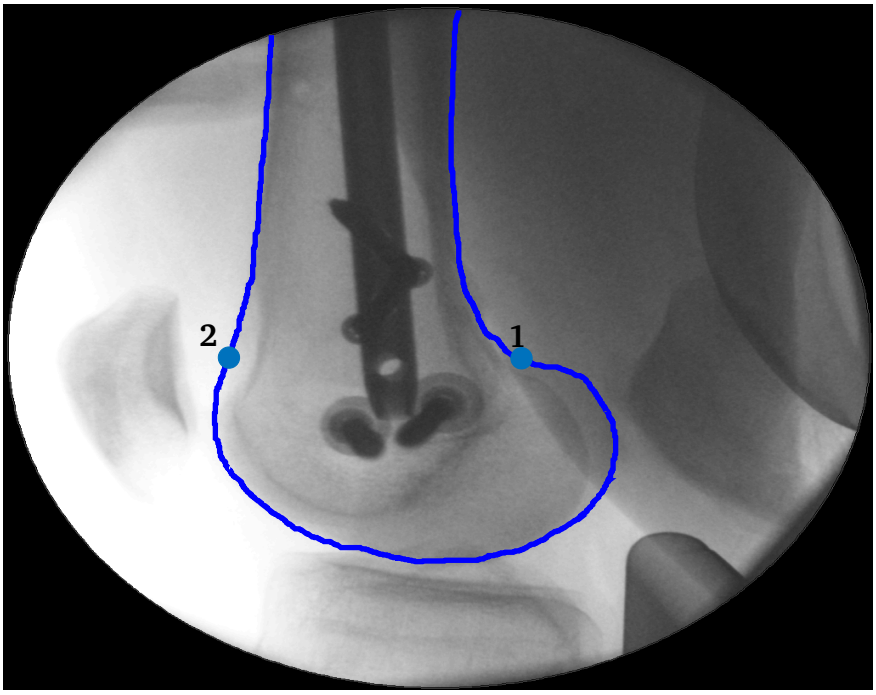


Figure A.8: The result of segmenting the distal femur contour and the described landmarks.

A.1.5 Correspondence and semi-landmarks

We consider a contour to be complete when all landmarks are present and no landmark is masked. In order to keep continuity in the boundary, we also consider as valid the small masked segmented guessed by the users as far as they are not at the extremes. In order to add more landmarks to the SSM, we also sample evenly spaced more contour points between the landmarks, the semi-landmarks. An example is shown in Fig. A.9, where a total of 40 landmarks are used considering the initial 9 landmarks.

A.2 Simulated shape data

In Chapter 5 we need simulated data in order to have a controled SSM model order in the tests. To this aim, we simulate realistic synthetic shapes, similarly to the simulated data in [103], and follow the model of a PDM, as described in Chapter 2, with

$$\mathbf{x} = \boldsymbol{\mu} + \mathbf{P}\mathbf{b} + \boldsymbol{\epsilon}. \tag{A.2}$$

We use the PDM eigenvectors $\mathbf{P}$, with corresponding eigenvalues $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_q]$, and shape mean $\boldsymbol{\mu}$. These parameters were obtained from the sample covariance matrix of the available real data, as well as the mean. We then choose values for $\mathbf{b}$ as random samples taken from $\mathbf{b} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\lambda})$. For the noise, we add white Gaussian samples $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ with different noise levels β. We define β as the ratio between

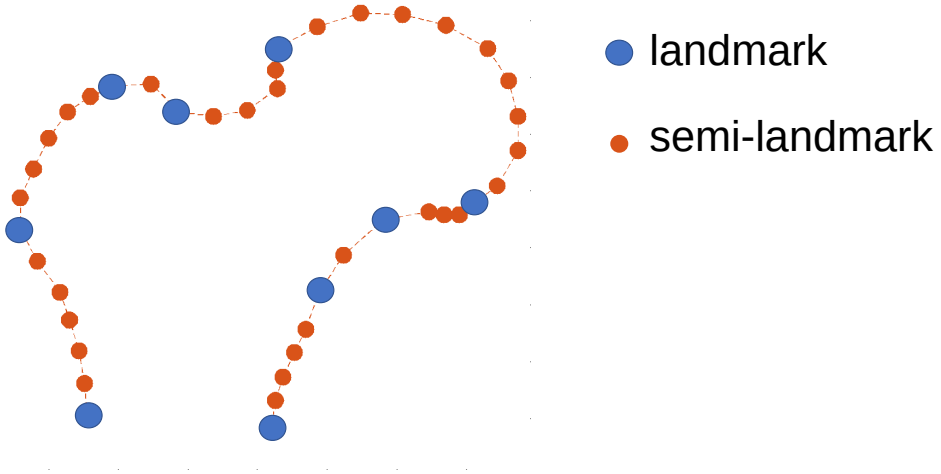


Figure A.9: An illustrative example of a femur shape with landmarks (blue) and semi-landmarks (orange).

the smallest kept signal eigenvalue and the noise variance, i.e., $\beta = \frac{\lambda_q}{\sigma^2}$. We choose β by visual inspection, so that the produced shape $\mathbf{x}$ looks realistic.

Then we randomly rotate, scale, and translate these synthetic shapes, that is, with random r and t, we compute

$$\boldsymbol{\kappa} = rV^{-1}(\mathbf{x}) + \mathbf{1}_N t. \tag{A.3}$$

After simulating M of these samples, we perform a generalized Procrustes analysis, as described in Chapter 2. The GPA typically colors the noise, and consequently it may not longer be white when the model order has to be selected in the tests. The resulting simulated shape vectors are the results from the generalized Procrustes analysis. We present a collection of artificially generated femur shapes in Fig. A.10.

A.3 Other freely available databases

One of the goals of the shape models and algorithms developed is this work was to avoid heuristic and ad-hoc solutions. In order to test that our approaches were not biased to work only in our fluroroscopic X-ray database, we searched for other freely available databases with anatomical shapes, and used them in our tests. This was a challenging tasks, since most of the databases used in medical imaging are not freely available. The following databases have been considered during the development of this thesis.

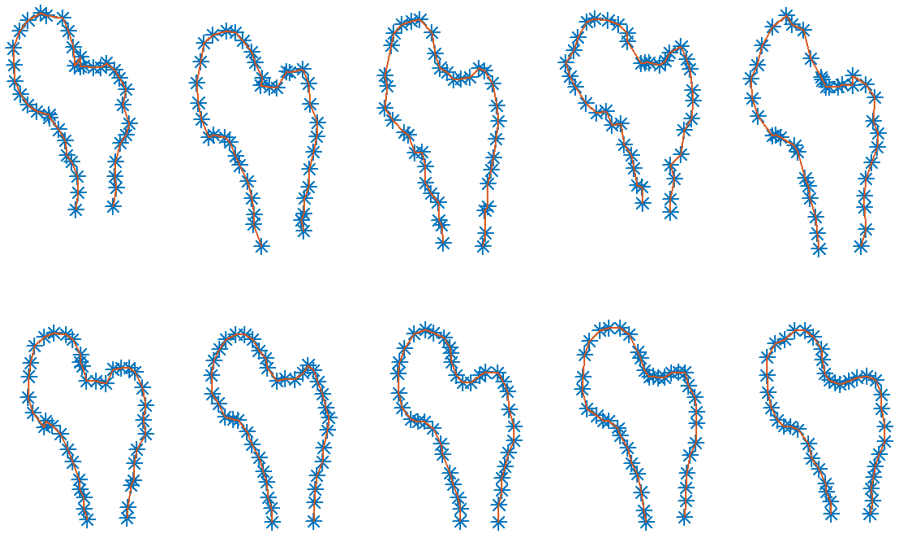


Figure A.10: A collection of artificially generated femur landmarks.

A.3.1 Chest X-ray diagnostic images

A group of radiology researchers in Japan shared a digital image database of 247 chest radiographs to study the presence of lung nodules [104]. The images were selected from 14 medical centers and were digitized by a laser digitizer with a 2048 $\times$ 2048 size (0.175 mm pixels) and a 12-bit gray scale.

Furthermore, the authors of [152] manually segmented (by two human observers) the lungs, heart and clavicle from the images. These observers placed 44 landmarks on each lung outline, 26 landmarks on the heart, and 24 on each clavicle, as shown in Fig. A.11.

A.3.2 Natural images from hands

We have also consider natural images, i.e., everyday regular photos, of hands to study the variability of the hand outline. We found two different data sources. First, a collection of 38 images of hands with 20 landmarks on the hand outline [31], as the example in Fig. A.12 (left). Second, a database containing 1000 scanned hands [133], as shown in Fig. A.12 (right), from which the contours have been automatically extracted with image processing techniques from MATLAB.

A.3.3 3D surfaces of the femur

A collection of 50 femur surfaces were collected from the Statistical Shape Modelling course in FutureLearn by the University of Basel [137]. These surfaces were segmented from Computed Tomography (CT) scans of the lower part of the body, from

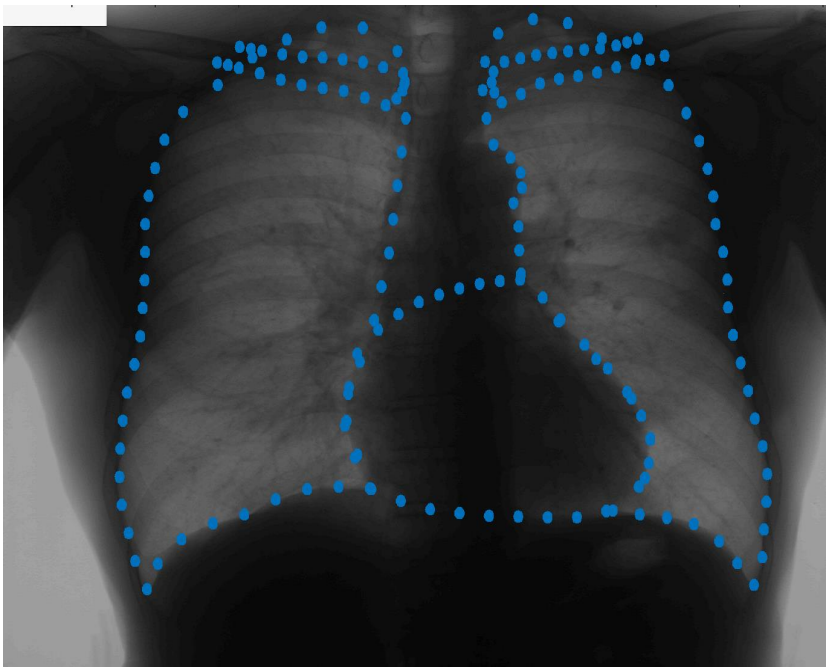


Figure A.11: One example of the chest database in [104]

25 different patients. They are described with triangle meshes. The dataset contains as well a set of 10 triangle meshes of partial femur shapes, as shown in Fig. A.13. [137]

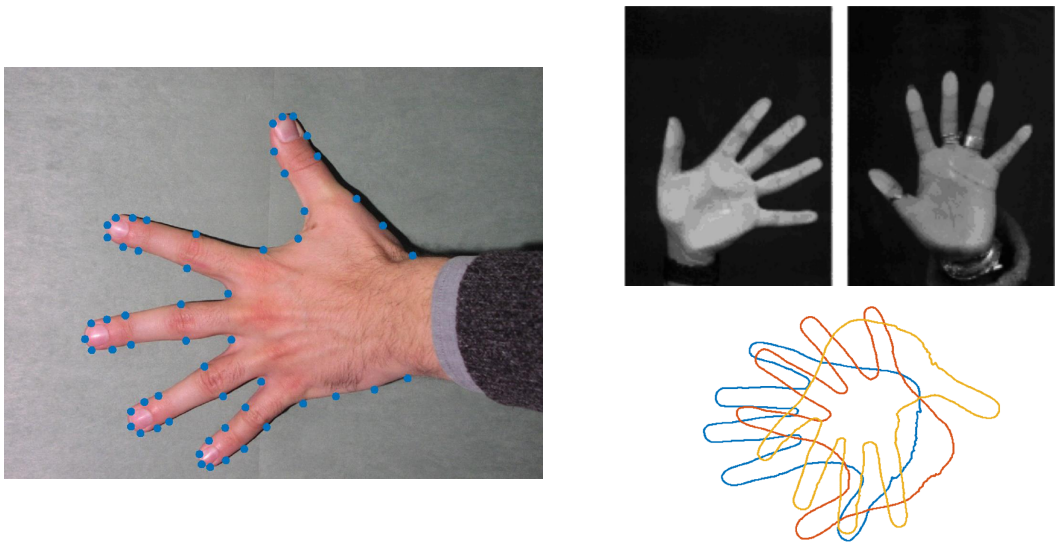


Figure A.12: Examples the databases of hands. Left: one example from [31] , with a natural image and the landmarks. Right: examples from [133], with the scanned hands and the automatically extracted contours.

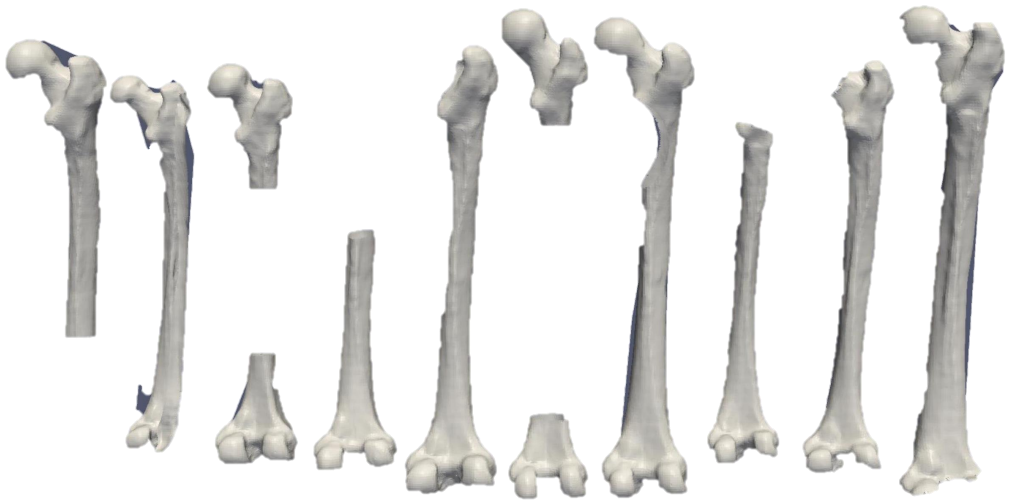


Figure A.13: Partial meshes from the dataset in [137]. Image source: [153]

Publications

Publications derived from this dissertation

The work developed in this thesis has given place to the following publications.

Journal articles

1. A. Eguizabal, P. Schreier, and J. Schmidt, "Procrustes registration of two-dimensional statistical shape models without correspondences," submitted to *IEEE Transactions on Image Processing* 2019, available on ArXiv.

2. A. Eguizabal, C. Lameiro, D. Ramirez and P. Schreier, "Source enumeration in the presence of colored noise," *IEEE Signal Processing Letters*, vol. 26, no. 3, 2019.

Conference contributions

1. A. Eguizabal, P. Schreier and D. Ramirez, "Model-order selection in Statistical Shape Models," in *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Aalborg, Denmark, Sept. 2018, pp. 475-479.

2. A. Eguizabal and P. Schreier, "A weighting strategy for Active Shape Models," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, Beijing, China, Sept. 2017, pp. 1-6.

Related work by the author

The author has also produced the following selected publications, which are not directly resulting from the main contributions described in the thesis, but are in the scope of medical image processing, machine learning and signal processing.

Journal articles

1. A. Eguizabal, A. M. Laughney, P.B. García-Allende, V. Krishnaswamy, W. A. Wells, K. D. Paulsen, B W. Pogue, J. M. Lopez-Higuera, and O. M. Conde, "Direct identification of breast cancer pathologies using blind separation of label-free localized reflectance measurements," *Biomedical Optics Express*, vol. 4, no. 7, 2013.

Conference contributions

1. S. Smirnov and A. Eguizabal, "Deep learning for object detection in fine-art paintings," in *Proc. of IEEE International Conference of Metrology for Archaeology and Cultural Heritage (MetrhoArchaeo)*, Cassino, Italy, Oct. 2018, (received award for best paper presented by a young researcher).

2. A. Eguizabal, E. Real, A. Ponton, M. Calvo, J. F. Val-Bernal, M. Mayorga, J. M. Revuelta, J. M. Lopez-Higuera and O. M. Conde, "Identification of vessel wall anomalies in thoracic aortic aneurysms through optical coherence tomography and gradient-based strategies," in *Proceedings of SPIE Europe, Photonic Solutions for Better Health Care IV*, Brussels, Belgium, May 2014, pp. 912943.

3. A. Eguizabal, A. M. Laughney, P.B. García-Allende, V. Krishnaswamy, W. A. Wells, K. D. Paulsen, B W. Pogue, J. M. Lopez-Higuera, and O. M. Conde, "ICA-guided delineation of breast cancer pathology," in *Proceedings of IEEE International Symposium on Biomedical Imaging (ISBI)*, Barcelona, Spain, May 2012, pp. 1611-1614.

Lebenslauf

Lebenslauf

Name	Eguizabal Aguado
Vorname	Alma
Wohnort	Paderborn
Geburtsdatum/-ort	28.02.1988
Nationalität	spanisch

Schulbildung

09/1994 – 06/2001	Grundschule (Primaria in Spanien)
09/2001 – 06/2004	Gymnasium (Secundaria in Spanien)
09/2004 – 06/2006	Bachhillerato in Spanien

Studium

09/2006 – 09/2011	Universidad de Cantabria, Spanien Ingeniero de Telecomunicaciones (Telekommunikationstechnik Diplom)
09/2011 – 09/2012	Universidad de Cantabria, Spanien Master TICRM (Master in Informationstechnologien und Kommunikation in Mobilfunknetzen)
09/2013 – 09/2019	Universität Paderborn Promotion In Elektrotechnik

Wissenschaftliche Tätigkeit

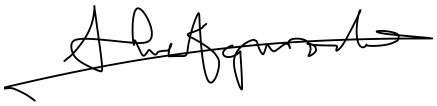
09/2010 – 09/2012	Universidad de Cantabria, Spanien Wissenschaftliche Hilfskraft
09/2012 – 11/2012	Helmholtz Zentrum München DAAD-Stipendiat
12/2012 - 05/2013	Universidad de Cantabria, Spanien Wissenschaftliche Mitarbeiterin

Beruf

Seit 06/2013	Universität Paderborn Wissenschaftliche Mitarbeiterin

Sprachkenntnisse	Spanisch Englisch Französisch Deutsch

Ort, Datum Paderborn, 01.08.2019

Unterschrift 

List of Figures

List of Algorithms

Bibliography

[1] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. Series in Artificial Intelligence, Prentice Hall, third ed., 2010.

[2] L. D. Rosemblum, *See what I am saying*. W. W. Norton and Company, 2010.

[3] F. I. Dretske, "Perception from an epistemological point of view," *The Journal of Philosophy*, vol. 68, no. 19, pp. 584–591, 1971.

[4] J. I. Eguizábal, *La huida de Perséfone: María Zambrano y el conflicto de la temporalidad*. PhD thesis, Universidad de Salamanca, 1993.

[5] R. C. Gonzalez and R. E. Woods, *Digital Image Processing (3rd Edition)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2006.

[6] A. Y. Ng, "Deep learning: What's next," in *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems, Singapore, May 9-13, 2016*, p. 1, 2016.

[7] D. A. Forsyth and J. Ponce, *Computer Vision: A Modern Approach*. Prentice Hall Professional Technical Reference, 2003.

[8] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, pp. 679–698, Nov 1986.

[9] D. G. Kendall, "Shape manifolds, procrustean metrics, and complex projective spaces," *Bulletin of the London Mathematical Society*, vol. 16, no. 2, pp. 81–121, 1984.

[10] L. Wang, M. Kohnen, O. Friman, and H. K. Hahn, "Fast automated segmentation of femoral heads in fluoroscopic x-ray images," in *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 984–988, March 2011.

[11] R. Davies, C. Twining, and C. Taylor, *Statistical Models of Shape, Optimisation and Evaluation*. Springer-Verlag London, 2008.

[12] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: application to breast MR images," *IEEE Transactions on Medical Imaging*, vol. 18, pp. 712–721, Aug 1999.

[13] L. H. Staib and J. S. Duncan, "Boundary finding with parametrically deformable models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, pp. 1061–1075, Nov 1992.

[14] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active Shape Models–their training and application," *Computer Vision and Image Understanding*, vol. 61, pp. 38 – 59, January 1995.

[15] S. Boyd and L. Vandenberghe, *Introduction to Applied Linear Algebra: Vectors, Matrices, and Least Squares*. Cambridge University Press, 2018.

[16] S. Wang and R. M. Summers, "Machine learning and radiology," *Medical Image Analysis*, vol. 16, no. 5, pp. 933 – 951, 2012.

[17] J. Glover, D. Rus, and N. Roy, "Probabilistic models of object geometry with application to grasping," *The International Journal of Robotics Research*, vol. 28, no. 8, pp. 999–1019, 2009.

[18] P. Garcia-Medrano, A. Pedergnana, L. Asryan, A. Olle, and A. Benito Calvo, "The quartzite cleavers form Atapuerca (Burgos, Spain: the standardization of shape," vol. 40, pp. 33–40, 12 2018.

[19] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 ed., 2009.

[20] C. Lindner, S. Thiagarajah, J. M. Wilkinson, T. arcOGEN Consortium, G. A. Wallis, and T. F. Cootes, "Fully automatic segmentation of the proximal femur using random forest regression voting," *IEEE Transactions on Medical Imaging*, vol. 32, pp. 1462–1472, August 2013.

[21] F. A. Mettler, W. Huda, T. T. Yoshizumi, and M. Mahesh, "Effective doses in radiology and diagnostic nuclear medicine: A catalog," *Radiology*, vol. 248, no. 1, pp. 254–263, 2008. PMID: 18566177.

[22] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, pp. 716–723, December 1974.

[23] D. G. Kendall, "A survey of the statistical theory of shape," *Statistical Science*, vol. 4, no. 2, pp. 87–99, 1989.

[24] M. Kähler, "Shape spaces and shape modelling analysis of planar shapes in a riemannian framework," Master's thesis, Mathematical Sciences at Durham University, 2012.

[25] "Enciclopedia de mitología griega y romana (y antigüedad clásica)." `http://www.alandalus-siglo21.org/mitog/griega.htm?/mitog/mitp9.htm`. Accessed: 2018-12-07.

[26] B. F. Green, "The orthogonal approximation of an oblique structure in factor analysis," *Psychometrika*, vol. 17, pp. 429–440, Dec 1952.

[27] P. H. Schönemann, "A generalized solution of the orthogonal procrustes problem," *Psychometrika*, vol. 31, pp. 1–10, Mar 1966.

[28] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, pp. 376–380, April 1991.

[29] R. Hartley, "Keynote lecture 2: Riemannian manifolds, kernels and learning," in *2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. XV–XV, Aug 2014.

[30] A. Srivastava and E. P. Klassen, *Functional and Shape Data Analysis*. Springer Series in Statistics, 2016.

[31] M. B. Stegmann and D. D. Gomez, "A brief introduction to statistical shape analysis," tech. rep., University of Denmark (DTU), 2002.

[32] M. Lüthi, T. Gerig, C. Jud, and T. Vetter, "Gaussian process morphable models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, pp. 1860–1873, Aug 2018.

[33] M. Bossa and S. Olmos, "Statistical linear models in Procrustes shape space," in *1st MICCAI Workshop on Mathematical Foundations of Computational Anatomy: Geometrical, Statistical and Registration Methods for Modeling Biological Shape Variability* (X. Pennec and S. Joshi, eds.), (Copenhagen, Denmark), pp. 102–111, Oct. 2006.

[34] P. T. Fletcher, C. Lu, S. M. Pizer, and S. Joshi, "Principal geodesic analysis for the study of nonlinear statistics of shape," *IEEE Transactions on Medical Imaging*, vol. 23, pp. 995–1005, Aug 2004.

[35] T. F. Cootes and C. Taylor, "Statistical models of appearance for computer vision," 2004.

[36] I. L. Dryden and K. V. Mardia, *Statistical Shape Analysis, with Applications in R. Second Edition.* Chichester: John Wiley and Sons, 2016.

[37] C. Goodall, "Procrustes methods in the statistical analysis of shape," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 53, no. 2, pp. 285–339, 1991.

[38] P. Schreier and L. Scharf, *Statistical Signal Processing of Complex-Valued Data: The Theory of Improper and Noncircular Signals*. Cambridge University Press, 2010.

[39] J. Alabort-i-Medina and S. Zafeiriou, "Unifying holistic and parts-based deformable model fitting," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3679–3688, June 2015.

[40] T. F. Cootes, C. J. Taylor, and A. Lanitis, "Multi-resolution search with active shape models," in *Proceedings of 12th International Conference on Pattern Recognition*, vol. 1, pp. 610–612 vol.1, Oct 1994.

[41] M. Jiang, G. He, and Z. Can, "Extending active shape models with color information for facial features localization," in *Proceedings of 2005 IEEE International Workshop on VLSI Design and Video Technology, 2005.*, pp. 183–186, May 2005.

[42] M. Rogers and J. Graham, "Robust active shape model search," in *Computer Vision — ECCV 2002*, (Berlin, Heidelberg), pp. 517–530, Springer Berlin Heidelberg, 2002.

[43] G. Behiels, F. Maes, D. Vandermeulen, and P. Suetens, "Evaluation of image features and search strategies for segmentation of bone structures in radiographs using active shape models," *Medical Image Analysis*, vol. 6, no. 1, pp. 47 – 62, 2002.

[44] D. Cristinacce and T. F. Cootes, "Feature detection and tracking with constrained local models," in *Proc. of British Machine Vision Conference*, vol. 3, pp. 929–938, 2006.

[45] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1, pp. I–I, Dec 2001.

[46] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2887–2894, June 2012.

[47] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3444–3451, June 2013.

[48] P. Martins, R. Caseiro, J. F. Henriques, and J. Batista, "Discriminative bayesian active shape models," in *Computer Vision – ECCV 2012*, (Berlin, Heidelberg), pp. 57–70, Springer Berlin Heidelberg, 2012.

[49] H. Ruppertshofen, C. Lorenz, G. Rose, and H. Schramm, "Discriminative generalized hough transform for object localization in medical images," *International Journal of Computer Assisted Radiology and Surgery*, vol. 8, pp. 593–606, Jul 2013.

[50] Y. Wang, S. Lucey, and J. F. Cohn, "Enforcing convexity for improved alignment with constrained local models," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, June 2008.

[51] S. Milborrow and F. Nicolls, "Active shape models with sift descriptors and mars," in *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, vol. 2, pp. 380–387, Jan 2014.

[52] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 886–893 vol. 1, June 2005.

[53] G. Rajamanoharan and T. F. Cootes, "Multi-view constrained local models for large head angle facial tracking," in *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pp. 971–978, Dec 2015.

[54] A. Zadeh, T. Baltrusaitis, and L. Morency, "Convolutional experts constrained local model for facial landmark detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2051–2059, July 2017.

[55] A. Arif, S. Rahman, K. Knapp, and G. Slabaugh, "Shape-aware deep convolutional neural network for vertebrae segmentation," in *Computational Methods and Clinical Applications in Musculoskeletal Imaging*, (Cham), pp. 12–24, Springer International Publishing, 2018.

[56] Y. Wu, T. Hassner, K. Kim, G. Medioni, and P. Natarajan, "Facial landmark detection with tweaked convolutional neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, pp. 3067–3074, Dec 2018.

[57] Z. Tang, X. Peng, S. Geng, L. Wu, S. Zhang, and D. Metaxas, "Quantized densely connected U-nets for efficient landmark localization," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 339–354, 2018.

[58] A. K. Davison, C. Lindner, D. C. Perry, W. Luo, and T. F. Cootes, "Landmark localisation in radiographs using weighted heatmap displacement voting," in *Computational Methods and Clinical Applications in Musculoskeletal Imaging - 6th International Workshop, MSKI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September, 2018*, pp. 73–85, 2018.

[59] F. Ambellan, A. Tack, M. Ehlke, and S. Zachow, "Automated segmentation of knee bone and cartilage combining statistical shape knowledge and convolutional neural networks: Data from the osteoarthritis initiative," *Medical Image Analysis*, vol. 52, pp. 109 – 118, 2019.

[60] D. Cristinacce and T. F. Cootes, "Boosted regression active shape models," in *Proc. of British Machine Vision Conference*, vol. 2, 01 2007.

[61] C. Santiago, J. C. Nascimento, and J. S. Marques, "2d segmentation using a robust active shape model with the em algorithm," *IEEE Transactions on Image Processing*, vol. 24, pp. 2592–2601, Aug 2015.

[62] J. M. Saragih, S. Lucey, and J. F. Cohn, "Face alignment through subspace constrained mean-shifts," in *2009 IEEE 12th International Conference on Computer Vision*, pp. 1034–1041, Sep. 2009.

[63] C. Lindner, P. A. Bromiley, M. C. Ionita, and T. F. Cootes, "Robust and accurate shape model matching using random forest regression-voting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, pp. 1862–1874, Sep. 2015.

[64] B. Shi, X. Bai, W. Liu, and J. Wang, "Face alignment with deep regression," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, pp. 183–194, Jan 2018.

[65] D. Marcos, D. Tuia, B. Kellenberger, L. Zhang, M. Bai, R. Liao, and R. Urtasun, "Learning deep structured active contours end-to-end," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[66] S. García, J. Luengo, and F. Herrera, "Tutorial on practical tips of the most influential data preprocessing algorithms in data mining," *Knowledge-Based Systems*, vol. 98, pp. 1 – 29, 2016.

[67] T. Albrecht, M. Lüthi, T. Gerig, and T. Vetter, "Posterior shape models," *Medical Image Analysis*, vol. 17, pp. 959–973, December 2013.

[68] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, ch. 12. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.

[69] C. M. Bishop, "Mixture models and EM," in *Pattern Recognition and Machine Learning (Information Science and Statistics)*, ch. 9, pp. 439–441, Berlin, Heidelberg: Springer-Verlag, 2006.

[70] T. Schneider, "Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values," *Journal of Climate*, vol. 14, no. 5, pp. 853–871, 2001.

[71] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, pp. 1–22, December 1977.

[72] A. Hill, T. Cootes, and C. Taylor, "Active shape models and the shape approximation problem," *Image and Vision Computing*, vol. 14, no. 8, pp. 601 – 607, 1996.

[73] M. Zhao, S. Z. Li, C. Chen, and J. Bu, "Shape evaluation for weighted active shape models," in *Proc. of the Asian Conference on Computer Vision*, pp. 1074–1079, 2004.

[74] G. Ruiz, E. Ramon, J. García, M. A. G. Ballester, and F. M. Sukno, "Weighted regularized ASM for face alignment," in *Proc. of the IEEE International Conference on Image Processing (ICIP)*, pp. 2906–2910, Sept 2016.

[75] K. Lekadir, R. Merrifield, and G. Z. Yang, "Outlier detection and handling for robust 3-d active shape models search," *IEEE Transactions on Medical Imaging*, vol. 26, pp. 212–222, Feb 2007.

[76] X. Yang, J. Cerrolaza, C. Duan, Q. Zhao, J. Murnick, N. Safdar, R. Avery, and M. G. Linguraru, "Weighted partitioned active shape model for optic pathway segmentation in MRI," in *Proc. of Clinical Image-Based Procedures. Translational Research in Medical Imaging*, pp. 109–117, Springer, 2014.

[77] A. M. Zoubir, "Introduction to robust signal processing." `https://www.spg.tu-darmstadt.de/media/spg/ieee_ssrsp/material/SummerSchool_Zoubir.pdf`, 2016.

[78] L. Breiman, "Statistical modeling: The two cultures (with comments and a rejoinder by the author)," *Statist. Sci.*, vol. 16, pp. 199–231, 08 2001.

[79] A. Charnes, E. L. Frome, and P. L. Yu, "The equivalence of generalized least squares and maximum likelihood estimates in the exponential family," *Journal of the American Statistical Association*, vol. 71, no. 353, pp. 169–171, 1976.

[80] L. Wang, M. Kohnen, O. Friman, and H. K. Hahn, "Fast automated segmentation of femoral heads in fluoroscopic x-ray images," in *Proc. of the IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 984–988, 2011.

[81] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, pp. 387–392, April 1985.

[82] A. M. Zoubir, M. Viberg, R. Chellappa, and S. Theodoridis, *Academic Press Library in Signal Processing: Volume 3 Array and Statistical Signal Processing*, ch. 20, pp. 859 – 935. Academic Press, 2014.

[83] S. Aouada, A. M. Zoubir, and C. M. S. See, "Source detection in the presence of nonuniform noise," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, pp. 165–168, May 2004.

[84] P. Stoica and M. Cedervall, "Detection tests for array processing in unknown correlated noise fields," *IEEE Transactions on Signal Processing*, vol. 45, pp. 2351–2362, September 1997.

[85] Q. T. Zhang and K. M. Wong, "Information theoretic criteria for the determination of the number of signals in spatially correlated noise," *IEEE Transactions on Signal Processing*, vol. 41, pp. 1652–1663, April 1993.

[86] L. C. Zhao, P. R. Krishnaiah, and Z. D. Bai, "On detection of the number of signals when the noise covariance matrix is arbitrary," *Journal of Multivariate Analysis*, vol. 20, no. 1, pp. 26 – 49, 1986.

[87] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 623–, 1948.

[88] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. New York, NY, USA: Wiley-Interscience, 2006.

[89] C. M. Bishop, *Neural Networks for Pattern Recognition*. New York, NY, USA: Oxford University Press, Inc., 1995.

[90] C. M. Bishop, *Pattern recognition and machine learning, 5th Edition*. Information science and statistics, Springer, 2007.

[91] T. Roos, P. Myllymäki, and J. Rissanen, "MDL denoising revisited," *IEEE Trans. Signal Processing*, vol. 57, no. 9, pp. 3347–3360, 2009.

[92] P. Stoica and Y. Selen, "Model-order selection: A review of information criterion rules," *IEEE Signal Processing Magazine*, vol. 21, pp. 36–47, July 2004.

[93] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Transactions on Signal Processing*, vol. 65, pp. 794–816, February 2017.

[94] R. R. Nadakuditi and A. Edelman, "Sample eigenvalue based detection of high-dimensional signals in white noise using relatively few samples," *IEEE Transactions on Signal Processing*, vol. 56, pp. 2625–2638, July 2008.

[95] A. L. Kraay and A. B. Baggeroer, "A physically constrained maximum-likelihood method for snapshot-deficient adaptive array processing," *IEEE Transactions on Signal Processing*, vol. 55, pp. 4048–4063, August 2007.

[96] E. J. Bedrick and C. L. Tsai, "Model selection for multivariate regression in small samples," *Biometrics*, vol. 50, pp. 226–231, 1994.

[97] A. Aubry, A. D. Maio, L. Pallotta, and A. Farina, "Maximum likelihood estimation of a structured covariance matrix with a condition number constraint," *IEEE Transactions on Signal Processing*, vol. 60, pp. 3004–3021, June 2012.

[98] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, March 2004.

[99] G. Stewart, "The efficient generation of random orthogonal matrices with an application to condition estimators," *SIAM Journal on Numerical Analysis*, vol. 17, no. 3, pp. 403–409, 1980.

[100] P. Viola and W. M. Wells, "Alignment by maximization of mutual information," *International Journal of Computer Vision*, vol. 24, pp. 137–154, Sep 1997.

[101] D. S. Zois and M. Raginsky, "Active object detection on graphs via locally informative trees," in *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, Sept 2016.

[102] R. H. Davies, C. J. Twining, T. F. Cootes, J. C. Waterton, and C. J. Taylor, "A minimum description length approach to statistical shape modeling," *IEEE Transactions on Medical Imaging*, vol. 21, pp. 525–537, May 2002.

[103] L. Mei, M. Figl, D. Rueckert, A. Darzi, and P. J. Edwards, "Statistical shape modelling: How many modes should be retained?," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2008.

[104] S. Junji, I. Shigehiko, K. Junpe, M. Tsuneo, K. Takeshi, K. Ken-ichi, M. Mitate, F. Hiroshi, K. Yoshie, and D. Kunio, "Development of a digital image database for chest radiographs with and without a lung nodule," *American Journal of Roentgenology*, vol. 174, pp. 71–74, 2000.

[105] A. Sotiras, C. Davatzikos, and N. Paragios, "Deformable medical image registration: A survey," *IEEE Transactions on Medical Imaging*, vol. 32, pp. 1153–1190, July 2013.

[106] M. Lüthi, C. Jud, and T. Vetter, "A unified approach to shape model fitting and non-rigid registration," in *Machine Learning in Medical Imaging*, (Cham), pp. 66–73, Springer International Publishing, 2013.

[107] A. Eguizabal, P. J. Schreier, and D. Ramírez, "Model-order selection in statistical shape models," in *Proc. of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, September 2018. `https://arxiv.org/abs/1808.00309`.

[108] I. L. Dryden and K. V. Mardia, *Statistical Shape Analysis, with Applications in R. Second Edition.* Chichester: John Wiley and Sons, 2016.

[109] H. Wei-En, Y. Ching-Hsiao, C. Chih-Ju, W. Hung-Kang, Y. Tsong-Han, and T. Ching-Shiow, "C-arm image-based surgical path planning method for distal locking of intramedullary nails," *Applied Bionics and Biomechanics*, May 2018.

[110] P. J. Besl and N. D. McKay, "A method for registration of 3-d shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, pp. 239–256, Feb 1992.

[111] A. Rangarajan, H. Chui, and F. L. Bookstein, "The softassign procrustes matching algorithm," in *Information Processing in Medical Imaging*, (Berlin, Heidelberg), pp. 29–42, Springer Berlin Heidelberg, 1997.

[112] B. Luo and E. Hancock, "A unified framework for alignment and correspondence," *Computer Vision and Image Understanding*, vol. 92, no. 1, pp. 26 – 55, 2003.

[113] A. Myronenko and X. Song, "Point set registration: Coherent point drift," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 2262–2275, Dec 2010.

[114] P. Bergström and O. Edlund, "Robust registration of surfaces using a refined iterative closest point algorithm with a trust region approach," *Numerical Algorithms*, vol. 74, pp. 755–779, March 2017.

[115] C. Scott and R. Nowak, "Robust contour matching via the order-preserving assignment problem," *IEEE Transactions on Image Processing*, vol. 15, pp. 1831–1838, July 2006.

[116] G. G. Demisse, D. Aouada, and B. Ottersten, "Deformation based curved shape representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, pp. 1338–1351, June 2018.

[117] M. Müller, *Information Retrieval for Music and Motion*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2007.

[118] I. Bartolini, P. Ciaccia, and M. Patella, "WARP: accurate retrieval of shapes using phase of fourier descriptors and time warping distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 142–147, Jan 2005.

[119] T. Rakthanmanon, B. Campana, A. Mueen, G. Batista, B. Westover, Q. Zhu, J. Zakaria, and E. Keogh, "Searching and mining trillions of time series subsequences under dynamic time warping," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, (New York, NY, USA), pp. 262–270, ACM, 2012.

[120] T. F. Cootes, "An introduction to active shape models," in *Image Processing and Analysis*, ch. 7, pp. 223–248, Oxford University Press, 2000.

[121] J. C. Gower, "Generalized procrustes analysis," *Psychometrika*, vol. 40, pp. 33–51, Mar 1975.

[122] F. L. Bookstein, "Landmark methods for forms without landmarks: Morphometrics of group differences in outline shape," *Medical Image Analysis*, vol. 1, pp. 225 – 243, 1997.

[123] D. Rueckert, A. F. Frangi, and J. A. Schnabel, "Automatic construction of 3-d statistical deformation models of the brain using nonrigid registration," *IEEE Transactions on Medical Imaging*, vol. 22, pp. 1014–1025, Aug 2003.

[124] T. F. Cootes, C. J. Twining, V. S. Petrovic, K. O. Babalola, and C. J. Taylor, "Computing accurate correspondences across groups of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 1994–2005, Nov 2010.

[125] Y. Chen and G. Medioni, "Object modelling by registration of multiple range images," *Image Vision Comput.*, vol. 10, pp. 145–155, Apr. 1992.

[126] S. Du, N. Zheng, S. Ying, Q. You, and Y. Wu, "An extension of the icp algorithm considering scale factor," in *2007 IEEE International Conference on Image Processing*, vol. 5, pp. V – 193–V – 196, Sep. 2007.

[127] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, pp. 43–49, Feb 1978.

[128] K. Paliwal, A. Agarwal, and S. Sinha, "A modification over sakoe and chiba's dynamic time warping algorithm for isolated word recognition," in *ICASSP '82. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 7, pp. 1259–1261, May 1982.

[129] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, pp. 67–72, February 1975.

[130] F. S. Hillier and G. J. Lieberman, *Introduction to Operations Research*. New York, NY, USA: McGraw-Hill, seventh ed., 2001.

[131] A. Mueen and E. J. Keogh, "Extracting optimal performance from dynamic time warping." `https://www.cs.unm.edu/~mueen/DTW.pdf`, 2016.

[132] A. Eguizabal and P. J. Schreier, "A weighting strategy for Active Shape Models," in *Proc. of the IEEE International Conference on Image Processing (ICIP)*, pp. 3610–3614, Sept 2017.

[133] E. Yoruk, E. Konukoglu, B. Sankur, and J. Darbon, "Shape-based hand recognition," *IEEE Transactions on Image Processing*, vol. 15, pp. 1803–1815, July 2006.

[134] W. Crum, O. Camara, and D. Hill, "Generalized overlap measures for evaluation and validation in medical image analysis," *IEEE Transactions on Medical Imaging*, vol. 25, pp. 1451–61, 12 2006.

[135] H. Edelsbrunner, "Alpha shapes – a survey," *Tessellations in the Sciences: Virtues, Techniques and Applications of Geometric Tilings*, January 2010.

[136] S. Uchida and H. Sakoe, "An efficient two-dimensional warping algorithm," *IEICE Transactions on Information and Systems*, June 1999.

[137] "Femur database from the course FutureLearn Statistical Shape Modelling. University of Basel, Switzerland." `https://www.smir.ch/courses/FutureLearnSSM`. [Online; accessed May-2019].

[138] J. Song, P. Babu, and D. P. Palomar, "Sparse generalized eigenvalue problem via smooth optimization," *IEEE Transactions on Signal Processing*, vol. 63, pp. 1627–1642, Apr. 2015.

[139] S. Rueda, J. K. Udupa, and Li Bai, "Landmark selection for shape model construction via equalization of variance," in *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 1271–1274, May 2008.

[140] R. R. Hagege and J. M. Francos, "Universal manifold embedding for geometrically deformed functions," *IEEE Transactions on Information Theory*, vol. 62, pp. 3676–3684, June 2016.

[141] L. C. Chen, A. G. Schwing, A. L. Yuille, and R. Urtasun, "Learning deep structured models," in *Proceedings of the 32Nd International Conference on Machine Learning - Volume 37*, ICML'15, pp. 1785–1794, JMLR.org, 2015.

[142] Z. Yavo, J. M. Francos, I. Santamaria, and L. L. Scharf, "Estimating the mean manifold of a deformable object from noisy observations," in *2016 IEEE 12th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, pp. 1–5, July 2016.

[143] R. Sharon, J. M. Francos, and R. R. Hagege, "Geometry and radiometry invariant matched manifold detection," *IEEE Transactions on Image Processing*, vol. 26, pp. 4363–4377, Sep. 2017.

[144] R. Bhalodia, S. Y. Elhabian, L. Kavan, and R. T. Whitaker, "DeepSSM: A deep learning framework for statistical shape modeling from raw images," in *Shape in Medical Imaging* (M. Reuter, C. Wachinger, H. Lombaert, B. Paniagua, M. Lüthi, and B. Egger, eds.), (Cham), pp. 244–257, Springer International Publishing, 2018.

[145] O. Oktem, C. Chen, N. O. Domanic, P. Ravikumar, and C. Bajaj, "Shape-based image reconstruction using linearized deformations," *Inverse Problems*, vol. 33, p. 035004, feb 2017.

[146] A. L'Heureux, K. Grolinger, H. F. Elyamany, and M. A. M. Capretz, "Machine learning with big data: Challenges and approaches," *IEEE Access*, vol. 5, pp. 7776–7797, 2017.

[147] C. H. Lee and H. J. Yoon, "Medical big data: promise and challenges," *Kidney research and clinical practice*, vol. 36, pp. 2211–9132, 2017.

[148] Z. Yaniv and L. Joskowicz, "Long bone panoramas from fluoroscopic x-ray images," *IEEE Transactions on Medical Imaging*, vol. 23, pp. 26–35, Jan 2004.

[149] H. Takai, M. Murayama, S. Kii, D. Mito, C. Hayai, S. Motohashi, and T. Takahashi, "Accuracy analysis of computer-assisted surgery for femoral trochanteric fracture using a fluoroscopic navigation system: Stryker ADAPT system," *Injury*, vol. 49, no. 6, pp. 1149 – 1154, 2018.

[150] Stryker, "Gamma3® long nail r1.5 and r2.0 - operative technique."

[151] Mathworks, "Creating apps with graphical user interfaces in matlab: http://de.mathworks.com/discovery/matlab-gui.html."

[152] B. van Ginneken, M. B Stegmann, and M. Loog, "Segmentation of anatomical structures in chest radiographs using supervised methods: A comparative study on a public database," *Medical image analysis*, vol. 10, pp. 19–40, 03 2006.

[153] "Probabilistic Shape Modelling: Projects and exercises. University of Basel, Switzerland." `https://dmi.unibas.ch/fileadmin/user_upload/dmi/Studium/Computer_Science/Vorlesungen_FS19/Probabilistic_Shape_Modelling/exercise-intro.pdf`. [Online; accessed July-2019].