

Imperial College London
Department of Mathematics
Statistics Section

Correspondence Sampling and Predictive Methodologies on Manifolds and Polyhedral Surfaces

Dimosthenis Tsagkrasoulis

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy in Mathematics of Imperial College London
September 2016

I certify that this thesis, and the research to which it refers, are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline.

Dimosthenis Tsagkrasoulis

Copyright

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives license. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the license terms of this work.

Abstract

In recent years, a significant increase of interest can be observed regarding the analysis of objects with inherently complex morphological structures. This is especially true in the fields of biomedical imaging and computer vision, where the advent of new three-dimensional (3D) data capturing technologies has enabled the relatively low-cost acquisition of detailed models representing the boundaries of physical and biological structures, such as cardiac, brain and facial surfaces. Within this framework, of large importance is the development of methodologies tailored for statistical analysis and predictive modeling with complex objects being either the input or the response variables.

In this thesis, we are concerned with aspects pertaining both the manipulation of complex surface objects, so as to enable their statistical analysis, as well as the construction of predictive methodologies that are suited to such data, and take advantage of the additional morphological structure to improve prediction accuracy. In more detail, we first concentrate on the problem of establishing dense point correspondence on collections of similar 3D polyhedral surfaces. This annotation is a prerequisite for any type of further data analysis. By considering these polyhedral objects as approximations of two-dimensional non-linear surfaces, a.k.a. manifolds, we are able to present a methodology for the automatic and dense point annotation, that can be accurately and efficiently employed on large datasets. We apply our algorithm on a detailed heritability and genome association study of the human face shape.

Subsequently, we deal with the problem of using morphological data, extracted from surface objects, as inputs for regression analysis. In particular, we extend our previous algorithm for point correspondence to the regression setting, by incorporating an adaptive procedure that identifies and annotates more densely subregions within the surfaces that are highly predictive of a related response variable. This procedure leads to significant improvements in prediction accuracy, as evaluated on an application of age prediction from facial surfaces.

Finally, we consider the opposite scenario of predictive modeling with complex objects constituting the responses. We identify as main problem the fact that the response space can no longer be considered Euclidean. To solve this, we construct a non-parametric regression methodology for manifold-valued objects. The method is versatile and can be applied in cases where the response space is a well-defined manifold, but also when such knowledge is not available. Model

fitting and prediction phases only require the definition of a suitable distance function on the response space. We apply our method in a variety of image completion problems, as well as the prediction of human faces from genotypic data.

Construction of suitable regression algorithms is necessary here, since most methodologies cannot deal with the issue of non-linearity in the response space. Existing manifold regression methods either require the rigorous definition of the underlying manifold or make use of kernel functions to implicitly project the responses on a very high-dimensional vectorial space, where standard regression methods can be applied. In the first case though, applicability is restricted to only very limited types of data, while in the second case, the issue of predicting on the response space can not be tackled easily or efficiently.

We present a non-parametric predictive methodology for manifold-valued objects, based on our previous work on distance-based Random Forest [153]. Predictions are made through a two-step approach, where we first find a point estimate on a Euclidean embedding of the responses, and then project that point back to the original space to acquire the manifold-valued prediction. The methodology can readily handle various types of data objects, since its only requirement is the definition of a meaningful distance function for the responses. Our predictive method was shown to outperform a number of different regression methods on various problems of image completion.

Acknowledgements

I would like to thank my supervisor Giovanni Montana, who gave me the opportunity to become a PhD student and guided me through the last four years of my studies, as well as supervisor Toni Bellotti, for his support and advice.

This work was made possible by the financial generosity of the EPSRC.

Many thanks go as well to my fellow colleagues.

To my family and friends

‘Happiness is not a continuous state of existence, but rather a collection of moments in time.’

M.F.

Contents

Abstract	ii
Acknowledgements	v
List of Tables	xv
List of Figures	xvii
List of Symbols	xxi
1 Introduction	1
1.1 Challenges in the Analysis of Surfaces and Manifold-Valued Data Objects	1
1.2 Summary of Contributions	3
1.2.1 Geodesic Correspondence Sampling for Ensembles of Polyhedral Surfaces	3
1.2.2 Regression-driven Adaptive Sampling for Ensembles of Polyhedral Surfaces	4
1.2.3 Non-parametric Predictive Methodology for Manifold-Valued Responses .	4
1.3 Thesis Structure	5
2 Background Material	6
2.1 Differential Geometry on Manifolds	6

2.2	Polyhedral Surfaces as Approximations of Manifold Surfaces	10
2.2.1	Shortest Polyhedral Geodesics	10
2.2.2	Straightest Polyhedral Geodesics	11
2.3	Random Forest	13
2.3.1	Training	14
2.3.2	Prediction	15
2.3.3	Variable Importance	15
2.4	Manifold Learning	16
2.4.1	Metric Multi-Dimensional Scaling	17
2.4.2	Isomap	18
3	Geodesic Correspondence Sampling for Ensembles of Polyhedral Surfaces	20
3.1	Introduction	20
3.1.1	Related Work	21
3.1.2	Motivation	22
3.2	Problem Definition	24
3.3	Geodesic Ensemble Surface Sampling Methodology	25
3.3.1	Dense Correspondence Sampling Through Entropy Optimization	25
3.3.2	Optimizing the Distribution of Points in Individual Surfaces	28
3.3.3	Point-Wise Correspondence Optimization	30
3.3.4	Initialization and Iterative Point Splitting	32
3.3.5	Algorithmic Details and Parameters	32
3.4	Validation Study	34

3.5	Application of GESSA on Face Shape Heritability and GWA Studies	42
3.5.1	Introduction and Motivation	42
3.5.2	Data Description, Preprocessing and Phenotyping	46
3.5.3	Curvature-based Morphological Traits	47
3.5.4	Heritability Study	51
3.5.5	Genome-Wide Association Study	58
3.5.6	Discussion	60
3.6	Conclusion	65
4	Regression-driven Adaptive Sampling for Ensembles of Polyhedral Surfaces	67
4.1	Introduction and Motivation	67
4.2	Problem Definition	68
4.3	Extension of GESSA for Adaptive Correspondence Sampling	69
4.3.1	Adaptive Distribution of Points on Surfaces	69
4.3.2	Regression-Driven Adaptive Correspondence Optimization	72
4.4	Application of Adaptive GESSA on the Prediction of Age	73
4.4.1	Introduction and Motivation	73
4.4.2	Data Description	74
4.4.3	Preliminary Study on Face Age Prediction	74
4.4.4	Age Prediction through Adaptive Correspondence Sampling	77
4.4.5	Discussion	83
4.5	Conclusion	84

5	Non-parametric Predictive Methodology for Manifold-Valued Responses	86
5.1	Introduction	86
5.1.1	Related Work	87
5.1.2	Motivation	92
5.2	Problem Definition	95
5.3	Distance Random Forest Regression	95
5.4	Predictive Methodology	96
5.4.1	RF-based Distance Prediction	97
5.4.2	Prediction on the Response Embedding	99
5.4.3	Projection from the Embedding to the Manifold	101
5.5	Simulation Studies	103
5.5.1	Prediction on a Circular Manifold	104
5.5.2	Prediction on a Spherical Manifold	107
5.5.3	Prediction on a Swiss-roll Manifold	112
5.6	Applications on Imaging and Facial Surface Data	117
5.6.1	Handwritten Digit Completion	118
5.6.2	2D Face Completion	121
5.6.3	3D Facial Surface Prediction from Genotypic Predictors	128
5.7	Conclusion	130
6	Conclusion	132
	Bibliography	134

A	Geodesic Correspondence Sampling for Ensembles of Polyhedral Surfaces	156
A.1	Alternative formulation for Point-Wise Correspondence Optimization	156
A.2	Sample Variability of Curvature Phenotypes in the TwinsUK Dataset	158
A.3	Structural Equation Modeling	161
A.4	Average SEM Statistics for the Heritability Study	164
A.5	Sparse Principal Component Analysis	167
A.6	Tuning the sPCA Sparsity Parameter	168
B	Application of Adaptive GESSA on the Prediction of Age	170
B.1	Feature Selection through Stability Selection for the various landmarking and phenotyping procedures	170
C	Non-parametric Predictive Methodology for Manifold-Valued Responses	172
C.1	Proof of Proposition 5.1	172
C.2	Out-Of-Sample Prediction	173

List of Tables

3.1	Mean Integrated Squared Errors between the validation surfaces' uniform density functions and density estimators computed using GESSA and Cates' single surface sampling methodologies.	37
3.2	Average distances and standard deviations (Std) between the 19 groundtruth landmarks and their respective nearest points, computed using GESSA, as well as the original method by Cates et al. [28]. Statistics were computed on the set of 11 validation faces. Normalized measurements were acquired by division with the mean facial width.	41
3.3	Abbreviations and related descriptions of selected landmarks and traits for distance-based heritability analysis.	57
3.4	Heritability estimates h^2 for 10 traits based on Euclidean Distances (EDTs) and 10 traits based on Geodesic Distances (GDTs). Models' Goodness-Of-Fit test p -values greater than 0.05 correspond to good fits of the observed data.	58
3.5	Comparison between previously reported heritable phenotypes, heritability maps and regional curvature traits.	64
4.1	Summary of age prediction results from 10-Fold Cross Validation, on the dataset of 400 unrelated subjects from the TwinsUK cohort.	82
4.2	Number of subjects and Mean Absolute Errors for 10-year age ranges and all models, on the dataset of 400 unrelated subjects from the TwinsUK cohort, based on 10-fold CV.	82

5.1	Covariance matrix of the 6 coordinate predictor variables in the dataset S used during the spherical manifold regression study.	107
5.2	Test mean squared errors of the various regression models for the simulated swiss-roll manifold dataset.	114
5.3	Test errors from the digit completion application. Classification Error was not applicable for RF and dRF. The number and percentage of badly reconstructed images was visually ascertained from Fig. 5.16.	119
A.1	Model-fitting average statistics for the curvature-based heritability studies. . . .	165
A.2	Model-fitting average statistics for the distance-based heritability study.	166

List of Figures

1.1	The challenge of computing dense point correspondence on a collection of surfaces.	2
1.2	The challenge of predicting manifold-valued objects.	3
2.1	Illustrative example of an unfolded shortest polyhedral geodesic.	12
2.2	Illustrative example of straightest polyhedral geodesics.	13
3.1	Example 3D facial surface with groundtruth landmarks for the Morphface Validation Dataset.	35
3.2	Computed GESSA corresponding point sets on validation faces using GESSA. .	38
3.3	Estimated Probability Density Function on a Validation Face.	39
3.4	Principal curvatures and shape characterization. General classification of shapes based on the signs of the two principal curvatures.	48
3.5	Topological characteristics of curvature indices.	50
3.6	Face curvature maps.	51
3.7	Heritability maps of the human face.	53
3.8	Eigenface maps of the top heritable regional traits.	56
3.9	Selected landmarks for the distance-based heritability analysis, depicted on the average TwinsUK facial surface.	57

3.10	Genomic Region plots showing genome-wide significant markers associated with the top heritable GC regional trait.	60
4.1	Uniform vs. curvature adaptive sampling of a single facial surface.	71
4.2	Preliminary Study - True vs. Predicted ages on the set of 952 twins, acquired from leave-2-out cross validation, using lasso regression.	76
4.3	Preliminary Study - Distributions of Predicted Absolute Age Differences between monozygotic (MZ) and Dizygotic (DZ) twin pairs.	76
4.4	Example faces with uniform dense point correspondense and estimated kernel density functions.	78
4.5	Example faces with age-driven dense point correspondense and estimated kernel density functions.	79
4.6	True vs. Predicted ages on the set of 400 unrelated subjects from the TwinsUK cohort, acquired from 10-fold cross validation, using lasso regression.	81
5.1	Illustration of various prediction methodologies on manifolds.	89
5.2	Distance Random Forest training methodology.	97
5.3	Illustration of the distance Random Forest prediction methodology.	98
5.4	Scatterplots of simulated input and response points for the first regression study on a circular manifold.	105
5.5	Scatterplots of simulated input and response points for the second regression study on a circular manifold.	106
5.6	Scatterplots of simulated response points for the second regression study on a circular manifold.	107
5.7	Scatterplots of the simulated input and response points for the regression study on a spherical manifold.	109

5.8	Scatterplot of $\{(y'_{i1}, y'_{i2}, y'_{i3}) i = 1, \dots, 250\}$, for dataset S with added noise ($\sigma = 0.2$).	110
5.9	Test MSE prediction errors for all models, noise settings and levels.	111
5.10	Test and predicted response observations for the regression study on a spherical Manifold.	111
5.11	Scatterplots of the simulated input and response points for regression study on a Swiss-roll manifold.	113
5.12	Pairwise input-output scatterplots for the constructed dataset S	113
5.13	True and predicted response test points of the various regression models, excluding 1NN, for the simulated swiss-roll manifold dataset with added response noise.	115
5.14	True and predicted response test points, projected on the y_1y_3 plane, of the various regression models, excluding 1NN, for the simulated swiss-roll manifold dataset with added response noise.	116
5.15	Test error vectors projected on the y_1y_3 plane of the various regression models, excluding 1NN, for the simulated swiss-roll manifold dataset with added response noise.	117
5.16	Test and reconstructed images from the digit completion application. Reconstructed digits were generated by concatenating the predicted responses with the upper -input- half of the test images. Bad reconstructions are enclosed in red squares. (a) Test digits. (b)-(e) Reconstructions using predictions from (b) 1-NN, (c) RF, (d) kRF and (e) dRF.	120
5.17	Five test set and corresponding reconstructed face images from the Olivetti dataset for 1-NN, 5-NN, LR and RR regression models.	123
5.18	Five reconstructed face images from the Olivetti dataset for RF and dRF regression models.	124
5.19	Six test set and corresponding reconstructed face images from the TwinsUK dataset for 1-NN, 5-NN and LR regression models.	126

5.20	Six reconstructed face images from the TwinsUK dataset for RF and dRF regression models.	127
5.21	Random subset of 9 test and predicted 3D faces. For each pair, the test face is on the left and the predicted face on the right.	129
5.22	Subset of 9 test 3D faces with sufficiently good predictions. For each pair, the test face is on the left and the predicted face on the right.	129
5.23	Subset of 9 test 3D faces with bad predictions. For each pair, the test face is on the left and the predicted face on the right.	130
A.1	Curvature variance maps illustrate the variance of curvature traits on all landmarks, computed from the full dataset of 952 TwinsUK subjects.	159
A.2	Mean and variance maps of absolute curvature differences between twins, for the full dataset of 952 TwinsUK subjects, as well as for MZ and DZ subsets.	160
A.3	Path diagram of the ACDE twin Structural Equation Model.	161
A.4	Effect of sPCA sparsity parameter on the heritability of regional traits.	169
B.1	Stability Selection facial maps for (a) traits based on uniformly positioned corresponding points and (b) traits based on adaptively positioned corresponding points.	171

List of Symbols

Roman Symbols

$\mathbf{a}_{RF}$ Vector of affinities between training response points and a new observation, computed using Random Forest

$\hat{d}_{\mathcal{M}}(\cdot, \cdot)$ Estimated manifold distance function

$d_{\mathcal{M}}(\cdot, \cdot)$ Manifold distance function

$\mathbf{D}_{\mathcal{M}}$ Matrix of manifold distances with elements $D_{\mathcal{M},ij} = d_{\mathcal{M}}(y_i, y_j)$

$\mathbf{D}_{\mathcal{M}}^{(2)}$ Matrix of squared manifold distances with elements $D_{\mathcal{M},ij}^{(2)} = d_{\mathcal{M}}^2(y_i, y_j)$

$d_{\mathcal{S}}(\cdot, \cdot)$ Distance function for a polyhedral surface

$e^{\mathcal{S}}$ Edge of polyhedral surface $\mathcal{S}$

$\exp_{y_{\mathcal{M}}}(\cdot)$ Exponential map from $\mathcal{T}_{y_{\mathcal{M}}}\mathcal{M}$ to $\mathcal{M}$

$\mathcal{F}_{bs}$ Space of functions used during backscoring from the embedding to the manifold

f_{bs} Backscoring function from the embedding to the manifold

$f_{\mathcal{M}}(\cdot)$ Manifold-valued function

$f^{\mathcal{S}}$ Face/triangle of polyhedral surface $\mathcal{S}$

$g_{bs}(\cdot, \cdot, \cdot, \cdot)$ Kernel function used during manifold regression's backscoring procedure

$G_{dRF}(\cdot)$ Split objective function of distance Random Forest

$G_{kRF}(\cdot)$ Split objective function of kernel Random Forest

$g_{kRF}(\cdot, \cdot)$ Gaussian kernel function used in kernel Random Forest

$g_{\mathcal{M}}(\cdot, \cdot)$ Kernel function for manifold-valued objects

g_{MDS} Kernel function used for out-of-sample prediction

$G_{RF}(\cdot)$ Split objective function of standard Random Forest

$g_{\mathcal{S}}(\cdot, \cdot)$ Geodesic kernel function on a polyhedral surface

g_{Sh}^{jk} Abbreviation defined as $g_{Sh}^{jk} := \frac{1}{h_{Sk}} g_{\mathcal{S}} \left(\frac{d_{\mathcal{S}}(\psi_j^i, \psi_k^i)}{h_{Sk}}, \sigma_{\mathcal{S}} \right)$

$g_{\mathcal{S}}^{jk}$ Abbreviation defined as $g_{\mathcal{S}}^{jk} := g_{\mathcal{S}} \left(d_{\mathcal{S}}(\psi_j^i, \psi_k^i), \sigma_{\mathcal{S}} \right)$

$g_{VI}(\cdot)$ gini Variable Importance function

$\hat{h}$ Heritability estimate

$\hat{H}_i(\psi_1^i, \dots, \psi_M^i)$ Estimated differential entropy of random variable Ψ_i on polyhedral surface $\mathcal{S}_i$ under observed data

h Heritability

$H_i(\Psi_i)$ Differential entropy of random variable Ψ_i on polyhedral surface $\mathcal{S}_i$

$\mathbf{h}_{\mathcal{S}}$ Weight vector for adaptive sampling

h_{Sk} k_{th} weight for adaptive sampling

(j_n, s_{j_n}) Input split dimension and split value on a Random Forest tree node n_{RF}

$\mathbf{K}_{bs}$ Data kernel matrix used during manifold regression's backscoring procedure

k_{CU} Curvedness

k_{GC} Gaussian curvature

$\mathbf{K}_{\mathcal{M}}$ Gramian Matrix for Metric multi-Dimensional Scaling

k_{max} Maximum curvature

k_{MC} Mean curvature

k_{min} Minimum curvature

k_{SI}	Shape Index
$L(\cdot)$	Length of a curve
l_G	Weight parameter for balancing entropy terms in GESSA
$\log_{y_{\mathcal{M}}}(\cdot)$	Logarithmic map from $\mathcal{M}$ to $\mathcal{T}_{y_{\mathcal{M}}}\mathcal{M}$
$\mathcal{M}$	Smooth manifold
n_{RF}	Random Forest tree node
$p^{aKDE}(\cdot)$	Adaptive Kernel Density Estimator
$p^{KDE}(\cdot)$	Kernel Density Estimator
$p(\cdot), q(\cdot)$	Probability density functions
$p^{\mathcal{S}}$	Path on a polyhedral surface $\mathcal{S}$
$\mathcal{S}$	Polyhedral surface/triangulated mesh
$\mathcal{S}$	Dataset
T_{RF}	Number of trees in Random Forest
t_{VI}, r_{VI}	Free parameters in the formula for computing $\mathbf{h}_{\mathcal{S}}$
$\mathcal{T}_{y_{\mathcal{M}}}\mathcal{M}$	Tangent space of $\mathcal{M}$ at point $y_{\mathcal{M}}$
$\mathbf{u}_{\mathcal{T}}$	Tangent vector
$\mathbf{v}^{\lambda}$	Eigenvector
$v^{\mathcal{S}}$	Vertex of polyhedral surface $\mathcal{S}$
$\mathbf{w}^{RF}$	Vector of gini variable importances
$\mathbf{X}$	Data matrix of multivariate input observations
X	Random input variable
x_i	i_{th} observation of an input/predictor variable
(x_i, y_i)	i_{th} observation of paired input-response variables

$x_{.k}$	k_{th} input feature/dimension of a predictor variable
x_{new}	New/Test input observation
$\mathbf{y}$	Data vector of univariate response observations
y_i	i_{th} observation of a response variable
$y_{\mathcal{M}}, w_{\mathcal{M}}, h_{\mathcal{M}}$	Points on manifold $\mathcal{M}$.
$\hat{y}_{new}$	Estimate of y_{new}
y_{new}	New/Test response observation
Y, Y_a, Y_b	Random response variables
z_i	i_{th} observation of a manifold-valued response variable embedded on a Euclidean space using Manifold Learning
$\hat{z}_{new}$	Estimate of z_{new}
z_{new}	New/Test response observation on the embedding space

Greek Symbols

α_{reg}	Regularization term to ensure non-singularity during point-wise correspondence optimization
$\Gamma(\Phi)$	Differential entropy of random variable Φ on shape space
$\hat{\Gamma}(\phi_1, \dots, \phi_N)$	Estimated differential entropy of random variable Φ on shape space under observed data
γ_{bs}	Smoothness interpolation parameter used during manifold regression's backscoring procedure
σ_{bs}	Standard deviation parameter used during manifold regression's backscoring procedure
γ_G	Learning rate parameter for the geodesic gradient descent algorithm
$\gamma_{\mathcal{M}}$	Smooth curve on manifold $\mathcal{M}$
λ	Eigenvalue

Ω	Data matrix of centered observations on shape space
Φ	Data matrix of observations on shape space
Φ	Random variable on shape space
ϕ_i	Representation of polyhedral surface $\mathcal{S}_i$ on the shape space
Ψ_i	Random variable on polyhedral surface $\mathcal{S}_i$
ψ_j^i	j th point on polyhedral surface $\mathcal{S}_i$
$\hat{\Sigma}_\Phi$	Estimated covariance matrix of random variable Φ under observed data
$\sigma_{\mathcal{S}}$	Standard deviation parameter for the geodesic kernel density function

Chapter 1

Introduction

1.1 Challenges in the Analysis of Surfaces and Manifold-Valued Data Objects

A continuously increasing array of biomedical and computer vision applications nowadays require the modeling and analysis of complex data objects, such as 3D polyhedral surfaces. Challenges can be identified with regards to data manipulation and processing, where the goal is to enable statistical data analysis, as well as the appropriate use of complex data for regression and predictive purposes. In this work, we concentrate, in particular, to three such challenges, outlined here.

The first challenge is the automatic annotation of corresponding points on sets of polyhedral surfaces (see Fig. 1.1 (a)). The increased availability of large collections of 3D data is making the use of semi or fully automated point correspondence methodologies nearly a necessity, since manual positioning of points in such scenarios is extremely time consuming and possibly inconsistent. A separate but related issue, concerns the ability to cover the surfaces densely and consistently with large sets of points, even in regions with no prominent morphological structure, which is also problematic in the case of manual annotation. A number of different approaches for point correspondence have already been presented in the literature, but can

still suffer from one or more of the following issues; dependence on some form of manual annotation, bias towards a template or mean surface, computational cost and non-scalability to large datasets, low landmarking accuracy.

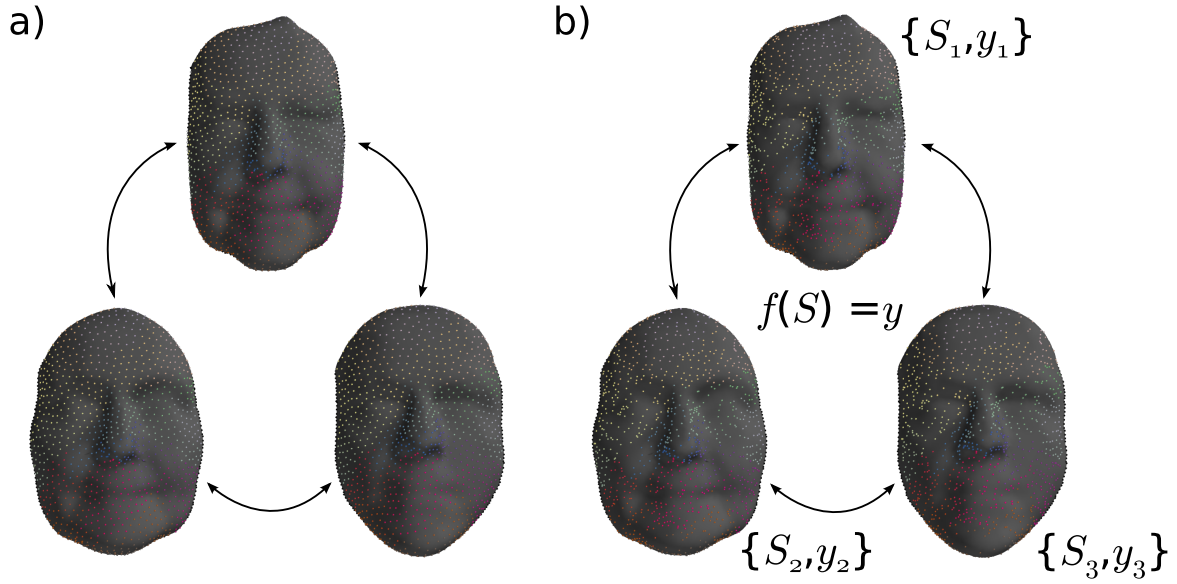


Figure 1.1: The challenge of computing dense point correspondence on a collection of surfaces. (a) With the advent of high-resolution, high-throughput 3D imaging technologies, automatic identification of dense correspondences between surfaces has become essentially necessary, if it is to capture coherently the rich morphological information of the data and use it for statistical analysis. (b) In the particular case of using morphological data to predict associated measurements, an extra challenge is how to distribute points on the surfaces, in order to achieve the best prediction performance.

The second challenge is related to the use of morphological surface data for prediction of response measurements associated with these surfaces. As mentioned previously, to enable regression analysis, surfaces must first be annotated with points corresponding to equivalent positions across the different observations. The common approach so far has been to disregard any additional information, which can be suboptimal in terms of predictive performance. A regression-guided positioning of points can capture better the relevant structural information and improve a regression model's predictive power (see Fig. 1.1 (b)). To our knowledge, this problem has not been tackled previously in the literature.

Finally, the third challenge we identify is the prediction of complex, manifold-valued responses. Such objects do not comply to the assumptions of Euclidean geometry and can occupy non-linear spaces (see Fig. 1.2). Construction of suitable regression algorithms is necessary here,

since most methodologies cannot deal with the issue of non-linearity in the response space. Existing manifold regression methods either require the rigorous definition of the underlying manifold or make use of kernel functions to implicitly project the responses on a very high-dimensional vectorial space, where standard regression methods can be applied. In the first case though, applicability is restricted to only very limited types of data, while in the second case, the issue of predicting on the response space can not be tackled easily or efficiently.

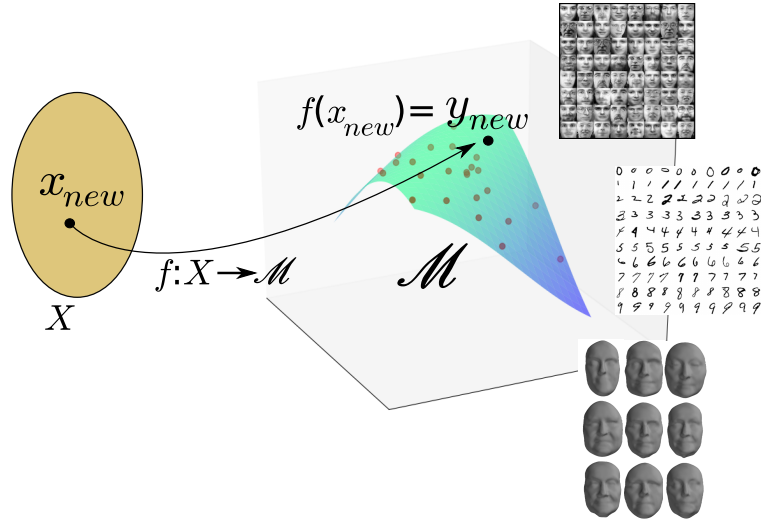


Figure 1.2: The challenge of predicting manifold-valued objects. A wide variety of data objects do not comply to the assumptions of Euclidean geometry. In order to create predictive models for such objects, it is necessary to take into consideration the constraints imposed by the shape of the response space.

Detailed background information on the aforementioned problems is presented respectively in Chapters 3, 4 and 5 where we further motivate and present our work.

1.2 Summary of Contributions

1.2.1 Geodesic Correspondence Sampling for Ensembles of Polyhedral Surfaces

In Chapter 3 we concentrate on the challenge of dense point correspondence and propose GESSA, a Geodesic Ensemble Surface Sampling Algorithm for the automatic and dense anno-

tation of collections of polyhedral surfaces. Our main contribution lies in the use of a kernel density estimator and a suitable gradient descent algorithm that employ geodesic distances and length minimizing paths between surface points to achieve, in an accurate and efficient way, a uniform distribution of points across the complete surfaces. We applied GESSA on a novel large scale heritability study of the human face, using as descriptors of shape various curvature traits. We were thus able to present, for the first time, detailed facial heritability maps. A website was created to present the results of the application at <http://heritabilitymaps.info/>. An article describing the GESSA methodology and the heritability application has been accepted for publication at Scientific Reports. A preprint is available at <https://arxiv.org/abs/1608.08199>.

1.2.2 Regression-driven Adaptive Sampling for Ensembles of Polyhedral Surfaces

In Chapter 4, we extend GESSA to allow non-uniform distribution of corresponding points based on, possibly unrelated to surface morphology, feature measurements, associated with the polyhedral surfaces. This is achieved by the use of an adaptive kernel density estimator with different weights assigned to each surface point. These weights are determined through a Random Forest model, which computes a variable ranking of individual point traits, based on their ability to predict the associated surface feature. To our knowledge, this is the first time that a regression-driven correspondence approach is presented in the literature. We apply this methodology on the prediction of age from morphological facial traits and show that predictive performance is enhanced, compared to a setting where uniform annotation of points is employed. A manuscript describing this work is currently under preparation.

1.2.3 Non-parametric Predictive Methodology for Manifold-Valued Responses

In Chapter 5, we present a non-parametric predictive methodology for manifold-valued objects, based on our previous work on distance-based Random Forest [153]. Predictions are

made through a two-step approach, where we first find a point estimate on a Euclidean embedding of the responses, and then project that point back to the original space to acquire the manifold-valued prediction. The methodology can readily handle various types of data objects, since its only requirement is the definition of a meaningful distance function for the responses. Our predictive method was shown to outperform a number of different regression methods on various problems of image completion. An article describing this work is currently under review for publication in Pattern Recognition Letters. A preprint can be found at <https://arxiv.org/abs/1701.08381>.

1.3 Thesis Structure

Chapter 2 includes a description of methods and mathematical constructions that are employed during the course of this work. In particular, we first present an overview of differential geometry on manifolds, followed by a description of polyhedral surfaces as approximations of two-dimensional manifolds. Subsequently, we introduce the Random Forest Regression method. Finally, the Manifold Learning techniques of Multi-Dimensional Scaling and ISOMAP are described. In Chapters 3, 4 and 5, we introduce, motivate, and formally define the three problems that constitute the main body of this thesis. In each of these Chapters, we subsequently describe our proposed methods for solving the problems and present related simulation and/or application studies that were performed to validate and illustrate our work. Finally, in Chapter 6, we conclude this thesis by providing a summary of our achievements and presenting directions for future work.

Chapter 2

Background Material

In this Chapter we present some necessary background material that is employed during the course of this work.

2.1 Differential Geometry on Manifolds

The notion of manifold permeates the complete length of this thesis. A manifold can be intuitively understood as a surface that locally resembles the Euclidean plane but has a different global structure. In chapter 3, we are going to present a dense point correspondence algorithm for ensembles of polyhedral surfaces, which can be considered approximations of manifolds, while in chapter 5, we are going to introduce a regression methodology for responses on manifold spaces. In the following, we present some necessary definitions and important results regarding differential geometry on manifolds, that are employed in this work. A manifold is formally defined as follows:

Definition 2.1 ([134]). *A subset $\mathcal{M} \subset \mathbb{R}^q$ is called a smooth m -dimensional manifold, or m -manifold, if every point $y_{\mathcal{M}} \in \mathcal{M}$ has an open neighborhood $U \subset \mathbb{R}^q$ such that $U \cap \mathcal{M}$ is diffeomorphic to an open subset in $\mathbb{R}^m$.*

It is possible to extend the notion of a vector from the Euclidean to the Manifold setting.

Definition 2.2 ([134]). Let $\mathcal{M} \subset \mathbb{R}^q$ be a smooth m -manifold and a point $y_{\mathcal{M}} \in \mathcal{M}$. A vector $\mathbf{u}_{\mathcal{T}} \in \mathbb{R}^q$ is called a tangent vector of $\mathcal{M}$ at $y_{\mathcal{M}}$, if there is a smooth curve $\gamma_{\mathcal{M}} : \mathbb{R} \rightarrow \mathcal{M}$ such that $\gamma_{\mathcal{M}}(0) = y_{\mathcal{M}}$, $\dot{\gamma}_{\mathcal{M}}(0) = \mathbf{u}_{\mathcal{T}}$, where $\dot{\gamma}_{\mathcal{M}}$ denotes the derivative of $\gamma_{\mathcal{M}}$.

The set of all tangent vectors of $\mathcal{M}$ at point $y_{\mathcal{M}}$, $\{\dot{\gamma}_{\mathcal{M}}(0) | \gamma_{\mathcal{M}} \text{ is smooth, } \gamma_{\mathcal{M}}(0) = y_{\mathcal{M}}\}$, is called the tangent space of $\mathcal{M}$ at point $y_{\mathcal{M}}$ and is denoted by $\mathcal{T}_{y_{\mathcal{M}}} \mathcal{M}$.

In section 3.3.2, we will show that the gradient of a function defined on a manifold belongs to the tangent bundle - the set of all tangent spaces - of that manifold.

We turn our attention to the concept of minimizing curves between points on the manifold, which allows the definition of a manifold distance metric.

Definition 2.3 ([134]). Let $\mathcal{M} \subset \mathbb{R}^q$ be a connected smooth m -manifold. For $y_{\mathcal{M}}, w_{\mathcal{M}} \in \mathcal{M}$, the set of smooth curves $[0, 1] \rightarrow \mathcal{M}$ connecting $y_{\mathcal{M}}$ and $w_{\mathcal{M}}$ is denoted as $\Omega_{y_{\mathcal{M}}, w_{\mathcal{M}}} := \{\gamma_{\mathcal{M}} : [0, 1] \rightarrow \mathcal{M} | \gamma_{\mathcal{M}} \text{ is smooth, } \gamma_{\mathcal{M}}(0) = y_{\mathcal{M}}, \gamma_{\mathcal{M}}(1) = w_{\mathcal{M}}\}$. Since $\mathcal{M}$ is connected, the set is nonempty for all pairs of points on $\mathcal{M}$.

The distance function $d_{\mathcal{M}} : \mathcal{M} \times \mathcal{M} \rightarrow [0, \infty]$ is defined by $d_{\mathcal{M}}(y_{\mathcal{M}}, w_{\mathcal{M}}) := \inf_{\gamma_{\mathcal{M}} \in \Omega_{y_{\mathcal{M}}, w_{\mathcal{M}}}} L(\gamma_{\mathcal{M}})$, where $L(\gamma_{\mathcal{M}}) = \int_0^1 |\dot{\gamma}_{\mathcal{M}}(t)| dt$ is the length of curve $\gamma_{\mathcal{M}}$.

Lemma 2.1 ([134]). The function $d_{\mathcal{M}} : \mathcal{M} \times \mathcal{M} \rightarrow [0, \infty]$ defines a metric on $\mathcal{M}$:

1. $y_{\mathcal{M}} = w_{\mathcal{M}} \Leftrightarrow d_{\mathcal{M}}(y_{\mathcal{M}}, w_{\mathcal{M}}) = 0 \quad \forall y_{\mathcal{M}}, w_{\mathcal{M}} \in \mathcal{M}$.
2. $d_{\mathcal{M}}(y_{\mathcal{M}}, w_{\mathcal{M}}) = d_{\mathcal{M}}(w_{\mathcal{M}}, y_{\mathcal{M}}) \quad \forall y_{\mathcal{M}}, w_{\mathcal{M}} \in \mathcal{M}$.
3. $d_{\mathcal{M}}(y_{\mathcal{M}}, w_{\mathcal{M}}) \leq d_{\mathcal{M}}(y_{\mathcal{M}}, h_{\mathcal{M}}) + d_{\mathcal{M}}(h_{\mathcal{M}}, w_{\mathcal{M}}) \quad \forall y_{\mathcal{M}}, w_{\mathcal{M}}, h_{\mathcal{M}} \in \mathcal{M}$.

It is important to highlight the fact that the Manifold distance function is not in general equal to the Euclidean distance function in the embedding space $\mathbb{R}^q$, as per the following remark, which is an important incentive for the work presented in this thesis.

Remark Every smooth curve $\gamma_{\mathcal{M}} : [0, 1] \rightarrow \mathcal{M} \subset \mathbb{R}^q$ with $\gamma_{\mathcal{M}}(0) = y_{\mathcal{M}}$ and $\gamma_{\mathcal{M}}(1) = w_{\mathcal{M}}$ satisfies the inequality $L(\gamma_{\mathcal{M}}) \geq \|y_{\mathcal{M}} - w_{\mathcal{M}}\|$, where $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^q$.

Thus, the distance between two points on the manifold will always be equal or greater to their distance, when the points are considered as belonging to $\mathbb{R}^q$.

Moving forward, we now present details connecting the manifold distance to tangent vectors, which are helpful for understanding the methodological details in section 3.3.

Definition 2.4 ([134]). *Let $I \subset \mathbb{R}$ be an open interval. A smooth curve $\gamma_{\mathcal{M}} : I \rightarrow \mathcal{M}$ is called a geodesic if $|\dot{\gamma}_{\mathcal{M}}(t)|$ is constant $\forall t \in I$.*

Geodesics can be seen as curves moving with a constant velocity on the manifold. It is easy to verify that in a Euclidean space, a geodesic corresponds to a straight line.

Definition 2.5 ([134]). *Let $I_{y, \mathbf{u}_{\mathcal{T}}}$ denote the union of all open intervals in $\mathbb{R}$ containing 0 and 1, such that there exist a geodesic $\gamma_{\mathcal{M}}$ satisfying $\gamma_{\mathcal{M}}(0) = y_{\mathcal{M}}$ and $\dot{\gamma}_{\mathcal{M}}(0) = \mathbf{u}_{\mathcal{T}}$. The set $V_y \subset \mathcal{T}_{y_{\mathcal{M}}} \mathcal{M}$ is defined by $V_y := \{\mathbf{u}_{\mathcal{T}} \in \mathcal{T}_{y_{\mathcal{M}}} \mathcal{M} | \gamma_{\mathcal{M}} \in I_{y, \mathbf{u}_{\mathcal{T}}}, \dot{\gamma}_{\mathcal{M}}(0) = \mathbf{u}_{\mathcal{T}}\}$. The exponential map $\exp_{y_{\mathcal{M}}} : V_y \rightarrow \mathcal{M}$ at point $y_{\mathcal{M}} \in \mathcal{M}$ assigns to every tangent vector $\mathbf{u}_{\mathcal{T}} \in V_y$ the point $\exp_{y_{\mathcal{M}}}(\mathbf{u}_{\mathcal{T}}) = \exp(y_{\mathcal{M}}, \mathbf{u}_{\mathcal{T}}) = \gamma_{\mathcal{M}}(1)$.*

The exponential map provides the necessary construct to move from a manifold point's tangent space to the manifold. Of equal importance is the exponential map's inverse function, defined as follows:

Corollary 2.2 ([134]). *The exponential map is smooth. As such there exist an open interval $B_y \subset V_y$ for which $\exp_{y_{\mathcal{M}}}$ is a diffeomorphism. The inverse function $\log_{y_{\mathcal{M}}} : \mathcal{M} \rightarrow B_y$, termed logarithmic map, is defined as*

$$\log_{y_{\mathcal{M}}}(w_{\mathcal{M}}) := \exp_{y_{\mathcal{M}}}^{-1}(w_{\mathcal{M}}) . \quad (2.1)$$

Theorem 2.3 ([134]). *Let $\mathcal{M} \subset \mathbb{R}^q$ be a smooth m -manifold. Fix a point $y_{\mathcal{M}} \in \mathcal{M}$ and let $\mathbf{u}_{\mathcal{T}} \in \mathcal{T}_{y_{\mathcal{M}}} \mathcal{M}$ with $|\mathbf{u}_{\mathcal{T}}|$ small enough so that $w_{\mathcal{M}} := \exp_{y_{\mathcal{M}}}(\mathbf{u}_{\mathcal{T}}) \in B_y$. Then $d_{\mathcal{M}}(y_{\mathcal{M}}, w_{\mathcal{M}}) = |\mathbf{u}_{\mathcal{T}}|$, and the length minimizing geodesic $\gamma_{\mathcal{M}}$, with $\gamma_{\mathcal{M}}(0) = y_{\mathcal{M}}$ and $\gamma_{\mathcal{M}}(1) = w_{\mathcal{M}}$, is unique.*

Theorem 2.3 states an important result with regards to the geometry of a smooth manifold. In essence, for a point $y_{\mathcal{M}} \in \mathcal{M}$ there exist a small enough neighborhood around $y_{\mathcal{M}}$, such that the minimizing length geodesic paths connecting $y_{\mathcal{M}}$ to all other points on the neighborhood are unique. Furthermore, the distance between $y_{\mathcal{M}}$ and another point in its neighborhood is equal to the norm of the tangent vector that connects the two points through the exponential map. Using the definition of the logarithmic map, we can write

$$d_{\mathcal{M}}(y_{\mathcal{M}}, w_{\mathcal{M}}) = |\log_{y_{\mathcal{M}}}(w_{\mathcal{M}})| . \quad (2.2)$$

The Hopf-Rinow theorem extends the above result on the complete manifold surface, provided that the manifold is geodesically complete:

Definition 2.6 ([134]). *Let $\mathcal{M} \subset \mathbb{R}^q$ be a smooth m -manifold and $y_{\mathcal{M}}$ a point in $\mathcal{M}$. $\mathcal{M}$ is said to be geodesically complete at $y_{\mathcal{M}}$, if for every $\mathbf{u}_{\mathcal{T}} \in \mathcal{T}_{y_{\mathcal{M}}}\mathcal{M}$, there is a geodesic curve $\gamma_{\mathcal{M}} : \mathbb{R} \rightarrow \mathcal{M}$ satisfying $\gamma_{\mathcal{M}}(0) = y_{\mathcal{M}}$ and $\dot{\gamma}_{\mathcal{M}}(0) = \mathbf{u}_{\mathcal{T}}$. $\mathcal{M}$ is called geodesically complete if the above stands for every point on the manifold.*

Theorem 2.4 (Hopf-Rinow [134]). *Let $\mathcal{M} \subset \mathbb{R}^q$ be a smooth, connected and geodesically complete m -manifold and $y_{\mathcal{M}}$ a point in $\mathcal{M}$. Then, for every point $w_{\mathcal{M}} \in \mathcal{M}$, there is a geodesic $\gamma_{\mathcal{M}} : \mathbb{R} \rightarrow \mathcal{M}$ such that*

$$\gamma_{\mathcal{M}}(0) = y_{\mathcal{M}} , \quad \gamma_{\mathcal{M}}(1) = w_{\mathcal{M}} , \quad L(\gamma_{\mathcal{M}}) = d_{\mathcal{M}}(y_{\mathcal{M}}, w_{\mathcal{M}}) . \quad (2.3)$$

Based on the above theorem, we see that equation (2.2) gives the connection between the manifold's distance function and the logarithmic map on the complete manifold. This will constitute the basis of our gradient descent algorithm presented in section 3.3.2.

In summary, for a smooth, connected and geodesically complete manifold, it is possible to measure and identify geodesic paths and distances by working on the manifold's tangent bundle, which is comprised of all tangent spaces in $\mathcal{M}$. In the length of this thesis, unless otherwise described, when we are referring to a manifold $\mathcal{M}$, we assume that $\mathcal{M}$ is smooth, connected

and geodesically complete.

2.2 Polyhedral Surfaces as Approximations of Manifold Surfaces

In Chapters 3 and 4, we are going to discuss correspondence sampling methodologies for polyhedral surfaces in $\mathbb{R}^3$. These surfaces can be considered piecewise planar approximations of 2-manifolds in $\mathbb{R}^3$. A polyhedral surface $\mathcal{S}$ can be represented, without loss of generality, as a triangular mesh, consisting of vertices and triangle faces.

2.2.1 Shortest Polyhedral Geodesics

An appropriate formulation of the manifold distance function for polyhedral surfaces has been given in [114]. Before providing the details, we introduce some necessary notions. Given edge-adjacent mesh faces $f_i^{\mathcal{S}}$ and $f_j^{\mathcal{S}}$, sharing edge $e^{\mathcal{S}}$, the planar unfolding of $f_i^{\mathcal{S}}$ onto $f_j^{\mathcal{S}}$ is defined as the projected image of points of $f_i^{\mathcal{S}}$ when these are rotated along edge $e^{\mathcal{S}}$ to become coplanar with $f_j^{\mathcal{S}}$ and lying on the opposite side of it. Planar unfolding can be extended to a sequence of edge-adjacent faces $f_1^{\mathcal{S}}, \dots, f_n^{\mathcal{S}}$, where $f_1^{\mathcal{S}}$ is unfolded onto $f_2^{\mathcal{S}}$, then $f_1^{\mathcal{S}}, f_2^{\mathcal{S}}$ are unfolded onto $f_3^{\mathcal{S}}$, and the same procedure is repeated until all faces have been unfolded and are coplanar with $f_n^{\mathcal{S}}$. The unfolded image of a path $p^{\mathcal{S}}$ between points belonging to $f_1^{\mathcal{S}}$ and $f_n^{\mathcal{S}}$ is taken to be the resulting path after the unfolding of the faces $f_1^{\mathcal{S}}, \dots, f_n^{\mathcal{S}}$ [114].

We now consider two vertex-adjacent faces $f_k^{\mathcal{S}}$ and $f_l^{\mathcal{S}}$, which share a common vertex $v^{\mathcal{S}}$. The angle of a path $p^{\mathcal{S}}$ passing from $f_k^{\mathcal{S}}$ to $f_l^{\mathcal{S}}$ through $v^{\mathcal{S}}$ is defined as follows. Let the path segments belonging to $f_k^{\mathcal{S}}$ and $f_l^{\mathcal{S}}$ respectively be $\overline{p_k^{\mathcal{S}}}$, $\overline{p_l^{\mathcal{S}}}$. The clockwise angle between $\overline{p_k^{\mathcal{S}}}$ and $\overline{p_l^{\mathcal{S}}}$ is taken to be the angle between these segments, after the sequence of faces containing $v^{\mathcal{S}}$, from $f_k^{\mathcal{S}}$ to $f_l^{\mathcal{S}}$ inclusive, moving on a clockwise direction, are unfolded onto $f_l^{\mathcal{S}}$. The anticlockwise angle is computed in the same fashion, by taking the angle between the segments $\overline{p_k^{\mathcal{S}}}$ and $\overline{p_l^{\mathcal{S}}}$, after unfolding the sequence of faces containing $v^{\mathcal{S}}$, from $f_k^{\mathcal{S}}$ to $f_l^{\mathcal{S}}$ inclusive in an anticlockwise

direction. The angle of path $p^{\mathcal{S}}$ through $v^{\mathcal{S}}$ is then defined as the minimum of these two angles [114].

The distance between two points in $\mathcal{S}$ is computed as the overall path length of piece-wise linear segments on the mesh faces, connecting the two points. Such a path, called hereafter as a shortest polyhedral geodesic is constructed as follows

Lemma 2.5 ([114]). *The shortest geodesic path between two points in a polyhedral mesh goes through an alternating sequence of vertices and possibly empty edge sequences (in which no edge can appear twice), such that the unfolded image of the path along any edge sequence is a straight line segment and the angle of the path passing through a vertex is greater than or equal to π .*

An illustrative example of an unfolded shortest polyhedral geodesic is provided in Fig. 2.1. The above formulation is optimal in the sense that, given two points on a polyhedral surface, it guarantees that a unique path can be constructed, whose length is the geodesic distance between these points. The unfolding of a shortest geodesic path equates to the logarithmic map, in the case of a polyhedral surface.

2.2.2 Straightest Polyhedral Geodesics

A drawback, of the shortest polyhedral path formulation, as given above, is that it cannot always be inverted in order to produce a unique formulation of the exponential map. In particular, given a point on the surface and a tangent vector representing the unfolding on the shortest geodesic path, if at some stage of the reverse folding process, the path passes through a mesh vertex, it is not possible to uniquely identify the path's direction without knowledge of the end point.

The construction of an exponential map in polyhedral surfaces is possible through the definition of straightest polyhedral geodesics given in [130]. Straightest geodesics are defined identically to the shortest geodesics, apart from cases where the path passes through a mesh vertex. Then, instead of preserving the minimum distance property, straightest geodesics preserve the

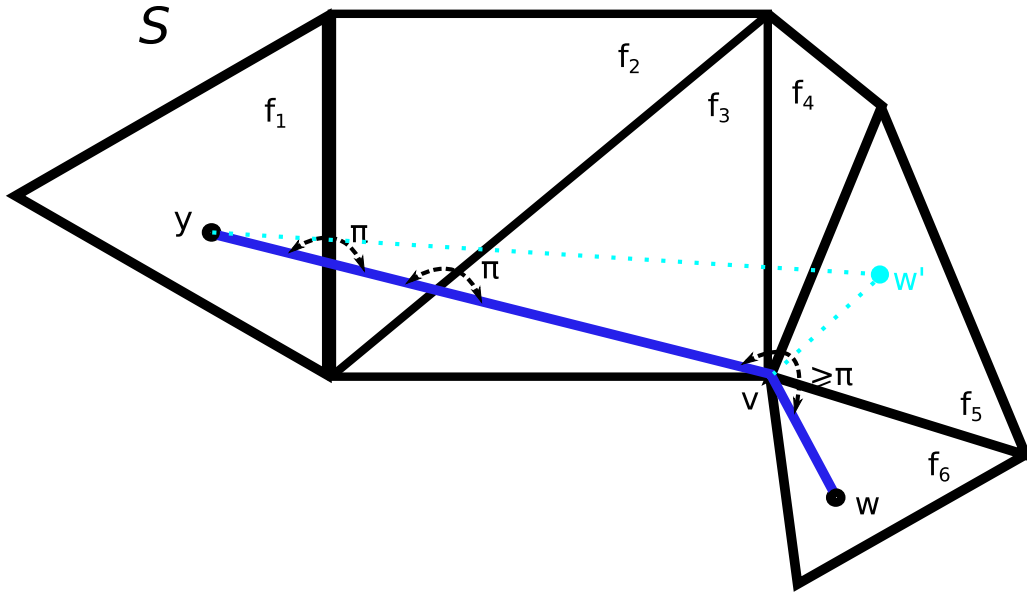


Figure 2.1: Illustrative example of an unfolded shortest polyhedral geodesic. The unfolded shortest polyhedral geodesic path between points y_S and w_S , belonging to faces f_1^S and f_6^S respectively, follows a straight line through faces f_2^S and f_3^S , because it passes through their adjacent edges. Since the path goes through vertex v^S , it is necessary that, when passing from that vertex, it creates an angle that is greater than or equal to π . It is easy to verify that, if the angle was less than π , then a different shortest path would be possible to be constructed, which does not pass through vertex v^S .

angular direction of the geodesic path to and from that vertex. In particular, the direction of a straightest geodesic through a mesh vertex is found by equating clockwise and anti-clockwise path angles through that vertex. See Fig. 2.2 (a) for an example of two straightest polyhedral geodesics. It must be noted that, due to this difference, straightest geodesics passing through spherical vertices (vertices whose sum of adjacent face angles is less than 2π) are not locally shortest. Furthermore, there exist regions in the local area around hyperbolic vertices (vertices whose sum of adjacent face angles is greater than 2π) in which two points cannot be connected by a straightest geodesic [130]. Figs. 2.2 (b),(c) illustrate these two cases.

By appropriately alternating between the two definitions of shortest and straightest polyhedral geodesics, it is possible to define logarithmic and exponential maps on polyhedral surfaces, as well as compute geodesic distances.

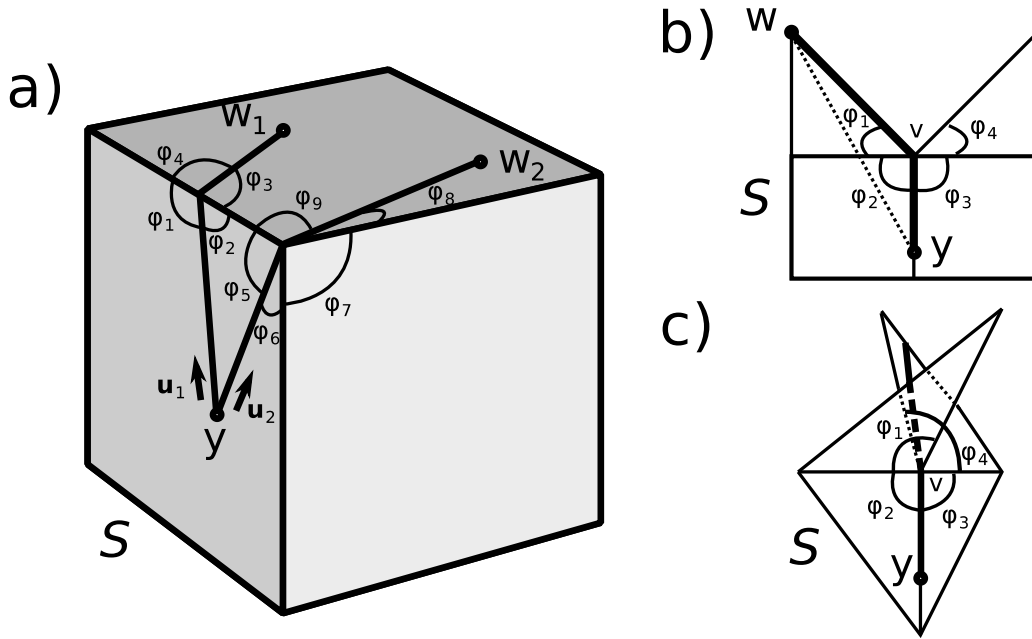


Figure 2.2: Illustrative example of straightest polyhedral geodesics. (a) Examples of two straightest polyhedral geodesics from one cube face to its adjacent face. Let the first geodesic from point y_S , towards direction $\mathbf{u}_1$, end in w_{S_1} . The path passes through an edge, as such it is equivalent to a shortest polyhedral geodesic, with $\phi_1 + \phi_4 = \phi_2 + \phi_3 = \pi$. Now consider the geodesic from point y_S , towards direction $\mathbf{u}_2$, that ends in w_{S_2} , which is not known a priori. Since the path passes through a vertex, the straightest geodesic direction through it is identified by equating clockwise and anticlockwise face angles, in particular for this example, $\phi_6 + \phi_7 + \phi_8 = \phi_5 + \phi_9$. (b) Example of an unfolded spherical vertex, with $\phi_1 + \phi_2 + \phi_3 + \phi_4 < 2\pi$. Let The straightest geodesic from y_S through vertex v^S end up in w_S (bold line). It is obvious that a shortest geodesic path exists between these points, which is not the same (dotted line). (c) Example of an unfolded hyperbolic vertex, with $\phi_1 + \phi_2 + \phi_3 + \phi_4 > 2\pi$. From a start point y_S , a straightest geodesic towards a point in the colored regions should pass through the vertex. Since the straightest geodesic is defined uniquely as the direction with equal clockwise and anti-clockwise path angles, only points lying on the straightest path, here illustrated by the bold line, are accessible, while points outside that line cannot be connected to y_S with a straightest geodesic.

2.3 Random Forest

Random Forest (RF) is a non-parametric, non-linear regression and classification algorithm which can be used for feature selection and prediction [23, 72]. In this work, we use RF for both purposes in Chapters 4 and 5 respectively.

RF comprises of a collection of Classification and Regression Trees (CARTs). A CART splits the input space recursively into small rectangular regions and then fits a simple model, commonly a constant value, in each one of them. In detail, training data are fed in the root node of

each tree. According to a predefined cost criterion, one univariate feature predictor and a split value are selected and the data are driven to the left or right children nodes. The process is repeated until a stop criterion is satisfied or all terminal nodes, called leafs, contain just one observation. Each leaf is associated with a subset of the training points. For prediction, a new input observation follows the split decision rules learned during training and reaches a leaf in the tree. It is then assigned a suitable value on the response space, most often the average response of that leaf's response training observations.

Individual trees are highly susceptible to overfitting and suffer from noise in the data. The ensemble method of RF overcomes these problems by fitting a large collection of uncorrelated trees to the data and making predictions by averaging the outcomes of all trees. Disassociation of trees is achieved by using different bootstrapped versions of the training set to build each CART. In addition, for each node in a tree, a random subset of features is considered as candidates for splitting the data in the children nodes. This procedure helps to alleviate the problems of the CART algorithm [72].

2.3.1 Training

For regression problems, a common split objective function of tree-based methods, and RF in particular, is the difference between a parent node's sample squared deviation from the mean (MSD) and the sum of MSDs for samples in its children nodes. Let a training dataset S of paired input-response points be $\{(x_i, y_i) | x_i \in \mathbb{R}^p, y_i \in \mathbb{R}^q, i = 1, \dots, N\}$. Then the splitting objective function can be written as:

$$G_{RF}(S_n) = \sum_{y_i \in S_n} \|y_i - \bar{y}_n\|^2 - \sum_{y_i \in S_{nl}} \|y_i - \bar{y}_{nl}\|^2 - \sum_{y_i \in S_{nr}} \|y_i - \bar{y}_{nr}\|^2, \quad (2.4)$$

where $\bar{y}_n$, $\bar{y}_{nl}$, and $\bar{y}_{nr}$ are mean responses of the data subsets S_n , S_{nl} , S_{nr} in node n_{RF} and its left and right children nodes respectively. Observations are split on left and right subsets according to one input feature x_{j_n} and a split value s_{j_n} for this predictor. The aim is to locate the optimal split (j_n, s_{j_n}) at each node n_{RF} , which maximizes the objective function resulting

from the particular partition, i.e.

$$(j_n, s_{j_n}) = \underset{(\{1, \dots, p\}, \mathbb{R})}{\operatorname{argmax}} G_{RF}(S_n) . \quad (2.5)$$

Intuitively, we are looking for that split in the input space that creates two data subsets with as low variances as possible in the response space.

2.3.2 Prediction

As outlined previously, prediction in RF is traditionally performed by an averaging procedure. A previously unseen input point x_{new} is fed to the root nodes of all trees in the RF. For each tree, the point traverses down through the nodes according to the learned optimal node splits until it reaches a terminal leaf. The average response of the training points in that leaf is the prediction acquired from that tree. A second average over all trees yields the response point estimate $\hat{y}_{new}$. In detail:

$$\hat{y}_{new} = \frac{1}{T_{RF}} \sum_{t=1}^{T_{RF}} \left(\frac{1}{N_t} \sum_{y_i \in S_t} y_i \right) , \quad (2.6)$$

where T_{RF} is the number of trees in the forest, S_t the subset of points in the leaf of tree t that the input observation ended in and N_t the cardinality of S_t .

2.3.3 Variable Importance

RF has the inherent ability to rank individual input variables according to how well they are associated to the responses. In the related literature, this attribute is commonly encountered under the term Variable Importance (VI) [63]. Various formulation of VI criteria have been presented, with the two most significant being Permutation VI (pVI) and Gini VI (gVI). pVI can be viewed as the predictive quality of individual input variables, since for each one, pVI is computed by permuting the observations of that variable and measuring the resulting increase in model prediction error. A variable that results in zero error increase should have no association with the responses. In contrast, large increase of error is translated as high VI.

gVI, on the other hand, is computed during model fitting. Below we present gVI in more detail, since, in chapter 4 we are going to utilize gVI for feature selection. This choice was made due to the more intrinsic nature of the gini index, which does not depend on model prediction. For each input feature $x_{.k}$, $k = 1, \dots, p$, its gVI is computed as follows

$$g_{VI}(k) = \frac{1}{T_{RF}} \sum_{t=1}^{T_{RF}} \sum_{l \in Q^{tk}} G_l^{tk} , \quad (2.7)$$

where Q^{tk} indicates the set of node indices in tree t , for which $x_{.k}$ was selected as the optimal feature, according to (2.5), and G_l^{tk} represents the result value of the objective function G_{RF} for the l_{th} such node.

In essence, for each tree, the importance of each feature $x_{.k}$ is computed as the sum of the cost criterion values on the nodes in which $x_{.k}$ was selected for splitting. By averaging over the trees in the forest, we acquire the RF gVI. In other words, gVI is the average reduction in MSD for which each feature is responsible for. A high gVI value corresponds to a bigger overall reduction in MSD, which in turn is translated as high importance.

2.4 Manifold Learning

In chapter 5 we employ Manifold Learning (ML) techniques for the purposes of constructing a manifold regression algorithm. Here we introduce ML and present the details of these techniques.

In many real-life applications, such as face recognition, image and network analysis, we deal with data objects that do not comply to the assumptions of Euclidean geometry, but, in addition, can not be represented as points on analytically defined manifolds [170]. It is also common for such objects to be simplistically represented as points on very high dimensional spaces, while the intrinsic dimensionality of the data - a.k.a. the number of parameters needed to account for the observed properties of that data - can be significantly lower.

ML techniques constitute a family of non-linear dimensionality reduction methodologies, con-

structured to deal with this problem [170]. In particular, given a dataset of observed data points, with dimensionality k , ML methods aim to project these points onto a space with lower dimensionality $q < k$ - ideally q being the intrinsic dimensionality of a dataset -, whilst retaining as well as possible the original geometrical structure of the data.

Let $S = \{y_i | y_i \in \mathcal{M} \subset \mathbb{R}^q, i = 1, \dots, N\}$ be a dataset of manifold-valued observations, with $\mathcal{M}$ being an unknown manifold of assumed dimensionality $m < q$, and a manifold distance function $d_{\mathcal{M}}(\cdot, \cdot)$. Below we describe two well-known ML methods, that are later employed in our work.

2.4.1 Metric Multi-Dimensional Scaling

Metric multi-Dimensional Scaling (MDS) [39, 22] computes an approximation $\{z_i | z_i \in \mathbb{R}^m, i = 1, \dots, N\}$ of the manifold-valued dataset that resides on a m -dimensional Euclidean space, by minimizing a stress function of the form:

$$\sum_{i=1}^N \sum_{j=1}^N (d_{\mathcal{M}}^2(y_i, y_j) - \|z_i - z_j\|^2) \quad , \quad (2.8)$$

where $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^m$. In essence, Metric MDS attempts to construct a Euclidean-valued dataset on which pairwise distances between points are as close as possible to their original manifold counterparts.

Theorem 2.6 ([64, 108]). *Let $\mathbf{D}_{\mathcal{M}}^{(2)}$ be the $N \times N$ matrix of squared manifold distances with elements $D_{\mathcal{M},ij}^{(2)} = d_{\mathcal{M}}^2(y_i, y_j)$, and*

$$\mathbf{K}_{\mathcal{M}} = -\frac{1}{2} \mathbf{H} \mathbf{D}_{\mathcal{M}}^{(2)} \mathbf{H} \quad , \quad (2.9)$$

where $\mathbf{H} = \mathbf{I}_N - \frac{1}{N} \mathbf{e} \mathbf{e}^T$, with $\mathbf{I}_N$ the $N \times N$ identity matrix and $\mathbf{e}$ an $N \times 1$ column vector of all ones.

If the distance function $d_{\mathcal{M}}$ is a metric, then $\mathbf{K}_{\mathcal{M}}$ is positive semi-definite.

Individual elements of matrix $\mathbf{K}_{\mathcal{M}}$ are given by

$$K_{\mathcal{M},ij} = -\frac{1}{2}(D_{\mathcal{M},ij}^{(2)} - \frac{1}{n}D_{\mathcal{M},\cdot j}^{(2)} - \frac{1}{n}D_{\mathcal{M},i\cdot}^{(2)} + \frac{1}{n^2}D_{\mathcal{M},\cdot\cdot}^{(2)}), \quad i, j = 1, \dots, N, \quad (2.10)$$

where $D_{\mathcal{M},\cdot j}^{(2)} = \sum_i D_{\mathcal{M},ij}^{(2)}$, $D_{\mathcal{M},i\cdot}^{(2)} = \sum_j D_{\mathcal{M},ij}^{(2)}$, $D_{\mathcal{M},\cdot\cdot}^{(2)} = \sum_{i,j} D_{\mathcal{M},ij}^{(2)}$ are the i^{th} row, j^{th} column and overall element-wise sums of $\mathbf{D}_{\mathcal{M}}^{(2)}$ respectively. Equation (2.10) is commonly referred to as the double centering formula.

Theorem 2.7 ([64, 108]). *Let $\mathbf{K}_{\mathcal{M}} = -\frac{1}{2}\mathbf{H}\mathbf{D}_{\mathcal{M}}^{(2)}\mathbf{H}$. Suppose that $\mathbf{K}_{\mathcal{M}}$ is a positive semi-definite matrix with rank p ($p \leq N-1$). Let $\lambda_1, \dots, \lambda_p$ be the p ordered non-zero eigenvalues of $\mathbf{K}_{\mathcal{M}}$ with corresponding eigenvectors $\mathbf{v}_1^\lambda, \dots, \mathbf{v}_p^\lambda$, and $\mathbf{v}_l^\lambda = (v_{l1}^\lambda, \dots, v_{lN}^\lambda)^\lambda$, $l = 1, \dots, p$. Then the solution to the MDS problem (2.8), assuming $m \leq p$, is given by*

$$z_{il} = \sqrt{\lambda_l} v_{li}^\lambda, \quad l = 1, \dots, m, \quad i = 1, \dots, N, \quad (2.11)$$

with $z_i = (z_{i1}, \dots, z_{im})$.

2.4.2 Isomap

The Isomap ML technique [165] is identical to metric MDS with the addition of a suitable algorithm for the estimation of the distance matrix $\mathbf{D}_{\mathcal{M}}^{(2)}$, when a manifold distance function is not known. Reminding that $y_i \in \mathcal{M} \subset \mathbb{R}^q$, the geodesic distances between pairs of data points are approximated as follows. A neighborhood graph is constructed, in which each point y_i is connected to its k nearest neighbors in $\mathbb{R}^q$. Graph edges are assigned weights equal to the Euclidean distances between the connected points in $\mathbb{R}^q$. For any two points y_i and y_j in S , $d_{\mathcal{M}}(y_i, y_j)$ is then estimated by the shortest path connecting y_i and y_j in that graph.

Attention is required when selecting the nearest neighbor parameter k , since it can significantly affect the quality of the approximations. Large k will result in a densely connected graph, with distance paths very similar to the Euclidean distances in $\mathbb{R}^q$. On the other hand, if k is selected to be too small, then the geodesic distances may be substantially overestimated,

while the graph may not be connected, meaning that it will not be possible to approximate all pairwise distances. In the rest of this work, we will refer to the distance formulation described here, as the Isomap metric.

Chapter 3

Geodesic Correspondence Sampling for Ensembles of Polyhedral Surfaces

3.1 Introduction

Statistical shape analysis of 3D surfaces play a key role in various computer vision and medical imaging applications [100]. It is an intuitive and efficient approach to characterizing the variability and structural changes in morphology and anatomy for ensembles of similar objects, for example cardiac, brain or facial surfaces [40, 171, 143, 6, 81]. Morphological traits have been used successfully in the past for recognition applications, as well as for studying and delineating causes and effects of particular diseases, biological processes such as aging, and the evolutionary interplay between genes and environment [100].

A primary requirement for studying shape variability in a data sample is the establishment of some type of accurate correspondence across surface objects, which most often take the form of landmark annotation or dense point correspondence. Landmarks are points positioned on all considered surfaces, in locations that must correspond in some way amongst the different subjects [41], and are anatomical features and points of interest. For example, commonly encountered facial surface landmarks are the eye and mouth corners. In the case of dense point correspondence, surface points, sometimes also referred to quasi-landmarks [35], are sampled

within the surfaces' boundaries [12].

An accurate correspondence of points across all surfaces is paramount for shape analysis. Until recently, the primary approach for achieving surface annotation was to manually localize a small number of landmarks on every data object. Manually choosing landmark points, though, can be not only cumbersome and restricted to small numbers of landmarks, but also hard to replicate consistently, and, furthermore, may suffer from potential investigators' bias. In the last years, a number of semi-automated and automated landmarking and point correspondence algorithms have appeared in the literature, which attempt to address these issues. In the following we concentrate specifically to methodologies that compute dense correspondences between surfaces, meaning that thousands of points are annotated in each object. This large number of corresponding annotated markers allows morphological variability to be captured and quantified in a much more granular level, compared to sparse annotations.

3.1.1 Related Work

Dense correspondence methods can be classified into one of two categories, based on whether annotation is performed in a pairwise fashion - each surface is registered to a common template, with ensemble correspondence extracted implicitly from all one-to-one surface annotations - or a groupwise fashion - all surfaces are considered simultaneously during annotation -. Pairwise methods constitute the majority in the existing literature. Within that family, the most prominent approach is based on fitting, or warping, a face template or a parametric surface model independently on each of the surfaces in the dataset, using algorithms such as Rigid/Non-rigid Iterative Closest Point [2, 21], Thin-Plate Spline point matching [33] or Markov Random Field labelling [142, 189]. By decomposition of the dense correspondences, models explaining the surfaces' data variability can be subsequently constructed, such as 3D Morphable Models [17, 21], Point Distribution Models [35, 62, 83].

The main bulk of these methodologies require an initial annotation of a number of landmarks in either a training set of surfaces, or even on the complete dataset, in order to initially guide the correspondence procedure. A large literature exists on extensions attempting to weaken

or alleviate this requirement by incorporating the use of additional preprocessing steps. In particular, a number of methodologies used the minimum description length principle or maximization of mutual information [41, 167, 136, 67, 91] to address the problem of learning a base facial model. Another approach depends on automatic extraction of few landmarks based on geometric principles, such as curvature [189, 195]. Finally, another commonly adopted procedure is to map 3D meshes to a planar domain, commonly through the use of harmonic or conformal maps [176, 179], thus transforming the 3D correspondence problem to one of 2D image landmark localization [4, 21, 65, 190].

The second class of dense correspondence methodologies attempts to optimize the location of surface points jointly on all considered surfaces. Although the result is the same as applying one-to-one correspondences between all surfaces in a group and a template or mean surface model, considering all data at once can have certain advantages, such as removing the need for the construction of a template and reducing the bias that can result by the multiple pairwise fits to one surface [171]. Groupwise functional correspondence methods reformulate the problem as one of finding correspondence between real-valued functional representations of the mesh surfaces, extracted for example using Heat or Wave Kernel Signatures [119, 191, 24]. In [28], the groupwise correspondence was achieved by optimizing the compactness of the surfaces' distribution, represented as vectors of their points' coordinates, subject to constraints enforcing uniform distribution of points on individual surfaces, while in [19], that methodology was extended to the case of multilinear facial models, accounting for multiple images and facial expressions from the same subjects. Finally, a number of algorithms, usually coined as groupwise methodologies, adopt an intermediate approach to the correspondence problem. They are based on bottom-up iterative pairwise alignments of similar surfaces, driven by an affinity graph connecting similar surfaces [164, 175].

3.1.2 Motivation

While extensive work has been done in the problem of identifying corresponding points on sets of surfaces, as outlined previously, it is possible to identify some key methodological issues that

are still prominent. In pairwise correspondence methods, the construction of a face model, or template, can be a tedious and error-prone procedure. Furthermore, the geometry of the final data objects after correspondence optimization, can be biased towards the mean or template surface, as has been highlighted in previous work [171, 19].

Apart from these shortcomings, a further consideration arises from the common need to use training datasets in order to build the surface models. While it is true that for particular objects, such as facial surfaces, some annotated databases do exist, see for example [188, 146], and can be used for this purpose, such requirement poses a significant restriction to wide applicability of the correspondence methodologies. On the other hand, most fully automated methodologies are primarily tailored towards a specific type of surface data, which allows the use of more sophisticated correspondence procedures, but comes at the cost of high specificity.

Iterative pairwise methods could address, to a certain extent, such problems, but have not been extensively used so far, probably due the extra computational cost they incur, as well as the difficulty in constructing meaningful affinities between unregistered surfaces. Functional groupwise methodologies may not suffer from the above issues, but are also associated with specific shortcomings. The most prominent, regarding the problem of point-to-point correspondence, being that the reverse mapping, from corresponding functions, to surface points, is not always easy to construct.

In [28], groupwise correspondence is achieved by manipulating the location of sampled points on the surfaces, such that an objective function comprised of two entropy-based terms is optimized. These terms are related to the uniform distribution of points in each surface and the overall correspondence among surfaces. Due to the fact that annotation under this formulation is equivalent to randomly sampling corresponding points from uniform distributions defined on the surfaces, the problem is also coined with the term ensemble surface sampling. The entropy formulation of the correspondence problem, and the associated optimization procedure, are attractive for a number of reasons. First, the methodology does not necessitate the construction of a template or any manual annotation. Second, the number of annotated points can be easily adjusted to the specific application requirements. Third, optimization can be easily tailored

to problems of adaptive sampling, as will be discussed in the next chapter. However, two key problems of the previously presented method could be easily pinpointed. The optimization of the uniform distribution of points on individual surfaces was controlled by a kernel density estimator that did not take into consideration the structure of the surfaces. As a result, the method does not provide an optimal uniform distribution of surface points and can not deal with highly curved surfaces. In addition, the gradient-descent optimization algorithm also ignored surface constraints. Point updates had to be recast on the surface after each iteration, which further impairs efficiency and performance.

Here, we build upon the work in [28] and present our Geodesic Ensemble Surface Sampling Algorithm (GESSA) for the automated identification of dense corresponding point sets across sets of similar polyhedral surfaces. We propose a suitable estimator for the probability density function of a variable defined on a manifold or polyhedral surface, and employ it in the construction of the objective function. Furthermore, a gradient descent algorithm is constructed, which enables the optimization to be performed directly on the surfaces. Incorporating these two key features inside the existing framework, we are able to deal with highly curved surfaces and improve upon computational space requirements. We validated GESSA on a small set of facial surfaces, before using it to perform a large-scale heritability study concerning the shape of the human face.

3.2 Problem Definition

Let us consider an ensemble of N polyhedral surfaces $\{\mathcal{S}_i | \mathcal{S}_i \subset \mathbb{R}^3, i = 1, \dots, N\}$. Our objective is to automatically distribute N collections of M points $\{\psi_j^i | \psi_j^i \in \mathcal{S}_i, j = 1, \dots, M\}$, with $\psi_j^i = (\psi_{j1}^i, \psi_{j2}^i, \psi_{j3}^i) \in \mathcal{S}_i \subset \mathbb{R}^3$, uniformly on each surface, and establish point-wise correspondence among points on the N surfaces, so that each set $\{\psi_j^i | i = 1, \dots, N\}$ defines the position of a quasi-landmark on all surfaces. An important aspect of the problem is that the developed methodology must be able to deal efficiently with high numbers of sampled points, where M can take values up to thousands.

3.3 Geodesic Ensemble Surface Sampling Methodology

The dense correspondence sampling problem can be broken down to two major components; point-wise correspondence optimization across the set of surfaces and uniform sampling of points on individual surfaces. We begin the methodological description by describing this general formulation of correspondence sampling, which is adopted by Cates et al. [28]. We then present each component individually, along with our extensions, before combining them to construct GESSA.

3.3.1 Dense Correspondence Sampling Through Entropy Optimization

Following the dense correspondence formulation first introduced in [28], the problem is broken down to two components. We first focus on the point-wise correspondence problem. Let us assume for the time being that sets of M points have been positioned in possibly random locations on the N surfaces. In subsection 3.3.4 we describe how the algorithm initially positions points on the surfaces. The coordinates of M points on a surface $\mathcal{S}_i$ can be aggregated to represent a point $\phi_i \in \mathbb{R}^{3M}$, with

$$\phi_i = (\psi_{11}^i, \psi_{12}^i, \psi_{13}^i, \dots, \psi_{M1}^i, \psi_{M2}^i, \psi_{M3}^i) . \quad (3.1)$$

The points ϕ_i can be thought of as representations of surfaces in $\mathbb{R}^{3M}$, which hereafter is referred to as the surfaces' shape space. Let $\Phi \in \mathbb{R}^{3M}$ be a random variable on shape space with probability density function $p(\Phi)$, and $\{\phi_i | i = 1, \dots, N\}$ realizations of that random variable. The differential entropy of Φ is $\Gamma(\Phi) = E[-\log p(\Phi)] = -\int_{\mathbb{R}^{3M}} p(\Phi) \log p(\Phi) dV$. Furthermore, let $\hat{\Gamma}(\phi_1, \dots, \phi_N)$ be an estimator of the differential entropy under the observed data.

Point-wise correspondence is optimized through the minimization of $\hat{\Gamma}$, w.r.t the points' positions. Under the assumption that Φ is normally distributed in $\mathbb{R}^{3M}$, the minimization of $\hat{\Gamma}$ is

equivalent to minimizing the sample variance. This increases the compactness of the surfaces' distribution in shape space, which equates to bringing equivalent points in the set of surfaces closer to each other.

The above minimization though would eventually lead all M points in each surface to be collapsed in the same location. Contrary to that, the goal is to distribute landmark points so that they 'cover' the complete surface areas. The solution is to balance the correspondence term with a component imposing a uniform distribution of points on individual surfaces [28].

Let Ψ_i be a random variable on surface $\mathcal{S}_i$ with probability density function $p_i(\Psi_i)$, and $\{\psi_j^i | j = 1, \dots, M\}$, realizations of Ψ_i . The differential entropy of Ψ_i is $H_i(\Psi_i) = E[-\ln p_i(\Psi)] = -\int_{\mathcal{S}_i} p_i(\Psi_i) \ln p_i(\Psi_i) d\mathcal{S}_i$. An estimator is given by the sample mean

$$\hat{H}_i(\psi_1^i, \dots, \psi_M^i) = -\frac{1}{M} \sum_{j=1}^M \ln p_i(\psi_j^i) . \quad (3.2)$$

Proposition 3.1. *The uniform distribution in a bounded surface $\mathcal{S}$ has maximum entropy among all continuous distributions with probability density functions supported in $\mathcal{S}$.*

Proof. Let $p(\psi), q(\psi)$ be continuous probability density functions on a surface $\mathcal{S}$, with $p > 0$, $q > 0$, $\int_{\mathcal{S}} p(\psi) d\mathcal{S} = 1$, $\int_{\mathcal{S}} q(\psi) d\mathcal{S} = 1$.

For any p, q and for all $\psi \in \mathcal{S}$

$$p(\psi) - p(\psi) \log p(\psi) \leq q(\psi) - p(\psi) \log q(\psi) , \quad (3.3)$$

with equality only when $p = q$. Indeed, inequality 3.3 can be written as $\log(\frac{q}{p}) \leq \frac{q}{p} - 1$, which is true for $\frac{q}{p} > 0$, with equality for $\frac{q}{p} = 1$.

By integrating 3.3 over $\mathcal{S}$, we get

$$\begin{aligned} \int_{\mathcal{S}} p(\psi) d\mathcal{S} - \int_{\mathcal{S}} p(\psi) \log p(\psi) d\mathcal{S} &\leq \int_{\mathcal{S}} q(\psi) d\mathcal{S} - \int_{\mathcal{S}} p(\psi) \log q(\psi) d\mathcal{S} \Leftrightarrow \\ &- \int_{\mathcal{S}} p(\psi) \log p(\psi) d\mathcal{S} \leq - \int_{\mathcal{S}} p(\psi) \log q(\psi) d\mathcal{S} . \end{aligned} \quad (3.4)$$

Let now $q(\psi) \equiv c$ be the constant density function with $\int_{\mathcal{S}} c d\mathcal{S} = 1$. We have

$$\begin{aligned}
& - \int_{\mathcal{S}} p(\psi) \log p(\psi) d\mathcal{S} \leq - \int_{\mathcal{S}} p(\psi) \log c d\mathcal{S} \Leftrightarrow \\
& - \int_{\mathcal{S}} p(\psi) \log p(\psi) d\mathcal{S} \leq - \log c \Leftrightarrow \\
& - \int_{\mathcal{S}} p(\psi) \log p(\psi) d\mathcal{S} \leq - \log c \int_{\mathcal{S}} c d\mathcal{S} \Leftrightarrow \\
& - \int_{\mathcal{S}} p(\psi) \log p(\psi) d\mathcal{S} \leq - \int_{\mathcal{S}} c \log c d\mathcal{S} ,
\end{aligned} \tag{3.5}$$

which, by the definition of the differential entropy, concludes the proof of proposition 3.1.

□

By proposition 3.1, the optimization term imposing a uniform distribution of points on an individual surface is selected to be $\hat{H}_i$.

We notice that, while the outcome of the second optimization is equivalent to sampling M points from a uniform distribution defined on the surface, the adopted approach follows the reverse path. In essence, assuming that $\psi_1^i, \dots, \psi_M^i$, have been sampled in some way from the surface $\mathcal{S}_i$, their positions is manipulated so that the density estimator corresponds to the uniform distribution in $\mathcal{S}_i$.

The combined cost function used in the dense correspondence sampling is given by [28]:

$$\hat{\Gamma}(\phi_1, \dots, \phi_N) - l_G \sum_{i=1}^N \hat{H}_i(\psi_1^i, \dots, \psi_M^i) , \tag{3.6}$$

where l_G is a weight parameter that balances the two terms. In the rest of this section we describe the GESSA algorithm for solving the combined optimization problem.

3.3.2 Optimizing the Distribution of Points in Individual Surfaces

The optimization problem for the uniform distribution of M points in surface $\mathcal{S}_i$ can be written as

$$\arg \min_{\phi_i \in \mathbb{R}^{3M}} -\hat{H}_j(\psi_1^i, \dots, \psi_M^i), \text{ s.t. } \psi_1^i, \dots, \psi_M^i \in \mathcal{S}_i, \quad (3.7)$$

with $\hat{H}(\psi_1^i, \dots, \psi_M^i) = -\frac{1}{M} \sum_{j=1}^M \ln p(\psi_j^i)$.

As in [28], in order to minimize the cost function, we employ a gradient descent algorithm, where points are iteratively moved proportionally to the negative gradient of the cost, until no significant improvement in cost can be achieved. The derivative of $-\hat{H}_i$ w.r.t a point ψ_j^i is

$$\frac{d(-\hat{H}_i)}{d\psi_j^i} = \frac{d(\frac{1}{M} \sum_{l=1}^M \ln p(\psi_l^i))}{d\psi_j^i} = -\frac{1}{M} \sum_{l=1}^M \frac{1}{p_i(\psi_l^i)} \frac{dp_i(\psi_l^i)}{d\psi_j^i} . \quad (3.8)$$

To proceed, estimates of $p_i(\psi_l^i)$ are needed. At this point, Cates et al. employed a standard Gaussian Kernel Density Estimator (KDE) in order to estimate the density. That estimator though depends on inaccurate Euclidean pairwise distances between surface points. Here, we construct a geodesic KDE (gKDE) as follows

$$p_i^{KDE}(\psi_l^i) = \frac{1}{M} \sum_{k=1}^M g_{\mathcal{S}}(d_{\mathcal{S}}(\psi_l^i, \psi_k^i), \sigma_{\mathcal{S}}) , \quad (3.9)$$

where $d_{\mathcal{S}}$ is the geodesic distance function on the polyhedral surface. As was detailed in section 2.2, the polyhedral surface can be seen as an approximation of a 2-manifold. The geodesic distance between two points on that surface can be computed as the length of the shortest polyhedral geodesic (SPG) [114], described in section 2.2.1. See section 3.3.5 for more details on how SPGs are computed. The isotropic geodesic kernel function $g_{\mathcal{S}} : \mathcal{S}_i \times \mathcal{S}_i \rightarrow \mathbb{R}$ is given by

$$g_{\mathcal{S}}(d_{\mathcal{S}}(\psi_l^i, \psi_k^i), \sigma_{\mathcal{S}}) = \frac{1}{2\pi\sigma_{\mathcal{S}}} e^{-(2\sigma_{\mathcal{S}})^{-1} d_{\mathcal{S}}^2(\psi_l^i, \psi_k^i)} , \quad (3.10)$$

where $\sigma_{\mathcal{S}}$ is a standard deviation parameter. $g_{\mathcal{S}}$ can be seen as the generalization of the gaussian kernel for manifolds and polyhedral surfaces, and is a semi-positive definite function, since $d_{\mathcal{S}}$

is a distance metric.

In order to reduce significantly the computational cost of the gradient descent algorithm, we make a further assumption. Within one iteration of the gradient descent optimization cycle, and during the computation of the update for point ψ_j^i , the remaining points are considered fixed in their positions. As such, $\frac{dp_i^{KDE}(\psi_l^i)}{d\psi_j^i} \equiv 0$, for $l \neq j$. This is discussed further in section 3.3.5. Utilizing the gKDE estimator and under this assumption, we now have

$$\begin{aligned} \frac{d(-\hat{H}_i)}{d\psi_j^i} &\approx -\frac{1}{M^2} \frac{1}{p_i^{KDE}(\psi_j^i)} \sum_{k=1}^M \frac{d(g_S(d_S(\psi_j^i, \psi_k^i), \sigma_S))}{d\psi_j^i} \\ &= -\frac{1}{M^2} \frac{1}{p_i^{KDE}(\psi_j^i)} \sum_{k=1}^M \frac{d\left((2\pi\sigma_S)^{-1} e^{-(2\sigma_S)^{-1} d_S^2(\psi_j^i, \psi_k^i)}\right)}{d\psi_j^i} \\ &= \frac{1}{2\sigma_S M} \sum_{k=1}^M \frac{g_S(d_S(\psi_j^i, \psi_k^i), \sigma_S)}{\sum_{l=1}^M g_S(d_S(\psi_j^i, \psi_l^i), \sigma_S)} \frac{d(d_S^2(\psi_j^i, \psi_k^i))}{d\psi_j^i}. \end{aligned} \quad (3.11)$$

We concentrate now on the derivative of the geodesic distance w.r.t a surface point. We remind that the geodesic distance between ψ_j^i and ψ_k^i can be equated to the norm of the logarithmic map $\log_{\psi_j^i}(\psi_k^i) \in \mathcal{T}_{\psi_j^i} \mathcal{S}_i$, which is a vector on the tangent space of $\mathcal{S}_i$ at point ψ_j^i (see equation (2.2)). In addition, under the assumption of smoothness and geodesical completeness, this logarithmic map is guaranteed to exist and be unique for any two points (see Hopf-Rinow Theorem 2.4). Let $b \in \mathbb{R}^3$ be the point in the tangent space of ψ_j^i such that $\log_{\psi_j^i}(\psi_k^i) = b - \psi_j^i$. By writing the logarithmic map as a vector in $\mathbb{R}^3$, it is easy to show that $\frac{d(\log_{\psi_j^i}(\psi_k^i))}{d\psi_j^i} = \frac{d(b - \psi_j^i)}{d\psi_j^i} = -1$. Based on the above, we have

$$\frac{d(d_S^2(\psi_j^i, \psi_k^i))}{d\psi_j^i} = \frac{d\|\log_{\psi_j^i}(\psi_k^i)\|^2}{d\psi_j^i} = 2\log_{\psi_j^i}(\psi_k^i) \frac{d(\log_{\psi_j^i}(\psi_k^i))}{d\psi_j^i} = -2\log_{\psi_j^i}(\psi_k^i). \quad (3.12)$$

Finally, we can write

$$\begin{aligned} \frac{d(-\hat{H}_i)}{d\psi_j^i} &\approx -\frac{1}{\sigma_S M} \sum_{k=1}^M \frac{g_S(d_S(\psi_j^i, \psi_k^i), \sigma_S)}{\sum_{l=1}^M g_S(d_S(\psi_j^i, \psi_l^i), \sigma_S)} \log_{\psi_j^i}(\psi_k^i) \\ &\approx -\frac{1}{\sigma_S M} \sum_{k=1}^M \frac{g_S^{jk}}{\sum_{l=1}^M g_S^{jl}} \log_{\psi_j^i}(\psi_k^i), \end{aligned} \quad (3.13)$$

with $g_S^{jk} := g_S(d_S(\psi_j^i, \psi_k^i), \sigma_S)$. The logarithmic map is given by the SPG [114].

Employing now the definition of the exponential map on manifolds 2.5, we get the following gradient descent update algorithm, which guarantees that the updated points will always lie on the surfaces:

$$\psi_{j,t+1}^i = \exp_{\psi_{j,t}^i} \left(-\gamma_G \frac{d(-\hat{H}_i)}{d\psi_j^i} \right), t \geq 0, \quad (3.14)$$

where γ_G is the gradient descent's learning rate parameter. In the case of a Euclidean (flat) surface, the above is equivalent to the standard gradient descent updates, since the exponential map reverts to addition. As described in section 2.2.2, the exponential map for a polyhedral surface can be computed using the shortest polyhedral geodesic path.

3.3.3 Point-Wise Correspondence Optimization

Having laid out the optimization procedure required for a uniform distribution of points on the individual surfaces, we concentrate on the point-wise correspondence optimization problem, which is written as follows:

$$\arg \min_{\phi_j \in \mathbb{R}^{3M}} \hat{\Gamma}(\phi_1, \dots, \phi_N), \quad i = 1, \dots, N, \quad \text{s.t. } \psi_1^i, \dots, \psi_M^i \in \mathcal{S}_i. \quad (3.15)$$

Here we adopt the same procedure as presented by [28]. Let again $\Phi \in \mathbb{R}^{3M}$ be a random variable in shape space with probability density function $p(\Phi)$. By assuming Φ is normally distributed, the differential entropy can be written as

$$\Gamma(\Phi) = \frac{1}{2} \ln \{ (2\pi e)^{3M} |\Sigma|_\Phi \}. \quad (3.16)$$

The sample covariance estimator under the observed data is

$$\hat{\Sigma}_\Phi = \frac{1}{N-1} \mathbf{\Omega}^T \mathbf{\Omega}, \quad (3.17)$$

where $\omega_i = \phi_j - \bar{\phi}$ are the centered observations, $\bar{\phi} = \frac{1}{N} \sum_{u=1}^N \phi_u$, and $\mathbf{\Omega}$ the $N \times 3M$ matrix

with ω_i as its rows. Furthermore, let Φ be the $N \times 3M$ matrix with ϕ_i as its rows. The entropy estimator $\hat{\Gamma}$ is then given by

$$\hat{\Gamma}(\phi_1, \dots, \phi_N) = \frac{1}{2} \ln \{ (2\pi e)^{3M} |\hat{\Sigma}_\Phi| \} = \frac{1}{2} \ln \left| \frac{1}{N-1} \Omega^T \Omega \right| + \frac{1}{2} \ln (2\pi e)^{3M} . \quad (3.18)$$

The matrix of partial derivatives of $\hat{\Gamma}$ w.r.t. elements of Φ is

$$\begin{aligned} \frac{d\hat{\Gamma}}{d\Phi} &= \frac{d\hat{\Gamma}}{d\Omega} \frac{d\Omega}{d\Phi} \\ &\approx \frac{d\hat{\Gamma}}{d\Omega} \\ &\approx \frac{1}{2(N-1)} \frac{d}{d\Omega} \ln |\Omega^T \Omega| \\ &\approx \frac{1}{(N-1)} (\Omega^+)^T \\ &\approx \frac{1}{N-1} ((\Omega^T \Omega)^{-1} \Omega^T)^T \\ &\approx \frac{1}{N-1} \Omega (\Omega^T \Omega)^{-1} . \end{aligned} \quad (3.19)$$

Above we made two assumptions. First, within an iteration of the gradient descent optimization cycle, the means $\bar{\phi}$ stay fixed, and as such $\frac{d\Omega}{d\Phi} \equiv \mathbf{I}$. Furthermore, we assumed that the sample covariance matrix is invertible. This is not always the case, since the matrix is only required to be positive semi-definite. To solve this issue, a regularization term is added to guarantee that the matrix of partial derivatives can be computed

$$\frac{d\hat{\Gamma}}{d\Phi} \approx \frac{1}{N-1} \Omega (\Omega^T \Omega + \alpha_{reg} \mathbf{I})^{-1} . \quad (3.20)$$

From this point on, we deviate from the method of Cates et al., in order to accommodate the point-wise correspondence updates (3.20) into our geodesic gradient descent method. We group partial updates into 3-dimensional vectors, corresponding to surface points ψ_j^i . Then each update vector is projected onto the respective point's tangent space, which is the plane of this point's mesh triangle. Let $\mathbf{n}_j^i$ be the perpendicular unit normal vector of the ψ_j^i 's mesh

triangle. The tangent update vector is given by

$$\frac{d\hat{\Gamma}}{d\psi_{j\parallel}^i} = \frac{d\hat{\Gamma}}{d\psi_j^i} - \mathbf{n}_j^i \frac{d\hat{\Gamma}}{d\psi_j^i} \quad (3.21)$$

Since $\frac{d\hat{\Gamma}}{d\psi_{j\parallel}^i}$ is a vector in the tangent space of ψ_j^i , it can be added to $\frac{d(-\hat{H}_i)}{d\psi_j^i}$, thus providing the combined gradient descent updates for the optimization of the dense correspondence cost function (3.6):

$$\psi_{j,t+1}^i = \exp_{\psi_{j,t}^i} \left(-\gamma_G \left(l_G \frac{d(-\hat{H}_i)}{d\psi_j^i} + \frac{d\hat{\Gamma}}{d\psi_{j\parallel}^i} \right) \right) , \quad (3.22)$$

with l_G controlling the relative significance of uniform sampling versus point-wise correspondence entropy terms.

3.3.4 Initialization and Iterative Point Splitting

Previously we worked under the premise that collections of M points were already in some way positioned in each surface. In GESSA, M is given as an input parameter. The algorithm initially positions randomly one point in each surface and optimizes their corresponding positions. Subsequently, each point is split into two and the overall optimization process runs again on the new point sets. This procedure is repeated iteratively until the required number of sampled surface points is reached and their positions are optimized.

3.3.5 Algorithmic Details and Parameters

In this part we discuss a number of details regarding the algorithmic implementation of GESSA. In section 3.3.2, we introduced the geodesic KDE (3.9), which makes use of distances defined on the surface of the polyhedral surfaces. During the length of this work, shortest polyhedral geodesics, and their lengths, were computed using the approximate SPG algorithmic implementation described in [161]. In particular, distances from one to the remaining, $(M - 1)$, surface points are computed by propagating a distance field, emanating from the source point, across

the mesh edges, in a Dijkstra-like fashion [161]. The complexity for computing the $(M - 1)$ distances is at worst $\mathcal{O}(v_S \log(v_S))$, where v_S is the number of vertices in the surface.

In the same section, a geodesic gradient descent was employed in order to manipulate the points' locations on the surfaces, so as to achieve their uniform distribution. Straightest polyhedral geodesics were used as exponential maps to update the position of the points on the surfaces. The complexity of computing each exponential map is $\mathcal{O}(f_S)$, where f_S is the number of faces in the surface. Practically point updates are computed very fast, especially as M increases, since they traverse only few faces each time. Updates are dependent on the density p_i^{KDE} . Without any prior assumption, it would be necessary to recompute the density function for each point update. This would impose significant computational cost and also restrict the ability to compute updates in parallel for all surface points. To deal with these issues, we assumed that $\frac{dp_i^{KDE}(\psi_l^i)}{d\psi_j^i} \equiv 0$, for $l \neq j$, within one iteration of the gradient descent optimization cycle. This assumption is reasonable, especially as the number of surface points M increases, since the standard deviation parameter in the gKDE is continually reduced (see below), and as the effect of other points is exponentially reduced according to their distance. Under the assumption, updates for all surface points within one iteration can be computed in parallel. A block coordinate descent algorithm, with coordinates of one point been optimized per iteration, could have been employed, but it would have impaired the ability to parallelize the algorithm. Furthermore, while it could be possible to perform multiple individual point updates per iteration, large movements could notably violate the above assumption.

We concentrate now on the point-wise correspondence optimization of section 3.3.3. During each iteration, we need to invert the $3M \times 3M$ matrix $\mathbf{\Omega}^T \mathbf{\Omega}$. The case of singularity was solved by adding a regularization term $\alpha_{reg} \mathbf{I}$. However, one potential problem is the computational complexity $\mathcal{O}((3M)^{2.373})$, which increases with the number of sampled surface points. For our validation and application studies, this was not a major issue. Nevertheless, this can be the case when more dense sampling need to be considered. In Appendix A.1, we present an alternative formulation of the optimization which, by employing the notion of a pseudo-determinant, necessitates the inversion of $\mathbf{\Omega} \mathbf{\Omega}^T$, rather than $\mathbf{\Omega}^T \mathbf{\Omega}$, and thus has a lower complexity when $3M > N$.

GESSA depends on four algorithmic parameters. First, the weight parameter l_G , which can be used to control the relative significance of uniform sampling versus point-wise correspondence entropy terms. In our experiments, within each point optimization cycle, l_G was initially set to 1.2, to encourage the uniform redistribution of split points, and reduced by 0.005 after each gradient descent update, down to the minimum value of 0.5. Another algorithmic parameter is the gradient descent learning rate γ_G . Again, within each point optimization cycle, γ_G was initially set to 1 and adapted after each gradient descent update according to the bold driver algorithm. In detail, after each point update, if the objective cost decreased, the learning rate was increased by a relative 5%, otherwise it was decreased by 25%. The third parameter is the standard deviation σ_S of the gKDE. In our studies, we initially set σ_S to be equal to an approximate radius of a surface, computed as $\frac{\sqrt{A_S}}{\pi}$, where A_S is the overall area of the polyhedral surface. After each point split, σ_S was reduced to half of its previous value. Since we dealt with ensembles of similar objects, σ_S was chosen to be the same for all surfaces. The final parameter is the regularization α_{reg} which guarantees that the sample covariance is invertible. In each cycle, α_{reg} was initially set to 1 and reduced after each point update down to the minimum value of 0.01.

3.4 Validation Study

Our validation study aimed at demonstrating that GESSA can produce a dense uniform coverage of surfaces, while consistently and accurately aligning points to achieve one-to-one correspondence across all surfaces. We concentrate in particular on facial polyhedral surfaces, since we latter apply GESSA to analyze a large dataset of faces (see section 3.5). We opted to use the publicly available Morphface dataset of 3D facial surfaces [122], which comprises faces with neutral pose, taken from 11 individuals of European ancestry. See section 3.6 for a discussion on the rationale behind using the specific dataset. All the faces were registered to the Basel Face Model (BFM) facial template. The template mesh consisted of 53,490 mesh vertices. The registration process identified the locations of these points on the 11 validation faces. For the purposes of this validation study, 19 landmarks, corresponding to prominent facial points,

were manually selected to constitute the groundtruth landmarks (GTLs). Figure 3.1 depicts these 19 landmarks on the facial surface. The Morphface surfaces were subjected to cropping in order to remove non relevant areas, such as neck and ears. Subsequently, the meshes were subsampled, yielding approximately 4,700 points per facial surface.

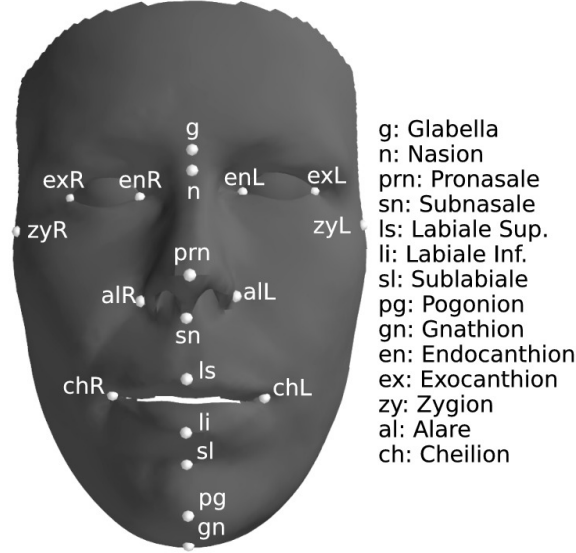


Figure 3.1: Example 3D facial surface with groundtruth landmarks for the Morphface Validation Dataset. The 19 groundtruth landmark positions were used during validation of our GESSA methodology.

The goals of our validation study were threefold. First, we aimed to evaluate how uniform the distribution of the sampled points on the surfaces would be. Second, we aimed to demonstrate that sampling approximately 4000 facial points produces a dense coverage of facial surfaces. Third, we aimed to show that GESSA is consistently and accurately aligning these points to achieve one-to-one correspondence across all surfaces. In all of the above, we compare results from GESSA and the original methodology by Cates et al [28], in order to highlight our contributions.

First, we focus on the evaluation of the facial points' surface distribution. The main differences between ours and Cates' methodology lie in the use of the geodesic KDE and geodesic gradient descent. These should result in a more uniform distribution of points. Both extensions have an influence not only when we deal with dense correspondence on multiple surfaces, but even more so when we want to resample a single polyhedral surface. Since for now we are interested in the

point distribution, we disentangle surface sampling from dense correspondence, and concentrate on each facial surface individually, from the Morphface dataset. From the methodological description, it is clear that both methodologies produce a KDE that should be close to the uniform distribution. To quantify the performance of these KDEs, we can compare them to the uniform density on the polyhedral surface, using the Mean Integrated Squared Error (MISE). Assuming that p_g^{KDE} , p_e^{KDE} are estimated kernel densities using GESSA and Cates' method, and a_u the fixed density with $\int_{\mathcal{S}} a_u d\mathcal{S} = 1$, then the integrated squared errors (ISE) are given respectively by $\int_{\mathcal{S}} (p_g^{KDE}(\psi) - a_u)^2 d\mathcal{S}$ and $\int_{\mathcal{S}} (p_e^{KDE}(\psi) - a_u)^2 d\mathcal{S}$. By repeating the surface resampling n_r times and averaging the resulted ISEs, we get the MISE for each method.

In detail, we compared the two methodologies based on the following procedure: Each facial surface from the Morphface dataset was resampled 10 times using GESSA and 10 times using the original method. The number of sampled points was fixed to 1024. The densities of the uniform distributions on the validation faces were computed as the reciprocals of the overall areas of the polyhedral surfaces. All ISE integrals were computed using a monte carlo integration approach: Sample points from all mesh triangles were drawn, in proportional number to the triangles' area, so as to acquire 50000 surface points, the integrals were then approximated by multiplying the surface area with the average density over the 50000 points. MISEs could be finally computed by averaging the 10 ISEs available for each surface and sampling method. Table 3.1 reports the MISEs for the 11 facial surfaces. For all of them, GESSA yielded consistently about an order of magnitude smaller MISEs, compared to the original methodology. It is also important to mention that the geodesic gradient descent was converging on average (overall for all splits and surfaces) after 254.34 steps, in contrast to 319.264 of the gradient decent algorithm used in Cates method, where updated points were computed by following the gradient in the embedding space $\mathbb{R}^3$, and had to be subsequently reprojected on the surface.

Having shown that our extensions have a positive impact on the uniform distribution of points on a polyhedral surface, we move on to the remaining objectives of our validation study. Using GESSA, we computed 4,096 GESSA-Generated Corresponding Points (GGCPs) in the validation surfaces. Figure 3.2 (a) depicts these on three example faces. Figure 3.3 shows the estimated surface density function of an example validation face. It can be observed here,

Subject ID	Cates MISE	GESSA MISE
1	0.0003543	0.0000317
2	0.0006208	0.0000519
3	0.000211	0.000064
4	0.0004745	0.0000832
5	0.0003729	0.0000445
6	0.0004272	0.0000291
7	0.0010851	0.0000903
8	0.0003543	0.0000148
9	0.000863	0.0000118
10	0.0006713	0.0000484
11	0.0012174	0.0001094

Table 3.1: Mean Integrated Squared Errors between the validation surfaces’ uniform density functions and density estimators computed using GESSA and Cates’ single surface sampling methodologies.

as well, that the density is almost uniform everywhere. In addition, examination of Fig. 3.2 (a) also shows clearly that the point sets are uniformly distributed across the complete facial surfaces.

For the second part of our validation, we needed to examine if the chosen number of sampled GGCPs is sufficient to densely cover the facial surfaces. A low number of sampled points would lead to extended facial areas having no or very few landmarks. On the other hand, very dense sampling would incur unnecessary computational costs. To evaluate this objective, we worked under the premise that, by sampling surfaces densely, we should be able to locate GGCPs sufficiently close to the preselected GTLs. We measured the average distance, over all faces in the validation set, from each GTL to its closest GGL. Figure 3.2 (b) shows the GTLs and respective closest GGCPs for three example validation faces. Table 3.2 reports the 19 average distances between GTLs and GGCPs over the set of 11 faces. Due to the fact that dimensional units on the Morphface dataset were unknown, we further divided each of the above values by the mean facial width - distance between the two zygomatic landmarks- in order to compute normalized distances, which are also included in Table 3.2. The average distances per landmark were always less than 3% of the mean facial width. By further averaging over the 19 landmarks considered, the overall average distance was 1.77% of the mean facial width. The results indicated that the number of points was sufficient to cover densely the surfaces and annotations extended over the complete surfaces.

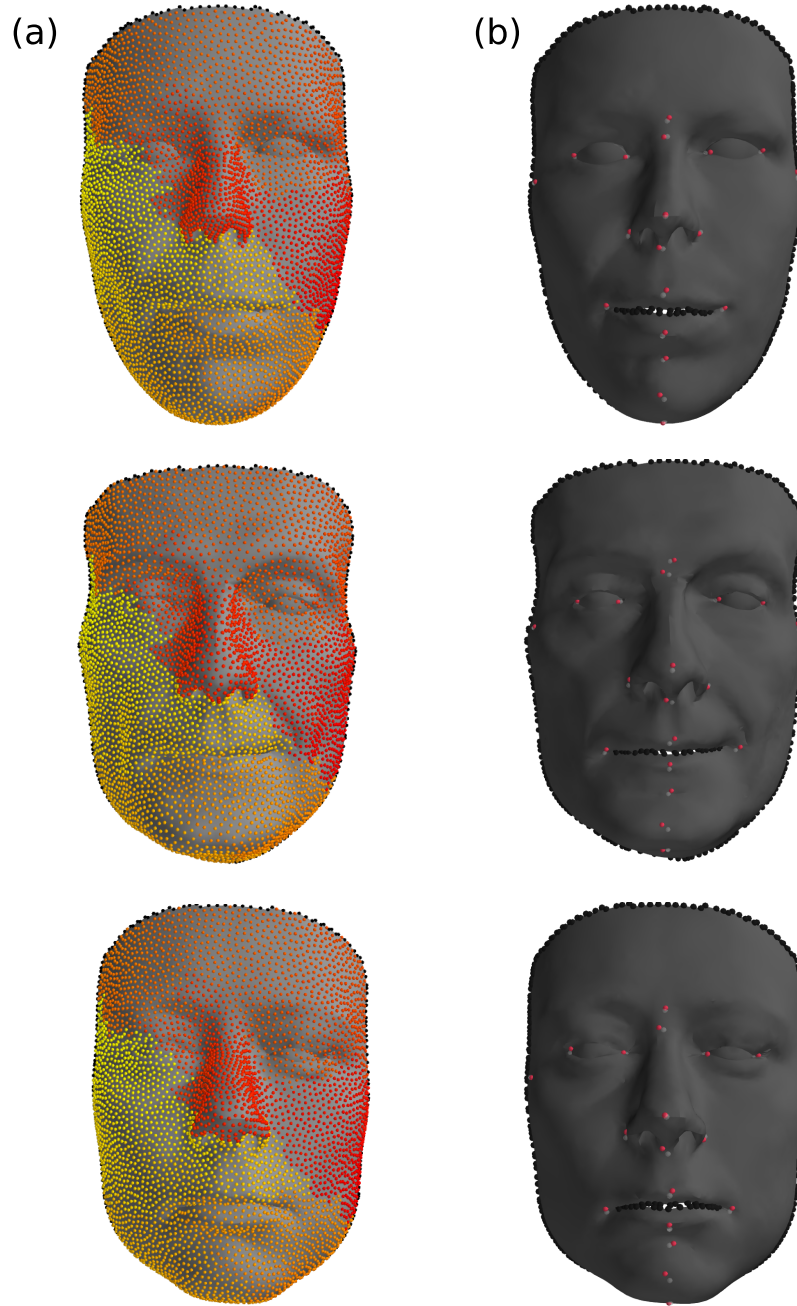


Figure 3.2: GESEA corresponding point sets on validation faces using GESEA. (a) 4,096 corresponding facial points were computed using our dense automated correspondence method. Results for three validation faces are shown here. Corresponding points are colored consistently among the different faces. (b) Preselected Groundtruth landmarks (GTLs) are shown in white. Red points represent the nearest GESEA-Generated Corresponding Points (GGCPs) to the GTLs for the same example faces. Dense annotation allowed selection of nearest landmarks with distances from GTLs being less than 3% of the mean facial width. Importantly, the algorithm was able to consistently place points across different faces.

Third, we concentrated on evaluating the ability of GESEA to place points in equivalent positions consistently across the complete set of surfaces. The standard deviations of the above

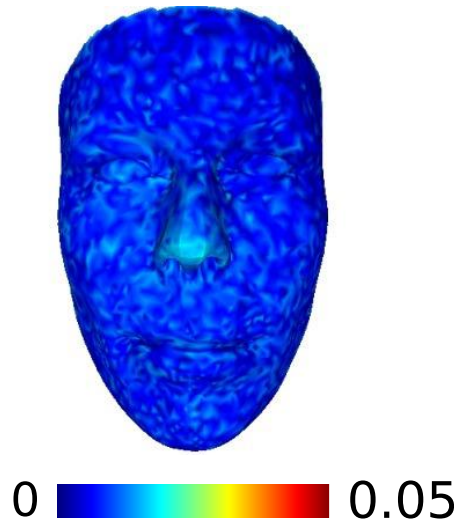


Figure 3.3: Estimated Probability Density Function on a Validation Face. In GESSA, points are considered samples of a random variable defined on the polyhedral surface. By maximizing the sample differential entropy, the density function becomes almost constant everywhere. The effect of this optimization is that surface points are effectively drawn from a uniform distribution.

GTL-GGCP distances can be used as measures of consistent placement, since small values indicate that GTL-GGCP distances are similar across all faces and, consequently, landmarks are positioned in equivalent positions across the dataset. The details are reported in Table 3.2. The standard deviations of the GTL-GGCP distances for all landmarks were always less than 1% of the mean facial width. By further averaging again over the 19 landmarks considered, the average standard deviation was 0.42%. The results indicated that the automatic dense point correspondence produced by the algorithm is highly accurate.

Finally, we repeated the process of dense correspondence, using, this time, the original methodology by Cates et al [28]. 4,096 points were again sampled from the facial surfaces, and closest points to the 19 preselected GTLs were identified. The respective distances and their standard deviations are also reported in Table 3.2. From the table it is clear that the range of the average distances to GTLs did not defer a lot between the two methodologies. This was expected since the same number of points were considered in both cases. Where our method exhibited a clear advantage was with regards to the standard deviations of these images. For Cates' methodology, we saw a big inconsistency with standard deviations. In particular, while with GESSA all normalized deviations ranged approximately from 0.0025 to 0.0065, with the original method normalized deviations were from 0.004 up to 0.027. This difference could be

attributed to the fact that GESSA, using the geodesic distances and maps, is more robust against shape variations on the different surfaces, which can through off correspondence in the original algorithm.

Landmarks	GESSA				Cates' Method [28]			
	Distance	Norm. Distance	Std	Norm. Std	Distance	Norm. Distance	Std	Norm. Std
Zygion (R)	1399.047	0.01077	422.24	0.00325	4050.393	0.031180	2586.878	0.019914
Labiale Inferius	1791.735	0.01379	472.386	0.00363	1318.361	0.010148	525.201	0.00404
Sublabiale	2162.848	0.01665	330.163	0.00254	2988.356	0.023	575.654	0.00443
Gnathion	2506.586	0.01929	480.526	0.00369	1983.285	0.015267	2169.251	0.016699
Glabella	1677.904	0.01291	555.001	0.00427	2552.905	0.019652	792.042	0.006097
Labiale Superius	2545	0.01959	594.255	0.00457	2669.003	0.020546	2717.417	0.020919
Cheilion (R)	2274	0.0175	542.486	0.00417	2987.45	0.022997	960.002	0.007390
Nasion	2896.955	0.0223	808.766	0.00622	3548.856	0.027319	1208.18	0.009300
Pronasale	1625.78	0.0125	557.833	0.00429	2868.173	0.022079	3408.408	0.026238
Subnasale	2389.583	0.01839	777.568	0.00598	2570.035	0.01978	935.089	0.007198
Alare (R)	3694.01	0.02844	678.1579	0.00522	2875.585	0.022136	1550.38	0.011935
Endocantion (R)	1526.617	0.01175	331.483	0.00255	2071.303	0.015945	911.103	0.007013
Exocanthion (R)	2152.691	0.01657	487.591	0.00375	1577.826	0.01214	2459.984	0.018937
Progonion	3008.975	0.02316	574.951	0.00442	2643.675	0.020351	2520.635	0.019404
Zygion (L)	3142.033	0.02419	522.777	0.00402	2996.398	0.023066	2269.696	0.017472
Cheilion (L)	2571.254	0.01979	583.14	0.00448	2526.366	0.019448	2143.37	0.016500
Alare (L)	2093.014	0.01611	514.262	0.00395	2064.655	0.015894	703.554	0.005416
Endocanthion (L)	2420.324	0.01863	367.685	0.00283	2740.083	0.021093	491.018	0.003779
Exocanthion (L)	1771.4	0.01363	736.372	0.00566	1404.446	0.010811	3616.366	0.027839

Table 3.2: Average distances and standard deviations (Std) between the 19 groundtruth landmarks and their respective nearest points, computed using GESSA, as well as the original method by Cates et al. [28]. Statistics were computed on the set of 11 validation faces. Normalized measurements were acquired by division with the mean facial width.

3.5 Application of GESSA on a Face Shape Heritability and Genome-Wide Association Study

3.5.1 Introduction and Motivation

The study of twin data has received significant attention in areas ranging from biology and psychology to behavioral genetics and epidemiology [20]. Twins are a valuable source for information and are extensively used to identify and delineate genetic and environmental causes for behavioral, disease and physiological traits. The power of twin designs arises from the fact that twins may be either monozygotic (MZ), developing from a single fertilized egg and therefore sharing all genetic material, or dizygotic (DZ), developing from two fertilized eggs and sharing on average 50% of their alleles. The variation in genetic similarity, together with the assumption of equal environmental influence for each twin pair creates the basis for exploring the effects of genetic and environmental variance [20, 118]. If a trait in question shows considerably better similarity within MZ than DZ twin pairs, this implies that genes play an important role in that phenotype.

The above relationship is quantified using the concept of heritability, a statistic that summarizes how much of the phenotypic variation is due to variation in genetic factors [187]. Different ways to estimate heritability have been presented, during the years, in the literature. In the simplest form, since MZ twins are on average twice as genetically similar as DZ twins, an estimate of heritability can be computed as approximately twice the difference in phenotypic correlation between MZ and DZ twins [53]. More recently, researchers have used Structural Equation Modelling (SEM) [133] to estimate heritability. In SEM, specific phenotypic variance-covariance structures are assumed for MZ and DZ twins and estimates of regression coefficients between latent genetic/environmental factors and phenotypic traits are learnt from the data at hand. Using the fitted model, heritability is estimated as the ratio of trait variance explained from the genetic latent factor divided by the variance explained from both genetic and environmental factors.

One trait that has attracted a lot of interest in twin, as well as parent-offspring, heritability studies is human facial morphology [36, 172, 107, 98, 82, 117, 104, 163, 73, 44]. The human face is an important interface of social interaction; communication, sensory input and expression in humans are to a large extent based on facial characteristics and traits [49]. Normal facial variation is associated with emotional expression [58], attractiveness [126] and even lifetime reproductive success [105]. Recent evidence suggests that evolution has contributed to increased diversity and complexity in human facial morphology, presumably due to the role of the face as a primary medium of individual identification and recognition [149]. The influence that facial features have in our life has spurred a long and ongoing interest in unraveling the roles that genes and environment play in the morphological characteristics of the human face.

Since the mid-twentieth century, anthropometric scientific research on parent-offspring resemblance and twin concordance has confirmed that variation in human face morphology is driven by genetics [172, 107, 98, 82, 117, 104, 163, 73, 44]. Heritability studies were carried out to quantify the extent of phenotypic variation that can be explained by genetic variability using, for instance, facial features extracted from cranial measurements. Moderate heritability, varying approximately between 0.35 and 0.65, was found for traits such as nasion-basion and nasion-sella distances, as well as the position of the lower jaw and the nasal height [87, 145, 25, 144, 96, 71, 27, 92, 150]. More recent studies used facial photographs instead, due to the simplicity in which the images can be obtained. However, common traits such as the upper lip height, as well as nasal breadth and vertical eye distance, extracted from standard photographs, were only found to be moderately heritable, with estimates between 0.4 and 0.53 [180, 26, 51, 13, 42, 95]. Given the almost perfect resemblance of identical twins, such heritability values appear surprising low. Attempts to replicate these findings across independent studies generated inconsistent evidence. A comparison of eight heritability studies reported low correlation (< 0.4) between heritability estimates for commonly examined traits such as head circumference, facial height and nose width [98]. Further examples include the heritability of facial width, which was reported to range from as low as 0.257 to 0.629 [27, 44, 158], nasal breadth heritability, varying between 0.352 and 0.639 [162, 5, 44], and cheek length heritability, ranging from 0.154 to 0.475 [27, 155].

Possible explanations for such unexpectedly low estimates and inconsistency across findings can be found in a number of experimental factors. First of all, in the way in which the traits have historically been measured. Radiographs and photographs are both flat, two-dimensional images. Their use to measure inherently three-dimensional (3D) objects, such as facial surfaces, limits the extent of shape variability that can be captured and constrains the range of facial morphological descriptors that can be extracted. Other issues can be identified in the process of constructing and measuring facial traits. Quantifying face variability is heavily dependent on establishing landmarks/corresponding points across faces. The number of points that can be manually annotated on a face is affected by the type of imaging modality used, and by the ability of a person to establish landmark locations in an unambiguous manner across samples. Consequently, it is common for studies to annotate only a few prominent facial markers, such as eye and mouth corners, nose tip and zygomatic bones [87, 145, 25, 144, 96, 71, 27, 92, 150, 180, 26, 51, 13, 42, 95], ranging in number between 10 up to, in exceptional cases, 40 landmarks. It becomes clear thus that manual landmarking poses an important constraint limiting the extend of facial variability that can be captured. Furthermore, most classical studies adopted facial traits derived from two-dimensional distances between landmark pairs [145, 25, 144, 71, 27, 92, 150, 180, 26, 42, 95]. More rarely, angles between connected landmark pairs have also been considered [87, 96, 51, 13]. The widespread adoption of such facial phenotypes could be justified by the small number of annotated landmarks, the relative simplicity in which these measurements can be acquired, as well as their ease of interpretation. On the other hand, they offer a perhaps oversimplified characterization of face morphology and fail to take into full account the geometric variability that can be observed across faces at a more granular level.

A separate limiting factor that has affected twin heritability studies is related to sample sizes employed and associated statistical modeling implications. The statistical power of heritability studies is defined as the probability of correctly rejecting the null hypothesis of zero heritability [181]. A multitude of factors affect power, including the combination of variance components used in the model, heritability effect size (ranging between 0 and 1), sample size and the proportion of MZ to DZ twins in the dataset. How to best optimize the experimental design in terms of sample size is still largely debatable [181]. By assuming the existence of only additive

genetic effects in the twin model, a previous simulation study reported minimum required sample sizes ranging between 75 and 2,000 twin pairs in order to achieve 95% statistical power, depending on heritability effect sizes ranging from 0.8 to 0.2 respectively, and keeping the MZ to DZ ratio close to one [110]. In practice, the large majority of twin studies to date have relied on sample sizes of 20 to 100 twin pairs [172, 117, 104, 73, 96, 71, 180, 26, 51, 13, 42], possibly due to difficulties in recruiting large cohorts of twins. The use of small to medium sized cohorts may have thus resulted in under-powered studies, especially for traits with low to moderate heritability.

In the following we present a novel heritability study of the human face shape that combines for the first time the use of 3D facial mesh models, dense automatic annotation of corresponding points and curvature-based phenotypes. 3D models were acquired using a system for high-resolution 3D photographic scanning, the 3dMD face imaging system. To capitalize on these representations, we applied GESSA on the dataset and automatically computed thousands of facial quasi-landmarks, giving us the ability to capture morphological variation across subjects at a much finer level of granularity, whilst removing human measurement errors. The availability of densely sampled corresponding positions on each face enabled a wider range of facial traits to be defined, each capturing a specific aspect of face-shape variability. In particular, we demonstrated that local curvature traits, computed at each three-dimensional position across the facial surface, provide highly informative quantitative measurements of facial geometry.

The face heritability study was performed on 952 British twins recruited from the TwinsUK adult twin registry [115]. All subjects were females and unselected for any disease, of which 197 were MZ and 279 DZ pairs. Each face was represented in the dataset as a 3D polyhedral mesh comprising of approximately 4,500 points. Using GESSA, we identified 4,096 corresponding points on each face, each one contributing a local curvature value whose heritability was independently assessed. Curvature-based heritability estimates of individual points were combined into face heritability maps highlighting in great detail, for the very first time, which facial parts are under high and low genetic control. A multivariate analysis further identified extended and well-defined facial regions sharing similar patterns of variability, with heritability estimates reaching or exceeding 0.7, including the chin, nasal regions, nasolabial folds, upper lips and

zygomatic bones. In addition, using a smaller set of the annotated facial points, we explored the heritability of more traditional distance-based facial traits. A number of facial lengths, including bizygomatic and nose width, had heritability estimates close to or greater than 0.7, values that are significantly higher than the ones previously encountered in the respective literature. Finally, we used the heritability findings as a guide for a preliminary Genome-Wide Association Study (GWAS), using a subset of 449 TwinsUK unrelated individuals. A GWA study focuses on finding associations between single-nucleotide polymorphisms (SNPs), located anywhere in the human genome, and phenotypic traits, which in our study are related to facial morphological features. The study identified significant genotypic markers on one chromosomal locus associated with a highly heritable curvature phenotype, which describes the shape around the philtrum facial area. To our knowledge, this study was the first to achieve such a detailed and comprehensive evaluation of facial shape heritability.

3.5.2 Data Description, Preprocessing and Phenotyping

Imaging data were retrieved for 1,547 participants of the TwinsUK Cohort [115], for which 3D facial models and age information was available. Research was done in accordance with the tenets of the Declaration of Helsinki. The sample consisted predominately of female twins unselected for any disease. 314 subjects were dropped due to artifacts in the images or lack of zygosity information, and 228 unrelated subjects were excluded from further analysis. An additional 56 subjects were removed due to mesh reconstruction errors during our phenotyping process. After preprocessing and quality control, the dataset consisted of 952 female subjects (59.31 mean age, 9.85 standard deviation). Of these, 197 pairs were MZ and 279 were DZ.

Due to large variation in the original pose and position of the original 3D meshes, we manually located outer eye corners and nasion in all faces. The three landmarks were used to impose a common orientation of the faces under the same coordinate space. These landmarks were discarded from any subsequent analysis step. Following that, the surfaces were cropped and trimmed to remove non-facial areas, such as neck and chest regions, hair and ears. Finally, the Iterative Closest Point (ICP) algorithm [16] was applied to align the cropped images. The

preprocessing pipeline was performed using the Meshlab software suite. Point correspondences for 4,096 corresponding points were automatically established on the 952 facial surfaces using GESSA.

3.5.3 Curvature-based Morphological Traits

The dense annotation of the facial surfaces allowed us to compute shape-related phenotypes, associated with the surfaces' curvature. The shape around a point on a surface can be characterized using curvature indices [97]. Curvature is a directional property and describes how bent the surface is around each point [135]. The curvature magnitude of a point in some direction is given by the reciprocal of the radius of the circle that best approximates the slice of surface in that direction [97]. Normal curvatures are defined on orthogonal planes to a surface point and for each such point, there exists a single normal curvature that has the largest absolute curvature magnitude. This is called the maximum curvature k_{max} . The curvature perpendicular to k_{max} is the minimum curvature k_{min} . These two surface attributes are collectively called principal curvatures and any normal curvature at a point on a surface can be derived as a combination of k_{max} and k_{min} [135]. Figure 3.4 shows a general characterization of shapes based on the signs of k_{max} and k_{min} .

Individually, each principal curvature does not provide a useful interpretation of local surface shape [97], as can be seen in Supplementary Fig. 3.4. As such, while principal curvatures, along with their directions, could be used as traits, they would lack visual interpretation and heritability maps would not be possible to be constructed. For this reason, we opted to utilize for our study more intuitive univariate curvature indices, which can be derived from the principal curvatures, to compute phenotypic traits. The four such measures were:

Mean Curvature:

$$k_{MC} = \frac{k_{max} + k_{min}}{2}, \quad (3.23)$$

Gaussian Curvature:

$$k_{GC} = k_{max} \times k_{min}, \quad (3.24)$$

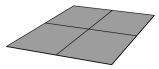
	$K_{\max} < 0$	$K_{\max} = 0$	$K_{\max} > 0$
$K_{\min} < 0$	 Bowl  Ellipsoid	 Cylinder  Valley	 Saddle 
$K_{\min} = 0$	X	 Plane	 Cylinder  Ridge
$K_{\min} > 0$	X	X	 Dome  Ellipsoid

Figure 3.4: Principal curvatures and shape characterization. General classification of shapes based on the signs of the two principal curvatures.

Curvedness:

$$k_{CU} = \sqrt{\frac{k_{\max}^2 + k_{\min}^2}{2}}, \quad (3.25)$$

Shape Index:

$$k_{SI} = \frac{\pi}{2} \times \arctan \left(\frac{k_{\max} + k_{\min}}{k_{\max} - k_{\min}} \right). \quad (3.26)$$

In order to explain the characteristics of the various indices, we first present two intuitive features that can be used to describe the shape of a surface patch. The first feature is the general shape morphology and is governed by the various sign combinations of $k_{\min}$ and $k_{\max}$, i.e. flat ($k_{\max} = k_{\min} = 0$), vs. convex cylindrical ($k_{\max} > 0, k_{\min} = 0$), vs. concave cylindrical ($k_{\max} = 0, k_{\min} < 0$), vs saddle structure ($k_{\max} > 0, k_{\min} < 0$), vs. convex cylindrical ($k_{\max} > 0, k_{\min} > 0$), vs. concave cylindrical ($k_{\max} < 0, k_{\min} < 0$). The second feature is curvature magnitude, i.e. how bent the surface is irrespective of shape morphology. Any univariate descriptor of curvature needs to be a compromise, since it cannot include all information provided by these two features [135, 97].

An illustration of the four indices' characteristics, including their domains and how shapes are differentiated is included in Fig. 3.5. The MC index provides a balanced measure between shape

morphology and curvature magnitude. It is strongly affected by directional shape differences (convex vs. concave shapes) but is also sensitive to curvature magnitude. GC distinguishes primarily between shape morphologies of the same or opposite principal curvature signs. It describes well variation between and within cylindrical and saddle-like structures, while being less sensitive to other shape characteristics. CU and SI indices are the most accurate quantitative measures of curvature magnitude and shape morphology respectively. While the four curvature indices are correlated, each one has its own merits as a facial trait descriptor, since, as discussed above, they accentuate varied morphological characteristics.

Calculations of local curvature on the 3D meshes was performed using a finite-differences algorithm [137]. Normal vectors perpendicular to the surface were computed in each facial point. Gradients across the surface were then approximated using finite normal differences of neighbor points. Principal curvatures were found through an eigenvalue decomposition of the normal gradients and curvature index values calculated from the respective formulas.

In our study, each corresponding point contributed four shape-related phenotypes, computed using the MC, GC, CU and SI formulas, resulting in 16,384 quantitative traits per face. For each one of the four measures, curvature maps can be obtained by color-coding all facial landmarks according to their curvature values, which provide easily interpretable representations of facial morphology. Figure 3.6 shows the curvature maps of the average face - constructed by averaging corresponding points' coordinates of all faces -.

In the MC curvature map, points belonging to protruding concave regions like the nose, chin and eyebrows had positive MC values, with higher measurements observed in the nasal surface. Flatter areas, such as cheeks and forehead, had MC values close to zero, while inner eye corners, ala of the nose, and to a lesser extent, the corner areas between lips and chin were comprised of negative-valued curvatures. In the GC map, facial features associated with positive values were the cylindrically structured inner eye corners and nose tip, while saddle-like regions such as nasion and base of the nose showed negative GC values. The CU curvature map highlighted facial parts with large overall curvature, for example nose and eyebrows, which confirms the intuitive observation that highly curved facial parts are mainly centrally located in the face.

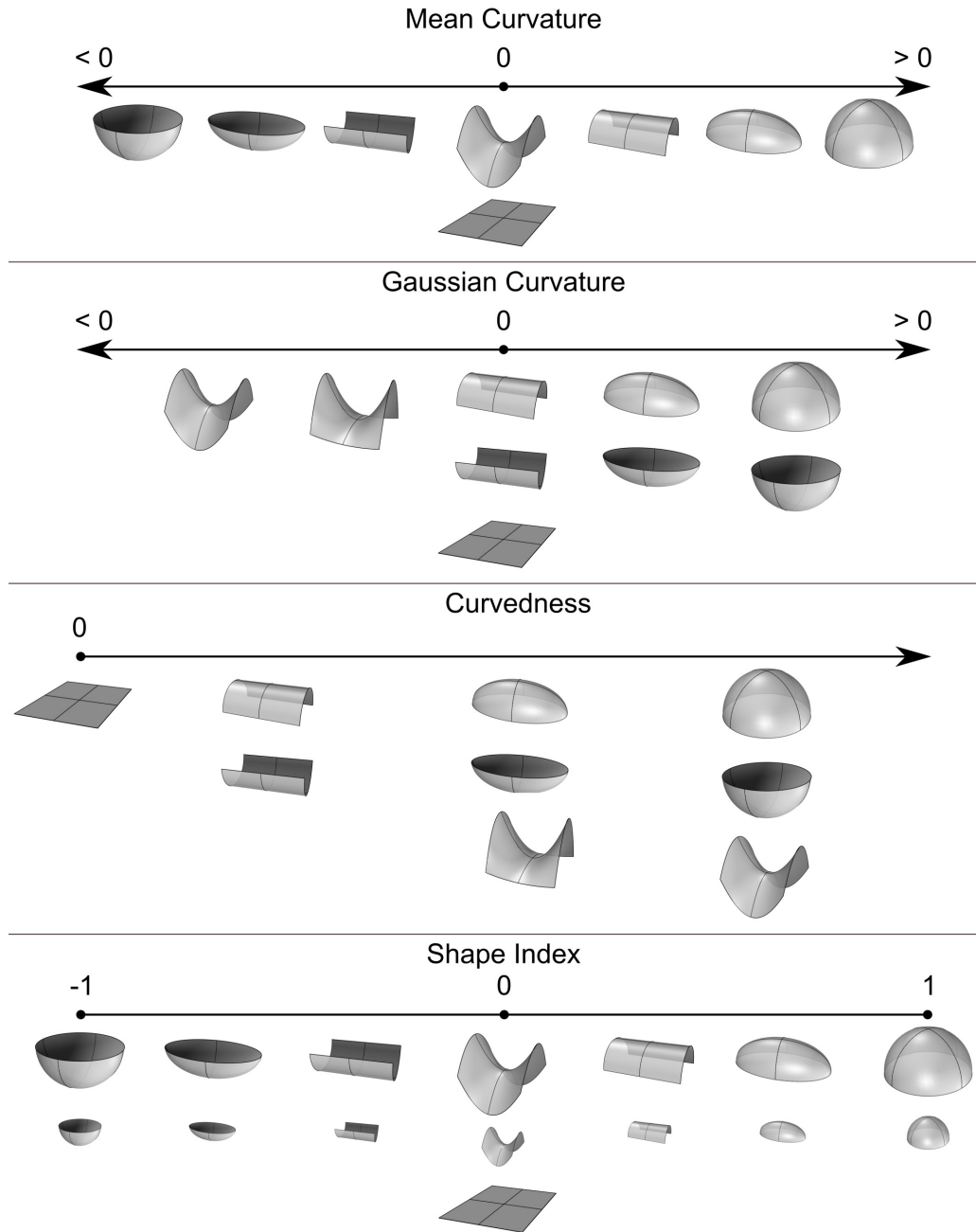


Figure 3.5: Topological characteristics of curvature indices. Each index highlights different attributes regarding the morphology of a surface patch. MC differentiates significantly areas of high and low curvature, as well as convex and concave shapes. GC discriminates well between spherical and saddle-like areas. CU is less representative of a particular morphology and reflects the absolute curvature magnitude in each point, irrespective of its specific shape. Finally, SI is scale-independent and able to differentiate between pure shape characteristics, e.g domes, ridges and saddles, regardless of their high or low CU.

The SI curvature map showed that that most facial areas have positive values corresponding to generally cylindrical structures. The sample variability of each curvature trait was further investigated. Results are presented in Appendix A.2.

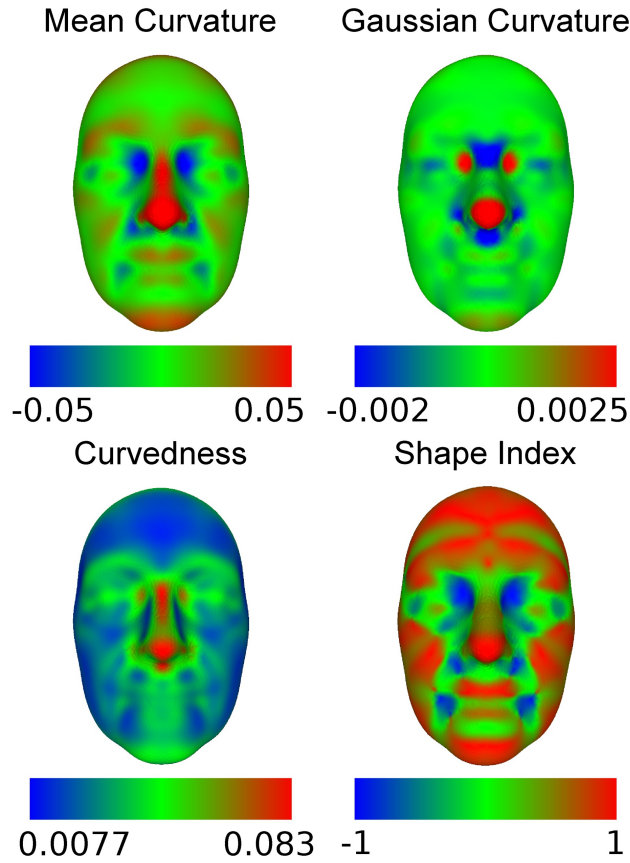


Figure 3.6: Face curvature maps. The figure shows the curvature maps of the average face from the TwinksUK dataset. The maps were created by computing values of curvature indices along all corresponding points on the face. Most facial parts have a convex topology, either cylindrical or spherical, with few saddle-like or flat areas that transition to concave regions localized in the eye, nose and mouth corners. Each map further highlighted the unique characteristics of its corresponding index. MC Curvature Map: Positive values corresponded to convex areas (nose, eyebrows, lips, chin), while negative values to concave ones (inner eye corners, subnasal region). GC Curvature Map: Many facial regions had very small values, due to being either flat or curved along only one direction. Exceptions were the spherical (nose tip, inner eye corners) and saddle-like (nasion, subnasal region) areas. CU Curvature Map: Consisting only of non-negative values, the map highlighted how curved the surface is without distinguishing between different shape morphologies. Large values were concentrated on the central part of the face. SI Curvature Map: Most facial areas had positive values, indicating the overall cylindrical structure of the human face.

3.5.4 Heritability Study

The curvature phenotypic data, as well as measurements of selected distance traits were used to perform various heritability analyses of the facial morphology. These analyses were done using Structural Equation Modeling (SEM) [133]. SEM evaluates which combination of additive (A) genetic, common (C) environmental and unique (E) environmental variance components can

best explain the observed phenotypic variance and covariance of MZ and DZ twin data. The importance of individual variance components was assessed by dropping parameters sequentially from the set of nested models $ACE \rightarrow AE \rightarrow E$. In choosing between models, variance components were excluded from the selection process if there was no significant deterioration in model fit after the component was dropped, based on the Akaike Information Criterion (AIC) [1]. The E component combined unique environmental results and random error and was retained in all models [133]. Heritability estimates for the AE models are calculated as $\frac{a^2}{a^2+e^2}$, where a and e are the path coefficients of the A and E variance components in the SEM model. ACE, AE and E structural equation models were fitted using the OpenMx software [18]. A detailed description of SEM is included in Appendix A.3.

Univariate Heritability Analysis

We performed univariate analyses of all 16,384 curvature traits - 4 traits per corresponding facial point - with the aim to combine local heritability results and produce global maps of heritability for the human face. The heritability of each trait was independently estimated using SEM. AE models were the best-fitting ones according to AIC. Summary statistics for all fitted models can be found in Appendix A.4. Model Goodness-of-fit was examined using a log-likelihood ratio test between the structured model and a fitted saturated model where no structure was imposed on the covariances. Test p -values above 0.05 translate to the structured model providing a better fit than the saturated one (see Appendix A.3 for further details). From the sets of 4,096 traits per index, 86.02%, 67.9%, 86.8% and 87.06% AE models for MC, GC, CU and SI respectively had goodness-of-fit p -values above 0.05.

For each curvature type, the 4,096 heritability estimates were visualized in a single heritability map plot. Each map provides a graphical representation of the extent by which the geometry of facial regions is controlled by genetic variability. Frontal and side facial views of these maps are shown in Fig. 3.7. For interactive viewing of the heritability maps, a website was created at <http://heritabilitymaps.info/>.

Several facial points with high heritability estimates (> 0.65) were localized on well defined

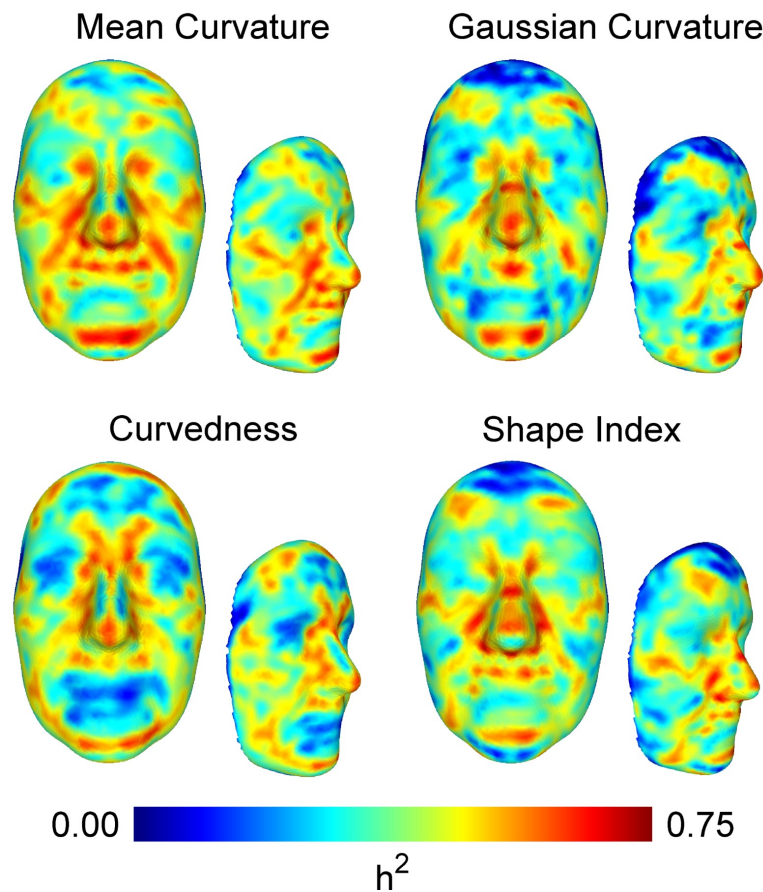


Figure 3.7: Heritability maps of the human face. Each heritability map consists of 4,096 heritability estimates. Facial areas with high heritability across all four maps were the mental region, philtrum, nasal tip, nasion, inner eye corners, nasolabial folds and frontal process of maxilla. MC Heritability Map: In addition to aforementioned areas, we observed genetic association along the lower parts of the eyes (zygomatic bones), the sides of mouth/chin complex (mental foramen), the complete upper lip region and the frontal eminences. GC Heritability Map: Compared to the previous map, zygomatic bones were not clearly distinguished and only the philtrum had high heritability in the upper lip region. A moderate genetic effect was observed around the whole nasion area. CU Heritability Map: The map included clear heritable lines in the upper and lower circumference of the facial surface. A further unique feature observed here was a flat heritable area around the ramus of the mandible. SI Heritability Map: Zygomatic lines were clearly visible in this map, along with the upper lip and the frontal eminences. The map highlighted primarily the mostly concave regions of transition between the lip and ala of the nose, the frontal process of maxilla and the lower more protruding parts of the nasal bones.

facial areas. Irrespective of the curvature index employed, points belonging to the mental region, philtrum, nasal tip, nasion, inner eye corners, nasolabial folds and frontal process of maxilla gave consistently high estimates. Amongst all curvature types, MC yielded the highest heritability estimates over extended facial regions. The MC heritability map highlighted further highly heritable areas, including the zygomatic lines around the eye sockets, side areas of the

mental foramen, the upper lip and frontal eminences. The results of the GC traits showed high heritability for a number of saddle-like facial structures, namely the whole nasion region, the philtral ridge, as well as the lower nasal bone. In the CU heritability map, a strong genetic influence is observed in the angular transition from the nasal bone and glabella towards the frontal eminences, as well as in the overall roundness of the facial circumference, highlighted by heritable lines across the upper part of the forehead and the lower part of the ramus of the mandible. The SI heritability map showed strong genetic control of the softly spherical flat regions of frontal eminences and upper lip, the cylindrical structure of the upper zygomatic bones and sides of the nose and finally the saddle-like areas of nasion, nasal bone and ala of the nose. The results indicated that local curvature is strongly determined by genes for large parts of the face.

Multivariate Heritability Analysis

Following the univariate heritability study, we aimed to identify larger facial areas showing common patterns of shape variation and produce single-valued heritability estimates for them. For each face and curvature index, we decomposed the aggregated single point phenotypes using sparse Principal Component Analysis (sPCA) [185, 186]. A detailed description of the sPCA methodology is included in Appendix A.5. sPCA automatically identified traits comprised of linear combinations of univariate phenotypes with similar curvature variability. We refer to these as regional traits. Sparsity affected the number of univariate traits comprising each regional phenotype, and was imposed in order to acquire measurements corresponding to extended but spatially consistent facial regions. The amount of sparsity was controlled through a single parameter. Different values were tested and results showed that the parameter had little effect in the heritability estimation process. Details on the effect of the sparsity parameter are given in Appendix A.6. The parameter was set to 7.5 for GC, 12.5 for SI and 15 for MC and CU. We estimated the heritability of 100 regional traits for each curvature index. Each set of regional traits was able to cumulatively explain 92.72%, 89.41%, 79.85% and 87.5% respectively of the MC, GC, CU and SI phenotypic curvature variance. Average heritability statistics for the regional trait analyses are given in Appendix A.4. We found again that the best models

were the AE as assessed by AIC. Goodness-of-fit p -values above 0.05 were acquired for 81, 75, 87 and 73 regional traits of MC, GC, CU and SI respectively.

The facial areas associated to regional traits were visualized by color-mapping the coefficients of the linear combinations - weights by which landmark traits contribute to the regional phenotypes - on the facial surface. We refer to these maps as Eigenface maps, since they correspond to the eigenvectors of the multivariate decomposition. Figure 3.8 shows the Eigenface maps of the top 5 heritable regional traits for each curvature index, along with their heritability values. The maps of the most heritable traits were mostly comprised of facial points closely located to each other, indicating patterns of common shape variation in the respective areas.

Heritability results for the regional and univariate traits were in good agreement. From Fig. 3.8 it is obvious that the morphologies of regions such as nasolabial folds, zygomatic bones, inner eye corners, mental region, frontal eminences and ala of the nose were highly heritable (> 0.65). Due to the good segmentation of the facial surface into clearly identifiable regional traits, we also identified heritable areas that were not as easily noticeable in the heritability maps. In particular, the mental foramen showed up as the second top heritable phenotype in the SI results, while the condyloid process of the mandible was highlighted in the fifth and third top heritable trait in MC and SI analyses, respectively.

Heritability of Distance-based Traits

An analysis of traditional distance-based facial traits was performed to gain insights about the relative merits of our phenotypes and also enable direct comparisons to previous heritability studies. Out of the 4,096 corresponding points, we located 17 landmarks for prominent fiducial points, by visual inspection of the average TwinsUK face. Figure 3.9 shows the selected landmarks. Using the computed correspondence, we were able to automatically locate the 17 landmarks on all 952 faces. Ten facial traits were derived from Euclidean distances (EDTs) between selected landmark pairs and ten equivalent traits were derived from Geodesic distances the same landmark pairs. The phenotypes are summarized in Table 3.3. We thus also examined whether the use of GDTs, only possible on 3D data, yielded any advantages compared to EDTs.

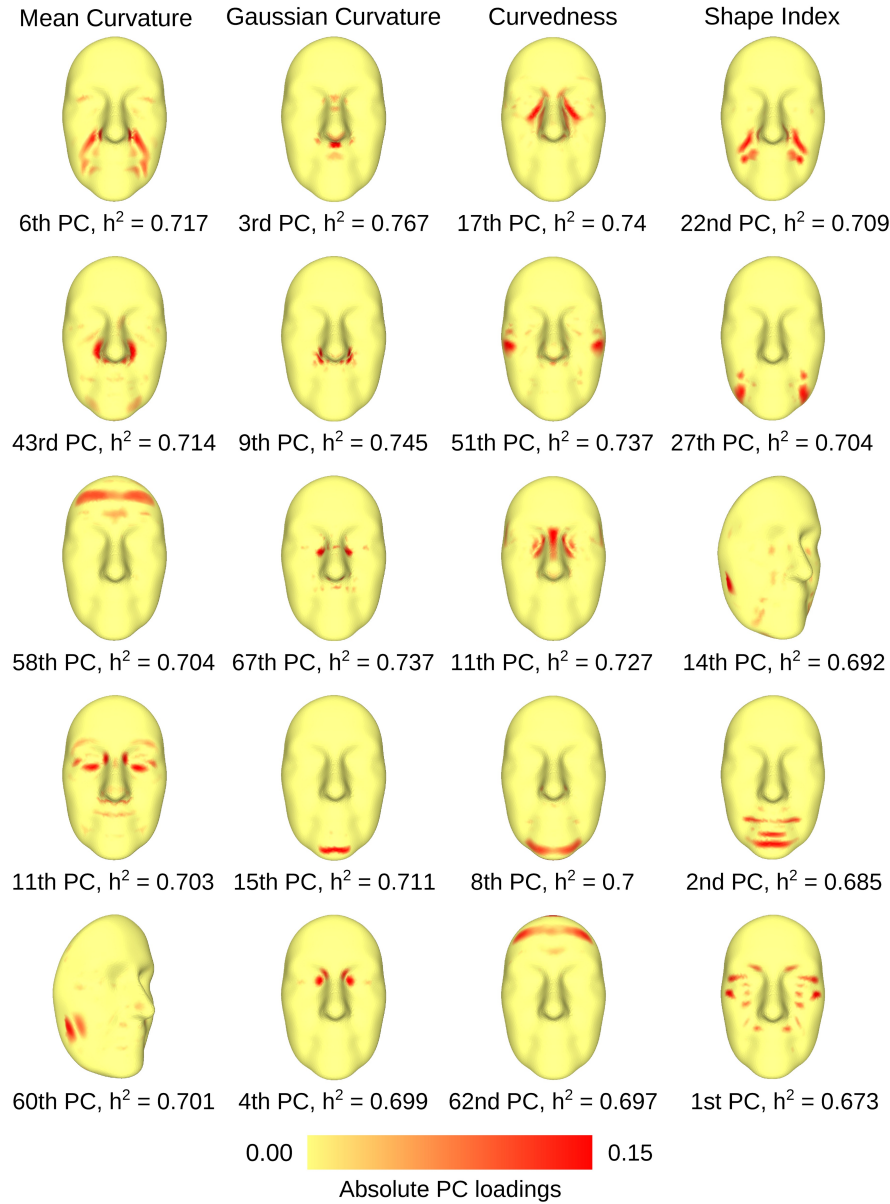


Figure 3.8: Eigenface maps of the top heritable regional traits. The maps depict the weights by which univariate phenotypes contribute to the regional traits. sPCA was utilized in order to achieve good spatial consistency of the Eigenface maps. MX Eigenface Maps: Distinctive facial characteristics that emerged as heritable were the nasolabial folds, transitions to the ala of the nose, frontal eminences, zygomatic areas between and below the eye sockets and the condyloid process of the mandible, with the latter been more clearly portrayed in the multivariate rather than the univariate heritability analysis. GC Eigenface Maps: The GC Eigenface maps were highly localized, compared to the rest of the curvature indices. Heritable regions included for the philtrum, ala transitions, inner eye corners and chin facial regions. CU Eigenface Maps: The top heritable traits highlighted mainly the zygomatic areas around the eye sockets, as well as the nasion, chin and upper forehead areas. SI Eigenface Maps: Highly heritable regional traits were located in the nasolabial folds, zygomatic areas, chin, condyloid process of the mandible and mental foramen regions.

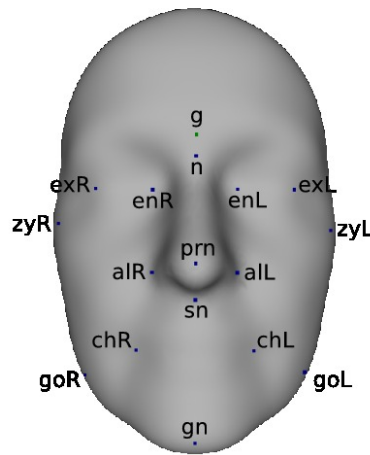


Figure 3.9: Selected landmarks for the distance-based heritability analysis, depicted on the average TwinsUK facial surface. 17 facial points corresponding to prominent fiducial markers were selected from the automatically computed set of 4,096 corresponding points. Landmark pairs were subsequently used to construct distance-based phenotypes. See Table 3.3 for naming conventions.

Landmarks		Distance Phenotypes	
Abbreviations	Description	Abbreviations	Description
g	Glabella	sn - gn	Height of lower face
n	Nasion	g - sn	Height of middle face
prn	Pronasale	n - sn	Nose Height
sn	Subnasale	sn - prn	Nasal protrusion
gn	Gnathion	alL - alR	Nose width
alL	Alare (L)	exL - exR	Intercanthal width
alR	Alare (R)	enL - enR	Biocular width
exL	Exocanthion (L)	zyL - zyR	Zygomatic width
exR	Exocanthion (R)	goL - goR	Mandible width
enL	Endocanthion (L)	chL - chR	Mouth width
enR	Endocanthion (R)		
zyL	Zygion (L)		
zyR	Zygion (R)		
chL	Cheilion (L)		
chR	Cheilion (R)		
goL	Gonion (L)		
goR	Gonion (R)		

Table 3.3: Abbreviations and related descriptions of selected landmarks and traits for distance-based heritability analysis.

Heritability estimation for the 10 EDTs and their equivalent GDTs was carried out using ACE, AE and E structural equation models, as before. AE models provided on average the best fits according to AIC. Average model statistics are given in Appendix A.4. Table 3.4 shows the

resulting heritability estimates as well as SEM Goodness-Of-Fit test p -values. In the remainder, we concentrate on traits whose models' Goodness-Of-Fit test p -values were greater than 0.05, thus providing good fits of the observed data. Four EDTs corresponding to horizontal facial measurements, namely nose, zygomatic, mandible and mouth widths were found to be highly heritable. Of these, estimates greater than 0.7 were acquired for the two upper/middle face EDTs, zygomatic and nose widths, while the heritabilities of the mandible and mouth widths were found to be slightly lower at approximately 0.62 and 0.67, respectively. The EDT corresponding to nasal protrusion was also found to be moderately heritable at approximately 0.55. Heritable GDTs included the mandible and mouth widths, nasal protrusion, lower face height and biocular width (i.e. the distance between inner eye corners). The first three GDTs had slightly lower heritability estimates compared to the corresponding EDTs. Mandible and mouth GDTs traverse the lip region, which had low heritability estimates in our previous analyses. On the other hand, the biocular GDT, which traverses a surface area with consistently high heritability, i.e the nasion, provided the highest estimate (0.789) among all distance phenotypes.

Phenotypic Traits	EDT		GDT	
	h^2	Goodness-Of-Fit p -value	h^2	Goodness-Of-Fit p -value
sn - gn	0.634	0.019	0.692	0.126
g - sn	0.75	0.00003	0.708	0.0004
n - sn	0.749	0.00001	0.699	0.0003
sn - prn	0.545	0.276	0.505	0.731
alL - alR	0.718	0.219	0.697	0.0002
exL - exR	0.706	0.005	0.651	0.0002
enL - enR	0.707	0.02	0.789	0.063
zyL - zyR	0.734	0.303	0.665	0.001
goL - goR	0.677	0.537	0.573	0.468
chL - chR	0.62	0.105	0.586	0.449

Table 3.4: Heritability estimates h^2 for 10 traits based on Euclidean Distances (EDTs) and 10 traits based on Geodesic Distances (GDTs). Models' Goodness-Of-Fit test p -values greater than 0.05 correspond to good fits of the observed data.

3.5.5 Genome-Wide Association Study

Following our heritability study, we evaluated the use of regional curvature and distance traits as phenotypes in a preliminary GWAS, using a subset of 449 unrelated individuals from the

TwinsUK cohort, for which genotypes were available. Subjects were genotyped using the HumanHap 300k Duo and HumanHap610-Quad array genotyping platforms from Illumina. Imputed genotypes were computed with reference to HapMap release 22 CEU population data using IMPUTE2 [79]. The raw genotypic data comprised of approximately 3,000,000 Single Nucleotide Polymorphisms (SNPs) per subject. The association analysis was performed using PLINK [131]. For a pair of univariate genotypic and phenotypic variables G and P respectively, PLINK assumes a linear causative model $P = sG + b$. The Null Hypothesis $H_0 : s = 0$ of no association (zero slope) between independent and dependent variables is tested against the alternative Hypothesis $H_1 : s \neq 0$ using a Wald style test based on the t -distribution. For an estimator $\hat{s}$ of parameter s , the t -statistic is $t_{\hat{s}} = \frac{\hat{s}}{\hat{SE}_{\hat{s}}}$, with $\hat{SE}_{\hat{s}}$ the estimated standard error of $\hat{s}$. Under H_0 , the sampling distribution of $t_{\hat{s}}$ follows the t -distribution with $N - 2$ degrees of freedom, where N is the sample size. Estimators $\hat{s}$ and $\hat{SE}_{\hat{s}}$ were computed using Ordinary Least Squares (OLS).

Based on the previous heritability results, different analyses were performed for the twenty most heritable regional traits, as well as five distance traits with heritability above 0.65, namely the alL - alR, zyL - zyR, goL - goR EDTs and sn - gn, enL - enR GDTs. After removal of SNPs with missingness ratio above 0.05, minor allele frequency below 0.05 and Hardy-Weinberg Equilibrium below 0.001, 1,337,124 genotypic markers were tested for association with the selected phenotypes. Genome-wide significance was ascertained if test p -values were lower than 0.5×10^{-8} .

In the GWAS, we identified four SNP markers reaching genome-wide significance, all associated with the top heritable GC curvature trait, which corresponds predominantly to the philtrum facial area (Fig. 3.8). The SNPs were located within a 100kb region at chromosomal location chr2p11.2. Genomic Region plots, centered around the significant markers, showing the resulted association test p -values, are included in Fig. 3.10. The r^2 correlation, also referred here as recombination rate, between the four significant SNPs was higher than 0.9, based on rates computed from the HapMap phase II, release 22, phased genotypic data [88]. The high recombination rate indicated the existence of a single genotypic loci in association with the regional trait.

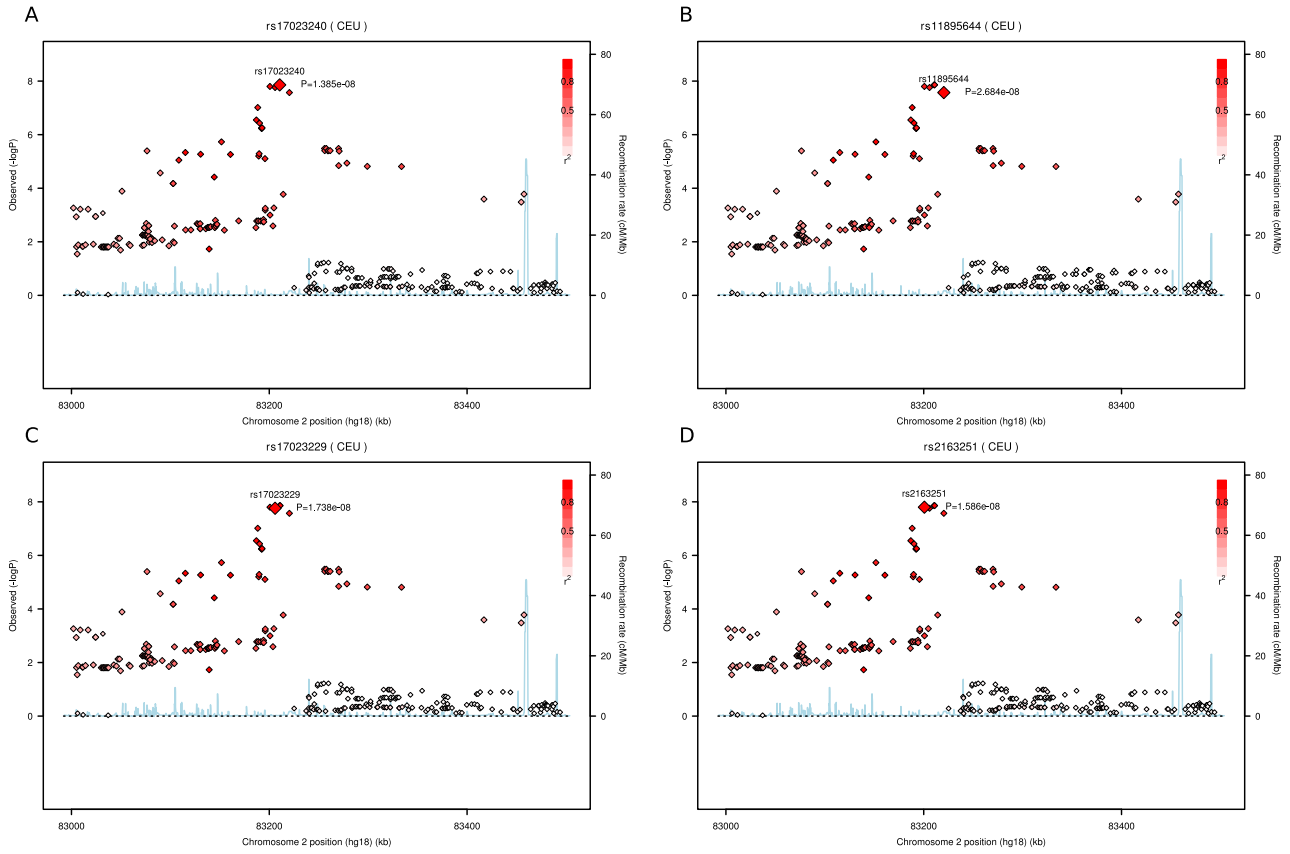


Figure 3.10: Genomic Region plots showing genome-wide significant markers associated with the top heritable GC regional trait. Observed genotype-phenotype association signals ($-\log p$) are plotted against physical positions of each SNP in a region centered around the four genome significant SNPs: (a) rs17023240, (b) rs11895644, (c) rs17023229, (d) rs2163251. The four SNPs showed very high correlation/recombination rate (HapMap Release 22, CEU) among them, indicating a single locus of significance.

3.5.6 Discussion

To our knowledge, the study presented above is the largest face heritability study ever done. Our landmarking and phenotyping pipeline was employed for the analysis of 456 female twin pairs from the TwinsUK cohort [115]. Compared to most commonly encountered sample sizes of between 20 and 100 pairs [172, 117, 104, 73, 96, 71, 180, 26, 51, 13, 42], the increased number of subjects improves the statistical power of our study to identify effects of heritability. Based on a previously published simulation study on the power of twin studies [110], our sample size surpasses the minimum requirements for having 95% power of rejection of the false - zero heritability - hypothesis at the 5% significance level even when the true heritability effect is as low as 0.3. Previous work has suggested that the TwinsUK sample is representative

of the general British population, where the sample were ascertained from [3]. For facial traits in particular, it may be expected that gender also plays a role in shaping up the facial characteristics we study. For this reason, caution is warranted before any of these results are generalized and extended to male subjects of the same ethnicity.

A salient aspect of this study, facilitated by the three-dimensional face models, was the use of curvature-based phenotypes. For each point in the facial surface, four different types of univariate measurements describing curvature for local patches centered around that point were considered. Each curvature type is able to highlight varying morphological structures. The use of these traits enabled us to characterize local variability in facial shape and identify its genetic content. Heritability estimates for these local traits were combined to provide detailed global maps of heritability for the human face. Furthermore, regional traits, defined as linear combinations of the single point phenotypes, were computed by a multivariate decomposition of the previous traits. Heritability analyses of the latter phenotypes allowed us to report on accurate heritability values for well-defined facial regions. In the literature, similar curvature traits have been successfully used for other applications such as face detection [37], recognition [84, 47], segmentation [141, 29] and affinity estimation [29, 192].

We were able to identify curvature-based facial traits that were highly heritable (> 0.65) in both of our curvature-based analyses. A direct comparison with previously reported heritable facial lengths and angles is not straightforward, due to the different nature of the measurements. Certain connections though were made between our heritability and Eigenface maps and related published findings. A heritability estimate of 0.53 was previously reported for nose width in a pedigree analysis of 229 Korean individuals [95]. The MC heritability map (Fig. 3.7) showed high heritability for the line between the left and right corners of the ala through the base of the nose. In the same study, inner eye corner distance had a heritability estimate of 0.61. The GC, CU and SI heritability maps all indicated that the shape of the nasion region, including the inner eye corners, was highly heritable. This inference was further supported by our multivariate analysis. The 3rd and 5th most heritable GC regional traits, as well as the 3rd top heritable CU trait (Fig. 3.8) had Eigenface maps concentrated on the nasion area (Fig. 3.8) and their respective heritability estimates ranged between 0.699 and 0.737. Moderate to

high heritability, ranging between 0.4 and 0.7, was identified in twin and family studies for the facial width [13, 150, 51, 92]. Our MC and SI heritability maps (Fig. 3.7) highlighted highly heritable lines following the curve of the zygomatic bones. Another phenotype that was reported with heritability estimates of 0.59 and 0.66 is head circumference [52, 92]. This result can be connected to the elevated heritability estimates regarding curvature magnitude in the periphery of the face, that we observed in the CU heritability map (Fig. 3.7). Finally, a number of twin and family studies identified moderate to high heritability estimates for various phenotypic traits relating to the position, length and angular structure of the jaw bone. Mandible ramus and body length heritabilities were reported to be 0.72 and 0.77 respectively in a study of 363 children and their parents [87], while the angle between the two lines was found to have moderate heritability 0.47 and 0.453 in the same analysis and in a different study of 77 twins [26]. Chin width was estimated to have heritability of 0.42 in a family study [95]. In our multivariate results, we observed high estimates in the chin area - 2nd SI, 8th CU and 15th GC regional traits, with respective trait heritabilities ranging from 0.685 to 0.711 (Fig. 3.8). All above comparisons are summarized in Table 3.5.

We also explored the heritability of facial traits based on Euclidean distances (EDTs), as well as equivalent phenotypes measured using Geodesic distances on the facial surfaces (GDTs). 10 EDTs and GDTs yielded reliable heritability estimates, ranging from 0.505 to 0.789. Evidence supporting our results were found in a number of family-based heritability analyses. A previously mentioned study reported its highest heritability estimates of 0.42, 0.44, 0.53 and 0.61 for mandible width, nasal protrusion, nose width and inner eye corner distance respectively [95]. Here, the nose width EDT and biocular width GDT yielded two of the three highest estimates -0.718 and 0.789 respectively-, while nasal protrusion and eye distance EDTs and GDTs were moderately to highly heritable, with heritability estimates ranging from 0.505 to 0.677. Nose width was also reported to be moderately to highly heritable in two studies of 125 Belgian and 342 Indian families with corresponding estimates of 0.639 and 0.498 [163, 5]. The same studies reported heritabilities of 0.606 and 0.605 for the bizygomatic width, while three further analyses, including large population samples from Russia and India and Europe, gave estimates of 0.52, 0.71 and 0.629 for the same trait [51, 92, 158]. Our most heritable EDT,

with $h^2 = 0.734$ corresponded to that width. In comparison to previously reported values, our heritability estimates were either similar or significantly higher.

Important inferences were made by exploring results from the various types of phenotypes used in this study. In the distance-based study, EDTs corresponding to mandible and mouth widths showed slightly higher heritability estimates in comparison to the respective GDTs. The latter traits had geodesic paths passing through the lip region, which, in the curvature analyses (Fig. 3.7), had consistently low heritability estimates. On the other hand, the GDT corresponding to inner eye corner distance gave the highest heritability estimate of 0.789. The geodesic path of that trait traverses the nasion region, which is a well known highly heritable area of the human face [121, 95, 180]. The results could indicate that GDTs are more sensitive than EDTs to the facial morphology between the considered landmarks. A comparison of summary statistics from the curvature and distance heritability analyses revealed that Structural Equation Models provided good data fits for approximately 80% of the curvature traits, but only 50% of the distance phenotypes, as assessed by Goodness-Of-Fit log-likelihood ratio tests. This observation could be a strong indicator that curvature phenotypes are more suitable for the study of facial morphology.

Finally, despite the small sample size for the needs of a GWAS, we were able to identify genome-wide significant genotypic markers at chromosomal location chr2p11.2, associated with the highest heritable GC regional trait. The respective Eigenface highlights the philtrum facial region, an area related to a well known genetic disorder, the syndromic Cleft of the lip and/or palate (CLP) [46]. The promising GWAS results indicate that the phenotyping approach and heritability results presented in this study can potentially inform and guide association studies relating the human face.

Phenotype	Ref.	Study	Sample Size	Ethnicity	Related Maps		
NW	[95]	FBH	229	Korean	MC Heritability Map		
IED	[95]	FBH	229	Korean	67th GC sPC	4th GC sPC	11th CU sPC
NP	[180]	TH	42	American	67th GC sPC	4th GC sPC	11th CU sPC
FW	[13]	TH	138	Asian	MC Heritability Map	SI Heritability Map	
FW	[150]	FBH	607	American	MC Heritability Map	SI Heritability Map	
FW	[51]	FBH	1406	European	MC Heritability Map	SI Heritability Map	
FW	[92]	FBH	373	Indian	MC Heritability Map	SI Heritability Map	
HC	[52]	FBH	1042	European	CU Heritability Map		
HC	[92]	FBH	373	Indian	CU Heritability Map		
MRL	[87]	FBH	363	European	CU Heritability Map	60th MC sPC	14th SI sPC
MBL	[87]	FBH	363	European	CU Heritability Map		
MR-MB	[87]	FBH	363	European	CU Heritability Map		
MR-MB	[26]	TH	77	European	CU Heritability Map		
CW	[95]	FBH	229	Korean	CU Heritability Map	15th GC sPC	8th CU sPC
NW. Node Width IED. Inner Eye Corner Distance NP. Nasion Protrusion FW. Face Width HC. Head Circumference MRL. Mandible Ramus Length MBL. Mandible Body Length MR-MB. Mandible Ramus - Mandible Body Angle CW. Chin Width							
FBH. Family-Based Heritability TH. Twin Heritability							

Table 3.5: Comparison between previously reported heritable phenotypes, heritability maps and regional curvature traits.

3.6 Conclusion

In this Chapter we presented GESSA for the automatic dense uniform sampling of correspondent points from a set of polyhedral surfaces. The optimization framework, first introduced in [28], was based on an entropy cost function which balances the uniform distribution of points in individual surfaces with point-wise correspondence across the surface ensemble. Here we built upon that framework and used geodesic distances between surface points, defined on the polyhedral surfaces, in order to increase precision during uniform point sampling. Furthermore, a suitable gradient descent optimization technique was developed to optimize the corresponding points' locations by operating directly on the surfaces. By incorporating these two key features, we were able to deal with highly curved surfaces and enhance the correspondence results.

Other dense 3D point correspondence methods for polyhedral data have been recently proposed, regarding applications mainly on face modeling and candidate association analyses, with the aim of studying genetic syndromes involving facial dysmorphisms and asymmetries [69, 35] and recognize genetic variants that explain normal face variation [124, 34]. In these works, pairwise correspondence methodologies were adopted and some form of manual landmarking was employed during facial annotation. In our methodology, given a the set of polyhedral surfaces roughly aligned in the same coordinate space - something that can be achieved using automated methods such as ICP - no manual annotation is required. Furthermore, since GESSA performs groupwise annotation, it does not suffer from potential bias towards a template or mean surface. Our algorithm is able to handle large datasets efficiently, as will be showcased in our application study.

The performance of GESSA was validated on the Morphface dataset, achieving very good landmarking accuracy. The TwinsUK dataset is not publicly available, thus we deemed it inappropriate for use in the validation study, since it would raise reproducibility issues. A number of other 3D face datasets do exist and have been commonly used in the literature, with prominent examples being: the BU-3DFE dataset [188], comprising of 100 subjects scanned during various facial expressions, the Bosphorus dataset [146], with 105 subjects scanned again with different expressions, the 3D-TEC dataset [173] of 107 pairs of twins, and the FRGCv2

dataset [128], which consists of approximately 4000 neutral and non-neutral expression 3D face images of 466 individuals. A number of reasons though lead us to not utilize the above datasets. The existence of many facial scans with various expressions in these datasets, apart from 3D-TEC, were not of particular interest in this work, since the TwinsUK dataset, that drove the construction of the methodology, includes only subjects with neutral expression. The BU-3DFE dataset comprises of subjects with a very large variety of ancestries, including White, Black, East-Asian, Middle-east Asian, Indian, and Hispanic Latino [188], while demographics of FRGCv2 could unfortunately not be found. The Bosphorus dataset does not readily provide polyhedral surface data, but rather 3D coordinate maps of acquired 2D images [146]. Finally, to the best of our knowledge, all mentioned datasets have strict license agreements, necessitating institutional support to acquire the data, with availability of 3D-TEC, in particular, made on a case by case basis [173].

We used GESSA to perform a large-scale twin heritability study, and an accompanying GWAS, of the human face shape. Heritability estimates were computed for local curvature phenotypes corresponding to single facial points and the results were combined to generate face heritability maps. Furthermore, regional curvature traits, corresponding to larger facial areas, were extracted through sPCA decomposition of the curvature data and their heritability was also estimated, yielding a number of highly heritable facial features. Heritability estimation was finally performed for a number of traditional facial length traits, with estimates been equivalent or higher to the ones found in the existing literature. In the application study, we provided a fresh perceptive in facial phenotyping and heritability analysis that could potentially inform future genome-wide studies and be useful in a variety of applications, ranging from population genetics and gene-mapping studies, to face modeling and reconstruction applications.

Chapter 4

Regression-driven Adaptive Sampling for Ensembles of Polyhedral Surfaces

4.1 Introduction and Motivation

In the areas of geometric modeling and computer-aided design, remeshing - computing new sets of vertices - of polyhedral surfaces is commonly employed for surface simplification and non-redundant reconstruction [166, 101, 43, 184, 112]. This simplification is predominately achieved by sampling new mesh vertices from a curvature-dependent surface distribution, with increased density around highly curved areas. The adaptive curvature sampling is based on the intuitive principle that, given a certain number of points, the geometry of a surface can be more accurately represented by this set of points, if curved surface areas are covered more densely than flat ones [101].

Contrary to the above, adaptive annotation of corresponding points on sets of surfaces has received minimal attention so far. This comes as no surprise, considering that prior to the development of dense correspondence methodologies, only sparse sets of preselected landmarks were considered, rendering the issue of such an adaptation non-existent. In the case of dense methodologies, we have observed an almost blind pursue for the increase in the number of annotated points, with no significant attention given to the question of how these are distributed

on the surfaces. While redundancy is unquestionably preferable to sparsity, it is reasonable to argue that a clever positioning of surface points offers certain advantages, both in terms of computational efficiency and in terms of data analysis.

Adaptive correspondence curvature sampling is a fairly straightforward extension for a number of dense methodologies, especially template-based ones, where the only difference would be the construction of a suitable template. In this work, we concentrate on a more complicated problem, that of adapting the points' distribution in order to optimize their locations according to feature measurements that are not intrinsic to the surfaces' structure. To showcase the usefulness of such a method, we present below an example in which such a procedure can be employed. In the field of medical imaging, a common application involves the study of brain structures with the goal of identifying regions representative of patients' status. Datasets are commonly split into case and control subjects. This classification can be used to guide the density of the landmarks' distribution, in essence combining the landmarking process with what would be a subsequent analysis step of feature selection.

In this Chapter, we present an extension of the GESSA correspondence algorithm that allows regression-driven adaptive sampling of ensembles of surfaces, hereafter referred to as adaptive GESSA. This is made possible through the use of an Adaptive geodesic Kernel Density Estimator (AgKDE), which ultimately dictates the distribution of landmarks, and a Random Forest Regression model for the feature selection process, which provides the weights required for AgKDE. Adaptive GESSA was used in an application of age prediction from facial surfaces. The results showed that denser annotation on age-related areas enhances the predictive performance.

4.2 Problem Definition

Regression-driven adaptive correspondence sampling extends the problem setting presented in the Chapter 3. Let us again consider a dataset of N polyhedral surfaces $\{\mathcal{S}_i | \mathcal{S}_i \subset \mathbb{R}^3, i = 1, \dots, N\}$. The main objective, of automatically distributing N collections of M points $\{\psi_j^i | \psi_j^i \in$

$\mathcal{S}_i, j = 1, \dots, M\}$, and establish point-wise correspondence among points on the N surfaces, still remains. In addition, let $\{y_i | y_i \in \mathbb{R}, i = 1, \dots, N\}$ be observations of a quantitative feature or attribute that is in some way associated to the polyhedral surfaces. Instead of uniformly distributing corresponding points on the surfaces, our second objective is to adapt the points' distribution so that surface regions better associated to that feature are covered more densely by them. It should be noted that while here we concentrate on scalar-valued traits, extension to the multivariate case is trivial.

4.3 Extension of GESSA for Adaptive Correspondence Sampling

The general formulation of the correspondence algorithm is adopted unchanged from Chapter 3. We again assume, during the methodological description, that sets of M points have been positioned in possibly random locations on the N surfaces. We begin by describing the necessary modifications in the kernel density estimator (3.9) and corresponding gradient descent updates (3.13), which allow for a non-uniform distribution of points, when such a distribution is desirable. These modifications apply not only in the correspondence setting, but also for point sampling of individual surfaces. Subsequently we describe how a RF regression model can be used to drive the adaptative positioning on corresponding points for an ensemble of surfaces.

4.3.1 Adaptive Distribution of Points on Surfaces

In Chapter 3, we constructed the gKDE (3.9) as a gaussian function on squared geodesic distances, with σ_S controlling the fixed bandwidth of the kernel. Let us now consider the AgKDE given by

$$p_i^{aKDE}(\psi_j^i) = \frac{1}{M} \sum_{k=1}^M \frac{1}{h_{S_k}} g_S \left(\frac{d_S(\psi_j^i, \psi_k^i)}{h_{S_k}}, \sigma_S \right), \quad (4.1)$$

where h_{S_k} are elements of a predefined weight vector $\mathbf{h}_S = (h_{S_1}, \dots, h_{S_M})$ associated with points $\psi_1^i, \dots, \psi_M^i$. Such an estimator is referred to in the literature as sample-point estimator,

and its use is advantageous, compared to a fixed bandwidth kernel, when the underlying distribution exhibits multimodality and/or skewness [138]. In more detail, AgKDE is a mixture of individually scaled kernels centered around the M surface points. Points with weights greater than one contribute over more extended areas during density estimation, while the inverse is true for points with weights below one. The effect of this scaling can be seen as a warping of the underlying surface according to the kernel bandwidths, with weights above/below one translating to a shrinkage/elongation of the regions around the equivalent points.

Let us now recompute the derivative of negentropy $-\hat{H}_i$ w.r.t point ψ_j^i , using the adaptive density function (4.1). Under the same assumptions as in Chapter 3, we have

$$\begin{aligned}
\frac{d(-\hat{H}_i)}{d\psi_j^i} &= -\frac{1}{M} \sum_{l=1}^M \frac{1}{p_i^{aKDE}(\psi_l^i)} \frac{dp_i(\psi_l^i)}{d\psi_j^i} . \\
&\approx -\frac{1}{M^2} \frac{1}{p_i^{aKDE}(\psi_j^i)} \sum_{k=1}^M \frac{1}{h_{Sk}} \frac{d\left(g_S\left(\frac{d_S(\psi_j^i, \psi_k^i)}{h_{Sk}}, \sigma_S\right)\right)}{d\psi_j^i} \\
&\approx \frac{1}{2\sigma_S M} \sum_{k=1}^M \frac{\frac{1}{h_{Sk}} g_S\left(\frac{d_S(\psi_j^i, \psi_k^i)}{h_{Sk}}, \sigma_S\right)}{\sum_{l=1}^M \frac{1}{h_{Sl}} g_S\left(\frac{d_S(\psi_j^i, \psi_l^i)}{h_{Sl}}, \sigma_S\right)} \frac{d\left(\frac{1}{h_{Sk}^2} d_S^2(\psi_j^i, \psi_k^i)\right)}{d\psi_j^i} \\
&\approx -\frac{1}{\sigma_S M} \sum_{k=1}^M \frac{1}{h_{Sk}^2} \frac{\frac{1}{h_{Sk}} g_S\left(\frac{d_S(\psi_j^i, \psi_k^i)}{h_{Sk}}, \sigma_S\right)}{\sum_{l=1}^M \frac{1}{h_{Sl}} g_S\left(\frac{d_S(\psi_j^i, \psi_l^i)}{h_{Sl}}, \sigma_S\right)} \log_{\psi_j^i}(\psi_k^i) \\
&\approx -\frac{1}{\sigma_S M} \sum_{k=1}^M \frac{1}{h_{Sk}^2} \frac{g_{Sh}^{jk}}{\sum_{l=1}^M g_{Sh}^{jl}} \log_{\psi_j^i}(\psi_k^i) ,
\end{aligned} \tag{4.2}$$

with $g_{Sh}^{jk} := \frac{1}{h_{Sk}} g_S\left(\frac{d_S(\psi_j^i, \psi_k^i)}{h_{Sk}}, \sigma_S\right)$. We see that the gradient descent updates are again weighted averages (according to the adaptive kernel values) of logarithmic maps. This time though, each term is further multiplied by the squared reciprocals of the weights h_{Sk} .

To clarify the effect of these weights on the distribution of surface points, it is helpful to consider that in the original formulation of GESSA, updates actually push points away from each other, so as to achieve the uniform coverage of a surface. Through the weighting of each logarithmic map by $\frac{1}{h_{Sk}^2}$, the repulsion force attributed to point ψ_k^i is lessened, when $h_{Sk} > 1$, and increased, when $h_{Sk} < 1$. The result is that entropy optimization leads to a non-uniform point distribution, controlled by $\mathbf{h}_S$.

An example of the different point distributions achieved through the original and adaptive GESSA formulations is visualized in Fig. 4.1. In detail, we sampled 1024 points from one facial surface, first using the uniform and then the adaptive optimization processes. In the adaptive case, weights h_{S_k} were computed to be the absolute Mean Curvature (aMC) values in each point (see subsection 3.5.3 for a description of Mean Curvature). In Fig. 4.1(b) we include a colormap of the face's aMC. The resulting point sets are depicted in Fig. 4.1(a). We see that adaptively sampled points congregate more densely in regions of high aMC. This desirable outcome is also obvious on the estimated density functions. Figure 4.1(c) depicts colormaps of the two density functions, measured on the original mesh vertices. The gKDE map shows almost constant values across the facial surface, while AgKDE mimics well the aMC map.

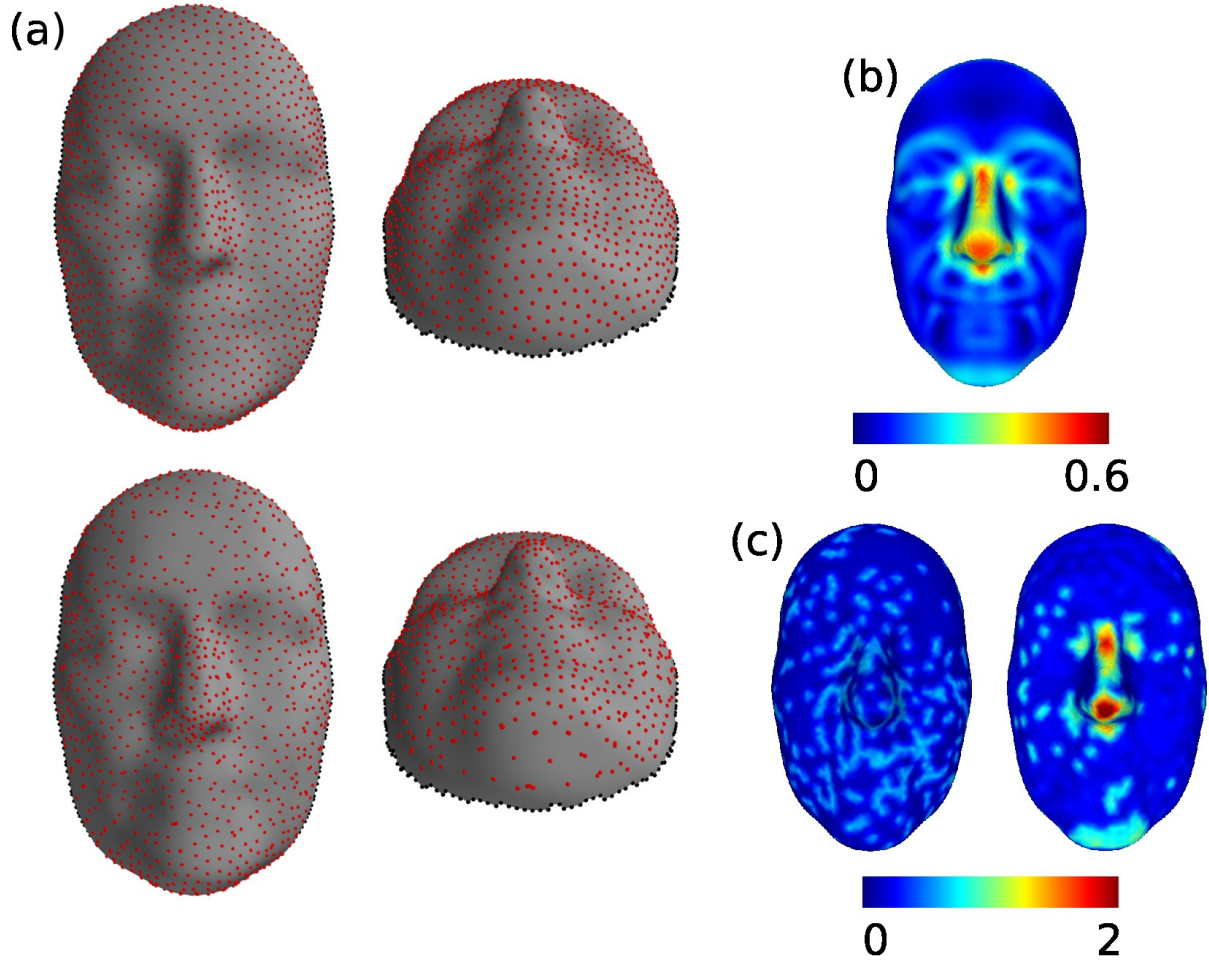


Figure 4.1: Uniform vs. curvature adaptive sampling of a single facial surface. (a) 1024 points sampled first uniformly and then adaptively, based on the facial mean curvature, as shown in (b). The resulted kernel density estimators for both cases are shown in (c).

4.3.2 Regression-Driven Adaptive Correspondence Optimization

We now concentrate on how to adaptively sample landmark points in a collection of surfaces, based on the associated measurements $\{y_i | y_i \in \mathbb{R}, i = 1, \dots, N\}$. Having seen how the AgKDE can be used for the adaptation of landmark positions, the problem reduces to one of appropriately selecting the weight vector $\mathbf{h}_S$. Let x_j^i be scalar-valued trait measurements associated with surfaces points ψ_j^i . Then we can construct the $N \times M$ trait matrix $\mathbf{X}$ and build a regression model of the form

$$\mathbf{y} = f(\mathbf{X}) + \boldsymbol{\epsilon} \ , \quad (4.3)$$

where $\mathbf{y} = (y_1, \dots, y_N)^T$ is the $N \times 1$ vector of feature observations, and $\boldsymbol{\epsilon}$ the $N \times 1$ vector of error terms.

A number of regression methodologies can be used to extract a ranking of significance for the column input variables x_k^i , which correspond to landmark positions ψ_k^i . In this work we employ the RF regression algorithm (see section 2.3), and the gini Variable Importance (gVI) index, as given in (2.7). By fitting the data to the RF model, we acquire a vector $\mathbf{w}^{RF} = (w_1^{RF}, \dots, w_M^{RF})$ of variable importances $w^{RF}(k) = g_{VI}(k)$, which can be associated to the M landmark locations on the surfaces. In order to compute the weights for the AgKDE, we subsequently use the formula

$$h_{Sk} = (t_{VI} + w_k^{RF})^{r_{VI}} \ , \quad (4.4)$$

where t_{VI} and r_{VI} are free parameters. t_{VI} can be used to control the ratio of adaptation weights that are above and below 1. Depending on the number of landmarks M , reasonable values for parameter t_{VI} range between 0.5 and 1. The power parameter $r_{VI} > 0$ can be used to scale up or down the adaptation effect in the landmark distribution. The exact form of this transformation is of course heavily dependent on the regression methodology, as well as the range and number of variable importance values. In this work, we opted for the formula (4.4), due to the fact that the RF gVIs were considered after normalization, with $\sum_{j=k}^M w_k^{RF} = 1$.

Two final notes are made here regarding the nature of trait measurements x_j^i and the gVI. Our methodology is based on the ability to associate each landmark position with a univariate

trait, for which an importance score can be computed through the regression model. Examples of such traits are the various curvature phenotypes presented in Chapter 3. Consideration of different types of traits, such as pairwise landmark distances or landmark coordinates is not impossible, but would require the construction of a concise technique in order to translate trait importances to individual landmark weights. Regarding the utilized importance measure, gVI is based on the values of the RF split objective function, thus it inherits the benefits of the bootstrapped RF methodology, providing a form of marginal importance for the traits.

4.4 Application of Adaptive GESSA on the Prediction of Age from 3D Facial Surfaces

4.4.1 Introduction and Motivation

Aging refers to the progressive decline of biological functionality and loss of homeostasis that leads to mortality [76]. Being a common process across the class of mammalian organisms, including human beings, aging is a highly individualized and influenced by a multitude of factors including genetic, epigenetic and environmental effects [120, 68, 32]. This combination of diversification and individualization results in a significant phenotypic heterogeneity among people with the same chronological age [120], which in turn has posed significant challenges in the research for identification of robust biomarkers of age. Most work so far has been concentrated at the cellular level, with biomarkers including, among others, chromatin changes [154], DNA mutations and polymorphisms [94], gene expression profiles [76], telomere length [157] and MiRNA signatures [50].

Despite the extensive work on the subject of aging, reliability of proposed biomarkers is still problematic [120, 31]. Recently, increased interest has been given in the evaluation of age in humans based on more readily available traits. One such phenotype is facial morphology, which is strongly associated to the aging process and shows significant changes during the course of human life [31]. Good accuracy was observed in two separate studies of chronological age

prediction using extended aging databases with 2D photos of human faces [57, 66]. To our knowledge, the best predictive performance of age from facial phenotypes so far was reported in a study that employed PCA to decompose landmark coordinates from 3D surfaces into facial traits, and used them as covariates in a Support Vector Regression model [31].

In this application, we make use of the TwinsUK dataset to perform an age prediction study based on curvature facial traits, which describe the shape of the facial surface. Whilst not the main focus of this work, we first show that our curvature phenotyping approach results in similar predictive performance, compared to the results presented in [31], whilst providing a more intuitive and straightforward understanding of the facial regions that are more affected by the aging process. Furthermore, we showcase an improvement in prediction accuracy when faces are annotated using the adaptive GESSA methodology.

In the following, we first present a preliminary study on face age prediction using the regional traits, as computed for the heritability study of Chapter 3.5 (see subsection 3.5.4 for a description of regional curvature phenotypes). Subsequently, we present a comparative study on age prediction, utilizing the uniform and adaptive variations of GESSA, and compare the performance of the two approaches.

4.4.2 Data Description

The same dataset, employed in the heritability study of Chapter 3.5, was used for the purposes of the age prediction application. See 3.5.2 for a detailed description of the initial data. In the preliminary study, all 952 facial surfaces were taken into consideration, while for the new comparative study, we concentrated on the subset of 476 unrelated individuals, for which sets of 1024 corresponding points were recomputed using both GESSA formulations.

4.4.3 Preliminary Study on Face Age Prediction

The purpose for this preliminary study was to establish a baseline prediction performance on the TwinsUK dataset and compare our curvature-based phenotyping approach to a previous

relevant study ([31]). We evaluated the prediction of age using as inputs the four sets of 100 curvature regional traits (one set for each curvature index), as used in the heritability study of Chapter 3.5. These were computed using sPCA on the four 952×4096 data matrices of respective univariate curvature phenotypes. For each subject, the 400 regional trait measurements were concatenated into an input vector, resulting in a 952×400 input matrix. Due to the different range of curvature index values, trait measurements were normalized prior to the regression analysis.

The prediction results presented below are based on a Leave-2-Out Cross Validation (L2O CV) procedure. In each fold, one pair of twin subjects was removed from the training dataset and was used for testing purposes. Lasso Linear Regression (LLR) [168] was employed as a regression model from the curvature inputs to the age response. Different values for the lasso sparsity parameter λ_{lasso} were considered. λ_{lasso} was kept the same in all CV rounds. Best results were acquired for $\lambda_{lasso} = 0.005$.

Figure 4.2 presents the combined prediction results from L2O CV. The Mean Absolute Error (MAE) over the complete dataset was 4.518 years, while the Pearson correlation between true and predicted ages was 0.817. Compared to the previous application presented in [31], our MAE was approximately 1.6 years lower. We were thus able to deduce that curvature phenotyping does provide informative traits for the problem of age prediction. An in depth comparison between the two studies is given in the Discussion section.

Due to the unique nature of our dataset, we were also able to assess the difference in prediction accuracy between monozygotic (MZ) and dizygotic (DZ) twin subsets. Figure 4.3 shows the distributions of predicted absolute age differences for the subsets of MZ and DZ twin pairs. We notice that larger values were more common in the DZ subset. While this result is by no means conclusive, it could hint towards a genetic influence in the rate of aging.

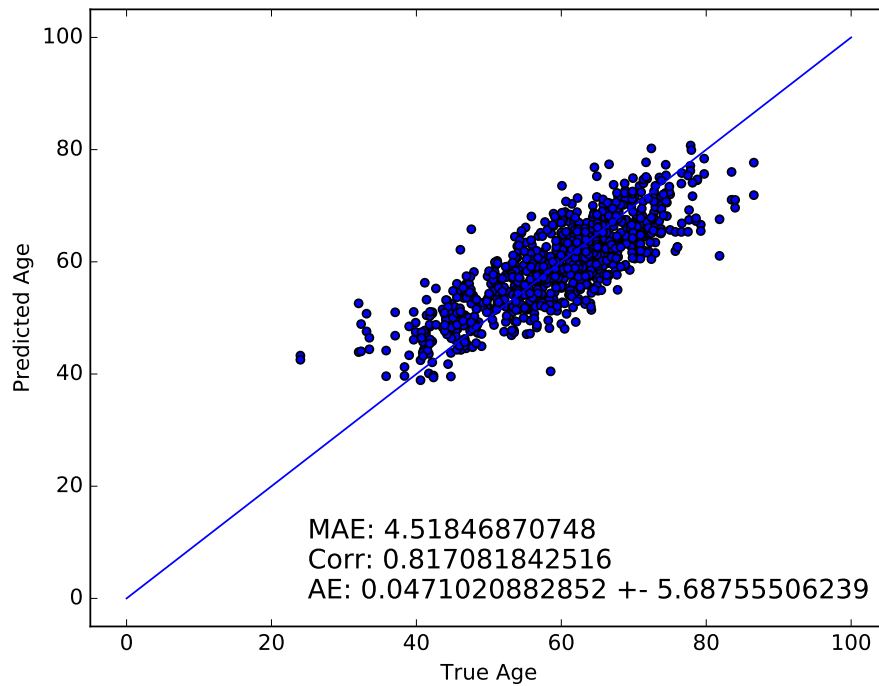


Figure 4.2: Preliminary Study - True vs. Predicted ages on the set of 952 twins, acquired from leave-2-out cross validation, using lasso regression. Average Error (AE), Mean Absolute Error (MAE) and the Pearson correlation coefficient (Corr) between the true and predicted sets are also shown in the figure.

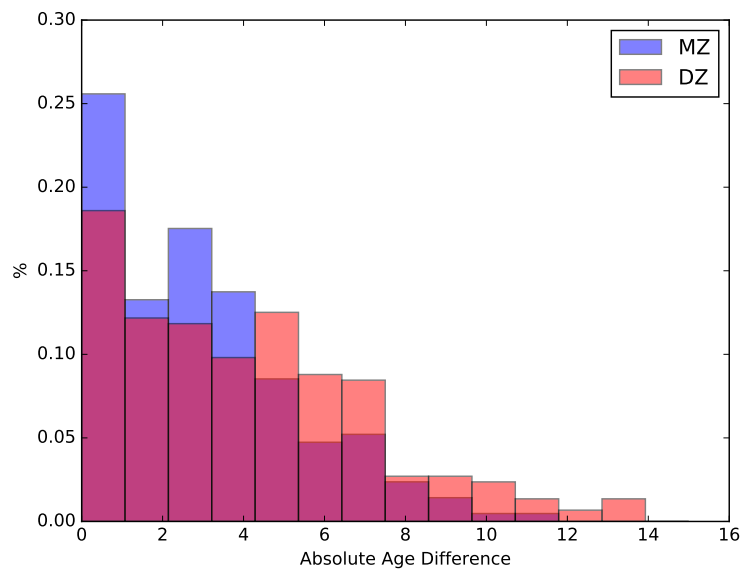


Figure 4.3: Preliminary Study - Distributions of Predicted Absolute Age Differences between monozygotic (MZ) and Dizygotic (DZ) twin pairs.

4.4.4 Age Prediction through Adaptive Correspondence Sampling

Having established that the morphological curvature traits are well suited to the application of age prediction from facial surfaces, we proceeded to evaluate whether adaptive sampling would be beneficial for the application on age prediction, compared to the uniform point annotation approach. In this study, we concentrated on a subset of 476 unrelated individuals from the TwinsUK cohort. First, we reapplied GESSA on the dataset of facial surfaces with the number of corresponding points set to $M = 1024$. In Fig. 4.4(a) we show the computed points on three example faces, while Fig 4.4(b) depicts the estimated density function on the face. Despite the decrease in the number of sampled points, compared to the application in Chapter 3.5, where M was chosen to be 4096, point coverage of the facial surfaces was sufficiently dense. To showcase the validity of this argument, we preselected 11 fiducial points (see Fig. 4.4(c)), and after GESSA was applied on the surfaces, we located closest points to these fiducial landmarks. These are shown in Fig. 4.4(d), and verified that a good landmarking precision was achieved.

Subsequently, we used adaptive GESSA to compute four different adapted point sets with cardinality of 1024, based on MC, GC, CU and SI curvature traits. In detail, for each curvature index, facial points were associated to their respective curvature phenotypes. A fixed subset of 76, out of the 476, subjects was selected for training of the RF regression models, with curvature traits as inputs and age as the response. Following the adopted nomenclature from section 4.3.2, the four 76×1024 trait matrices $\mathbf{X}_{MC}$, $\mathbf{X}_{GC}$, $\mathbf{X}_{CU}$ and $\mathbf{X}_{SI}$, were used as inputs in different RF models, with the same response vector $\mathbf{y}$, comprised of the ages of the 76 subjects. All RF models were unpruned and built with 100 trees. As described in the previous section, this allowed us to compute four variable ranking vectors $\mathbf{w}_{MC}^{RF}$, $\mathbf{w}_{GC}^{RF}$, $\mathbf{w}_{CU}^{RF}$, $\mathbf{w}_{SI}^{RF}$ for the sampled facial points. These weights were subsequently employed to adapt the distributions of the corresponding points on the remaining 400 subjects. The parameter t_{VI} in the formula (4.4) was modified each time so that 50% of points had weights above one, while the other half had weights below one. r_{VI} was set to 100. Three examples of age-adapted annotated faces, along with the estimated densities are included in Fig 4.5. We notice that the adapted point sets still provided a good coverage of the complete facial surfaces, while specific regions, especially

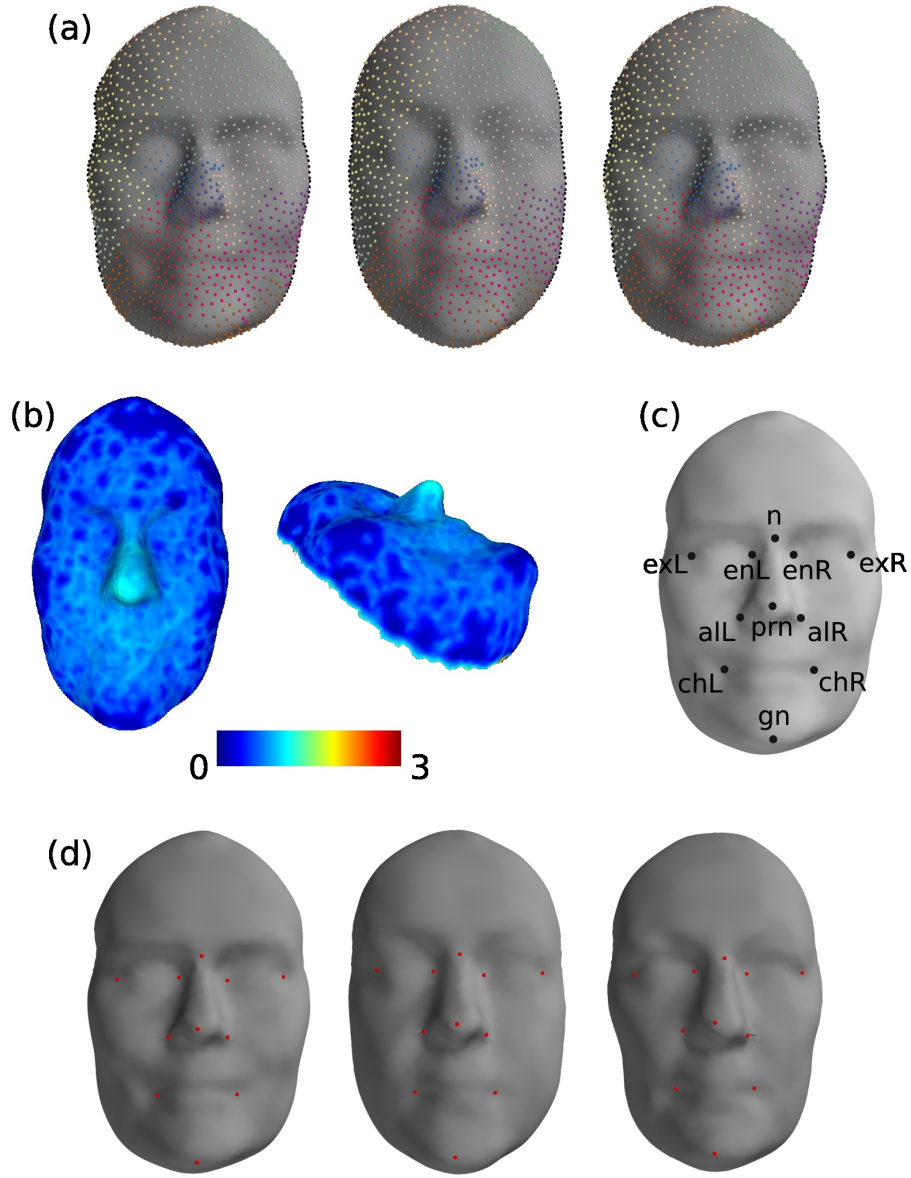


Figure 4.4: Age Prediction - 1024 corresponding points were located, using GESSA on a set of 476 facial surfaces. Three example faces with the computed correspondence are shown in (a), while (b) depicts the estimated density function. A further affirmation of the good landmarking accuracy was given by visual inspection of closest GESSA points to 11 predefined fiducial landmarks, shown in (c). (d) Depicts the identified points for the same three example faces.

around the mouth, nose and eyes were more densely annotated. From this point onwards, the 76 subjects that were used as training observations to guide the adaptation, were not included in any further regression analysis.

After the completion of both uniform and adaptive procedures and exclusion on the subjects used for training the RF models, we ended up with eight 400×1024 data matrices of phenotypic observations, one for each combination of sampling approach and curvature index. Before

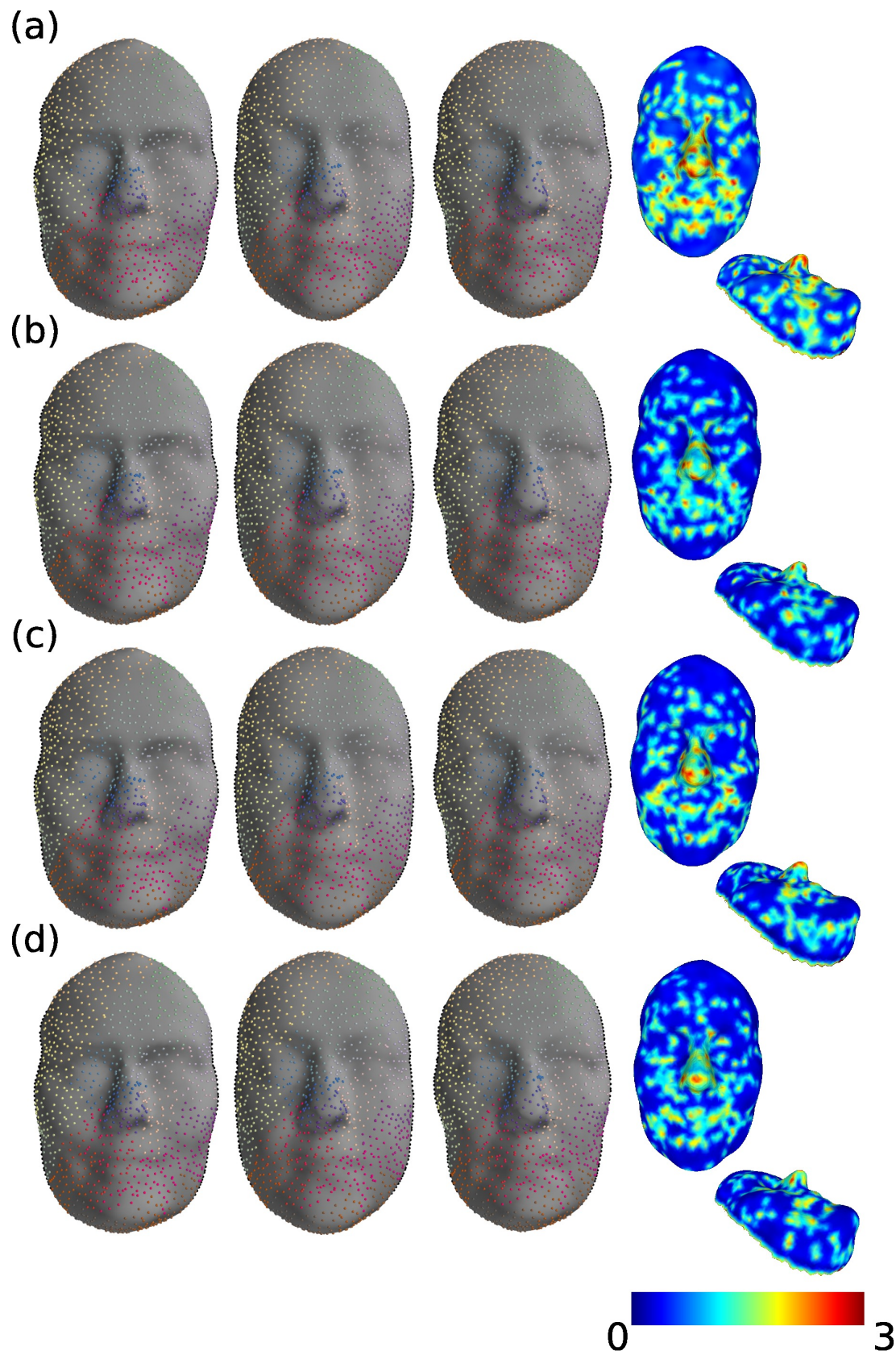


Figure 4.5: Age Prediction - Examples of three faces with computed age-driven adapted point sets and the estimated density functions for (a) MC (b) GC (c) CU and (d) SI curvature traits.

proceeding to the study of age prediction, we employed Stability Selection (SS) in order to investigate which facial parts are highly associated with age. For all eight input data, very similar results were acquired. Informative traits were mostly localized around the mouth (nasolabial folds), under the chin, and to a lesser extent, close to the zygomatic bones. We also noticed that stability maps of respective curvature indices were very similar for both landmarking approaches, with somewhat increased localization of high values observed for the adaptive case. Further details on the above can be found in Appendix B.1. It must be noted that SS results were not used for feature selection during the main application of age prediction, which will be presented below.

The ability to predict age from facial phenotypes was assessed again using LLR. We performed a number of regression analyses, using as inputs each of the curvature index input data independently, as well as combined curvature datasets for each landmarking approach. In more detail, 10 regression problems were set up. In all of them, the response vector $\mathbf{y}$ was the same, comprising of 400 subject ages (range: 32.37 – 86.57, mean value: 59.664, standard deviation: 9.6123). Regarding the predictor data, four regression models were built using as inputs the 400×1024 matrices of MC, GC, CU and SI traits, based on the uniformly distributed point sets. Equivalently, another four models used trait matrices based on the adapted point sets. Finally, we concatenated traits from all curvature predictors into two 400×4096 matrices, for the uniform and adaptive cases respectively, and examined the predictive performance of models trained on these as well.

Prediction results presented below were based on 10-Fold CV. For each fold, 40 subjects were excluded during training and were used to acquire test prediction errors. The results for all folds were combined and are presented in Table 4.1 (Mean Absolute Prediction Error, Pearson correlation coefficient and p -value between true and predicted age responses). Scatterplots of true vs. predicted ages for all models are shown in Fig. 4.6.

Irrespective of the curvature index considered, CV mean absolute errors for models using input traits based on the adapted landmark sets were consistently lower, compared to their uniform-based counterparts. As expected, by combining the four curvature input datasets, we were able

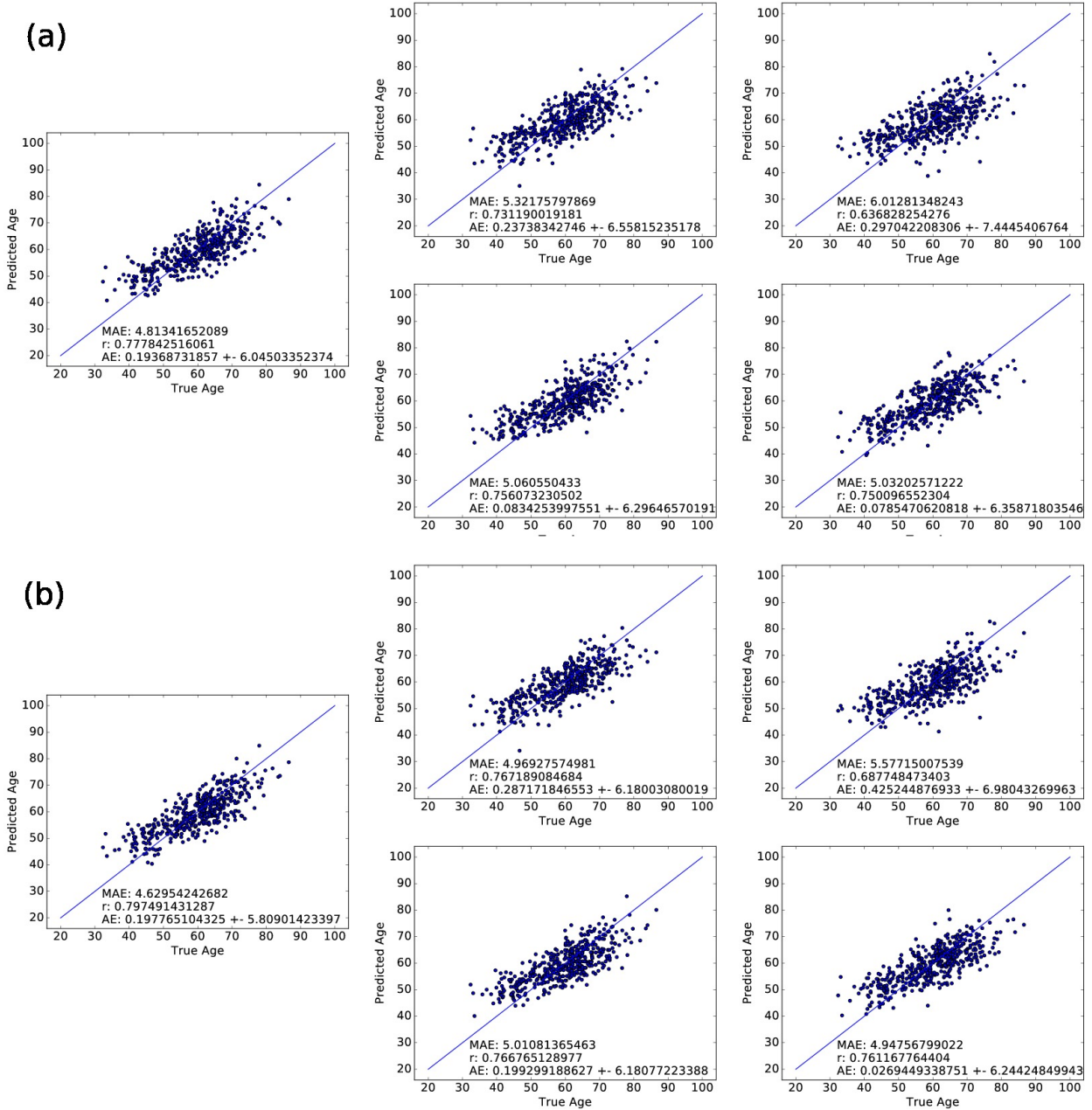


Figure 4.6: True vs. Predicted ages on the set of 400 unrelated subjects from the TwinsUK cohort, acquired from 10-fold cross validation, using lasso regression. (a) Predictions were made using as inputs the phenotypic measurements computed on the uniform point sets and combined traits from all curvature indices (left), MC only (top middle), GC only (top right), CU only (bottom middle) and SI only (bottom right). (b) Similar to above, for phenotypic measurements computed on the adapted point sets. Average Error (AE), Mean Absolute Error (MAE) and the Pearson correlation coefficient (r) between the true and predicted ages are also shown in the figure.

to get better prediction accuracy. In the adaptive case, MAE was approximately 0.183 years lower than in the uniform case, while correlation between true and predicted ages was also higher by 0.019. We also investigated the predictive performance of the models in various age

GESSA	Curvature Index	No of traits Considered	CV MAE	CV AE	Pearson r	Pearson p -value
Uniform	MC	1024	5.322	0.237 ± 6.558	0.731	$2.9e - 76$
	GC	1024	6.013	0.297 ± 7.445	0.637	$1.93e - 52$
	CU	1024	5.061	0.083 ± 6.296	0.756	$2.54e - 84$
	SI	1024	5.032	0.079 ± 6.358	0.75	$4.31e - 86$
	Combined	4096	4.813	0.194 ± 6.045	0.778	$3.27e - 92$
Adaptive	MC	1024	4.969	0.287 ± 6.18	0.767	$3.05e - 88$
	GC	1024	5.577	0.425 ± 6.98	0.688	$3.66e - 64$
	CU	1024	5.011	0.199 ± 6.181	0.767	$4.34e - 88$
	SI	1024	4.948	0.027 ± 6.244	0.761	$2.67e - 82$
	Combined	4096	4.63	0.198 ± 5.81	0.797	$3.82e - 100$

Table 4.1: Summary of age prediction results from 10-Fold Cross Validation, on the dataset of 400 unrelated subjects from the TwinsUK cohort.

ranges. In Table 4.2, we report Mean Absolute prediction Errors for data subsets with true age ranging as follows: 30 – 40, 40 – 50, 50 – 60, 60 – 70, 70 – 80, 80 – 90. Best predictive performance was acquired for ranges 50 – 60 and 60 – 70. This is not surprising, considering the age’s mean and standard deviation of our dataset (mean value: 59.664, standard deviation: 9.6123), as well as the fact that 311 subjects, out of the 400, fall within these age groups. Regarding the different curvature indices employed in the study, GC was the less suited shape descriptor in terms of prediction performance, while the remaining three indices were very closely matched to each other. This outcome could be justified by the more localized nature of GC, in comparison to the other curvature descriptors.

Age Range	30-40	40-50	50-60	60-70	70-80	80-90
No of Subjects	9	71	123	188	54	4
Uniform MC MAE	13.618	6.615	4.114	4.451	7.447	12.993
Uniform GC MAE	14.145	7.478	4.651	4.938	8.904	15.01
Uniform CU MAE	12.723	6.246	3.783	4.506	6.684	10.131
Uniform SI MAE	12.518	5.818	3.951	4.0629	7.567	9.303
Uniform Combined MAE	12.528	5.695	3.646	4.113	6.984	11.25
Adaptive MC MAE	13.332	6.051	4.031	3.903	7.339	13.879
Adaptive GC MAE	13.383	6.896	4.291	4.56	8.571	11.501
Adaptive CU MAE	12.367	6.085	3.635	4.54	6.773	9.985
Adaptive SI MAE	12.855	5.678	4.016	4.178	7.638	12.089
Adaptive Combined MAE	12.294	5.715	3.461	3.955	6.642	8.55

Table 4.2: Number of subjects and Mean Absolute Errors for 10-year age ranges and all models, on the dataset of 400 unrelated subjects from the TwinsUK cohort, based on 10-fold CV.

4.4.5 Discussion

We presented here a detailed evaluation of age prediction from morphological curvature traits describing the shape of the human face. To our knowledge, this is the second study that uses 3D facial surface data. The first study, presented in [31], employed a set of 332 3D facial surfaces from subjects of Chinese origin. In that work, a dense correspondence procedure was also used, followed by identification of morphological phenotypes through PCA decomposition on the dataset of vectorized landmark coordinates. Using L1O CV and support vector regression, the authors reported a CV MAE of 6.105, with correlation between true and predicted ages being 0.855. In our preliminary age prediction study, we employed the complete TwinsUK dataset of 952 faces and used as inputs regional curvature traits. Using L2O CV, due to the twin structure of our data, and lasso regression, we were able to acquire a significantly lower CV MAE, of 4.518. True vs Predicted age correlation was lower in our study (0.817). We attribute this result to the smaller range of ages in our data, which was 54.2 years ($32.37 - 86.57$), in contrast to 60 ($17 - 77$) in [31], as well as the fact that our age data were more concentrated around the mean value, with a mean age of 59.664 and standard deviation 9.6123, as opposed to mean age 42.959 and 15.06 standard deviation [31]. Due to the different datasets used in the two studies, it is of course not possible to unequivocally state that the use of curvature phenotypes is more advantageous. Nevertheless, a definitive benefit of our approach is that curvature measures are more intuitive as facial traits, as was highlighted in the heritability study of section 3.5.

In the main application we explored whether our age-driven adaptive point correspondence approach improves predictive performance, compared to uniform annotation of faces. We used both uniform and adaptive formulations of GESSA to compute new sets of 1024 facial points on 400 unrelated subjects from the TwinsUK cohort. Morphological traits were computed using four curvature indices and constituted the inputs in the regression models. Five different input matrices were considered, four for data derived by each curvature descriptor, as well as an additional input matrix where traits from all curvature indices were concatenated. We performed 10-fold cross validation to predict ages for all subjects. Irrespective of the input, the use of adaptive-based traits as inputs resulted in better prediction accuracy, compared to regression

with uniform-based traits.

Finally, we elaborated on the suitability of each individual curvature index to capture age-related morphological traits. SI phenotypes led to the best predictive performance, followed closely by the MC and CU indices. GC performance was significantly worse than the rest. This can be attributed to the highly localized nature of differentiation that GC provides, as shown in Figures 3.5 and 3.6.

4.5 Conclusion

In this Chapter we presented an extension to the GESSA methodology for adaptive correspondence sampling of polyhedral surfaces, driven by feature measurements that are associated to the surfaces. The algorithm makes use of a locally adaptive kernel density estimator, in order to enforce denser allocation of points in surface regions with traits that are highly predictive of the associated feature measurements. Adaptive kernel density estimation is a well-known technique with established advantages, compared to fixed-bandwidth kernel density estimation, predominantly when the underlying distribution exhibits skewness and multimodality [138, 140, 139]. Two variations of adaptive estimators exist, namely balloon and sample-point estimators. The first formulation varies the bandwidth of the kernel with the estimation point. Contrary to that, in sample-point estimators, the bandwidth is varied with each data point and not with the estimation point. In this work we employ an estimator of the second type, since it lends itself naturally to the requirement of weighting each point according to its predictive performance. Furthermore, it has been shown that sample-point estimators exhibit better asymptotic performance than both fixed-bandwidth and balloon estimators [90].

As described previously, adaptive KDE is used in our method as a means to achieve the manipulation of the points' distribution on the surfaces. Kernel bandwidths are selected according to predictive performance weights of traits corresponding to surface points. These weights are computed by means of a RF Regression model which was chosen due to the algorithms' good performance in variable ranking [59], as well as its ability to handle multivariate data. While RF

variable importance can be susceptible to high correlation among the predictor variables (see for example [106]), we have not noticed any detrimental effects during our application study. Furthermore, in cases when such correlation is recognized as an important issue, different RF variable importance measures, such as conditional VI [160, 7] can be utilized.

To the best of our knowledge, this is the first time that a regression-driven adaptive correspondence methodology is presented in the literature. The seamless combination of correspondence sampling with the feature selection process provides advantages both in terms of computational efficiency and data analysis. To showcase these benefits, we performed a predictive analysis of age from facial curvature phenotypes. The adaptive correspondence approach provided dense coverage of facial areas that are clearly associated to the aging process, while, by employing the adapted point set to construct phenotypic traits, we were able to get improved age predictions.

Chapter 5

Non-parametric Predictive Methodology for Manifold-Valued Responses

5.1 Introduction

Predictive modeling is an integral part of data analysis, and encompasses the process of developing models which can accurately predict yet-to-be-seen data. A multitude of regression models exist that can be used for prediction of univariate, or multivariate vectorial, responses. Nevertheless, less work has been done on the difficult problem of modeling and predicting more complex objects that possess additional structure, be it morphological, directional, or otherwise [132].

Images [74], shapes [156, 151], graphs [169], deformation tensors [99, 193] are examples of such complex data types that appear naturally as responses in image analysis, computer vision, medical imaging and other application areas. While such objects are usually represented as points on very high-dimensional spaces, a well known observation is that meaningful measurements actually lie on smooth non-linear hyper-surfaces of lower dimensionality, embedded in these

representation spaces [38, 129, 30].

Consequently, any suitable predictive modeling methodology should not treat these objects as points on Euclidean spaces, but rather work under the additional constraints imposed by the data's inherent structure. Major difficulties of course arise by forgoing the assumption of a Euclidean response space, such as lack of coordinates, vectors and no analytical or unique definitions of expected values [55].

In this work, we are particularly interested in experimental settings where the input observations may also be very-high dimensional objects, something that poses further requirements regarding performance and computational efficiency in the construction of a predictive method. In the following, we describe a number of predictive methodologies for structured response data, identify existing shortcomings and motivate our non-parametric predictive methodology.

5.1.1 Related Work

Three main families of methodologies concerning regression with manifold-valued objects can be found in the literature, namely intrinsic manifold regression, kernel-based and ML-based methods. Below we give an account of each of these approaches.

Intrinsic Manifold Regression

The majority of intrinsic manifold regression models are generalizations of linear regression for manifold-valued responses. Given a dataset $\{(x_i, y_i) \mid x_i \in \mathbb{R}^p, y_i \in \mathbb{R}, i = 1, \dots, N\}$ of N paired input-response points, the multiple linear regression model takes the form

$$y_i = \sum_{j=1}^p x_{ij} \beta_j + \alpha + \epsilon_i, \quad i = 1, \dots, N, \quad (5.1)$$

with $x_i = (x_{i1}, \dots, x_{ip})$, $\alpha \in \mathbb{R}$ the intercept, $\beta_1, \dots, \beta_p$ the regression coefficients and $\epsilon_1, \dots, \epsilon_N \in \mathbb{R}$ the residual error terms. Least squares estimators for the intercept and regression coefficients

are given by

$$(\hat{\alpha}, \hat{\beta}_1, \dots, \hat{\beta}_p) = \underset{(\alpha, \beta_1, \dots, \beta_p)}{\operatorname{argmin}} \sum_{i=1}^N \|y_i - \alpha - \sum_{j=1}^p x_{ij}\beta_j\|^2 . \quad (5.2)$$

In the case of a manifold-valued response variable now, with $y_i \in \mathcal{M}$, $i = 1, \dots, N$, the proposed generalization of the aforescribed multiple linear regression model is taken to be [55, 56, 75, 159, 193, 151]:

$$y_i = \exp \left(\exp(\alpha, \sum_{j=1}^p x_{ij}\beta_j), \epsilon_i \right), \quad i = 1, \dots, N , \quad (5.3)$$

with $\exp$ denoting the exponential map on $\mathcal{M}$ (see definition 2.5), $\alpha \in \mathcal{M}$ and $\epsilon_i, \sum_{j=1}^p x_{ij}\beta_j \in \mathcal{T}_{y_i}\mathcal{M}$. Figure 5.1(a) provides an illustration of the intrinsic regression model. When the manifold is a Euclidean space, then the exponential maps revert to vector addition operations and models (5.1) and (5.3) are identical. Least squares estimators for the intercept and regression coefficients can be computed by [55, 56, 75, 159, 193, 151]:

$$\begin{aligned} (\hat{\alpha}, \hat{\beta}_1, \dots, \hat{\beta}_p) &= \underset{(\alpha, \beta_1, \dots, \beta_p)}{\operatorname{argmin}} \sum_{i=1}^N d_{\mathcal{M}}(y_i, \exp(\alpha, \sum_{j=1}^p x_{ij}\beta_j))^2 \\ &= \underset{(\alpha, \beta_1, \dots, \beta_p)}{\operatorname{argmin}} \sum_{i=1}^N |\log_{y_i}(\exp(\alpha, \sum_{j=1}^p x_{ij}\beta_j))|^2 , \end{aligned} \quad (5.4)$$

where $d_{\mathcal{M}}$ is the distance metric on the manifold and $\log$ denotes the logarithmic map on $\mathcal{M}$ (see corollary 2.2).

In [55], a gradient descent algorithm was constructed to solve (5.4), necessitating the analytical formulation of the exponential map's partial derivatives with respect to α and β , as well as their respective adjoint operators. The described manifold regression method was applied on the analysis of medical shape data, more specifically in the prediction of the 2D shape of corpus callosum in the human brain using age as the input variable [55, 56, 54]. In that application, each shape was represented by a set of points in the 2D plane and the Kendall's shape space [156] was used as the underlying response manifold. A similar methodology, termed Grassmannian geodesic regression [78], was applied on the same problem, as well as for traffic speed prediction and crowd counting. An intrinsic linear manifold regression model was also applied on the

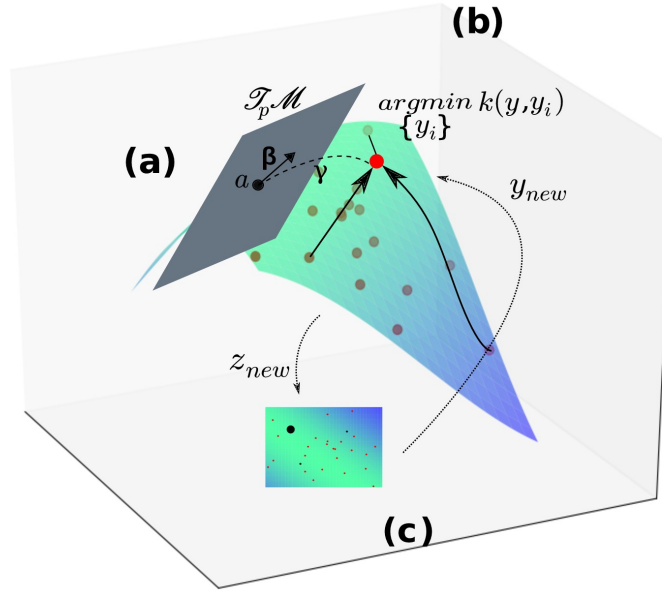


Figure 5.1: Various prediction methodologies on manifolds. (a) Intrinsic regression methods fit a geodesic curve $\gamma_{\mathcal{M}}$ on the manifold, parametrized by the point α and a velocity vector β , defined on α 's tangent space. (b) Kernel methods formulate prediction as a minimization problem over $\mathcal{M}$, which is usually simplified by solving over the set of training responses. (c) ML-based methods first make predictions on an embedding of $\mathcal{M}$, which then need to be mapped back to the original space.

prediction of diffusion tensors in brain MRI images from clinical and diagnostic measurements, with the response manifold been the space of 3×3 positive-definite matrices [193, 151]. More recently, a framework for polynomial regression on Riemannian manifolds was presented in [77]. The authors proposed the use of Riemannian polynomials, constructed as generalizations of manifold geodesics, and a gradient descent algorithm, similar to the one presented in [55]. The included application again focused on the prediction of the corpus callosum shape from age.

A non-parametric intrinsic manifold regression method was proposed by Wang et al. [178, 177]. Specifically, the authors presented a non-parametric bayesian regression model for Riemannian manifold spaces using Brownian Motion (BM) priors. In the following we present the details of that method, which we later use for comparisons to our own methodology. Let X be an input variable that take values in $[0, 1]$, and Y the response variable in a manifold $\mathcal{M}$. We denote the underlying function $f_{\mathcal{M}} : [0, 1] \rightarrow \mathcal{M}$ that relates the two variables by determining the density function $p_{f_{\mathcal{M}}(X)}(Y)$ so that $Y|X \sim p_{f_{\mathcal{M}}(X)}(Y)$. Let X have a distribution with support in

$[0, 1]$ and density $p(X)$. Finally, assume N paired observations $\{(x_i, y_i) | x_i \in [0, 1], y_i \in \mathcal{M}, i = 1, \dots, N\}$ with a joint density $\prod_{i=1}^N p(x_i) p_{f_{\mathcal{M}}(x_i)}(y_i)$. In [178], it was shown that the density of the posterior distribution, $p(\hat{f}_{\mathcal{M}}(X) | \{(x_i, y_i)\})$, for an estimator $\hat{f}_{\mathcal{M}}$ of $f_{\mathcal{M}}$, given the observed data can be computed as:

$$p(\hat{f}_{\mathcal{M}}(X) | \{(x_i, y_i) | i = 1, \dots, N\}) \propto \int_{f \in [0,1] \times \mathcal{M}} \prod_{i=1}^N p(x_i) p_{f_{\mathcal{M}}(x_i)}(y_i) d\Pi(f_{\mathcal{M}}) . \quad (5.5)$$

The authors restricted the density function $p_{f_{\mathcal{M}}(X)}(Y)$ to be a heat/gaussian kernel on the manifold, similar in form to (3.10). Regarding the prior distribution $\Pi(f_{\mathcal{M}})$, they presented two alternative Brownian Motion (BM) priors, with heat kernels expressing the Brownian transition probabilities. In particular, the density of the Continuous BM (CBM) was defined as $p_{f_{\mathcal{M}}}^C(x) \propto \int g_{x_n - x_{n-1}}(y_n, y_{n-1}) \dots g_{x_1 - x}(y_1, y_d) \mu x_n \dots d\mu(x_1)$, with $g_{x_i - x_j}$ a heat kernel with bandwidth $(x_i - x_j)$ and μ a probability measure on $\mathcal{M}$, while the density of the Discrete MB (DBM) was $p_{f_{\mathcal{M}}}^D(x) \propto \prod_{k=1}^h g_h(f(kh - h), f(kh))$, with h a bandwidth parameter chosen in the interval $(0, 1]$, such that $\frac{1}{h}$ is an integer, and $k \in \mathbb{N}$ a free parameter. Maximum a posteriori estimators were obtained through a simulated annealing (SA) algorithm on the corresponding posterior distributions. It is important to notice that the random moves in SA necessitate the analytical definition of exponential maps on $\mathcal{M}$. The methodology was demonstrated on the problem of regression from the interval $[0, 1]$ to the circle manifold $\mathcal{S}^1$.

Kernel-based Methods

While kernel methods have not originated from, and are not uniquely associated with, manifold regression, they are well-suited to the problem at hand. Kernel methods first appeared in the literature of machine learning as regression and classification methodologies tailored for complex input objects, such as trees or graphs [148]. The basic idea behind these methods is that, if the input data lie on a non-linear space, then they can be mapped on a very high (or infinite) dimensional inner-product space, usually called feature space, in which standard regression methodologies can be applied [148]. This mapping is implicit, meaning that data representations on the feature space are not explicitly computed. Rather, the geometry of the

initial space is "transferred" to the feature space by the use of a suitable kernel function that defines the objects' inner-product in that ambient feature space. This indirect mapping of data to an inner-product space, through the use of suitable kernel functions constitutes the so-called kernel trick [148]. The usefulness of the kernel trick becomes clear when considering that model fitting for a number of well known methodologies, such as Ridge Regression and Support Vector Machines (SVM), is easily formulated to depend only on inner products between the training inputs [148].

Following the construction of kernel-based methods for complex inputs, a number of methodologies were proposed to extend the kernel trick on the space of responses [169, 147, 85, 61, 8]. A prominent example of such a method was presented in [169], where the authors proposed an SVM algorithm that used joint input-response kernels in order to analyze tree-structured data. Prediction in kernel-based regression methods can be described as a form of kernel minimization problem over the response space (see Fig. 5.1(b)), but this minimization is rarely solved over the complete response space, but rather over the discrete set of training responses. Recently, Lin et al. [103], as well as Banerjee et al. [10, 9], employed kernel methodologies for manifold regression, with illustrative examples using responses on the spaces of a 2-sphere, as well as the Kendall space of planar shapes (represented as ordered collections of points). Analytical formulas for prediction were constructed in these cases, which though, are specific for these manifolds and do not constitute a general predictive methodology.

A kernel modification of the RF algorithm, Kernel Random Forest (kRF) [60], is of particular interest in our work. kRF utilizes a modified split objective function during the training stage, that necessitates only the computation of pairwise kernel values between the training responses

$$\begin{aligned}
 G_{kRF}(S_w) = & \sum_{i=1}^{N_w} g_{kRF}(y_i, y_i) - \frac{1}{N_w} \sum_{i=1}^{N_w} \sum_{j=1}^{N_w} g_{kRF}(y_i, y_j) \\
 & - \sum_{i=1}^{N_{wl}} g_{kRF}(y_i, y_i) + \frac{1}{N_{wl}} \sum_{i=1}^{N_{wl}} \sum_{j=1}^{N_{wl}} g_{kRF}(y_i, y_j) \\
 & - \sum_{i=1}^{N_{wr}} g_{kRF}(y_i, y_i) + \frac{1}{N_{wr}} \sum_{i=1}^{N_{wr}} \sum_{j=1}^{N_{wr}} g_{kRF}(y_i, y_j) ,
 \end{aligned} \tag{5.6}$$

where $g_{kRF}(y_i, y_j) = \exp\left(-\frac{\|y_i - y_j\|^2}{2\sigma^2}\right)$, N_w, N_{wl}, N_{wr} are cardinalities of the data subsets S_w, S_{wl}, S_{wr} in node w and its left and right children nodes respectively. In that respect, the methodology is similar to our training algorithm, as will be detailed below. However, the problem of predicting new responses was not tackled sufficiently in that work, since predictions were drawn just from the set of training responses. In section 5.5, we present simulation experiments to compare our predictive methodology against kRF, as well as other regression methods.

ML-based Methods

Under the term ML-based Methods, we encompass a family of methods that first apply Manifold Learning on the initial response data, and subsequently use standard regression models trained on the learned response embedding. If prediction on the original response space is required, then a suitable methodology must be formulated to map predicted points from the embedding to the manifold, a process that is usually referred to as backscoring (see Fig 5.1(c)) [48].

5.1.2 Motivation

Intrinsic manifold regression methods provide a natural framework for the analysis and prediction of manifold-valued responses. The assumption of an underlying manifold drives the choice of geometric elements employed during model definition and parameter estimation. A prerequisite in this case is that response objects have to be represented as points on analytically defined manifolds, with examples being 2D shapes, represented as collections of points [156, 55, 77], and deformation tensors, represented as semi-positive definite 3×3 matrices [151, 193]. This requirement is rarely easily satisfied, either due to the nature of data, the inability to define the manifold, or even the difficult preprocessing of data to generate suitable representations. Furthermore, most models presented so far are generalizations of multiple linear or polynomial regression, and may not be ideal for regression analyses with highly dimensional input observations. Due to these reasons, the scope of intrinsic models can be limited.

A non-parametric intrinsic regression method was presented in [178, 177]. A disadvantage of that methodology lies in the necessity for constructing of a prior distribution over the space of functions from the input space to the manifold. The authors proposed two BM priors, a continuous one, drawing upon previous work on measures on Riemannian manifolds [11], as well as a discretized prior, constructed by mapping subintervals of $[0, 1]$ to geodesic curves on the manifold. In the first case, the theoretical background does not account for higher dimensionality of the input space, while in the second case, the complexity of the prior's construction grows exponentially with the number of input dimensions, and thus the methodology can quickly become computationally prohibitive. In addition, the requirement for an analytical definition of the manifold is still applicable here, again constraining the application of this method to cases of well-described manifold spaces.

Kernel-based methods are the most extensively used techniques that deal with complex data objects [60, 70, 86, 169, 182, 103, 9], nevertheless they also suffer from a number of known issues. First, depending on the data at hand, a meaningful kernel function must be found or constructed. While a rich literature on kernel functions for various data types exists [148], it is by no means exhaustive. Furthermore, kernel construction is not always intuitive and remains a largely subjective process. Second, a common problem with kernel methods is the formulation of a prediction methodology. Usually prediction is described as a kernel minimization problem over the response space, which can be solved only by reducing the search space to the objects in the training dataset. While this approach may be sufficient in some cases, it evidently lacks precision.

ML-based methodologies constitute a convenient alternative over both intrinsic and kernel-based regression models, since they decouple the process of model definition from the shape of the response space. For any specific application, appropriate selection of the ML technique is largely based on intuition and previous experience, and a decrease in model fitting accuracy, compared, for example, with an equivalent intrinsic manifold model, is to be expected, since the response embedding is not guaranteed to completely capture the structure of the underlying manifold. ML-based methods make predictions on the embedding space, which then must be mapped back to the response space. Unfortunately, ML techniques generate discrete (defined

on the training responses) and one way maps from the manifold to the embedding, and the existence of an inverse and continuous map cannot always be guaranteed [48].

In this Chapter we present a new methodology for predictive modeling of manifold-valued responses, which addresses a number of key issues listed above. Our objective was to propose a unified framework for regression and prediction of complex objects that is accurate, computationally efficient and can be readily deployed in a wide variety of applications. In the training phase, We employ a modified RF regression algorithm that can be trained using only pairwise distances between response observations. The non-parametric RF methodology is efficient and can handle high dimensional input spaces. No analytical definition of the underlying response manifold is required, apart from the construction of an appropriate distance function for the responses, a requirement akin to finding a suitable inner product kernel. Comparing our distance-based approach to kernel methods, we identify a number of advantages that stem from the use of distances. First, a vast library of readily available distance metrics exists in the literature (see for example [45]). Second, in cases where the manifold metric is not known, geodesic distances can be approximated, using for example the Isomap distance metric (see subsection 2.4.2). Third, it allows the formulation of a standardized prediction methodology, as outlined below.

In the prediction phase, a new response point estimate is found as follows. Pairwise distances between the unseen response and all training observations are predicted. Using this set of distances, the response is predicted on a euclidean embedding computed using metric MDS, and finally is backscored to the response manifold. Our prediction methodology follows a two-step approach akin to ML-based methods, with two prominent advantages, compared to these techniques. First, our regression model is fitted on the manifold distances, which enhances the quality of the fitted model. Second, due to the fact that geodesic distances are known, metric MDS is employed for ML and predictions can be analytically computed and backscored to the response manifold.

5.2 Problem Definition

Consider a dataset S of N observed input-response points $\{(x_i, y_i) | i = 1, \dots, N\}$, with input $x_i = (x_{i1}, \dots, x_{ip}) \in \mathbb{R}^p$ corresponding to response $y_i \in \mathcal{M} \subset \mathbb{R}^q$. $\mathcal{M}$ is a possibly unknown manifold, equipped with a distance metric $d_{\mathcal{M}}(\cdot, \cdot)$ (see definition 2.3 and lemma 2.1). We will refer to $\mathbb{R}^q$ as the response representation space. Also, let $\mathbf{D}_{\mathcal{M}}$ be the $N \times N$ matrix of pairwise distances between the observed response points, with elements $D_{\mathcal{M},ij} = d_{\mathcal{M}}(y_i, y_j)$, for $i, j = 1, \dots, N$. The problem consists of defining a suitable predictive regression methodology that utilizes information about the structure of the manifold during model fitting, and, given a new input point x_{new} , is capable of predicting a response point estimate $\hat{y}_{new}$ of $y_{new} \in \mathcal{M}$.

5.3 Distance Random Forest Regression

In the standard RF regression methodology (see Section 2.3), the objective function employed to split training responses into different tree nodes is based on the maximization of the difference between a parent node's response Mean Squared Deviation (MSD) and the sum of response MSDs for samples in its children nodes, as given in (2.4). This function though is ill-suited for manifold-valued responses, since it assumes the existence of a vector representation and a Euclidean norm for the responses. To solve this issue, we make use of the modified distance Random Forest (dRF) algorithm, which we introduced in [153]. In dRF, the split criterion is formulated to depend only on pairwise distances between responses. The modification is based on the fact that each summation in (2.4) can be rewritten in terms of pairwise distances between the observations' responses.

Proposition 5.1. *Let $\{y_i | y_i \in \mathbb{R}^q, i = 1, \dots, N\}$ be observed points on a q -dimensional Euclidean space. The sum of squared deviations from the mean $\bar{y} = \frac{1}{N} \sum_{i=1}^N y_i$ can be written in terms of pairwise distances between points as follows:*

$$\sum_{i=1}^N \|y_i - \bar{y}\|^2 = \frac{1}{2N} \sum_{i=1}^N \sum_{j=1}^N d_E^2(y_i, y_j) , \quad (5.7)$$

where $d_E(\cdot, \cdot)$ denotes the Euclidean distance function.

The proof of proposition 5.1 can be found in Appendix C.1. The key advantage from 5.1 is that the computation of MSDs can be decoupled from the use of Euclidean norms and means, and depend only on pairwise distances. For manifold-value responses, We define the split objective function for dRF as follows:

$$G_{dRF}(S_w) = \frac{1}{2N_w} \sum_{y_i \in S_w} \sum_{y_j \in S_w} d_{\mathcal{M}}^2(y_i, y_j) - \frac{1}{2N_{wl}} \sum_{y_i \in S_{wl}} \sum_{y_j \in S_{wl}} d_{\mathcal{M}}^2(y_i, y_j) - \frac{1}{2N_{wr}} \sum_{y_i \in S_{wr}} \sum_{y_j \in S_{wr}} d_{\mathcal{M}}^2(y_i, y_j) , \quad (5.8)$$

with $y_i \in \mathcal{M}$ and N_w, N_{wl}, N_{wr} cardinalities of the data subsets S_w, S_{wl}, S_{wr} in node w and its left and right children nodes respectively.

The objective (5.8) can be understood as a generalization of the cost function G_{RF} (2.4), enabling the RF algorithm to be applied in more general metric spaces. In previous work, we used the dRF methodology on regression applications with graph and covariance-based response objects [153]. dRF lends itself naturally to manifold-valued data, whether the manifold in question is analytically defined or implied by the use of a specific distance metric. A schematic illustration of the dRF regression algorithm can be seen in Fig. 5.2.

5.4 Predictive Methodology

Having presented a suitable regression method for manifold-valued responses, we turn our attention to the construction of a prediction methodology. As described in section 2.3, prediction in RF models is traditionally performed by the averaging formula (2.6). This approach though is not valid under the assumption of non-vector manifold response spaces [125].

We propose a prediction methodology, which can be summarized as follows. For a new input point x_{new} , we use the trained RF model to predict all pairwise distances between the unknown y_{new} and each training response y_i . We show that by using this set of distances, it is then

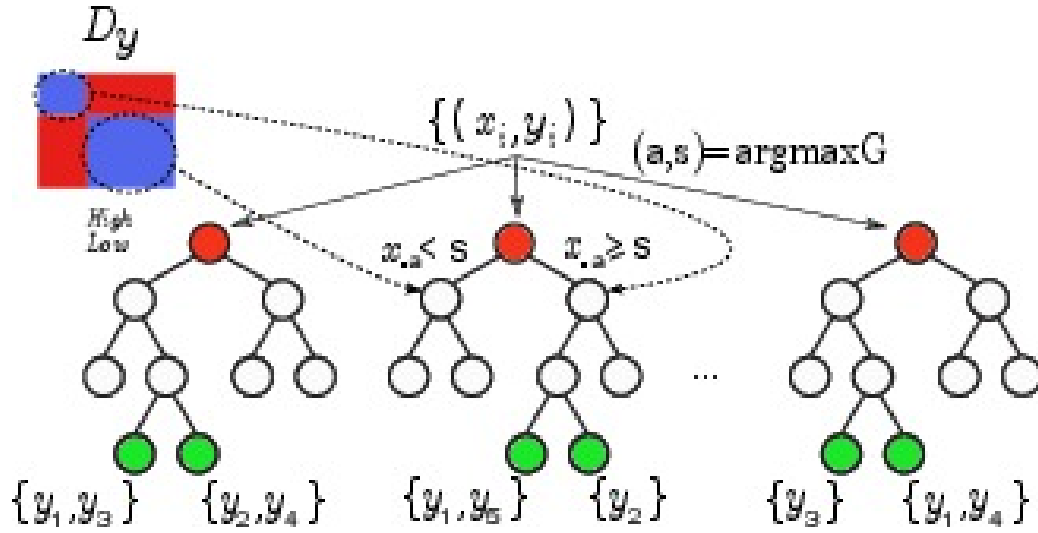


Figure 5.2: Distance Random Forest training methodology. The data are recursively divided according to selected input variables and split values for each node. The cost function G_{dRF} drives this selection. In dRF, the aim is to create subgroups within which the points are close together, according to the predefined manifold's distance metric.

possible to predict the response on an Euclidean embedding of the manifold, learned from the training response set using metric MDS. Subsequently, a backscoring method is employed to project the point back to the original response manifold $\mathcal{M}$. Figure 5.3 provides an illustration of the proposed prediction methodology.

5.4.1 RF-based Distance Prediction

At the first stage of our prediction method, the objective is to predict the set of distances $\{\hat{d}_{\mathcal{M}}(y_{new}, y_i) | i = 1, \dots, N\}$. To achieve this, we exploit the inherent ability of the dRF model to provide a measure of similarity between pairs of response observations, hereafter referred to as dRF-based similarity. When the new input point x_{new} is 'dropped' through each tree in the forest, it reaches a particular leaf associated with a subset of the tree's training observations. The unknown response y_{new} can be considered similar to that leaf's training responses (see Fig. 5.3(b)).

In order to compute the dRF-based similarities between the new point and each observation in the training set, a similarity vector $\mathbf{a}_{RF} = (a_{RF,1}, \dots, a_{RF,N})$ is used, with each element $a_{RF,i}$ corresponding to the dRF-based similarity between y_{new} and y_i . $\mathbf{a}_{RF}$ is constructed as follows.

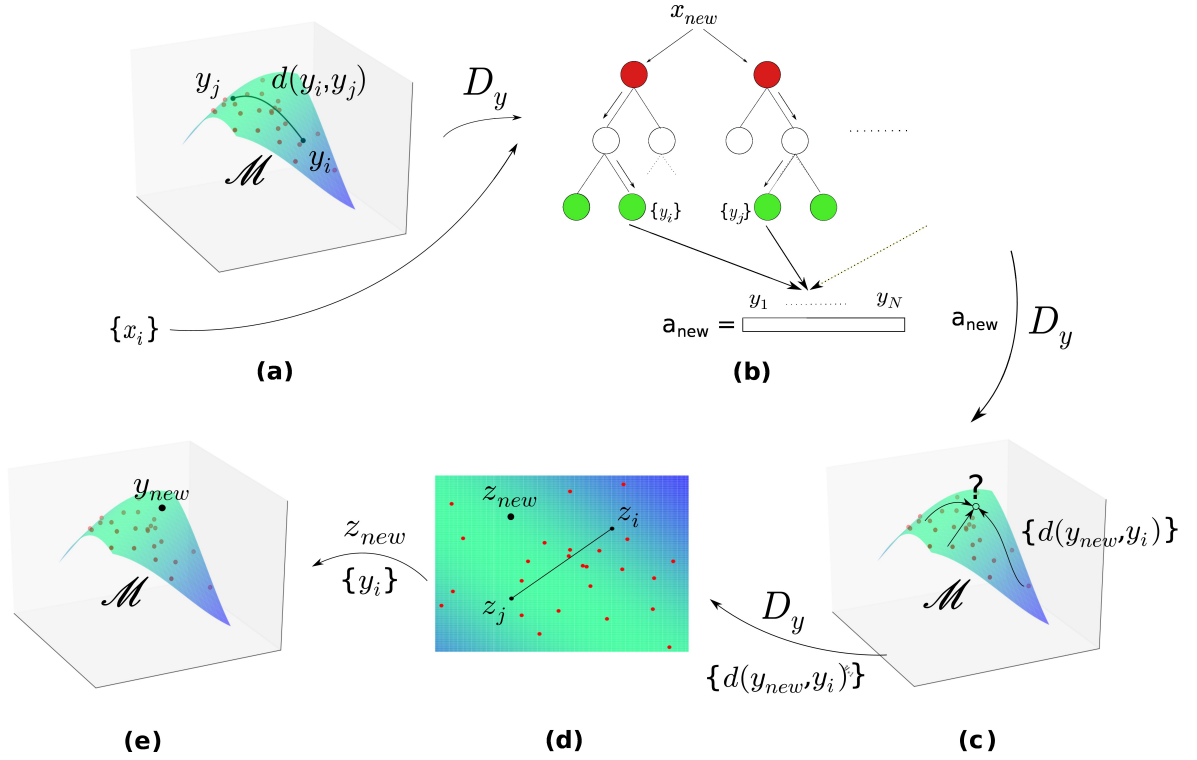


Figure 5.3: Illustration of the dRF prediction methodology. (a) The dRF model is fitted on the dataset S of N observed input-response points, with $y_i \in \mathcal{M}$, and the distance matrix $\mathbf{D}_{\mathcal{M}}$ of pairwise manifold distances between the observed response points. (b) When a new input y_{new} is observed, the point is passed through the forest and is associated with subsets of training response points in the trees' leaves. A similarity vector $\mathbf{a}_{RF}$ between the training responses and the still unknown response point is extracted from the model. (c) Based on $\mathbf{a}_{RF}$, the set of distances between the new point and the training responses on the manifold are predicted. (d) Knowledge of these distances allows prediction of the response value on a Euclidean embedding of the manifold, extracted through metric Multi-Dimensional Scaling. (e) Finally, a kernel-based backscoring method is used to project the predicted point back to the original manifold and acquire $\hat{y}_{new}$.

Initially $a_{RF,i}$ is set to zero, for all $i = 1, \dots, N$. For each tree in the forest, $a_{RF,i}$ is incremented by one, each time x_{new} ends in the same node as x_i . Normalization of the similarity values is performed by division with the total number of trees.

Based on the similarity vector $\mathbf{a}_{RF}$ and the training distance matrix $\mathbf{D}_{\mathcal{M}}$, the response distances are predicted using the algorithmic procedure 1 (Fig. 5.3(c)). In detail, the algorithm initially identifies the closest training object to the new response, according to $\mathbf{a}_{RF}$, and assigns as distance between these two responses the minimum distance observed in the training set. By the construction of the algorithm, this is also the minimum value that can be assigned to any distance hereafter. Subsequently, it iteratively identifies the furthest response from the closest

observation and assigns a value for the distance between that and the new point as follows: For each triplet including the considered point, the new point and a point for which the distance to y_{new} has been already estimated, it keeps the maximum of the two known distances. Finally, it assigns as the new distance the minimum of all identified maximums. The algorithm guarantees that the new point will lie in close proximity to at least its closest neighbor based on the dRF affinities, and that predicted distance values respect the triangular inequality property of metric $d_{\mathcal{M}}$.

Algorithm 1 Prediction of $\{\hat{d}_{\mathcal{M}}(y_{new}, y_i) | i = 1, \dots, N\}$ using the dRF similarity vector $\mathbf{a}_{RF}$ and the training distance matrix $\mathbf{D}_{\mathcal{M}}$.

Require: $\mathbf{a}_{RF}$ and $\{D_{\mathcal{M},ij} | i, j = 1, \dots, N\}$
 $U = \{1, \dots, N\}, \quad Q = \emptyset$
 $l \leftarrow \underset{i \in U}{\operatorname{argmax}} \quad a_{RF,i}$
 $\hat{d}_{\mathcal{M}}(y_{new}, y_l) \leftarrow \min\{D_{\mathcal{M},ij} | i, j = 1, \dots, N\}$
 $U \leftarrow U \setminus \{l\}, \quad Q \leftarrow Q \cup \{l\}$
repeat
 $p \leftarrow \underset{i \in U}{\operatorname{argmax}} \quad D_{\mathcal{M},il}$
 $\hat{d}_{\mathcal{M}}(y_{new}, y_p) \leftarrow \min\{\max\{\hat{d}_{\mathcal{M}}(y_{new}, y_i), D_{\mathcal{M},ip}\} | i \in Q\}$
 $U \leftarrow U \setminus \{p\}, \quad Q \leftarrow Q \cup \{p\}$
until $U = \emptyset$
return $\{\hat{d}_{\mathcal{M}}(y_{new}, y_i) | i = 1, \dots, N\}$

5.4.2 Prediction on the Response Embedding

We show now that, given the distances $\{\hat{d}_{\mathcal{M}}(y_{new}, y_i) | i = 1, \dots, N\}$, it is possible to predict a point estimate of the response on a Euclidean embedding of the manifold, learned through Multi-Dimensional Scaling (MDS) [39] (see Fig. 5.3(d)).

Let $\{z_i \in \mathbb{R}^m, i = 1, \dots, N\}$ be m -dimensional Euclidean representations of the original training response points $\{y_i | i = 1, \dots, N\}$, learned through metric MDS, as described in Section 2.4.1. The result of the MDS decomposition is a discrete and one-way mapping $z_i = f_{MDS}(y_i), z_i \in \mathbb{R}^m, y_i \in \mathcal{M}$, defined on the training data set (see (2.11)).

A number of Out-Of-Sample (OOS) methods exist which allow mapping of a new manifold observation on the learned space of the embedding. In [15], the authors constructed an OOS

method for MDS, and showed that the precision of the predictions was comparable to intrinsic perturbations of the embedding algorithm.

Let $Y_a, Y_b \in \mathcal{M}$ be random variables defined on the manifold surface, with the same probability density function, and $y_a, y_b \in \mathcal{M}$ two realizations of Y_a and Y_b respectively. Then a kernel function g_{MDS} can be defined as follows:

$$g_{MDS}(y_a, y_b) = -\frac{1}{2}(d^2(y_a, y_b) - E[d^2(Y_a, y_a)] - E[d^2(Y_b, y_b)] + E[d^2(Y_a, Y_b)]) , \quad (5.9)$$

where $E(\cdot)$ denotes expectation. It is easy to verify that under observed data $\{y_i | i = 1, \dots, N\}$, if expectations in equation (5.9) are substituted by the respective sample means, then the estimator of g_{MDS} generates the MDS decomposition matrix $\mathbf{K}_{\mathcal{M}}$ as given in (2.10).

Reminding that λ_k is the k^{th} eigenvalue and u_{ik}^λ the i^{th} element of the k^{th} eigenvector of matrix $\mathbf{K}_{\mathcal{M}}$, the coordinates of an OOS prediction $\hat{z}_{new} = (\hat{z}_{new,1}, \dots, \hat{z}_{new,m})$ are given by [15]:

$$\hat{z}_{new,k} = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^N u_{ik}^\lambda \hat{g}_{MDS}(y_{new}, y_i), \quad k = 1, \dots, m , \quad (5.10)$$

with $\hat{g}_{MDS}$ denoting the mean estimator of g_{MDS} under the augmented dataset $\{y_i | i = 1, \dots, N\} \cup \{y_{new}\}$:

$$\begin{aligned} \hat{g}_{MDS}(y_{new}, y_i) = & -0.5d_{\mathcal{M}}^2(y_{new}, y_i) \\ & + \frac{0.5}{N+1} \left(\sum_{j=1}^N d_{\mathcal{M}}^2(y_j, y_{new}) + d_{\mathcal{M}}^2(y_{new}, y_{new}) \right) \\ & + \frac{0.5}{N+1} \left(\left(\sum_{j=1}^N d_{\mathcal{M}}^2(y_j, y_i) \right) + d_{\mathcal{M}}^2(y_{new}, y_i) \right) \\ & - \frac{0.5}{(N+1)^2} \left(\left(\sum_{j,l=1}^N d_{\mathcal{M}}^2(y_j, y_l) \right) + 2 \sum_{j=1}^N d_{\mathcal{M}}^2(y_{new}, y_j) \right) . \end{aligned} \quad (5.11)$$

Details regarding the OOS formulation (5.10) are given in the Appendix C.2. We observe that the OOS projection does not depend on the actual value of y_{new} , but rather on the distances between the new response and points on the training dataset. In our methodology,

y_{new} is unknown, but since the set of estimated distances $\{\hat{d}_{\mathcal{M}}(y_{new}, y_i) | i = 1, \dots, N, \}$ is known, the formula (5.10) can be employed to predict a point estimate of our response $\hat{z}_{new}$ on the embedding space $\mathbb{R}^m$.

5.4.3 Projection from the Embedding to the Manifold

Finally, we deal with the issue of projecting, or backscoring, the predicted point back to the manifold (see figure 5.3(e)). We formulate the problem of backscoring to the manifold as an interpolation problem in the representation space $\mathbb{R}^q$. Specifically, we are looking for a smooth continuous function $f_{bs} : \mathbb{R}^m \rightarrow \mathbb{R}^q$ from the space of the embedding to $\mathbb{R}^q$, such that $y_i \approx f_{bs}(z_i)$ for $i = 1, \dots, N$:

$$\operatorname{argmin}_{f_{bs} \in \mathcal{F}_{bs}} \gamma_{bs} \sum_{i=1}^N (f_{bs}(z_i) - y_i)^2 + \|f_{bs}\|_{\mathcal{F}_{bs}}^2, \quad (5.12)$$

where γ_{bs} is a weight parameter balancing the smoothness of the interpolation and the adherence to the data and $\mathcal{F}_{bs}$ is a space of smooth functions equipped with the norm $\|\cdot\|_{\mathcal{F}_{bs}}$, which will be defined below.

An optimal solution to the optimization problem (5.12) was constructed in [174]. Let $\mathcal{Z}_{bs}$ be a closed subset of $\mathbb{R}^m$. We define g_{bs} to be a kernel function of the form

$$g_{bs}(r, z_a, t, z_b), \quad r, t \in T^q = \{1, \dots, q\}, z_a, z_b \in \mathcal{Z}_{bs}. \quad (5.13)$$

Also, let g_{bs} be semi-positive definite on $(T^q \times \mathcal{Z}_{bs}) \times (T^q \times \mathcal{Z}_{bs})$, i.e

$$\sum_{k=1}^N \sum_{l=1}^N a_k a_l \sum_{r=1}^q \sum_{t=1}^q g_{bs}(r, z_k, t, z_l) \geq 0, \quad (5.14)$$

for any finite set of points $\{z_k | z_k \in \mathcal{Z}_{bs}, k = 1, \dots, N\}$ and real numbers $\{a_k | a_k \in \mathbb{R}, k = 1, \dots, N\}$.

For a fixed (r, z_a) , equation (5.13) defines a function from $\mathbb{R}^m$ to $\mathbb{R}^q$ by the formula

$$f_{rz_a}(z_b) = (g_{bs}(r, z_a, 1, z_b), \dots, g_{bs}(r, z_a, q, z_b))^T . \quad (5.15)$$

Based on the above, let $\mathcal{F}_{bs}$ be the space of all finite linear combinations of functions of the form (5.15), as (r, z_a) varies in $T^q \otimes \mathcal{Z}_{bs}$, and its closure w.r.t the following scalar inner product:

$$\langle f_{rz_a}, f_{tz_b} \rangle = g_{bs}(r, z_a, t, z_b) , \quad (5.16)$$

with $\|f_{rz_a}\|_{\mathcal{F}_{bs}} = \sqrt{\langle f_{rz_a}, f_{rz_a} \rangle}$.

Then we can formulate the minimizer of (5.12) as follows. Given our training responses $\{y_i | i = 1, \dots, N\}$ and their equivalent embedded points $\{z_i | i = 1, \dots, N\}$, let $\mathbf{Y}$ be the $N \times q$ matrix of training responses with rows $Y_i = y_i$, and $\mathbf{K}_{bs}$ the $Nq \times Nq$ matrix of kernel values $K_{bs, N(r-1)+i, N(t-1)+j} = g_{bs}(r, z_i, t, z_j)$. The minimizing function can be estimated as [174]:

$$\hat{f}_{bs}(\hat{z}_{new}) = \sum_{i=1}^N \sum_{r=1}^q C_{ir}^{bs} f_{rz_i}(\hat{z}_{new}) , \quad (5.17)$$

where C_{ir} are elements of the $N \times q$ coefficient matrix $\mathbf{C}^{bs}$ given by

$$vec(\mathbf{C}^{bs}) = (\mathbf{K}_{bs} + \frac{N}{\gamma_{bs}} \mathbf{I})^{-1} vec(\mathbf{Y}) , \quad (5.18)$$

with $vec(\cdot)$ denoting the vectorization of a matrix into a column vector. From equations (5.17) and (5.18), we observe that the value of the function $\hat{g}$ at point $\hat{z}_{new}$ requires just the knowledge of pairwise kernel values between the $N+1$ points $z_1, \dots, z_N, \hat{z}_{new}$. As such, the point estimate of the response is taken to be $\hat{y}_{new} = \hat{f}_{bs}(\hat{z}_{new})$.

The computational complexity of inverting the matrix $\mathbf{K}_{bs} + \frac{N}{\gamma_{bs}} \mathbf{I}$ is $\mathcal{O}((Nq)^{2.373})$. While this is high, it has to be performed once, since the matrix depends only on the pairwise kernel values between the training points and the smoothing parameter γ_g . Nevertheless, in our studies, we opted to simplify the minimizer (5.17), in favor of having one tuning parameter, as well as

improving the efficiency of the backscoring algorithm. We chose

$$g_{bs}(r, z_a, t, z_b) = \begin{cases} \exp\left(-\frac{\|z_a - z_b\|^2}{\sigma_{bs}}\right), & r, t = 1, \dots, q, \quad r = t \\ 0, & r \neq t \end{cases}, \quad (5.19)$$

where σ_{bs} is the single free parameter, adjusting the bandwidth of the kernels. This simplification has an additional benefit in the computational efficiency of the algorithm. $\mathbf{K}_{bs} + \frac{N}{\gamma_{bs}} \mathbf{I}$ becomes a diagonal block matrix, with all diagonal elements consisting of the same $N \times N$ gram matrix. As a result, the complexity of the matrix inversion that is necessary for the backscoring methodology is reduced to $\mathcal{O}(N^{2.373})$.

We notice that the backscoring formulation does not take into explicit consideration the structure of the response manifold and $\hat{g}(\hat{z}_{new})$ is not guaranteed to lie exactly on $\mathcal{M}$. Nevertheless, we justify our choice by pointing that since $\hat{g}$ is a smooth interpolating function from the embedding to the training responses, predictions close to the training objects should also comply fairly well to the manifold structure.

5.5 Simulation Studies

In this section we present a series of simulation studies to assess the performance of our predictive methodology. Comparisons were performed against Nearest Neighbor (NN) and regular RF models, which do not take into consideration the special form of the response space, kRF [60], which employs a Radial Basis Function (RBF) kernel to capture the structure of the response space during training, as well as the non-parametric intrinsic regression methodology by Wang et al. [178]. Three sets of simulation studies were considered. First, we performed regression on a one-dimensional circular manifold. Second, we carried out simulations on a spherical manifold. In both these cases, geodesic distances between the response observations can be analytically computed. Finally, we performed a regression analysis on a swiss-roll manifold, for which no analytical formula for computing the pairwise distances between the response points exists. Distances were thus approximated using the Isomap metric (see section 2.4.2).

5.5.1 Prediction on a Circular Manifold

In the first experiment, the goal was to compare our methodology against the CBM/DBM intrinsic manifold regression method presented in [178] (see 5.1.1 for details on that method). We adopted the same simulation framework from Wang et al. [178]. In particular, we simulated two training datasets of $N = 100$ paired input-response points $\{(x_i, y_i) | i = 1, \dots, N\}$, with inputs $x_i \in [0, 1]$ corresponding to responses $y_i = (y_{i1}, y_{i2}) \in \mathcal{S}^1 \subset \mathbb{R}^2$ on the circumference of a circle, a 1-manifold embedded in $\mathbb{R}^2$. In more detail, our first dataset was constructed as follows: 100 input observations were sampled from the uniform distribution $\mathcal{U}(0, 1)$. These were mapped on the interval $[-\pi, \pi]$ using the function

$$\theta_i = f_{\mathcal{S}^1}^a(x_i) = (x_i + 0.5)^2 . \quad (5.20)$$

Gaussian noise with $\sigma = 0.1$ was subsequently added on θ_i . Finally, we took $y_{i1} = \cos(\theta_i)$, $y_{i2} = \sin(\theta_i)$. The second dataset was constructed similarly, but this time using

$$\theta_i = f_{\mathcal{S}^1}^b(x_i) = \sin(6x_i) . \quad (5.21)$$

Regarding the intrinsic methods of CBM and DBM, regression was performed using the set $\{\theta_i | i = 1, \dots, N\}$ as responses, since they represent a 1-dimensional angular parametrization of $\mathcal{S}^1$. The distance metric, exponential and logarithmic maps on the manifold can be analytically computed as:

$$d_{\mathcal{S}^1}(\theta_i, \theta_j) = |\theta_i - \theta_j|, \quad \exp_{\theta_i}(\mathbf{u}_{\mathcal{T}}) = \theta_i + \mathbf{u}_{\mathcal{T}}, \quad \log_{\theta_i}(\theta_j) = \theta_j - \theta_i , \quad (5.22)$$

where $\mathbf{u}_{\mathcal{T}} \in \mathcal{T}_{\theta_i} \mathcal{S}^1 \subset \mathbb{R}$. The standard deviation parameter for both priors was 0.1. In our methodology, the same distance function was used for training the dRF model and learning a 1-dimensional embedding of $\{y_i\} | i = 1, \dots, N\}$. The forest was built with 1000 trees and backscoring was performed using $\sigma_{bs} = 0.5$ and $\gamma_{bs} = 10$.

The testing datasets were constructed by taking 50 equidistant points in the interval $[0, 1]$,

which were considered the input test points, with angular responses computed using (5.20) and (5.21). Regarding the first simulated dataset, Fig. 5.4 (a) shows the graph of the training and test observations, based on the angular responses. Figures 5.4 (b), (c) and (d) present the estimated angular responses for the test inputs, based on the CBM and DBM intrinsic regression models, as well as our dRF predictive methodology. Since dRF casts predictions in the original representation space ($\mathbb{R}^2$), angular responses were acquired using $\theta_i = \arctan(\frac{y_{i2}}{y_{i1}})$. Mean prediction errors, computed based on the manifold's distance metric are also included in the figure. The lowest prediction error was acquired from the CBM model, but dRF was very close in terms of performance, and DBM was the worst model, due to oversmoothing of the estimated function. Similar results were also acquired for the second simulated dataset, as can be seen in Fig. 5.5.

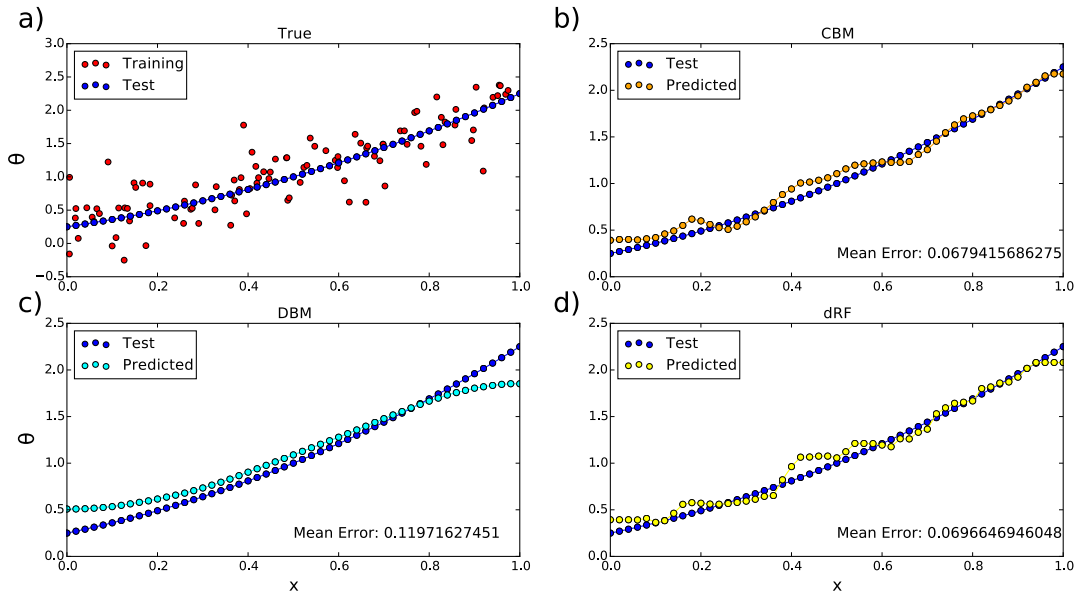


Figure 5.4: Scatterplots of simulated input and response points for the first regression study on a circular manifold, with $\theta_i = f_{\mathcal{S}^1}^a(x_i) = (x_i + 0.5)^2$. (a) Input vs. angular response train and test observations, (b) Input vs. test and predicted angular responses for the CBM model, (c) Input vs. test and predicted angular responses for the DBM model, (d) Input vs. test and predicted angular responses for the dRF model.

Figure 5.6 includes the scatterplots of training, test and predicted points of the second simulation experiment, in the $\mathbb{R}^2$ representation space, for all models. Since the DBM and CBM are intrinsic manifold regression models, working directly on the angular representation of $\mathcal{S}^1$ and utilizing the exponential and logarithmic maps of the manifold, there is no prediction error

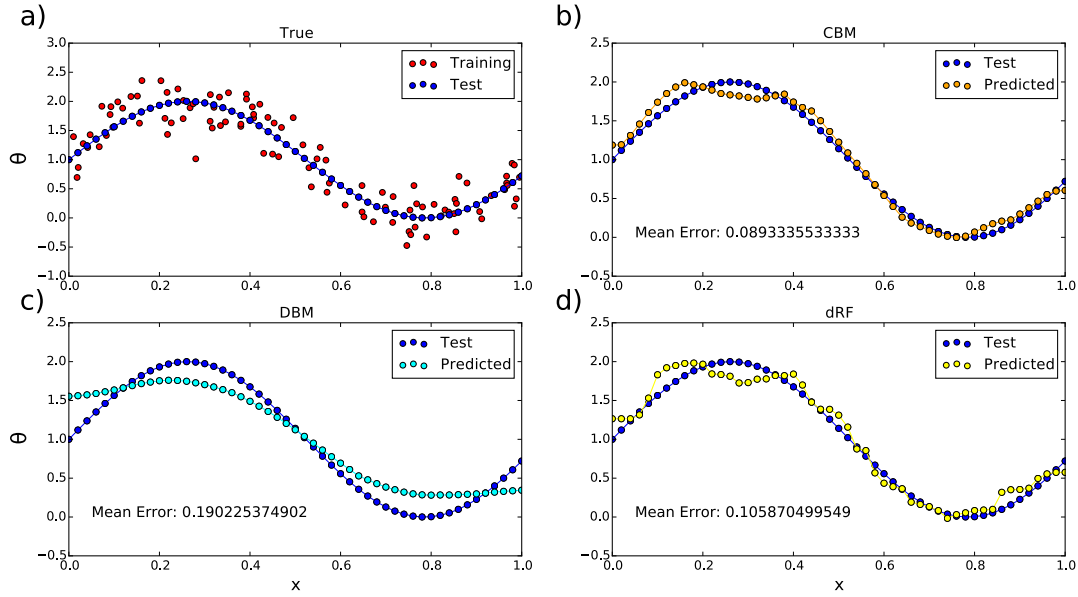


Figure 5.5: Scatterplots of simulated input and response points for the second regression study on a circular manifold, with $\theta_i = f_{\mathcal{S}^1}^b(x_i) = \sin(6x_i)$. (a) Input vs. angular response train and test observations, (b) Input vs. test and predicted angular responses for the CBM model, (c) Input vs. test and predicted angular responses for the DBM model, (d) Input vs. test and predicted angular responses for the dRF model.

perpendicular to the manifold's structure. dRF, on the other hand, does not utilize and is not dependent on the analytical definition of the manifold. As a consequence, predicted responses are not guaranteed to lie directly on the manifold. Nevertheless, it can be clearly noticed from Fig 5.6 (d) that dRF predictions respect very well the manifold constraints.

While in the particular examples CBM performs slightly better than our predictive methodology, it is important to highlight that the intrinsic regression methods are associated with significant limitations from which dRF does not suffer. In particular, CBM and DBM models can only be constructed in cases of well defined response manifolds, as is $\mathcal{S}^1$, that was used here. As was discussed in section 5.1.2, this requirement reduces their applicability significantly. Furthermore, both models necessitate univariate predictors. Finally, the continuous prior distribution used in CBM often does not have a density function [178]. On the other hand, our dRF predictive approach, whilst providing very similar predictive performance in a problem where both methods can be applied, is also free from the aforementioned limitations. This allows for its application on a much larger pool of problems, as will be shown in the consequent simulation and application studies.

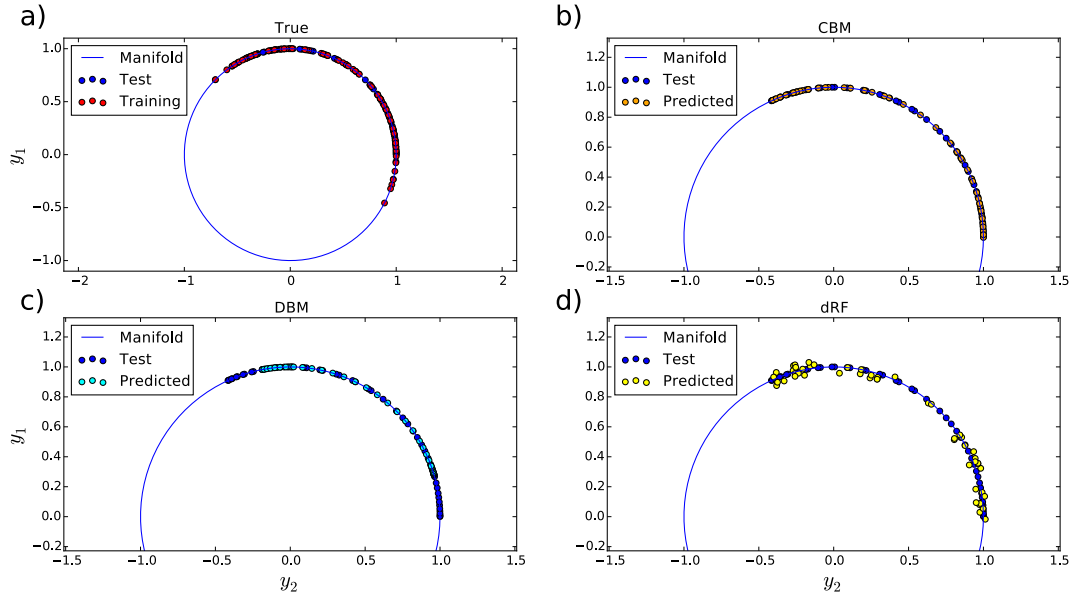


Figure 5.6: Scatterplots of simulated response test and predicted points in the $\mathbb{R}^2$ representation space for the second regression study on a circular manifold. (a) Train and test response observations, (b) Test and predicted response observations for the CBM model, (c) Input vs. test and predicted response observations for the DBM model, (d) Input vs. test and predicted response observations for the dRF model.

5.5.2 Prediction on a Spherical Manifold

In this study, we concentrate on the case of multivariate predictors. Let S denote our simulated dataset of $N = 250$ paired input-response points $\{(x_i, y_i) | i = 1, \dots, N\}$, with inputs $x_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}, x_{i6}) \in \mathbb{R}^6$ corresponding to responses $y_i = (y_{i1}, y_{i2}, y_{i3}) \in \mathcal{S}^2 \subset \mathbb{R}^3$. Input coordinate variables were grouped in pairs $(x_{.1}, x_{.2})$, $(x_{.3}, x_{.4})$ and $(x_{.5}, x_{.6})$ and, for each pair, N observations were drawn uniformly from the unit disk. For completeness, table 5.1 shows the sample covariance of the coordinate variables from the complete simulated dataset.

	$x_{.1}$	$x_{.2}$	$x_{.3}$	$x_{.4}$	$x_{.5}$	$x_{.6}$
$x_{.1}$	0.262	-0.016	-0.009	-0.022	-0.004	-0.024
$x_{.2}$	-0.016	0.249	0.018	-0.008	0.015	-0.01
$x_{.3}$	-0.009	0.018	0.234	-0.006	0.002	0.015
$x_{.4}$	-0.022	-0.008	-0.006	0.259	0.02	0.005
$x_{.5}$	-0.004	0.015	0.002	0.02	0.239	0.016
$x_{.6}$	-0.024	-0.01	0.015	0.005	0.016	0.258

Table 5.1: Covariance matrix of the 6 coordinate predictor variables in the dataset S used during the spherical manifold regression study.

Only the first two input dimensions were constructed to be predictive of the response. Figure

5.7(a) depicts the scatterplot of $\{(x_{i1}, x_{i2}) | i = 1, \dots, 250\}$. In order to generate the response points, x_{i1} and x_{i2} were initially transformed to lie on the unit square (see Fig. 5.7(b)):

$$\begin{aligned} x'_{i1} &= \frac{\sqrt{x_{i1}^2 + x_{i2}^2}}{\max(\|x_{i1}\|, \|x_{i2}\|)} x_{i1}, \\ x'_{i2} &= \frac{\sqrt{x_{i1}^2 + x_{i2}^2}}{\max(\|x_{i1}\|, \|x_{i2}\|)} x_{i2} . \end{aligned} \quad (5.23)$$

An affine transformation (scaling $\rightarrow$ rotation $\rightarrow$ translation) was then applied on (x'_{i1}, x'_{i2}) (Fig. 5.7c):

$$\begin{aligned} x''_{i1} &= \cos(\theta) s_1 x'_{i1} + \sin(\theta) s_2 x'_{i2} + t_1, \\ x''_{i2} &= -\sin(\theta) s_1 x'_{i1} + \cos(\theta) s_2 x'_{i2} + t_2 , \end{aligned} \quad (5.24)$$

where s_1, s_2 are scaling factors, θ the rotation angle and t_1, t_2 the translation factors. In particular: $s_1 = 2, s_2 = 1, \theta = \frac{\pi}{3}, t_1 = 0, t_2 = 0.65$.

Finally, (x''_{i1}, x''_{i2}) were projected on the 2-sphere with $r = 3$ to acquire the responses (see Fig. 5.7(d)):

$$\begin{aligned} y_{i1} &= x''_{i1}, \\ y_{i2} &= x''_{i2}, \\ y_{i3} &= \sqrt{r^2 - x''_{i1}^2 - x''_{i2}^2} . \end{aligned} \quad (5.25)$$

Various types of Gaussian noise were added on the response points. Let $y'_i = y_i + n_i$, where the noise $n_i = n_i^{per} + n_i^{man}$ could be decomposed to two components. n_i^{per} was taken to be always perpendicular to the spherical manifold, and n_i^{man} always parallel to the manifold. The study was performed using four different noise settings: No noise, noise only perpendicular to the manifold, noise only on manifold, combined noise including both perpendicular and on manifold components.

Furthermore, for each of the noise settings, four different noise levels were used. The level of noise was controlled by the standard deviation of the Gaussian distribution employed to draw

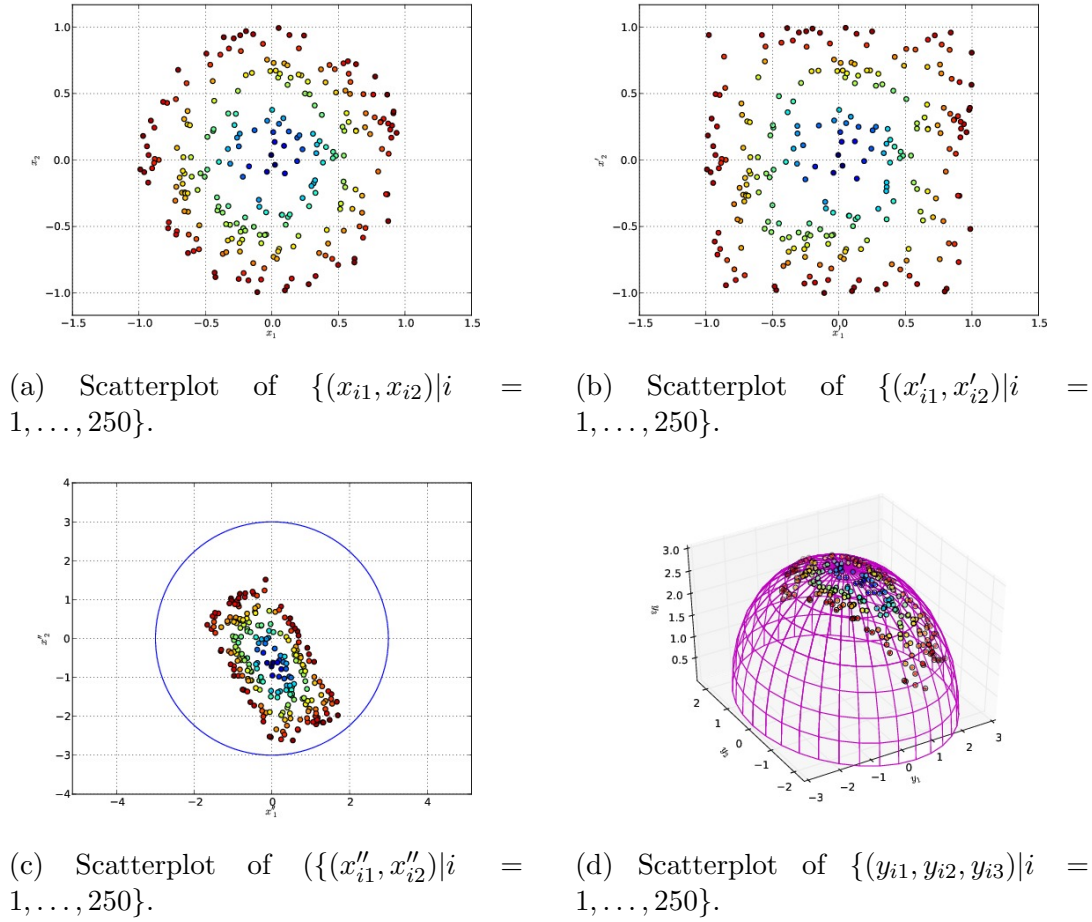


Figure 5.7: Scatterplots of the simulated input and response points for the regression study on a spherical manifold. Uniformly sampled points from the unit circle (a) were mapped to the unit square (b), scaled, rotated, translated (c) and finally projected on the upper hemisphere of radius $r = 3$ (d), to form the response observations. All plots use the same colormap of the distance between points (x_{i1}, x_{i2}) and the center $(0, 0)$.

the noise observations with four levels corresponded to 0.05, 0.1, 0.15, 0.2 standard deviation respectively. For example, Fig. 5.8 shows the response observations of S with combined noise, for $\sigma = 0.2$. Overall, 13 combinations of noise-setting and noise-level were considered.

The simulated dataset S was split into S_{train} , comprising of 200 input-response pairs and S_{test} , consisting of the remaining 50 observations. 5 different regression methodologies were compared during this study, namely 1-NN Regression, k -NN regression, regular RF, kRF and dRF. For dRF, the Great Circle Distance (GCD) [152] was used to compute the training response distance matrix $\mathbf{D}$. GCD is the shortest distance between two points on the surface of a sphere, and is given by:

$$d_{S^2}(y_i, y_j) = r \cdot \arccos(y_i y_j) , \quad (5.26)$$

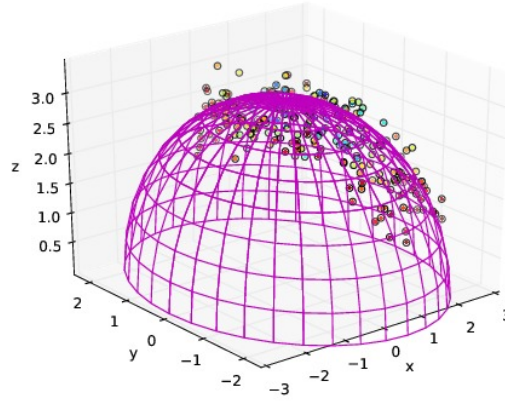


Figure 5.8: Scatterplot of $\{(y'_{i1}, y'_{i2}, y'_{i3}) | i = 1, \dots, 250\}$, for dataset S with added noise ($\sigma = 0.2$).

with $r = 3$.

For kRF, the training gram matrix was calculated using the Gaussian kernel with $\sigma = 0.5$. We followed the prediction methodology as described in [60]:

$$\hat{y}_{new} = \arg \min_{y \in S_{train}} \left(g_{kRF}(y, y) - 2 \sum_{i=1}^N a_{RF}(x_{new}, x_i) g_{kRF}(y, y_i) \right), \quad (5.27)$$

where $a_{RF}(\cdot, \cdot)$ is the RF-based affinity. The minimization problem was solved over S_{train} . As such, all predicted responses were drawn from the training set.

4-fold cross validation on S_{train} with no added noise was employed to tune the various model parameters. For k -NN regression, the number of neighbors k was selected to be 3. All RFs were built with 500 trees, no pruning and the number of candidate split predictors in each node was set to 3. Specifically for the dRF prediction, a 2-dimensional embedding space was used and backscoring was performed using $\sigma_{bs} = 80$ and $\gamma_{bs} = 50$.

Figure 5.9 shows the test mean squared Euclidean errors for the various models, noise settings and noise levels. Figure 5.10(a) depicts the embedding of the training responses (combined noise setting - medium noise level). Test points were projected on the embedding using their original values and the OOS formula, while predicted points were acquired through dRF. Finally, Fig. 5.10(b) shows the test and predicted points after backscoring the predictions on the spherical

manifold. Adherence of the predicted responses to the manifold's structure was good for all RF models. We observed better predictive performance when using our methodology compared to the nearest neighbor methods, as well as regular RF and kRF.

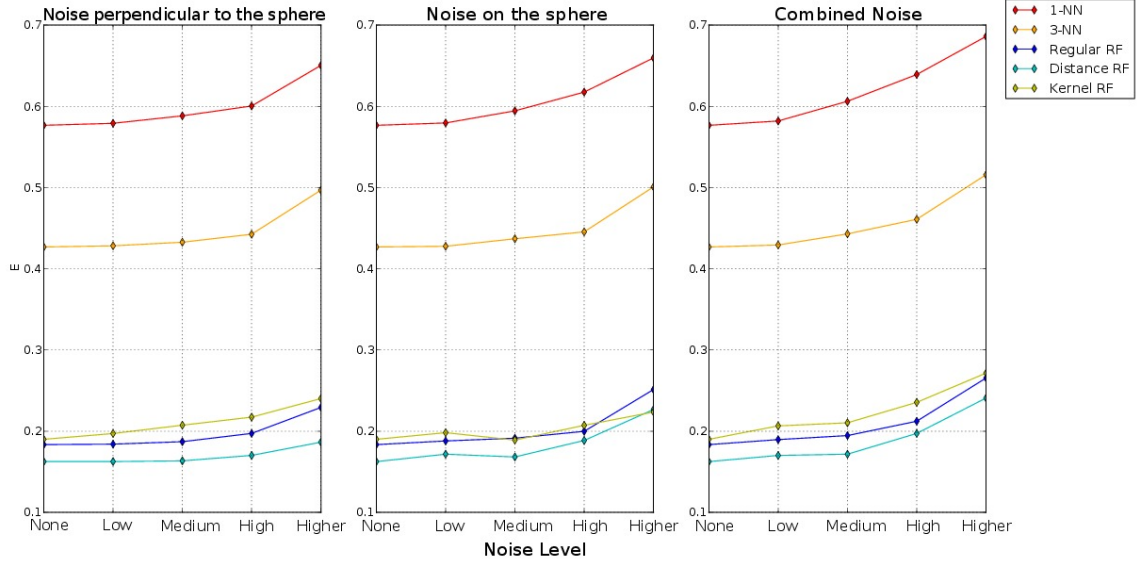


Figure 5.9: Test MSE prediction errors for all models, noise settings and levels.

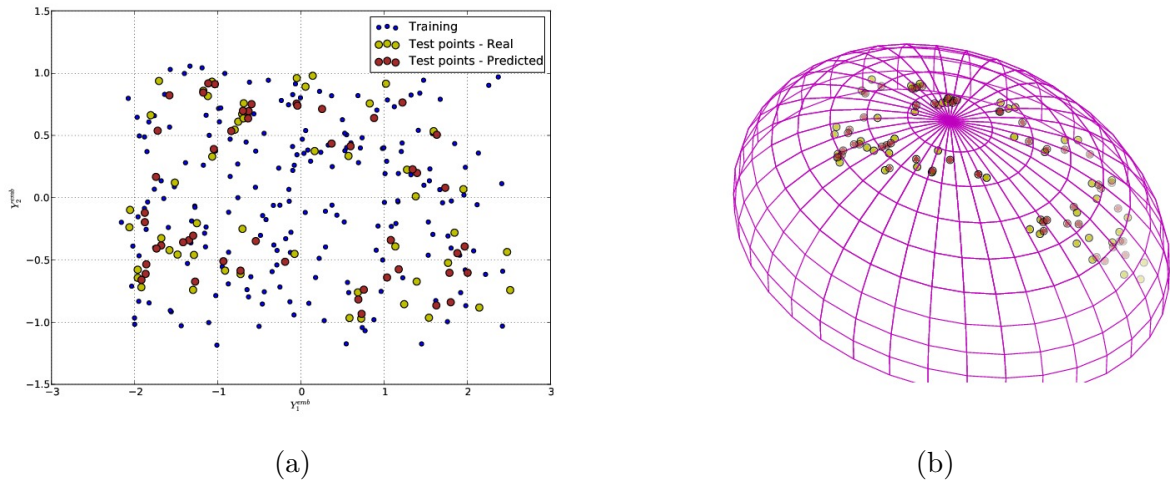


Figure 5.10: Test and predicted response observations (Combined noise setting - Medium level noise). (a) Embedding of 200 Training response observations (blue) with 50 OOS projected test points from (Yellow) and corresponding predicted points using dRF (brown). (b) Test and predicted points on the spherical manifold.

5.5.3 Prediction on a Swiss-roll Manifold

For the third simulation study, we have further forgone with the assumption of an analytically defined manifold. Let S denote the simulated dataset of $N = 900$ paired input-response points $\{(x_i, y_i) | i = 1, \dots, N\}$, with input $x_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}, x_{i6}) \in \mathbb{R}^6$ corresponding to response $y_i = (y_{i1}, y_{i2}, y_{i3}) \in \mathcal{M}_{sr} \subset \mathbb{R}^3$, where $\mathcal{M}_{sr}$ denotes the 2-dimensional swiss roll manifold embedded in $\mathbb{R}^3$. Again only the first two input dimensions were built to be predictive of the output. The dataset was constructed as follows. We first sampled 900 values $\{t_i : i = 1, \dots, 250\}$ from the uniform distribution $\mathcal{U}(\pi, 3\pi)$ and 900 values $\{u_i : i = 1, \dots, 250\}$ from the uniform distribution $\mathcal{U}(0, 21)$. The swiss-roll response points were acquired by:

$$y_{i1} = t_i \cos t_i, y_{i2} = u_i, y_{i3} = t_i \sin t_i. \quad (5.28)$$

In order to generate the data for the input coordinate variables $x_{.1}$ and $x_{.2}$, we first centered and scale t_i and u_i to $[-1, 1]$:

$$t'_i = \frac{(t_i - \bar{t}_i)}{\max t_i}, u'_i = \frac{(u_i - \bar{u}_i)}{\max u_i} \quad (5.29)$$

and finally mapped them in the unit circle:

$$x_{i1} = t'_i \sqrt{1 - \frac{u_i'^2}{2}}, x_{i2} = u'_i \sqrt{1 - \frac{t_i'^2}{2}}. \quad (5.30)$$

The values for the remaining four input coordinate variables were drawn from the standard normal Gaussian distribution. Gaussian Noise, with $\Sigma = 0.5\mathbf{I}_3$, was again added on the response points. Due to the complexity of the space, we were not able to distinguish between perpendicular and on manifold directions for the noise. Figure 5.11(a) shows the scatterplot of $\{(t_i, u_i) | i = 1, \dots, 900\}$ from which we generated our dataset, while Figs. 5.11(b), 5.11(c) and 5.11(d) show the input and response data points without and with added noise. In order to clarify the constructed dependence between input and response points, Fig. 5.12 includes

scatterplots for each of the relevant input/response dimensions.

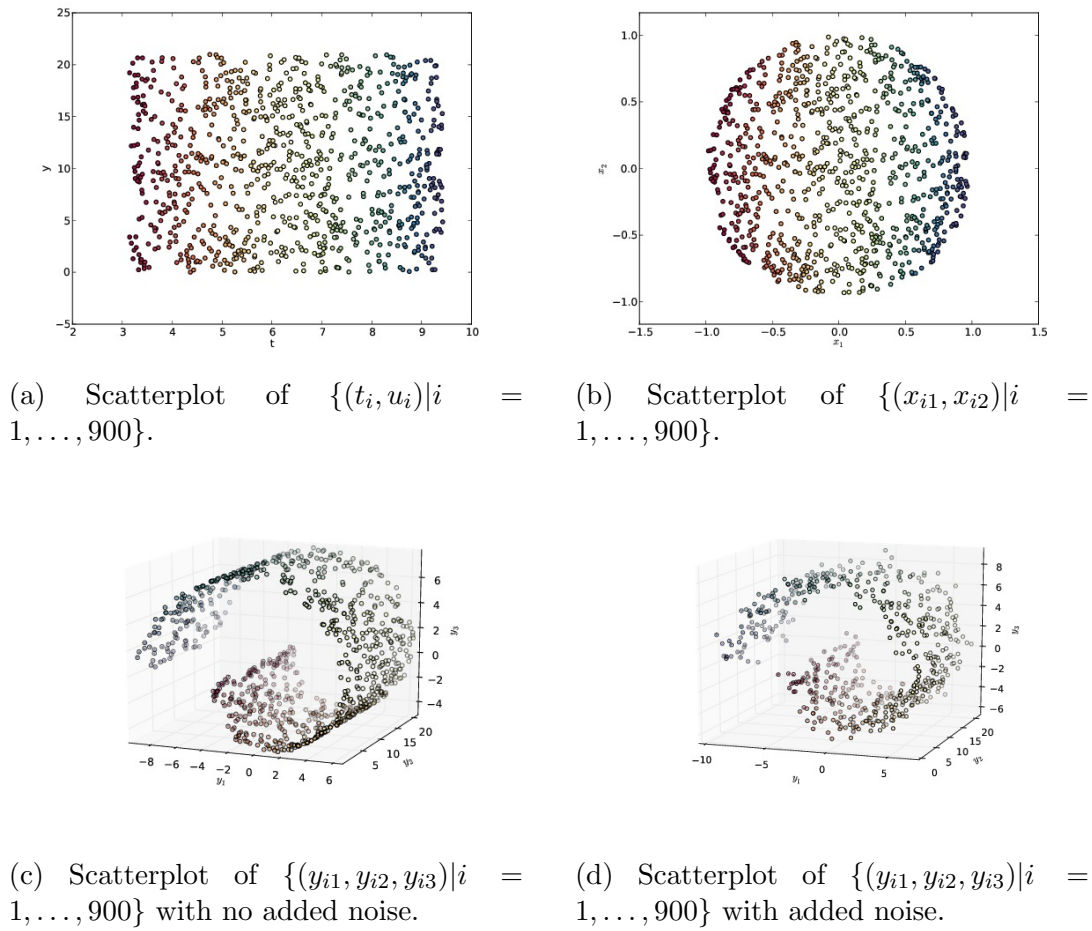


Figure 5.11: Scatterplots of the simulated input and response points for regression study on a Swiss-roll manifold. All plots use as colormap the values t_i .

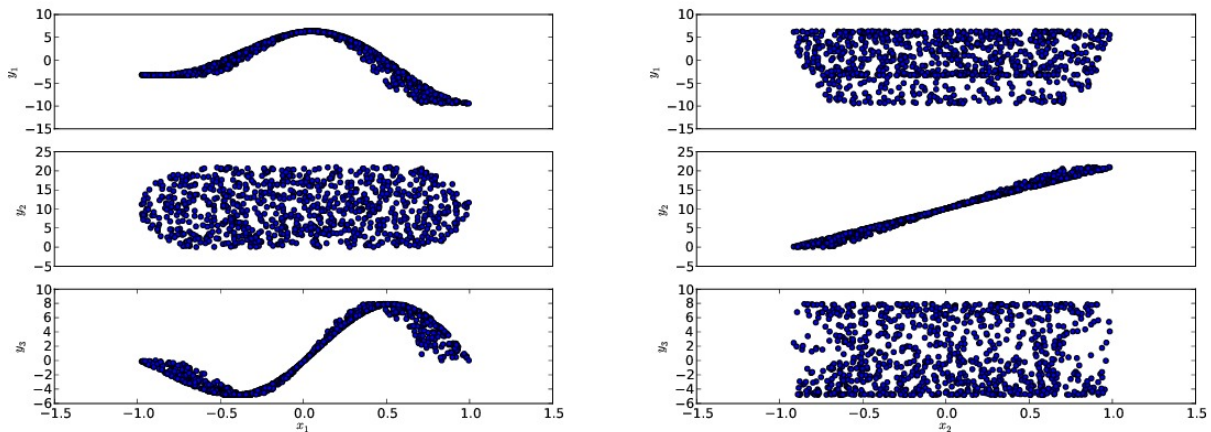


Figure 5.12: Scatterplots of $\{(x_{i1}, y_{i1}) | i = 1, \dots, 900\}$, $\{(x_{i1}, y_{i2}) | i = 1, \dots, 900\}$, $\{(x_{i1}, y_{i3}) | i = 1, \dots, 900\}$, $\{(x_{i2}, y_{i1}) | i = 1, \dots, 900\}$, $\{(x_{i2}, y_{i2}) | i = 1, \dots, 900\}$, $\{(x_{i2}, y_{i3}) | i = 1, \dots, 900\}$ for the constructed dataset S .

In this study, the manifold's structure diverges more from linearity than the spherical manifold

used in 5.5.2. In addition, there is no analytical formula for computing geodesic distances between the response points. As such, we approximated these distances using the Isomap distance function with the number of neighbors set to 5.

The simulated dataset S was split into S_{train} , comprising of 600 input-response pairs and S_{test} , consisting of the remaining 300. The same set of predictive models as in the previous study were employed. 2-fold cross validation on S_{train} without added noise on the responses was used to tune model parameters. For k -NN regression, the value of k was selected to be 5. All RFs were built with 150 trees, no pruning and the number of candidate split features in each node was set to 3. For dRF, the backscoring parameters used were $\sigma_{bs} = 100$ and $\gamma_{bs} = 200$. For kRF, the training gram matrix was calculated using the Gaussian kernel with $\sigma = 2.5$.

Table 5.2 presents the test mean squared errors for the two considered noise settings. We notice a big error difference between the NN and the RF-based methods, which can be attributed to the non-linear association between input and response points, as well as the added noisy dimensions in the inputs. dRF performed better than the other methods, yielding a relative reduction in error of 9.874% and 3.757%, respectively for the two noise settings, compared to the regular RF, as well as 6.717% and 9.293%, compared to kRF. Figure 5.13 shows the test and predicted response points 5-NN, RF, kRF and dRF for the setting with added noise on the responses.

	MSE No noise	MSE noise
1-NN	4.699 ± 0.071	4.945 ± 0.0957
5-NN	3.75 ± 0.188	3.839 ± 0.116
RF	1.202 ± 0.069	1.256 ± 0.007
kRF	1.161 ± 0.1	1.332 ± 0.0257
dRF	1.083 ± 0.0512	1.209 ± 0.0291

Table 5.2: Test mean squared errors of the various regression models for the simulated swiss-roll manifold dataset.

We focus now on the compliance of the predicted points to the underlying response manifold. We stress out that the euclidean prediction error can be somewhat misleading with regards to this matter. Concentrating first on the setting without noise on the training dataset, it is clear that methods which cast predictions from the pool of training responses have no prediction error related to confinement withing the manifold surface. From the compared methods, 1-NN

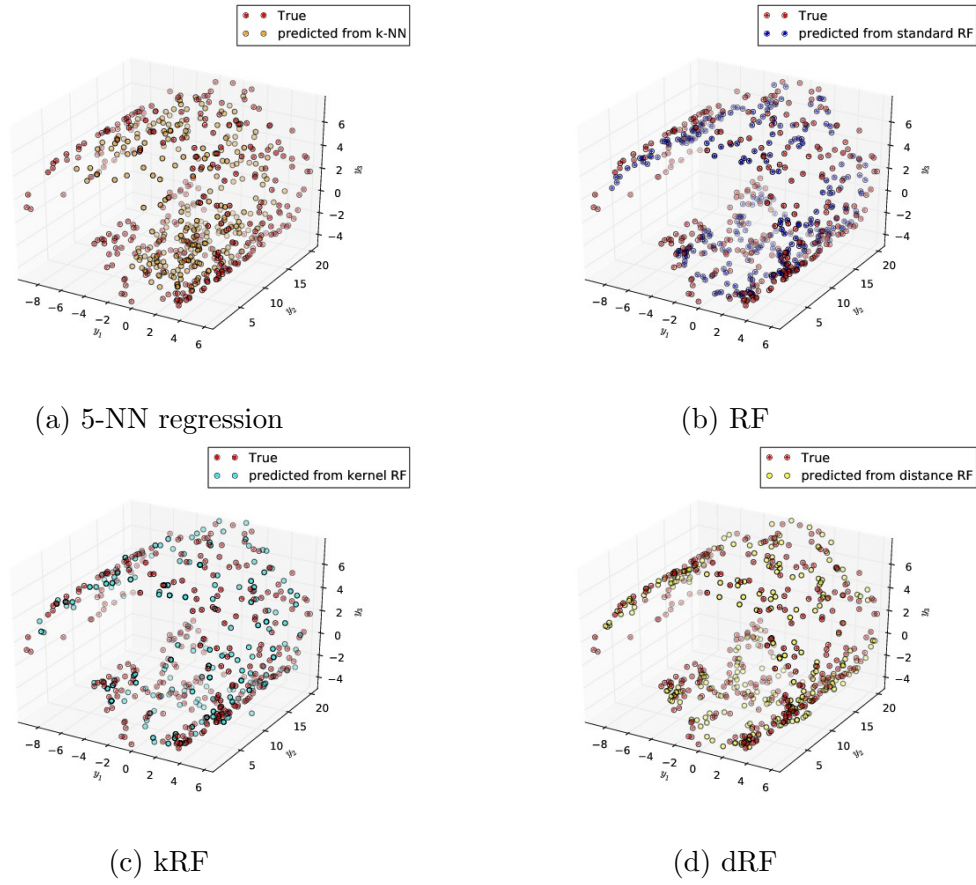


Figure 5.13: True and predicted response test points of the various regression models, excluding 1NN, for the simulated swiss-roll manifold dataset with added response noise.

and kRF fall into that category. The NN approach could not provide a good predicting model. On the other hand, training data were sufficiently dense in order to allow kRF to yield good results, despite the relatively simple prediction method.

For the rest of the methods, compliance to the manifold is not guaranteed, since they are based on averaging procedures to predict the responses. k -NN reduced the prediction error compared to 1-NN, but it failed to keep the predicted points on the manifold. RF performed fairly well, despite the use of a Euclidean-based split function during the training process. This can be attributed to the non-linear nature of the algorithm, in association to its robustness regarding the additional noisy input dimensions. dRF yielded the best results, compared to these two methods.

The validity of the above remark is also true for the setting with noise added in the responses. This becomes clearer by visually inspecting the test and predicted responses of the various

models. Figure 5.14 depicts test and predicted response points, similar to Fig. 5.13, projected on the y_1y_3 plane, while Fig. 5.15 shows the projected error vectors of the test dataset. We can clearly see from Figs 5.14 and 5.15 that k -NN missed the goal of regressing on the manifold. For RF, which was the most impervious method to the addition of noise based on the prediction error, it is clear that it constantly slightly underestimated the radius of the predicted points around the y_1 axis - the same is true for the no noise setting - plot not shown here-. This can be reasoned after taking into consideration that prediction in this model involves the calculation of average points on the euclidean space $\mathbb{R}^3$, unaware of a possible structure in the response space. kRF preserved the manifold structure of the predicted points better than RF, but, due to the simplified prediction method, it suffered significantly from the added noise in the responses, which is also evident from the larger increase in the prediction error, compared to RF. dRF performed consistency in all three fronts, namely low prediction error, compliance to the underlying response manifold and good robustness to the addition of noise.

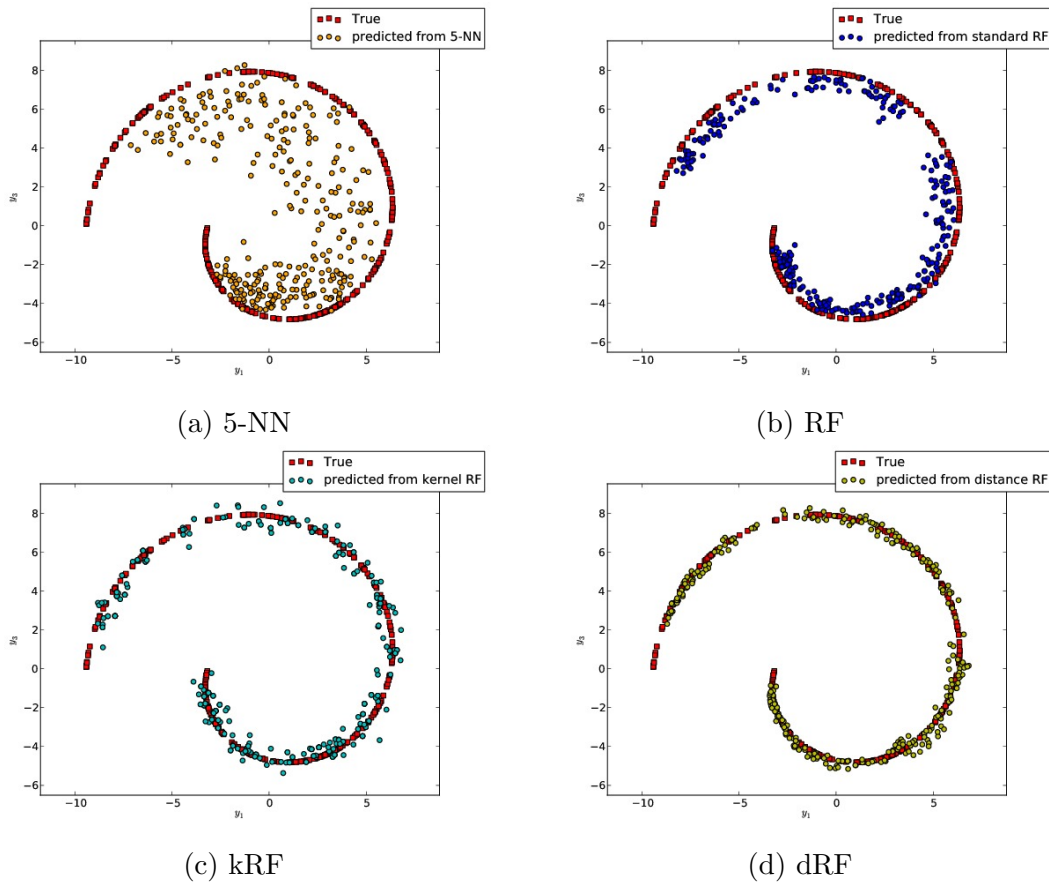


Figure 5.14: True and predicted response test points of the various regression models, excluding 1NN, for the simulated swiss-roll manifold dataset with added response noise. Points are projected on the y_1y_3 plane to clearly show the compliance to the response manifold.

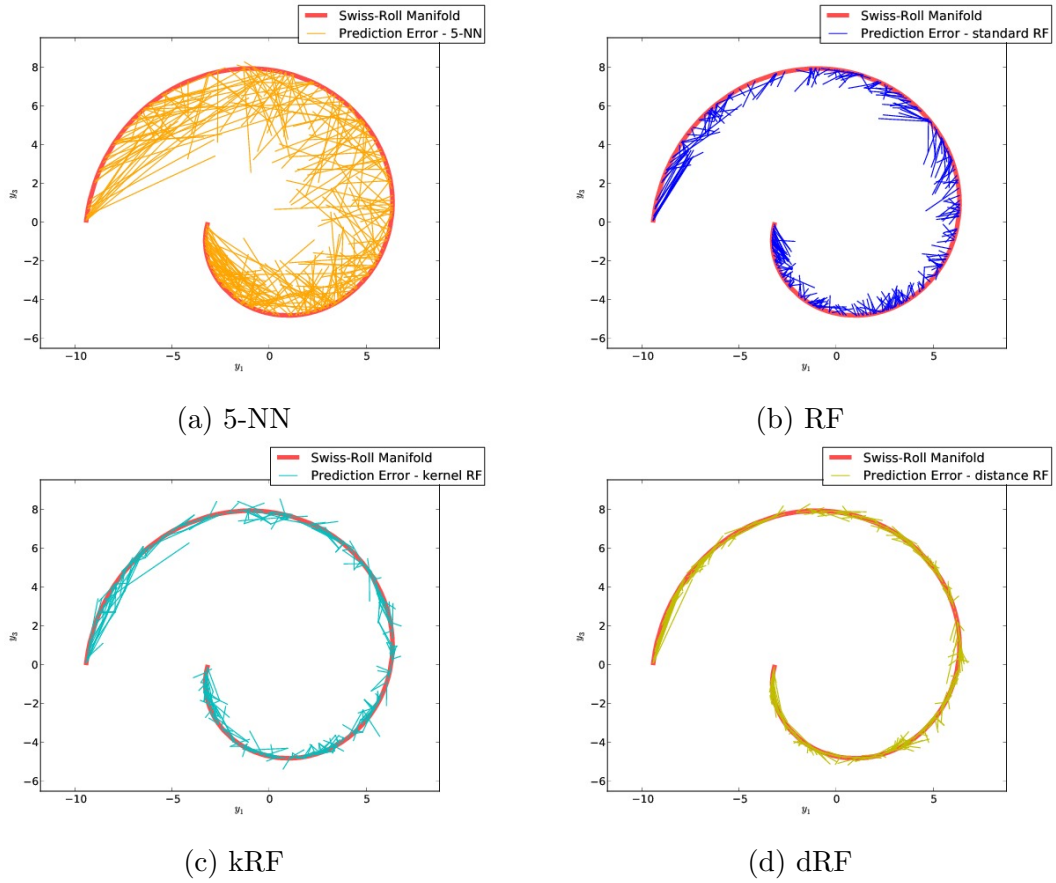


Figure 5.15: Test error vectors projected on the y_1y_3 plane of the various regression models, excluding 1NN, for the simulated swiss-roll manifold dataset with added response noise.

5.6 Applications on Imaging and Facial Surface Data

In imaging analysis, it is common to assume that a collection of similar images lies on a manifold [127]. This assumption is guided by the complex nature of images as data objects, as well as the observation that the Euclidean metric and the corresponding geometric structure that is imposed on the space often do not bode well with the human perception of similarity and difference between images [127]. The analytical definition of imaging manifolds is restricted to only simple cases, such as for images showing plain 2D geometrical shapes, for which shape manifolds can be well defined [156, 56]. In most applications though, the approximate imaging manifold structure is learned through a Manifold Learning technique. As we have seen previously, this approach does not come problem-free. In the case of predictive modeling, in particular, the problem of backscoring predictions to the original space, a.k.a reconstructing images from their embedding representation, can be challenging.

In the following, we use our manifold regression methodology for various image completion applications, where the objective is to predict one half of an image from the other half. A number of methodologies are compared, with our dRF predictive algorithm outperforming other methods. In addition, we apply our methodology on the prediction of 3D facial surfaces from genotypic data, with promising results.

5.6.1 Handwritten Digit Completion

For this application, we extracted 1000 gray-scale images of handwritten digits, from the UCI Machine Learning Repository [102]. Each digit class, from 0 to 9, was represented in the dataset by 100 8×8 pixel images. Input data were constructed by vectorization of the 8×4 upper half pixel intensities. The dataset was split into training and testing subsets with $N_{train} = 800$ and $N_{test} = 200$. The upper part of each images was taken as input for the predictive models. Responses comprised of the images' bottom parts. The test set can be seen in Fig. 5.16(a).

We compared predictions from 1-NN, RF, kRF and dRF models. All RFs were built with 300 trees and 5 candidate split features in each node. For dRF, the distance matrix was constructed using the Isomap distance, with the number of neighbors set to 5. A 25-dimensional embedding space was used and the backscoring parameters were $\sigma_{bs} = 3$ and $\gamma_{bs} = 20$. For kRF, the training gram matrix was calculated using the Gaussian kernel with $\sigma = 5$. The reconstructed test digits for the various models are shown in Fig. 5.16(b)-(e).

Table 5.3 summarizes the prediction results for the test data. In the first column we include the test Euclidean Mean Squared Errors (EMSE) for all models. In the case of 1-NN and kRF, which draw predictions from the training dataset, we are also able to report misclassification errors, a.k.a the percentage of predicted lower parts that mismatched the ground truth, which are shown on the second column of Table 5.3. Finally, in the last two columns, we report on the number and percentage of badly reconstructed test images from visual inspection, based on the following criteria: blurriness of the reconstructed image, smooth transition between the upper and bottom image parts, correct digit reconstruction. This qualitative performance measure

is important due to the non-Euclidean nature of the data, which, as will be discussed below, makes the EMSE unsuitable for judging the predictive performance.

Table 5.3: Test errors from the digit completion application. Classification Error was not applicable for RF and dRF. The number and percentage of badly reconstructed images was visually ascertained from Fig. 5.16.

	MSE	Clas. Error	Bad Rec. No	Bad Rec. %
1-NN	3.3665	0.165	40	20
RF	2.7651	-	54	27
kRF	3.2675	0.21	48	24
dRF	3.3287	-	37	18.5

1-NN and kRF cast predictions drawn directly from the training images. As such, no blurriness existed on the reconstructed digits of Fig. 5.16(b) and 5.16(d). Bad reconstructions were either misclassifications or reconstructions where the transition between the upper (input) and lower (predicted) image parts was not smooth. Surprisingly, 1NN outperformed kRF in terms of classification error, as well as upon visual inspection of the images.

It is obvious from Fig. 5.16(c) that standard RF is ill-suited for the specific application. The RF prediction approach of averaging pixel intensities from various images resulted in a high number of blurry and nonsensical digit reconstructions. It is important to notice that while RF performed the worst, based on visual assessment of the reconstructed images, it had the lowest test EMSE, as shown in table 5.3. This observation highlights the unsuitability of EMSE as a measure of performance, in the case of manifold-valued responses.

Reconstructed digits from our dRF model are shown in Fig. 5.16(e). Our predictions were not drawn directly from the training set. Nevertheless, we notice that the large majority of reconstructions had no fuzziness and the transition from upper to lower parts was smooth. For cases where some blurriness could be noticed in the reconstructions, its effect was significantly less severe than in the RF predictions, resulting in the overall lowest number of badly reconstructed digits, based on visual inspection.



Figure 5.16: Test and reconstructed images from the digit completion application. Reconstructed digits were generated by concatenating the predicted responses with the upper -input-half of the test images. Bad reconstructions are enclosed in red squares. (a) Test digits. (b)-(e) Reconstructions using predictions from (b) 1-NN, (c) RF, (d) kRF and (e) dRF.

5.6.2 2D Face Completion

Similar to the handwritten digit completion presented previously, the second set of applications again focused on image completion, this time though the datasets comprised of human face images.

Olivetti Dataset

We first employed the Olivetti faces dataset, as included in the scikit-learn machine learning python package [123]. It consists of ten different gray-scale 64×64 pixel images for each of 40 distinct subjects, with same subject images taken at different times and with varying lighting conditions, pose and facial expressions. Input data were constructed by vectorization of the 64×32 upper half pixel intensities. The dataset was split into training and testing subsets S_{train} and S_{test} , with $N_{train} = 300$ and $N_{test} = 100$. Images of the same subject were only allocated either to the training or the testing set.

The methodologies compared in this application were k -NN regression, linear Regression (LR), Ridge Regression (RR), RF and dRF. In particular for k -NN, we fitted 1-NN and 5-NN regression models. Regarding dRF, the training distance matrix was constructed using the Isomap metric with the number of neighbors set to 5. 3-fold cross validation on S_{train} was used to tune the regularization parameter for RR, whose value was set to 1 and the RF model parameters. RF and dRF were built with 300 trees, no pruning and 8 candidate split features in each node. For dRF, a 25-dimensional embedding space was used and the backscoring parameters were $\sigma_{bs} = 9$ and $\gamma_{bs} = 50$.

Five test images can be seen in the first column of Fig. 5.17, with the corresponding reconstructed images - comprising of the test input upper half and predicted response lower half - for 1-NN, 5-NN, LR and RR are also shown in the same figure, while Fig. 5.18 presents the equivalent results for RF and dRF. As we saw in the previous application, Euclidean MSEs do not provide a suitable descriptor of performance. As such we rely again on visual inspection of the reconstructed images to compare the different methodologies and draw conclusions.

1-NN exhibits a hits and miss behavior, with some predictions being relatively similar to the original face, whilst others, such as in the third and fifth depicted faces being completely different. In addition, there is minimal smoothness in the transitions from the input to the predicted half images, accentuated specifically on the nasal and zygomatic edges. 5NN demonstrates the expected blurriness in its predictions, an observation that is even more noticeable for the LR model. RR and RF predictions also suffered from a large amount of blurriness, although transitions between input and response are much smoother. Finally, dRF reconstructions exhibit the best transition smoothness of all methods, with the predicted half-images being well aligned to their inputs. Although the predictions are not completely free of blur artifacts, their effect is less severe and facial details, such as nasolabial folds and lips, are clearly portrayed.



Figure 5.17: Five test set and corresponding reconstructed face images from the Olivetti dataset for 1-NN, 5-NN, LR and RR regression models. The predicted responses were concatenated with the upper -input- half of the test images to generate the reconstructed face images.



Figure 5.18: Five reconstructed face images from the Olivetti dataset for RF and dRF regression models. The predicted responses were concatenated with the upper -input- half of the test images to generate the reconstructed face images.

TwinsUK Dataset

The second 2D face completion application was performed on a more extended dataset of 815 color frontal facial images from individual subjects of the TwinsUK cohort. The 128×128 pixel images were converted to gray-scale and 800 were assigned to the training dataset S_{train} , with the remaining 15 constituting S_{test} . The same methodologies as in 5.6.1 were considered, although results from the RR model are not presented here, due to them being almost identical to the LR results. For dRF, the distance matrix was constructed using the Isomap metric on the training responses, with the number of neighbors set to 7. Parameter tuning was performed through 4-fold cross validation on S_{train} . RF and dRF were built with 500 trees, no pruning and 25 candidate split features in each node. A 36-dimensional embedding space was used and the backscoring parameters were $\sigma_{bs} = 7$ and $\gamma_{bs} = 36$.

Figures 5.19 and 5.20 depict six test images with the corresponding reconstructed images from the fitted models. Despite the increased sample size of our dataset, we could not notice any significant improvement in the quality of the predicted images, for any of the employed methodologies. Visual inspection of the images resulted in the same conclusions as in 5.6.2.



Figure 5.19: Six test set and corresponding reconstructed face images from the TwinsUK dataset for 1NN, 5-NN and LR regression models. The predicted responses were concatenated with the upper -input- half of the test images to generate the reconstructed face images.



Figure 5.20: Six reconstructed face images from the TwinsUK dataset for RF and dRF regression models. The predicted responses were concatenated with the upper -input- half of the test images to generate the reconstructed face images.

5.6.3 3D Facial Surface Prediction from Genotypic Predictors

Lastly, we applied our dRF methodology on the problem of predicting 3D human facial surfaces from genotypic inputs. The dataset comprised of a subset of 449 images of unrelated subjects from the TwinsUK cohort, for which 3D frontal facial meshes and whole genome SNP data were available, and dense facial landmark correspondences were computed using GESSA (see Chapter 3). For a detailed description of the TwinsUK imaging and genotypic dataset see sections 3.5.2 and 3.5.5. Each genotype comprised of approximately 1.3 million SNPs. In order to reduce the training computational cost, PCA was applied on the genotypic input matrix to reduce the dimensionality of the data. All 449 principal components were retained and used as input variables. 49 subjects were kept for the testing set S_{test} , and the remaining 400 were used as training data. The procrustes distance metric [93] was employed in order to define distances between facial responses. For two faces with computed correspondence, the procrustes distance is defined as the root of summed Euclidean distances between the respective facial landmarks, after the two point sets are superimposed. The dRF regression model was built with 750 trees, no pruning and 20 candidate split features in each node. A 50-dimensional embedding space was used and the backscoring parameters were set to $\sigma_{bs} = 5$ and $\gamma_{bs} = 100$.

The average prediction error, defined as the mean of Procrustes distances between test and predicted faces was 0.03521. In order to provide a better comprehension of the results, we report that the average pairwise Procrustes distance between all faces in the dataset was 0.05174. Figure 5.21 shows 9 randomly selected pairs of test and predicted faces. For each pair, on the left is depicted the test face and on the right the corresponding prediction. The predicted facial surfaces do not show any significant reconstruction artifacts. Regarding prediction performance, the genotypic inputs do not appear to be highly predictive of the facial surfaces. Predicted faces tend to be very similar to each other, especially regarding details such as eye, nose and mouth shapes. For some faces, notable similarity could be observed on the peripheral structure and size of the face. Figure 5.22 includes 9 test faces for which predictions were sufficiently good in terms of general face shape and size, while 5.23 includes 9 test faces for which the predicted faces did not bare any noticeable similarity. Average prediction errors for the subset of faces

shown in Figs 5.22 and 5.23 were 0.02856 and 0.039754 respectively.



Figure 5.21: Random subset of 9 test and predicted 3D faces. For each pair, the test face is on the left and the predicted face on the right.



Figure 5.22: Subset of 9 test 3D faces with sufficiently good predictions. For each pair, the test face is on the left and the predicted face on the right.

While it cannot be denied that the shape of the human face is dependent on the genetic content, as was ascertained in our work regarding facial heritability (see section 3.5), it was



Figure 5.23: Subset of 9 test 3D faces with bad predictions. For each pair, the test face is on the left and the predicted face on the right.

not possible in this application to demonstrate ideal predictive performance. We believe that this lack of predictive power should not to be attributed to the methodology employed, since the predicted faces were realistic, albeit very similar to each other. The application was by no means exhaustive. Different preprocessing approaches for the genotypic inputs, as well as consideration of environmental and aging input variables may prove to significantly improve performance, and is something considered for future work.

5.7 Conclusion

In this work we addressed the problem of setting up a regression model when the response variable to be predicted lies on a manifold. Our objective was to employ learning and prediction algorithms that can readily utilize information about the shape of the manifold through an appropriate distance metric on that space. We built upon our previous work on distance Random Forest [153], and proposed a suitable prediction algorithm for manifold-valued responses.

In contrast to intrinsic manifold regression methods, which require analytically defined response

manifolds, our methodology only necessitates the definition of a meaningful distance metric in the response space. This can be especially useful for real-life applications, such as image analysis, where the underlying manifolds are usually not known. Our distance-based regression algorithm draws similarities to the family of kernel-based methodologies, which employ suitable kernel functions to define inner products between data objects. One advantage over kernel methods is the vast library of readily available distance metrics for almost any considered type of data. Furthermore, our methodology presents a unified framework, which deals with backscoring to the original response space, an issue that a lot of presented kernel methods do not tackle sufficiently well.

Apart from the standard parameters of the RF algorithm, our predictive methodology depends on the following hyperparameters, the dimensionality of the MDS embedding m , the smoothing parameter γ_{bs} and the backscoring kernel bandwidth σ_{bs} . All parameters can be tuned independently of the tuning procedure for the RF model. Additionally, in cases where the dimensionality of the manifold is known or assumed, m should ideally be fixed to that value. Two simulation studies were performed to showcase the predictive performance of our methodology. In these, we observed that predicted points adhered well to the underlying response manifold. At the same time, the dRF method showcased good robustness to the addition of noise on the observations.

A number of simulation studies and varied applications were performed to showcase the predictive performance of our methodology. An unfortunate issue was the inability to define universally objective error functions. As such, comparisons and conclusions were made mainly by visual inspection of the predicted objects. In the simulation studies, we observed that predicted points adhered well to the underlying response manifold. At the same time, the dRF models showcased good robustness to the addition of noise on the observations. dRF provided the best prediction performance out of all compared methodologies in the ensuing image completion applications as well. On the final application of face prediction from genotypic input data, predictions were realistic, even if not ideal in terms of similarity to the test faces.

Chapter 6

Conclusion

The content of this thesis revolved around methodologies concerning complex data objects that can either reside in non-Euclidean manifold spaces, or be represented as closed manifolds, or polyhedral approximations thereof. In each of Chapters 3, 4 and 5, we proposed methods tailored towards utilizing the complex nature of such data objects, and showcased the key benefits in a varied array of studies.

In Chapter 3, we presented GESSA, a dense point correspondence sampling methodology for ensembles of polyhedral surfaces. By incorporating the use of geodesic distances, logarithmic and exponential maps in the formulation of a kernel density estimator and related gradient descent updates, we achieved an accurate uniform annotation of points across the complete surfaces' areas, whilst increasing the efficiency and performance of the optimization process. These additions also allow application of the methodology to surfaces with a highly non-linear structure. In the future, variations of the point-wise correspondence optimization part could be considered. At the moment, this optimization is performed on a space of vectorized point coordinates, which does not take into explicit consideration the surfaces' constraints, and necessitates an initial automatic alignment of the surfaces, e.g. using ICP. Feature-based point-wise correspondence is a possible extension that might remove the dependence of the method on this initial alignment. We used our correspondence methodology to perform a large-scale heritability study of the human face shape, based on a dataset of 952 3D facial meshes of twins. The

dense point correspondence enabled the computation of various local curvature phenotypes, whose combined heritability estimates provided, for the first time, detailed heritability maps of the human face. The heritability results were subsequently used to guide a preliminary genomic association study of the human face, which identified genome-wide significant polymorphisms associated to the shape of the nasion. A manuscript detailing the GESSA methodology and associated heritability study have been accepted for publication in Scientific Reports. A full-scale GWAS, using a larger cohort of people, is under consideration.

In Chapter 4, we extended GESSA to enable regression-driven adaptive dense annotation of ensembles of surfaces. Modifications included the use of an adaptive kernel density estimator, with bandwidths determined through feature selection, performed with the use of an RF regression model. To our knowledge, this is the first time that such a methodology is presented in the literature. Due to fact that the adaptive algorithm is a direct extension of GESSA, any further advancements of the initial methodology are relevant in the adaptive setting as well. The method can also be used with multivariate surface associated features, since RF is capable of handling high-dimensional responses. Future research could additionally focus on the development of weighting schemes that would allow one-to-many or many-to-one relationships between surface points and phenotypic traits used as inputs in the regression model. This will necessitate the construction of suitable weight assigning functions that can map variable rankings to individual landmark locations. Adaptive sampling was employed on an application of face age prediction, where we achieved significant improvements in prediction performance, compared to an equivalent age prediction study based on a uniform distribution of corresponding surface points. A manuscript detailing adaptive GESSA and the age prediction study is under preparation for publication. The methodology could be further proven useful in a number of other applications. For example, in medical imaging, patient status could be used to guide landmark annotation of organs and tissues, while in face recognition applications, landmarks could be driven towards facial regions more indicative of subject identity.

Finally, in Chapter 5, we presented a predictive modeling approach for response objects occupying non-linear manifold spaces. The regression methodology was based on a distance modification of the RF algorithm that we previously published [153], which decouples the model's

training from the problem of response representation. For prediction purposes, the method of point averaging, employed by standard RF, was clearly unsuitable here. We constructed a predictive framework in which response point estimates are first predicted on a Euclidean embedding of the response manifold, learned from the training dataset, and then projected back on the original space. While the theoretical background is built on manifold spaces, the algorithm only necessitates the definition of a suitable distance function in the response space, thus showcasing good flexibility with regards to the types of data on which it can be applied. The methodology showed very good performance in a number of simulations and application studies, primarily focusing of image completion. In the future, further research could be concentrated on process optimization. In particular, one area of refinement is the estimation of the Euclidean embedding's dimensionality. The number of dimensions used affects the prediction accuracy, and tuning through cross validation can be computationally expensive in some cases. Dimensionality estimators do exist in the literature [116], and it would be interesting to evaluate their use in our methodology. Another potential research topic is the use of different kernel functions in the backscoring formulation. In this work, we opted for an isotropic Gaussian kernel, favoring the simplicity associated with the single standard deviation hyper-parameter. Nevertheless, the theoretical procedure allows the use of kernels with more complex structure, as well as the combined use of multiple kernel functions, and it would be of particular interest to evaluate whether potential performance benefits can be achieved with such use. An article describing this work is currently under review for publication in *Pattern Recognition Letters*.

Bibliography

- [1] Hirotugu Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974.
- [2] Brian Amberg, Sami Romdhani, and Thomas Vetter. Optimal step nonrigid ICP algorithms for surface registration. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.
- [3] Toby Andrew, Deborah J Hart, Harold Snieder, Marlies de Lange, Tim D Spector, and Alex J MacGregor. Are twins and singletons comparable? a study of disease-related and lifestyle characteristics in adult women. *Twin research*, 4(06):464–477, 2001.
- [4] Epameinondas Antonakos, Joan Alabort-i Medina, Georgios Tzimiropoulos, and Stefanos Zafeiriou. Hog active appearance models. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 224–228. IEEE, 2014.
- [5] Rector Arya, Ravindranath Duggirala, Anthony G Comuzzie, Sobha Puppala, Saileela Modem, Bhaskara R Busi, and Michael H Crawford. Heritability of anthropometric phenotypes in caste populations of Visakhapatnam, India. *Human biology*, pages 325–344, 2002.
- [6] Michel A Audette, Frank P Ferrie, and Terry M Peters. An algorithmic overview of surface registration techniques for medical imaging. *Medical image analysis*, 4(3):201–217, 2000.
- [7] Lidia Auret and Chris Aldrich. Empirical comparison of tree ensemble variable importance measures. *Chemometrics and Intelligent Laboratory Systems*, 105(2):157–170, 2011.

- [8] Gökhan Hasan Bakır. *Predicting structured data*. The MIT Press, 2007.
- [9] Monami Banerjee, Rudrasis Chakraborty, Edward Ofori, Michael S Okun, David E Vaillancourt, and Baba C Vemuri. A nonlinear regression technique for manifold valued data with applications to medical image analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4424–4432, 2016.
- [10] Monami Banerjee, Rudrasis Chakraborty, Edward Ofori, David Vaillancourt, and Baba C Vemuri. Nonlinear regression on riemannian manifolds and its applications to neuro-image analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 719–727. Springer, 2015.
- [11] Christian Bär and Frank Pfäffle. Wiener measures on riemannian manifolds and the feynman-kac formula. *arXiv preprint arXiv:1108.5082*, 2011.
- [12] Philip G Batchelor, Andrew P King, et al. 3D medical imaging. In *3D Imaging, Analysis and Applications*, pages 445–495. Springer, 2012.
- [13] Bülent Baydaş, Abdulvahit Erdem, Ibrahim Yavuz, and Ismail Ceylan. Heritability of facial proportions and soft-tissue profile characteristics in Turkish Anatolian siblings. *American journal of orthodontics and dentofacial orthopedics*, 131(4):504–509, 2007.
- [14] Yoshua Bengio, Olivier Delalleau, Nicolas Le Roux, Jean-François Paiement, Pascal Vincent, and Marie Ouimet. Learning eigenfunctions links spectral embedding and kernel PCA. *Neural Computation*, 16(10):2197–2219, 2004.
- [15] Yoshua Bengio, Jean-François Paiement, Pascal Vincent, Olivier Delalleau, Nicolas Le Roux, and Marie Ouimet. Out-Of-Sample extensions for LLE, isomap, MDS, eigenmaps, and spectral clustering. *Advances in neural information processing systems*, 16:177–184, 2004.
- [16] Paul J Besl and Neil D McKay. Method for registration of 3D shapes. In *Robotics-DL tentative*, pages 586–606. International Society for Optics and Photonics, 1992.

- [17] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194. ACM Press/Addison-Wesley Publishing Co., 1999.
- [18] Steven Boker, Michael Neale, Hermine Maes, Michael Wilde, Michael Spiegel, Timothy Brick, Jeffrey Spies, Ryne Estabrook, Sarah Kenny, Timothy Bates, et al. OpenMx: An open source extended structural equation modeling framework. *Psychometrika*, 76(2):306–317, 2011.
- [19] Timo Bolkart and Stefanie Wuhrer. A groupwise multilinear correspondence optimization for 3D faces. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3604–3612, 2015.
- [20] Dorret Boomsma, Andreas Busjahn, and Leena Peltonen. Classical twin studies and beyond. *Nature reviews genetics*, 3(11):872–882, 2002.
- [21] James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah, and David Dunaway. A 3D morphable model learnt from 10,000 faces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5543–5552, 2016.
- [22] Ingwer Borg and Patrick JF Groenen. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.
- [23] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- [24] Alexander M Bronstein, Michael M Bronstein, and Ron Kimmel. *Numerical geometry of non-rigid shapes*. Springer Science & Business Media, 2008.
- [25] PJ Byard, Kartavya Sharma, James Russell, and DC Rao. A family study of anthropometric traits in a Punjabi community: II. An investigation of familial transmission. *American journal of physical anthropology*, 64(2):97–104, 1984.
- [26] Carine Carels, Nathalie Van Cauwenberghe, I Savoye, Guy Willems, Ruth Loos, Cathérine Derom, and Robert Vlietinck. A quantitative genetic study of cephalometric variables in twins. *Clinical orthodontics and research*, 4(3):130–140, 2001.

- [27] Elizabeth Carson. Maximum likelihood estimation of human craniometric heritabilities. *American Journal of Physical Anthropology*, 131(2):169–180, 2006.
- [28] Joshua Cates, Thomas P Fletcher, Martin Styner, Martha Shenton, and Ross Whitaker. Shape modeling and analysis with entropy-based particle systems. In *Information Processing in Medical Imaging*, pages 333–345. Springer, 2007.
- [29] Alexander Ceron, Augusto Salazar, and Flavio Prieto. Relevance analysis of 3D curvature-based shape descriptors on interest points of the face. In *Image Processing Theory Tools and Applications (IPTA), 2010 2nd International Conference on*, pages 452–457. IEEE, 2010.
- [30] Isaac Chavel. *Riemannian geometry: a modern introduction*, volume 98. Cambridge university press, 2006.
- [31] Weiyang Chen, Wei Qian, Gang Wu, Weizhong Chen, Bo Xian, Xingwei Chen, Yaqiang Cao, Christopher D Green, Fanghong Zhao, Kun Tang, et al. Three-dimensional human facial morphologies as robust aging markers. *Cell research*, 2015.
- [32] Weiyang Chen, Xian Xia, Yi Huang, Xingwei Chen, and Jing-Dong J Han. Bioimaging for quantitative phenotype analysis. *Methods*, 102:20–25, 2016.
- [33] Haili Chui and Anand Rangarajan. A new point matching algorithm for non-rigid registration. *Computer Vision and Image Understanding*, 89(2):114–141, 2003.
- [34] Peter Claes, Denise K Liberton, Katleen Daniels, Kerri Matthes Rosana, Ellen E Quillen, Laurel N Pearson, Brian McEvoy, Marc Bauchet, Arslan A Zaidi, Wei Yao, et al. Modeling 3D facial shape from DNA. *PLoS Genet*, 10(3):e1004224, 2014.
- [35] Peter Claes, M Walters, and J Clement. Improved facial outcome assessment using a 3D anthropometric mask. *International journal of oral and maxillofacial surgery*, 41(3):324–330, 2012.
- [36] Philip J Clark. The heritability of certain anthropometric characters as ascertained from measurements of twins. *American journal of human genetics*, 8(1):49, 1956.

- [37] Alessandro Colombo, Claudio Cusano, and Raimondo Schettini. 3D face detection using curvature analysis. *Pattern recognition*, 39(3):444–455, 2006.
- [38] Dietmar Cordes and Rajesh R Nandy. Estimation of the intrinsic dimensionality of fMRI data. *Neuroimage*, 29(1):145–154, 2006.
- [39] Trevor F Cox and Michael AA Cox. *Multidimensional scaling*. CRC Press, 2010.
- [40] William R Crum, Thomas Hartkens, and DLG Hill. Non-rigid image registration: Theory and practice. *The British Journal of Radiology*, 2014.
- [41] Rhodri H Davies, Carole J Twining, Tim F Cootes, John C Waterton, and Camillo J Taylor. A minimum description length approach to statistical shape modeling. *Medical Imaging, IEEE Transactions on*, 21(5):525–537, 2002.
- [42] Cecar G Demayo, Mark Anthony J Torres, Astrid L Sinco, and Marites L Bonachita-Sanguila. Geometric morphometric analyses of facial shape in twins. *The Internet Journal of Biological Anthropology*, 4(1), 2010.
- [43] Mathieu Desbrun, Nicolas Tsingos, and Marie-Paule Gascuel. Adaptive sampling of implicit surfaces for interactive modelling and animation. In *Computer Graphics Forum*, volume 15, pages 319–325. Wiley Online Library, 1996.
- [44] Eric J Devor. Transmission of human craniofacial dimensions. *Journal of craniofacial genetics and developmental biology*, 7(2):95–106, 1986.
- [45] Michel Marie Deza and Elena Deza. Encyclopedia of distances. In *Encyclopedia of Distances*, pages 1–583. Springer, 2009.
- [46] Michael J Dixon, Mary L Marazita, Terri H Beaty, and Jeffrey C Murray. Cleft lip and palate: Understanding genetic and environmental influences. *Nature Reviews Genetics*, 12(3):167–178, 2011.
- [47] Chitra Dorai and Anil K Jain. Cosmos - A representation scheme for 3D free-form objects. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(10):1115–1130, 1997.

- [48] Nicolas Duchateau, Mathieu De Craene, Gemma Piella, and Alejandro F Frangi. Constrained manifold learning for the characterization of pathological deviations from normality. *Medical Image Analysis*, 2012.
- [49] Paul Ekman, Wallace V Friesen, and Phoebe Ellsworth. *Emotion in the human face: Guidelines for research and an integration of findings*. Elsevier, 2013.
- [50] Abdou ElSharawy, Andreas Keller, Friederike Flachsbart, Anke Wendschlag, Gunnar Jacobs, Nathalie Kefer, Thomas Brefort, Petra Leidinger, Christina Backes, Eckart Meese, et al. Genome-wide mirna signatures of human longevity. *Aging cell*, 11(4):607–616, 2012.
- [51] Sergey Ermakov, Eugene Kobylansky, and Gregory Livshits. Quantitative genetic study of head size related phenotypes in ethnically homogeneous chuvasha pedigrees. *Annals of human biology*, 32(5):585–598, 2005.
- [52] Sergey Ermakov, Michael G Rosenbaum, Ida Malkin, and Gregory Livshits. Family-based study of association between ENPP1 genetic variants and craniofacial morphology. *Annals of human biology*, 37(6):754–766, 2010.
- [53] Douglas Scott Falconer. *Introduction to quantitative genetics*. DS Falconer, 1960.
- [54] P Thomas Fletcher and Miaomiao Zhang. Probabilistic geodesic models for regression and dimensionality reduction on riemannian manifolds. In *Riemannian Computing in Computer Vision*, pages 101–121. Springer, 2016.
- [55] Thomas Fletcher. Geodesic regression on riemannian manifolds. In *Proceedings of the Third International Workshop on Mathematical Foundations of Computational Anatomy-Geometrical and Statistical Methods for Modelling Biological Shape Variability*, pages 75–86, 2011.
- [56] Thomas Fletcher. Geodesic regression and its application to shape analysis. In *Innovations for Shape Analysis*, pages 35–52. Springer, 2013.

- [57] Xin Geng, Zhi-Hua Zhou, and Kate Smith-Miles. Automatic age estimation based on facial aging patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(12):2234–2240, 2007.
- [58] Shawn N Geniole, Amanda E Keyes, Justin M Carré, and Cheryl M McCormick. Fearless dominance mediates the relationship between the facial width-to-height ratio and willingness to cheat. *Personality and Individual Differences*, 57:59–64, 2014.
- [59] Robin Genuer, Jean-Michel Poggi, and Christine Tuleau-Malot. Variable selection using random forests. *Pattern Recognition Letters*, 31(14):2225–2236, 2010.
- [60] Pierre Geurts, Louis Wehenkel, and Florence d’Alché Buc. Kernelizing the output of tree-based methods. In *Proceedings of the 23rd international conference on Machine learning*, pages 345–352. Acm, 2006.
- [61] Pierre Geurts, Louis Wehenkel, and Florence d’Alché Buc. Gradient boosting for kernelized output spaces. In *Proceedings of the 24th international conference on Machine learning*, pages 289–296. ACM, 2007.
- [62] Deboshmita Ghosh, Andrei Sharf, and Nina Amenta. Feature-driven deformation for dense correspondence. In *SPIE Medical Imaging*, pages 726136–726136. International Society for Optics and Photonics, 2009.
- [63] Benjamin A Goldstein, Eric C Polley, and Farren Briggs. Random forests for genetic association studies. *Statistical applications in genetics and molecular biology*, 10(1), 2011.
- [64] John C Gower. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3-4):325–338, 1966.
- [65] Carl Martin Grewe and Stefan Zachow. Fully automated and highly accurate dense correspondence for facial surfaces. In *European Conference on Computer Vision*, pages 552–568. Springer, 2016.

- [66] Guodong Guo, Yun Fu, Charles R Dyer, and Thomas S Huang. Image-based human age estimation by manifold learning and locally adjusted robust regression. *Image Processing, IEEE Transactions on*, 17(7):1178–1188, 2008.
- [67] Jianya Guo, Xi Mei, and Kun Tang. Automatic landmark annotation and dense correspondence registration for 3D human facial images. *BMC bioinformatics*, 14(1):232, 2013.
- [68] Pavel Hamet and Johanne Tremblay. Genes of aging. *Metabolism*, 52:5–9, 2003.
- [69] Peter Hammond and Michael Suttie. Large-scale objective phenotyping of 3d facial morphology. *Human mutation*, 33(5):817–825, 2012.
- [70] Sam Hare, Amir Saffari, and Philip HS Torr. Struck: Structured output tracking with kernels. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 263–270. IEEE, 2011.
- [71] Edward F Harris and Michelle G Johnson. Heritability of craniometric and occlusal variables: A longitudinal sib analysis. *American Journal of Orthodontics and Dentofacial Orthopedics*, 99(3):258–268, 1991.
- [72] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*, volume 1. Springer New York, 2001.
- [73] Roland C Hauspie, Susanne Charles, and Elisabeth Defrise-Gussenhoven. Testing for the presence of genetic variance in factors of face measurements of Belgian twins. *Annals of human biology*, 12(5):429–440, 1985.
- [74] Xiaofei He, Wei-Ying Ma, and Hong-Jiang Zhang. Learning an image manifold for retrieval. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 17–23. ACM, 2004.
- [75] Matthias Hein. Robust nonparametric regression with metric-space valued output. In *Advances in Neural Information Processing Systems*, pages 718–726, 2009.

- [76] Stephen L Helfand and Sharon K Inouye. Rejuvenating views of the ageing process. *Nature Reviews Genetics*, 3(2):149–153, 2002.
- [77] Jacob Hinkle, P Thomas Fletcher, and Sarang Joshi. Intrinsic polynomials for regression on riemannian manifolds. *Journal of Mathematical Imaging and Vision*, 50(1-2):32–52, 2014.
- [78] Yi Hong, Roland Kwitt, Nikhil Singh, Brad Davis, Nuno Vasconcelos, and Marc Niethammer. Geodesic regression on the grassmannian. In *European Conference on Computer Vision*, pages 632–646. Springer, 2014.
- [79] Bryan N Howie, Peter Donnelly, and Jonathan Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*, 5(6):e1000529, 2009.
- [80] Joop J Hox. An introduction to structural equation modeling. *Family Science Review*, 11:354–373.
- [81] Di Huang, Jia Sun, Xudong Yang, Dawei Weng, and Yunhong Wang. 3D face analysis: Advances and perspectives. In *Chinese Conference on Biometric Recognition*, pages 1–21. Springer, 2014.
- [82] Stuart W Hunter, Daniel R Balbach, and Donald E Lamphiear. The heritability of attained growth in the human face. *American journal of orthodontics*, 58(2):128–134, 1970.
- [83] Timothy James Hutton. *Dense surface models of the human face*. PhD thesis, Citeseer, 2004.
- [84] Tolga Inan and Ugur Halici. 3D face recognition with local shape descriptors. *Information Forensics and Security, IEEE Transactions on*, 7(2):577–587, 2012.
- [85] Ke Jia, Lei Wang, and Nianjun Liu. Efficient structured support vector regression. In *Computer Vision–ACCV 2010*, pages 586–598. Springer, 2011.

- [86] Thorsten Joachims, Thomas Hofmann, Yisong Yue, and Chun-Nam Yu. Predicting structured objects with support vector machines. *Communications of the ACM*, 52(11):97–104, 2009.
- [87] Berglind Johannsdottir, Freyr Thorarinsson, Arni Thordarson, and Thordur Eydal Magnusson. Heritability of craniofacial characteristics between parents and offspring estimated from lateral cephalograms. *American Journal of Orthodontics and Dentofacial Orthopedics*, 127(2):200–207, 2005.
- [88] Andrew D Johnson, Robert E Handsaker, Sara L Pulit, Marcia M Nizzari, Christopher J O'Donnell, and Paul IW De Bakker. SNAP: A web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics*, 24(24):2938–2939, 2008.
- [89] Ian T Jolliffe, Nickolay T Trendafilov, and Mudassir Uddin. A modified principal component technique based on the lasso. *Journal of computational and Graphical Statistics*, 12(3):531–547, 2003.
- [90] Chris M Jones. Variable kernel density estimates and variable kernel density estimates. *Australian Journal of Statistics*, 32(3):361–371, 1990.
- [91] Ioannis A Kakadiaris, Georgios Passalis, George Toderici, Mohammed N Murtuza, Yunliang Lu, Nikos Karampatziakis, and Theoharis Theoharis. Three-dimensional face recognition in the presence of facial expressions: An annotated deformable model approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4):640–649, 2007.
- [92] Biswajit Karmakar, Sergey Ermakov, Konstantin Yakovenko, and Eugene Kobylansky. Genetic determination of head-size-related anthropometric traits in an ethnically homogeneous sample of 373 Indian pedigrees of west Bengal. *Human biology*, 79(5):501–514, 2007.
- [93] David G Kendall. A survey of the statistical theory of shape. *Statistical Science*, pages 87–99, 1989.

- [94] Omar F Khabour and J M Barnawi. Association of longevity with il-10-1082 g/a and tnf- α -308 g/a polymorphisms. *International journal of immunogenetics*, 37(4):293–298, 2010.
- [95] Hyun-Jin Kim, Sun-Wha Im, Ganchimeg Jargal, Siwoo Lee, Jae-Hyuk Yi, Jeong-Yeon Park, Joohon Sung, Sung-Il Cho, Jong-Yeol Kim, Jong-Il Kim, et al. Heritabilities of facial measurements and their latent factors in Korean families. *Genomics & informatics*, 11(2):83–92, 2013.
- [96] Lisa King, Edward F Harris, and Elizabeth A Tolley. Heritability of cephalometric and occlusal variables as assessed from siblings with overt malocclusions. *American Journal of Orthodontics and Dentofacial Orthopedics*, 104(2):121–131, 1993.
- [97] Jan J Koenderink and Andrea J van Doorn. Surface shape and curvature scales. *Image and vision computing*, 10(8):557–564, 1992.
- [98] Luci Kohn. The role of genetics in craniofacial morphology and growth. *Annual Review of Anthropology*, pages 261–278, 1991.
- [99] Christophe Lenglet, Mikaël Rousson, Rachid Deriche, and Olivier Faugeras. Statistics on the manifold of multivariate normal distributions: Theory and application to diffusion tensor MRI processing. *Journal of Mathematical Imaging and Vision*, 25(3):423–444, 2006.
- [100] Shuo Li and João Manuel RS Tavares. *Shape analysis in medical image analysis*. Springer, 2014.
- [101] Stan Z Li. Adaptive sampling and mesh generation. *Computer-Aided Design*, 27(3):235–240, 1995.
- [102] Moshe Lichman. Uci machine learning repository, 2013.
- [103] Lizhen Lin, Brian St Thomas, Hongtu Zhu, and David B Dunson. Extrinsic local regression on manifold-valued data. *Journal of the American Statistical Association*, (just-accepted), 2016.

- [104] William K Lobb. Craniofacial morphology and occlusal variation in monozygous and dizygous twins. *The Angle Orthodontist*, 57(3):219–233, 1987.
- [105] John Loehr and Robert B O’Hara. Facial morphology predicts male fitness and rank but not survival in second world war Finnish soldiers. *Biology letters*, 9(4):20130049, 2013.
- [106] Feihan Lu and Eva Petkova. A comparative study of variable selection methods in the context of developing psychiatric screening instruments. *Statistics in medicine*, 33(3):401–421, 2014.
- [107] Anders Lundström and John McWilliam. Comparison of some cephalometric distances and corresponding facial proportions with regard to heritability. *The European Journal of Orthodontics*, 10(1):27–29, 1988.
- [108] Kanti V Mardia. Some properties of classical multi-dimensional scaling. *Communications in Statistics-Theory and Methods*, 7(13):1233–1241, 1978.
- [109] Nicholas G Martin and Lindon J Eaves. The genetical analysis of covariance structure. *Heredity*, 38(1):79, 1977.
- [110] Nicholas G Martin, Lindon J Eaves, Michael J Kearsey, and Peter Davies. The power of the classical twin study. *Heredity*, 40(1):97–116, 1978.
- [111] Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.
- [112] Miriah D Meyer, Pierre Georgel, and Ross T Whitaker. Robust particle systems for curvature dependent sampling of implicit surfaces. In *Shape Modeling and Applications, 2005 International Conference*, pages 124–133. IEEE, 2005.
- [113] Thomas Minka. Inferring a gaussian distribution. Technical report, 2008.
- [114] Joseph SB Mitchell, David M Mount, and Christos H Papadimitriou. The discrete geodesic problem. *SIAM Journal on Computing*, 16(4):647–668, 1987.

- [115] Alireza Moayyeri, Christopher J Hammond, Ana M Valdes, and Timothy D Spector. Cohort profile: TwinsUK and healthy ageing twin study. *International journal of epidemiology*, 42(1):76–85, 2013.
- [116] Philippos Mordohai and Gérard Medioni. Dimensionality estimation, manifold learning and function approximation using tensor voting. *Journal of Machine Learning Research*, 11(Jan):411–450, 2010.
- [117] Minoru Nakate, Pao-Lo Yu, Bailey Davis, and Walter E Nance. Genetic determinants of cranio-facial morphology: A twin study. *Annals of human genetics*, 37(4):431–443, 1974.
- [118] Michael Neale and Lon Cardon. *Methodology for genetic studies of twins and families*. Number 67. Springer Science & Business Media, 1992.
- [119] Maks Ovsjanikov, Mirela Ben-Chen, Justin Solomon, Adrian Butscher, and Leonidas Guibas. Functional maps: A flexible representation of maps between shapes. *ACM Transactions on Graphics (TOG)*, 31(4):30, 2012.
- [120] Athanasios G Pallis, Sigrid Hatse, Barbara Brouwers, Graham Pawelec, Claire Falandry, Ulrich Wedding, Lissandra Dal Lago, Lazzaro Repetto, Alistair Ring, and Hans Wildiers. Evaluating the physiological reserves of older patients with cancer: The value of potential biomarkers of aging? *Journal of geriatric oncology*, 5(2):204–218, 2014.
- [121] Lavinia Paternoster, Alexei I Zhurov, Arshed M Toma, John P Kemp, Beate St Pourcain, Nicholas J Timpson, George McMahon, Wendy McArdle, Susan M Ring, George Davey Smith, et al. Genome-wide association study of three-dimensional facial morphology identifies a variant in PAX3 associated with nasion position. *The American Journal of Human Genetics*, 90(3):478–485, 2012.
- [122] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3D face model for pose and illumination invariant face recognition. In *Advanced video and signal based surveillance, 2009. AVSS’09. Sixth IEEE International Conference on*, pages 296–301. IEEE, 2009.

- [123] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [124] Shouneng Peng, Jingze Tan, Sile Hu, Hang Zhou, Jing Guo, Li Jin, and Kun Tang. Detecting genetic association of common human facial morphological variation using high density 3D image registration. *PLoS computational biology*, 9(12), 2013.
- [125] Xavier Pennec. Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25(1):127–154, 2006.
- [126] David I Perrett, Keith A May, and Sau Yoshikawa. Facial shape and judgements of female attractiveness. *Nature*, 368(6468):239–242, 1994.
- [127] Gabriel Peyré. Manifold models for signals and images. *Computer Vision and Image Understanding*, 113(2):249–260, 2009.
- [128] P Jonathon Phillips, Patrick J Flynn, Todd Scruggs, Kevin W Bowyer, Jin Chang, Kevin Hoffman, Joe Marques, Jaesik Min, and William Worek. Overview of the face recognition grand challenge. In *Computer vision and pattern recognition, 2005. CVPR 2005. IEEE computer society conference on*, volume 1, pages 947–954. IEEE, 2005.
- [129] Nathalie Pochet, Frank De Smet, Johan AK Suykens, and Bart LR De Moor. Systematic benchmarking of microarray data classification: Assessing the role of non-linearity and dimensionality reduction. *Bioinformatics*, 20(17):3185–3195, 2004.
- [130] Konrad Polthier and Markus Schmies. *Straightest geodesics on polyhedral surfaces*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1998.
- [131] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

- [132] Inam Ur Rahman, Iddo Drori, Victoria C Stodden, David L Donoho, and Peter Schröder. Multiscale representations for manifold-valued data. *Multiscale Modeling & Simulation*, 4(4):1201–1232, 2005.
- [133] Frühling V Rijsdijk and Pak C Sham. Analytic approaches to twin data using structural equation models. *Briefings in bioinformatics*, 3(2):119–133, 2002.
- [134] Joel W Robbin and Dietmar A Salamon. Introduction to differential geometry. *ETH, Lecture Notes, preliminary version, January*, 2011.
- [135] Andy Roberts. Curvature attributes and their application to 3D interpreted horizons. *First break*, 19(2):85–100, 2001.
- [136] Daniel Rueckert, Luke I Sonoda, Carmel Hayes, Derek LG Hill, Martin O Leach, and David J Hawkes. Nonrigid registration using free-form deformations: application to breast MR images. *IEEE transactions on medical imaging*, 18(8):712–721, 1999.
- [137] Szymon Rusinkiewicz. Estimating curvatures and their derivatives on triangle meshes. In *3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004. Proceedings. 2nd International Symposium on*, pages 486–493. IEEE, 2004.
- [138] Stephan R Sain. *Adaptive kernel density estimation*. PhD thesis, Citeseer, 1994.
- [139] Stephan R Sain. Multivariate locally adaptive density estimation. *Computational Statistics & Data Analysis*, 39(2):165–186, 2002.
- [140] Stephan R Sain and David W Scott. On locally adaptive density estimation. *Journal of the American Statistical Association*, 91(436):1525–1534, 1996.
- [141] Augusto Salazar, Alexander Cerón, and Flavio Prieto. 3D curvature-based shape descriptors for face segmentation: an anatomical-based analysis. In *Advances in Visual Computing*, pages 349–358. Springer, 2010.
- [142] Augusto Salazar, Stefanie Wuhler, Chang Shu, and Flavio Prieto. Fully automatic expression-invariant face correspondence. *Machine Vision and Applications*, 25(4):859–879, 2014.

- [143] Georgia Sandbach, Stefanos Zafeiriou, Maja Pantic, and Lijun Yin. Static and dynamic 3D facial expression recognition: A comprehensive survey. *Image and Vision Computing*, 30(10):683–697, 2012.
- [144] Shelley R Saunders, Frank Popovich, and Gordon W Thompson. A family study of craniofacial dimensions in the Burlington growth centre sample. *American journal of orthodontics*, 78(4):394–403, 1980.
- [145] Irene Savoye, Ruth Loos, Carine Carels, Catherine Derom, and Robert Vlietinck. A genetic study of anteroposterior and vertical facial proportions using model-fitting. *The Angle Orthodontist*, 68(5):467–470, 1998.
- [146] Arman Savran, Neşe Alyüz, Hamdi Dibeklioglu, Oya Çeliktutan, Berk Gökberk, Bülent Sankur, and Lale Akarun. Bosphorus database for 3D face analysis. In *European Workshop on Biometrics and Identity Management*, pages 47–56. Springer, 2008.
- [147] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *Artificial Neural Networks ICANN’97*, pages 583–588. Springer, 1997.
- [148] John Shawe-Taylor and Nello Cristianini. *Kernel methods for pattern analysis*. Cambridge university press, 2004.
- [149] Michael J Sheehan and Michael W Nachman. Morphological and population genomic evidence that human faces have evolved to signal individual identity. *Nature communications*, 5, 2014.
- [150] Richard J Sherwood, Dana L Duren, Ellen W Demerath, Stefan A Czerwinski, Roger M Siervogel, and Bradford Towne. Quantitative genetics of modern human cranial variation. *Journal of human evolution*, 54(6):909–914, 2008.
- [151] Xiaoyan Shi, Martin Styner, Jeffrey Lieberman, Joseph G Ibrahim, Weili Lin, and Hongtu Zhu. Intrinsic regression models for manifold-valued data. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2009*, pages 192–199. Springer, 2009.

- [152] Bryan P Shumaker and Roger W Sinnott. Astronomical computing: 1. computing under the open sky. 2. virtues of the haversine. *Sky and telescope*, 68:158–159, 1984.
- [153] Aaron Sim, Dimosthenis Tsagkraloulis, and Giovanni Montana. Random forests on distance matrices for imaging genetics studies. *Statistical applications in genetics and molecular biology*, 12(6):757–786, 2013.
- [154] David A Sinclair and Philipp Oberdoerffer. The ageing epigenome: Damaged beyond repair? *Ageing research reviews*, 8(3):189–198, 2009.
- [155] Torstein Sjøvold. A report on the heritability of some cranial measurements and non-metric traits. In *Multivariate statistical methods in physical anthropology*, pages 223–246. Springer, 1984.
- [156] Christopher Small. The statistical theory of shape. 1996.
- [157] Zhangfa Song, Guido Von Figura, Yan Liu, Johann M Kraus, Chad Torrice, Patric Dillon, Masami Rudolph-Watabe, Zhenyu Ju, Hans A Kestler, Hanna Sanoff, et al. Lifestyle impacts on the aging-associated expression of biomarkers of dna damage and telomere dysfunction in human blood. *Aging cell*, 9(4):607–615, 2010.
- [158] Corey S Sparks and Richard L Jantz. A reassessment of human cranial plasticity: Boas revisited. *Proceedings of the National Academy of Sciences*, 99(23):14636–14639, 2002.
- [159] Florian Steinke, Matthias Hein, and Bernhard Schölkopf. Nonparametric regression between general riemannian manifolds. *SIAM Journal on Imaging Sciences*, 3(3):527–563, 2010.
- [160] Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC bioinformatics*, 9(1):307, 2008.
- [161] Vitaly Surazhsky, Tatiana Surazhsky, Danil Kirsanov, Steven J Gortler, and Hugues Hoppe. Fast exact and approximate geodesics on meshes. In *ACM transactions on graphics (TOG)*, volume 24, pages 553–560. ACM, 2005.

- [162] Charles Susanne. Genetic and environmental influences on morphological characteristics. *Annals of Human Biology*, 2(3):279–287, 1975.
- [163] Charles Susanne. Heritability of anthropological characters. *Human Biology*, pages 573–580, 1977.
- [164] Zhenyu Tang, Di Jiang, and Yong Fan. Image registration based on dynamic directed graphs with groupwise image similarity. In *Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on*, pages 492–495. IEEE, 2013.
- [165] Joshua B Tenenbaum, Vin de de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319+, 2000.
- [166] Demetri Terzopoulos and Manuela Vasilescu. Sampling and reconstruction with adaptive meshes. In *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR’91., IEEE Computer Society Conference on*, pages 70–75. IEEE, 1991.
- [167] Hans Henrik Thodberg. Minimum description length shape and appearance models. In *Information Processing in Medical Imaging*, pages 51–62. Springer, 2003.
- [168] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.
- [169] Ioannis Tsochantaridis, Thomas Hofmann, Thorsten Joachims, and Yasemin Altun. Support vector machine learning for interdependent and structured output spaces. In *Proceedings of the twenty-first international conference on Machine learning*, page 104. ACM, 2004.
- [170] Laurens Van der Maaten, Eric Postma, and HJ Van Den Herik. Dimensionality reduction: A comparative review. *Journal of Machine Learning Research*, 10:1–41, 2009.
- [171] Oliver Van Kaick, Hao Zhang, Ghassan Hamarneh, and Daniel Cohen-Or. A survey on shape correspondence. In *Computer Graphics Forum*, volume 30, pages 1681–1707. Wiley Online Library, 2011.

- [172] Steven G Vandenberg and Herluf H Strandkov. A comparison of identical and fraternal twins on some anthropometric measures. *Human biology*, pages 45–52, 1964.
- [173] Vipin Vijayan, Kevin Bowyer, and Patrick Flynn. 3D twins and expression challenge. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 2100–2105. IEEE, 2011.
- [174] Grace Wahba. Multivariate function and operator estimation, based on smoothing splines and reproducing kernels. In *Santa Fe Institute studies in the sciences of complexity - Proceedings*, volume 12, pages 95–95. Addison -Wesley Publishing CO, 1992.
- [175] Qian Wang, Liya Chen, Pew-Thian Yap, Guorong Wu, and Dinggang Shen. Groupwise registration based on hierarchical image clustering and atlas synthesis. *Human brain mapping*, 31(8):1128–1140, 2010.
- [176] Sen Wang, Yang Wang, Miao Jin, Xianfeng David Gu, and Dimitris Samaras. Conformal geometry and its applications on 3D shape matching, recognition, and stitching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(7):1209–1220, 2007.
- [177] Xu Wang. *Searching, Clustering and Regression on non-Euclidean Spaces*. PhD thesis, UNIVERSITY OF MINNESOTA, 2015.
- [178] Xu Wang and Gilad Lerman. Nonparametric bayesian regression on manifolds via brownian motion. *arXiv preprint arXiv:1507.06710*, 2015.
- [179] Yang Wang, Mohit Gupta, Song Zhang, Sen Wang, Xianfeng Gu, Dimitris Samaras, and Peisen Huang. High resolution tracking of non-rigid motion of densely sampled 3D data using harmonic maps. *International Journal of Computer Vision*, 76(3):283–300, 2008.
- [180] Seth M Weinberg, Trish E Parsons, Mary L Marazita, and Brion S Maher. Heritability of face shape in twins: a preliminary study using 3D stereophotogrammetry and geometric morphometrics. *Dentistry 3000*, 1(1), 2013.
- [181] Christopher J Westland. Lower bounds on sample size in structural equation modeling. *Electronic Commerce Research and Applications*, 9(6):476–487, 2010.

- [182] Jason Weston, Olivier Chapelle, Vladimir Vapnik, André Elisseeff, and Bernhard Schölkopf. Kernel dependency estimation. In *Advances in neural information processing systems*, pages 873–880, 2002.
- [183] Christopher Williams and Matthias Seeger. The effect of the input density distribution on kernel-based classifiers. In *Proceedings of the 17th international conference on machine learning*, pages 1159–1166, 2000.
- [184] Andrew P Witkin and Paul S Heckbert. Using particles to sample and control implicit surfaces. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, pages 269–277. ACM, 1994.
- [185] Daniela Witten, Rob Tibshirani, Sam Gross, Balasubramanian Narasimhan, and Maintainer Daniela Witten. Package PMA. *Genetics and Molecular Biology*, 8(1):28, 2013.
- [186] Daniela M Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, page kxp008, 2009.
- [187] Naomi Wray and P Visscher. Estimating trait heritability. *Nature Education*, 1(1):29, 2008.
- [188] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J Rosato. A 3D facial expression database for facial behavior research. In *Automatic face and gesture recognition, 2006. FGR 2006. 7th international conference on*, pages 211–216. IEEE, 2006.
- [189] Yun Zeng, Chaohui Wang, Xianfeng Gu, Dimitris Samaras, and Nikos Paragios. A generic deformation model for dense non-rigid surface registration: A higher-order MRF-based approach. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3360–3367, 2013.
- [190] Yun Zeng, Chaohui Wang, Xianfeng Gu, Dimitris Samaras, and Nikos Paragios. Higher-order graph principles towards Non-Rigid surface registration. *IEEE transactions on pattern analysis and machine intelligence*, 38(12):2416–2429, 2016.

- [191] Chao Zhang, William AP Smith, Arnaud Dessein, Nick Pears, and Hang Dai. Functional faces: Groupwise dense correspondence using functional maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5033–5041, 2016.
- [192] Junli Zhao, Cuiting Liu, Zhongke Wu, Fuqing Duan, Minqi Zhang, Kang Wang, and Taorui Jia. 3D facial similarity measure based on geodesic network and curvatures. *Mathematical Problems in Engineering*, 2014, 2014.
- [193] Hongtu Zhu, Yasheng Chen, Joseph G Ibrahim, Yimei Li, Colin Hall, and Weili Lin. Intrinsic regression models for positive-definite matrices with applications to diffusion tensor imaging. *Journal of the American Statistical Association*, 104(487), 2009.
- [194] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.
- [195] Syed Zulqarnain Gilani, Faisal Shafait, and Ajmal Mian. Shape-based automatic detection of a large number of 3D facial landmarks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4639–4648, 2015.

Appendix A

Geodesic Correspondence Sampling for Ensembles of Polyhedral Surfaces

A.1 Alternative formulation for Point-Wise Correspondence Optimization

Here we present an alternative formulation for the Point-Wise Correspondence Optimization (see section 3.3.3). The problem (3.15) remains the same. Similarly, let $\Phi \in \mathbb{R}^{3M}$ be a normally distributed random variable in shape space. This time, though, we utilize the pseudo-determinant of the covariance matrix in the density function, resulting in the differential entropy

$$\Gamma(\Phi) = \frac{1}{2} \ln\{(2\pi e)^{3M} |\Sigma_\Phi|_+\} . \quad (\text{A.1})$$

The only difference between equations (3.16) and (A.1) is the use of the pseudo-determinant $|\Sigma_\Phi|_+$, which is defined as $\lim_{k \rightarrow 0} \frac{|\Sigma_\Phi + k\mathbf{I}|}{k^{3M - \text{rank}(\Sigma_\Phi)}}$ [113]. The pseudo-determinant equals the product of all non-zero eigenvalues of the matrix [113]. If the covariance matrix is full-rank, then the pseudo-determinant is the same as the standard determinant. Substituting the sample

covariance estimator (3.17), we get

$$\hat{\Gamma}(\phi_1, \dots, \phi_N) = \frac{1}{2} \ln \{ (2\pi e)^{3M} |\hat{\Sigma}_\Phi|_+ \} = \frac{1}{2} \ln \left| \frac{1}{N-1} \mathbf{\Omega}^T \mathbf{\Omega} \right|_+ + \frac{1}{2} \ln (2\pi e)^{3M} , \quad (\text{A.2})$$

with $\mathbf{\Omega}$ being again the $N \times 3M$ matrix of centered observations.

Let now $\mathbf{\Omega} = \mathbf{U}_\Omega \mathbf{S}_\Omega \mathbf{V}_\Omega^T$ be the singular value decomposition of $\mathbf{\Omega}$, with $\mathbf{U}_\Omega, \mathbf{V}_\Omega^T$ being $N \times N$ and $3M \times 3M$ unitary matrices, and $\mathbf{S}_\Omega$ the $N \times 3M$ rectangular diagonal matrix of singular values. We can then write

$$\begin{aligned} \mathbf{\Omega}^T \mathbf{\Omega} &= \mathbf{V}_\Omega \mathbf{S}_\Omega^T \mathbf{S}_\Omega \mathbf{V}_\Omega^T \\ \mathbf{\Omega} \mathbf{\Omega}^T &= \mathbf{U}_\Omega^T \mathbf{S}_\Omega \mathbf{S}_\Omega^T \mathbf{U}_\Omega . \end{aligned} \quad (\text{A.3})$$

The pseudo-determinants of the above are

$$\begin{aligned} |\mathbf{\Omega}^T \mathbf{\Omega}|_+ &= |\mathbf{V}_\Omega|_+ |\mathbf{S}_\Omega^T \mathbf{S}_\Omega|_+ |\mathbf{V}_\Omega^T|_+ = |\mathbf{S}_\Omega^T \mathbf{S}_\Omega|_+ \\ |\mathbf{\Omega} \mathbf{\Omega}^T|_+ &= |\mathbf{U}_\Omega^T|_+ |\mathbf{S}_\Omega \mathbf{S}_\Omega^T|_+ |\mathbf{U}_\Omega|_+ = |\mathbf{S}_\Omega \mathbf{S}_\Omega^T|_+ . \end{aligned} \quad (\text{A.4})$$

It is easy to verify that $|\mathbf{S}_\Omega^T \mathbf{S}_\Omega|_+$ and $|\mathbf{S}_\Omega \mathbf{S}_\Omega^T|_+$, and consequently $|\mathbf{\Omega}^T \mathbf{\Omega}|_+$ and $|\mathbf{\Omega} \mathbf{\Omega}^T|_+$, are equal, since both are the products of the squared non-zero singular values of $\mathbf{\Omega}$. By utilizing this result, we can now write the objective function as

$$\hat{\Gamma}(\phi_1, \dots, \phi_N) = \frac{1}{2} \ln \left| \frac{1}{N-1} \mathbf{\Omega} \mathbf{\Omega}^T \right|_+ + \frac{1}{2} \ln (2\pi e)^{3M} . \quad (\text{A.5})$$

Following the same assumptions as in section 3.3.3, The matrix of partial derivatives of $\hat{\Gamma}$ is

$$\begin{aligned} \frac{d\hat{\Gamma}}{d\mathbf{\Phi}^T} &\approx \frac{d\hat{\Gamma}}{d\mathbf{\Omega}^T} \\ &\approx \frac{1}{2(N-1)} \frac{d}{d\mathbf{\Omega}^T} \ln |(\mathbf{\Omega}^T)^T \mathbf{\Omega}^T|_+ \\ &\approx \frac{1}{(N-1)} ((\mathbf{\Omega}^T)^+)^T \\ &\approx \frac{1}{N-1} ((\mathbf{\Omega} \mathbf{\Omega}^T)^{-1} \mathbf{\Omega})^T \\ &\approx \frac{1}{N-1} \mathbf{\Omega}^T (\mathbf{\Omega} \mathbf{\Omega}^T)^{-1} . \end{aligned} \quad (\text{A.6})$$

In comparison with the derivative matrix of (3.19), in (A.6) we have the inverse of the gramian matrix $\mathbf{\Omega}\mathbf{\Omega}^T$, with computational complexity $\mathcal{O}((N)^{2.373})$, instead of $\mathcal{O}((3M)^{2.373})$, as in the original formulation. The use of the transpose $\mathbf{\Omega}^T$ for the matrix derivatives does not affect the formulation, since partial updates are subsequently grouped into 3-dimensional vectors, corresponding to surface points ψ_j^i . We again assumed that the sample grammian matrix is invertible. A regularization term can be added, as in (3.20), to guarantee its non-singularity. Subsequent steps are identical to those described in section 3.3.3.

A.2 Sample Variability of Curvature Phenotypes in the TwinsUK Dataset

In order to examine the sample variability of the curvature traits, curvature variance maps, depicted in Fig. A.1, were constructed. Depending on the curvature index, various facial areas showed increased variability. High variance of the MC measurements was observed in areas such as eye sockets, ala of the nose and mouth. GC and CU variance was located mainly in the nose, eye and philtrum regions, while the SI captured increased variation in the zygomatic and mouth areas. It is important to notice that particular facial areas, namely the mouth and eye regions showed consistently high variability, which could relate to the increased motional ability of the specific structures. Irrespective of the curvature index used, average phenotypic variability was always higher in the subset of DZ, compared to the MZ pairs. As heritability is based on differences in similarity between MZ and DZ pairs [20], means and variances of absolute trait differences between pairs were also computed and the results were combined into the facial maps of Fig. A.2. It is clear that higher mean and variance values were measured in the DZ subset, compared to MZ.

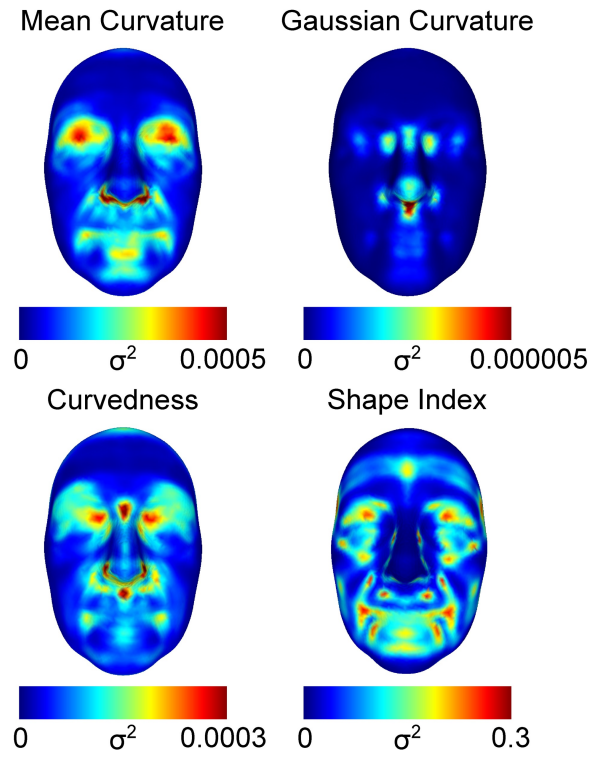


Figure A.1: Curvature variance maps illustrate the variance of curvature traits on all landmarks, computed from the full dataset of 952 TwinsUK subjects.

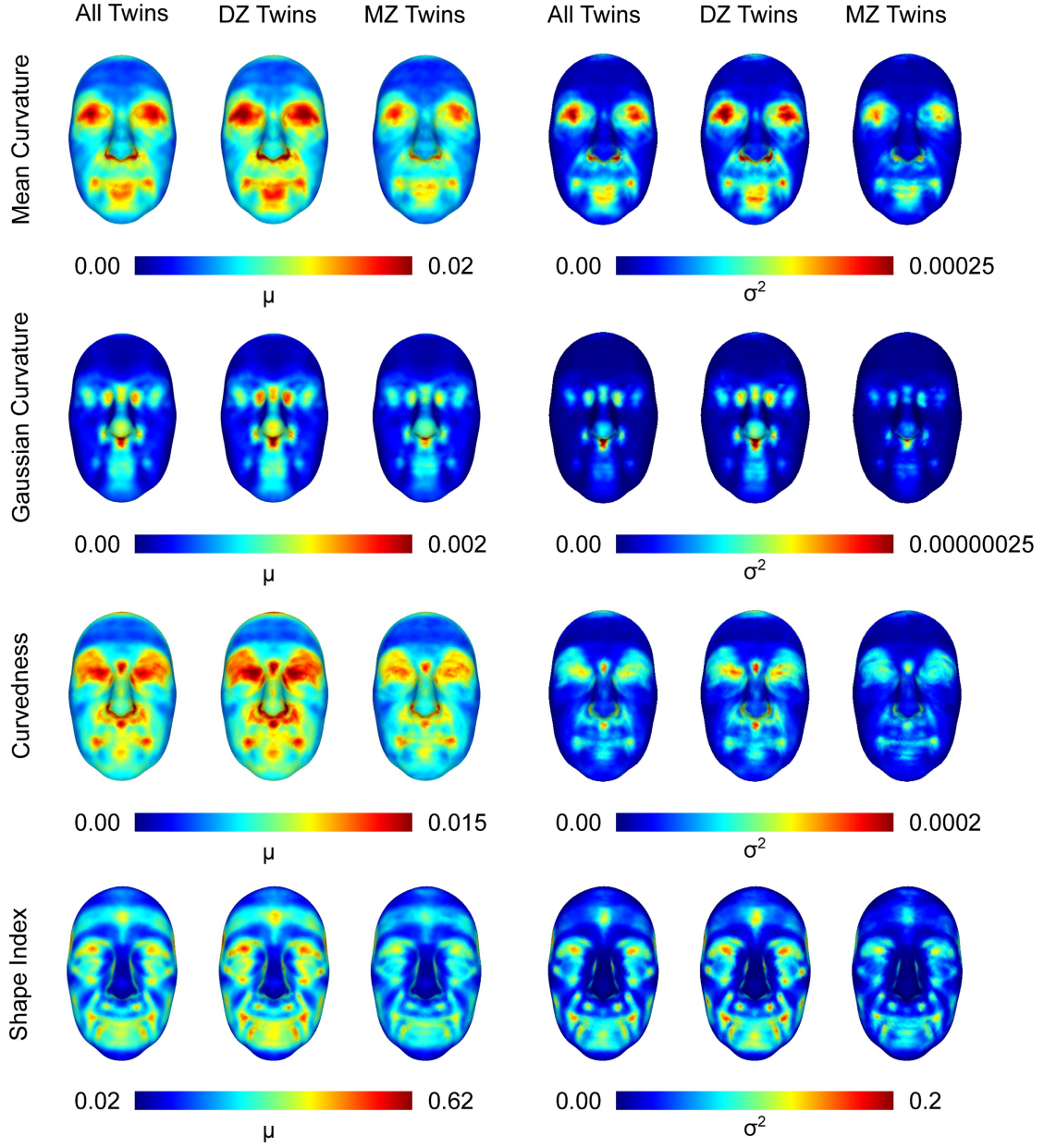


Figure A.2: Mean and variance maps of absolute curvature differences between twins. For each pair of twins, the absolute difference in curvature values was computed on all facial points. The maps show the mean and variance of the differences for all twins, as well as for MZ and DZ subsets.

A.3 Structural Equation Modeling

In this work, we estimated heritability as the proportion of phenotypic variance explained by genetic factors. Since the genetic and environmental variables are unobserved (latent), their effects are inferred from twin resemblance using Structural Equation Modeling (SEM). SEM encompasses a broad family of statistical modeling techniques and can be viewed as a combination of path and latent factor analysis [80]. SEM is also widely referred to as covariance structure modeling, since an SE model implies a structure for the covariances between observed variables.

In heritability studies, observed phenotypic variation can be partitioned into variance components from the following latent factors: additive (A) genetic, dominant (D) genetic, common (C) environmental and unique (E) environmental, with the latter component also including measurement error. A path diagram of a SEM including all of the above latent factors can be seen in Fig. A.3. The structure of covariance is implied by latent factor correlations between twin pairs. Additive and dominant genetic effects are correlated 1 between MZ pairs, while only 0.5 and 0.25 respectively between DZ pairs. Common environmental effects have correlation 1 for both types of twins while unique environmental factors are uncorrelated.

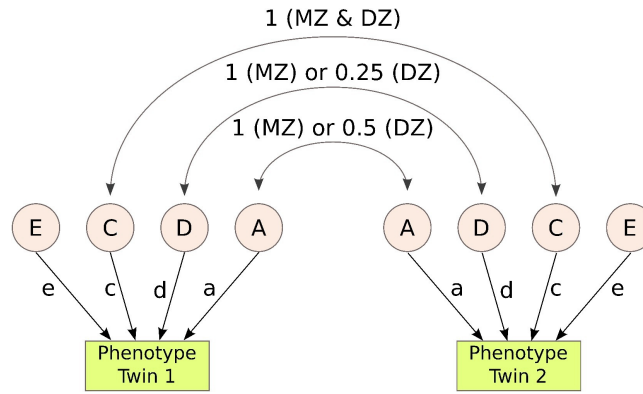


Figure A.3: Path diagram of the ACDE twin Structural Equation Model. Latent factors A, D, C and E correspond to additive and dominant genetic, common and unique environmental effects respectively. Double arrows represent the latent factor correlations between pairs of twins. Additive and dominant genetic effects are correlated 1 for MZ pairs, 0.5 and 0.25 respectively for DZ pairs. Common environmental effects have correlation 1 for both types of twins while unique environmental factors are uncorrelated. a , d , c and e are regression path coefficients of the respective latent factors. Heritability in a ACDE SEM model is computed as

$$h_{ACDE}^2 = \frac{a^2 + d^2}{a^2 + d^2 + c^2 + e^2}.$$

For twin studies in particular, the model ADCE is over-specified and cannot be estimated using twin data alone [133]. Submodels that can be examined are ACE, ADE, AE, CE and E. Models composed of the components DE are not considered biologically plausible. In this work, we considered ACE, AE and E models. As such, below we focus and describe in detail the definition of the ACE SE model.

A univariate ACE model can be expressed as

$$P = aA + cC + eE, \quad (\text{A.7})$$

where for simplicity and without loss of information, P is an observed zero-centered, continuous phenotypic variable and A , C , E are unobserved latent factors with fixed unit variances and covariances that depend on the type of twin. Finally, a , c , e are regression coefficients expressing the effects of the latent variables in the phenotype. Now Let $\mathbf{P}_{MZ}$ be a $N_{MZ} \times 2$ matrix of phenotypic observations with each row coming from one pair of MZ twins and N_{MZ} the number of MZ paired observations. Respectively $\mathbf{P}_{DZ}$ denotes a $N_{DZ} \times 2$ matrix of phenotypic observations with each row coming from one pair of DZ twins. Furthermore let

$$\mathbf{L} = \begin{bmatrix} a & 0 \\ c & 0 \\ e & 0 \\ 0 & a \\ 0 & c \\ 0 & e \end{bmatrix}, \quad (\text{A.8})$$

be the matrix of regression coefficients and finally $\mathbf{\Lambda}_{MZ}$, $\mathbf{\Lambda}_{DZ}$ be $N_{MZ} \times 6$ and $N_{DZ} \times 6$ matrices of unobserved A, C, E factors for MZ and DZ twin subsets respectively.

Expressing the ACE twin model in matrix notation, we have

$$\mathbf{P}_{MZ} = \mathbf{\Lambda}_{MZ}\mathbf{L}, \quad \mathbf{P}_{DZ} = \mathbf{\Lambda}_{DZ}\mathbf{L}. \quad (\text{A.9})$$

The expected phenotypic covariances from the ACE model are:

$$\begin{aligned}\Sigma_{MZ} &= E[\mathbf{P}_{MZ}^T \mathbf{P}_{MZ}] = \mathbf{L}^T E[\Psi_{MZ}] \mathbf{L} \\ \Sigma_{DZ} &= E[\mathbf{P}_{DZ}^T \mathbf{P}_{DZ}] = \mathbf{L}^T E[\Psi_{DZ}] \mathbf{L},\end{aligned}\tag{A.10}$$

where the expectations of the correlation matrices Ψ_{MZ} , Ψ_{DZ} for the latent factors are derived from the SEM path diagram. In particular, $E[\Psi_{MZ}]$ has two off diagonal elements equal to 1, corresponding to $\text{corr}(A_1, A_2)$ and $\text{corr}(C_1, C_2)$, while $E[\Psi_{DZ}]$ has $\text{corr}(A_1, A_2) = 0.5$ and $\text{corr}(C_1, C_2) = 1$ (see Figure A.3).

The structured covariance matrices, as modeled by SEM, can now be easily computed:

$$\Sigma_{MZ} = \begin{bmatrix} a^2 + c^2 + e^2 & a^2 + c^2 \\ a^2 + c^2 & a^2 + c^2 + e^2 \end{bmatrix}\tag{A.11}$$

$$\Sigma_{DZ} = \begin{bmatrix} a^2 + c^2 + e^2 & 0.5a^2 + c^2 \\ 0.5a^2 + c^2 & a^2 + c^2 + e^2 \end{bmatrix}.\tag{A.12}$$

Maximum Likelihood is used to estimate the regression coefficients a , c , e . Let $\mathbf{S}_{MZ}$ and $\mathbf{S}_{DZ}$ be the observed sample covariances. Assuming that the phenotypic response is normally distributed, the probabilities of observing $\mathbf{S}_{MZ}$ and $\mathbf{S}_{DZ}$ given estimates $\hat{\Sigma}_{MZ}$ and $\hat{\Sigma}_{DZ}$ follow the Wishart distribution with N_{MZ} and N_{DZ} degrees of freedom respectively. The log-likelihood functions can be written as follows, after the omission of constant terms [109]:

$$\begin{aligned}-2LL_{MZ} &\approx N_{MZ} \left[\ln |\hat{\Sigma}_{MZ}| + \text{tr}(\hat{\Sigma}_{MZ}^{-1} \mathbf{S}_{MZ}) \right] \\ -2LL_{DZ} &\approx N_{DZ} \left[\ln |\hat{\Sigma}_{DZ}| + \text{tr}(\hat{\Sigma}_{DZ}^{-1} \mathbf{S}_{DZ}) \right].\end{aligned}\tag{A.13}$$

Estimates $\hat{a}, \hat{c}, \hat{e}$ are obtained by maximizing the combined likelihood function $-2(LL_{MZ} + LL_{DZ})$. Model fit can be assessed using a log-likelihood ratio test between the structured model and a fitted saturated model where no structure is imposed on the covariances. The ratio statistic is distributed approximately as a chi-squared distribution with degrees of freedom (df)

equal the difference in df between the structured and the saturated model.

At this point we can also define the heritability estimate $\hat{h}_{ACE}^2$ from the ACE model as

$$\hat{h}_{ACE}^2 = \frac{\hat{a}^2}{\hat{a}^2 + \hat{c}^2 + \hat{e}^2}. \quad (\text{A.14})$$

An important aspect of SEM in twin studies is that the significance of individual variance components can be assessed by dropping parameters sequentially from nested models; here ACE→AE→E. In choosing between models, variance components are excluded in the selection process if there is no significant deterioration in model fit, assessed commonly by the Akaike Information Criterion (AIC) [1], after a component is dropped. The E component represents unique environmental results and random error and is always retained [133]. Heritability is estimated from the AE model as

$$\hat{h}_{AE}^2 = \frac{\hat{a}^2}{\hat{a}^2 + \hat{e}^2}. \quad (\text{A.15})$$

A.4 Average SEM Statistics for the Heritability Study

Average SEM statistics for the curvature-based and distance-based heritability studies of Chapter 3.5 are shown in tables A.1 and A.2 respectively.

		Univariate Curvature Phenotypes				Regional Curvature Phenotypes			
		MC	GC	CU	SI	MC	GC	CU	SI
ACE	df	5	5	5	5	6	6	6	6
	χ^2	-0.376	12.348	-2.385	-0.078	9.552	10.263	9.25	10.927
	p	0.573	0.442	0.576	0.603	0.29	0.273	0.29	0.251
	$-2\log L$	2169.791	-4186.128	1919.943	6730.955	4850.608	4859.602	4862.932	4868.236
	AIC	2179.791	-4176.128	1929.943	6740.955	4858.608	4867.602	4870.932	4876.236
AE	df	6	6	6	6	7	7	7	7
	χ^2	0.15	12.954	-1.99	0.274	10.029	11.062	9.681	11.254
	p	0.592	0.459	0.599	0.63	0.326	0.303	0.33	0.292
	$-2\log L$	2170.317	-4185.521	1920.339	6731.308	4851.084	4860.402	4863.362	4868.563
	AIC	2178.317	-4177.521	1928.338	6739.308	4857.084	4866.402	4869.362	4874.563
E	df	7	7	7	7	8	8	8	8
	χ^2	70.223	66.951	56.491	41.964	99.493	91.209	86.867	83.24
	p	0.0195	0.03	0.073	0.111	0.002	0.002	0.002	0.01
	$-2\log L$	2240.391	-4131.524	1978.82	6772.997	4940.549	4940.549	4940.549	4940.549
	AIC	2246.391	-4125.524	1984.82	6778.997	4944.549	4944.549	4944.549	4944.549

Table A.1: Model-fitting average statistics for the curvature-based heritability studies.

		Euclidean Distance Traits	Geodesic Distance Traits
ACE	df	6	6
	χ^2	13.458	15.757
	p	0.1941	0.1678
	$-2 \log L$	4775.921	4834.57
	AIC	2879.921	2938.569
AE	df	7	7
	χ^2	16.4	17.444
	p	0.1489	0.184
	$-2 \log L$	4778.863	4836.256
	AIC	2878.862	2938.256
E	df	8	8
	χ^2	179.8	163.724
	p	0	0
	$-2 \log L$	4942.263	4982.537
	AIC	3042.263	3082.536

Table A.2: Model-fitting average statistics for the distance-based heritability study.

A.5 Sparse Principal Component Analysis

Different sparse PCA methods have been presented in the literature [194, 186, 89]. In this work, we used Penalized Matrix Decomposition (PMD), as proposed in [186], which has been shown to be similar but more computationally efficient than the SCoTLASS sPCA formulation [186, 89].

Without loss of generality, let $\mathbf{P}$ be a column-wise zero-mean $N \times M$ data matrix. Standard PCA seeks unit loading vectors $\mathbf{v}_k$ so that linear transformations - principal components - $\mathbf{P}\mathbf{v}_k$ have successively maximum variance. The first PC loading vector is thus computed as

$$\mathbf{v}_1 = \arg \max_{\mathbf{v}} \mathbf{v}^T \mathbf{P}^T \mathbf{P} \mathbf{v}, \text{ s.t. } \mathbf{v}^T \mathbf{v} = 1 \quad (\text{A.16})$$

Consequent loading vectors can be computed by repeating the same process on the deflated data matrices. Given $\mathbf{P}_k$ and $\mathbf{v}_k$, the deflated data matrix $\mathbf{P}_{k+1} = \mathbf{P}_k - \mathbf{P}_k \mathbf{v}_k \mathbf{v}_k^T$, with $\mathbf{P}_1 = \mathbf{P}$.

The PDM and SCoTLASS procedures for sPCA modify the optimization problem (A.16) with an additional L_1 regularization constraint on the loading vectors. It has been shown that for the first PC, both methods are equivalent to the following penalized matrix decomposition problem [186]:

$$\mathbf{v}_1 = \arg \max_{\mathbf{u}, \mathbf{v}} \mathbf{u}^T \mathbf{P} \mathbf{v}, \text{ s.t. } \|\mathbf{v}\|_1 \leq s_p, \|\mathbf{v}\|_2^2 \leq 1, \|\mathbf{u}\|_2^2 \leq 1, \quad (\text{A.17})$$

where s_p is the sparsity parameter, with lower values leading to sparser loading vectors $\mathbf{v}$. The above problem is biconvex and can be optimized by iteratively alternating between maximization with respect to $\mathbf{u}$ and $\mathbf{v}$. SCoTLASS imposes orthogonality constraints between subsequent loading vectors, which makes optimization very difficult. PMD does not utilize such constraints. Subsequent components in PMD are computed by applying the same procedure on the deflated data matrices.

A.6 Tuning the sPCA Sparsity Parameter

We opted to use sPCA since a main objective in our decomposition analysis was to construct composite phenotypes corresponding to spatially coherent facial areas, something that could not be achieved through standard PCA. This coherency was imposed by controlling the sparsity parameter s_p . Tuning the parameter is commonly performed through cross validation, by selecting the value that leads to minimum average CV reconstruction error of the data [185]. This process though could have led to spatially extended loading vectors for which biological interpretation would be difficult. Furthermore, the parameter would need tuning for each principal component independently, which would be computationally expensive.

Since our primary objective was to estimate the heritability of principal components, we evaluated the effect of sparsity by comparing heritability estimates of 100 sPCs for each curvature descriptor and 7 different parameter values. Results are presented in Fig. A.4. Heritability estimates for all components were subsequently computed. We notice that the sorted heritability estimates show similar behavior across curvature descriptors for all s_p values tested.

Based on the fact that heritability estimation would not be significantly affected by the particular value of s_p , we selected constant values for each descriptor after visual inspection of eigenface maps - constructed by mapping the sPCs' loadings on the facial surface - with the primary focus being to have eigenmaps suitable for biological interpretation. In particular, the parameter was set to 7.5 for GC, 12.5 for SI and 15 for MC and CU.

We retained for further analysis all 100 sPCs for each curvature descriptor.

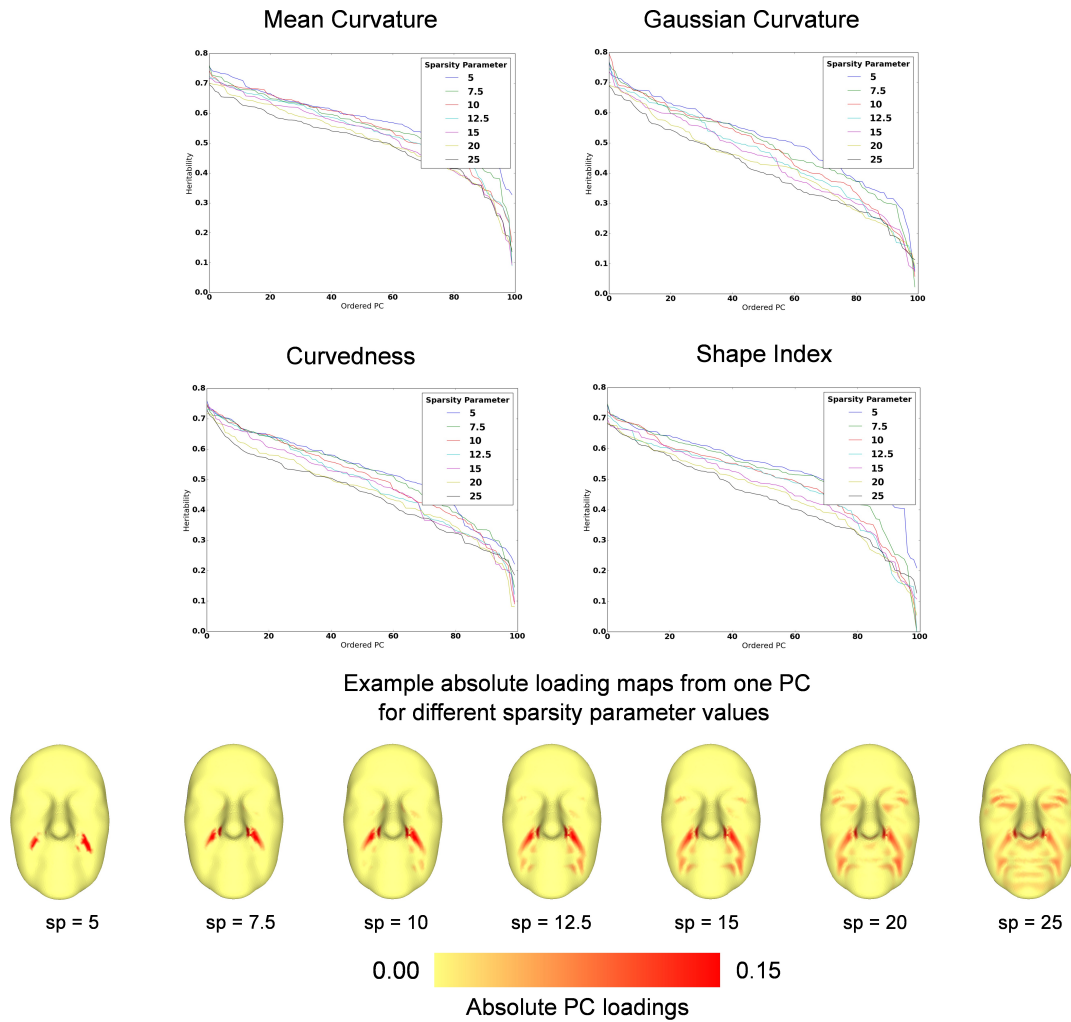


Figure A.4: Effect of sparsity parameter on the heritability of regional traits. (Top) Different sparsity parameters were tested in order to assess the parameter's contribution on the heritability study. Each plot shows the sorted heritability values for the first 100 regional traits per curvature index and for 7 different sparsity parameter values. Heritability estimates showed similar behavior irrespective of how sparse the components were. (Bottom) Eigenface maps of an example composite trait (from the Mean curvature index) for different sparsity parameter values.

Appendix B

Application of Adaptive GESSA on the Prediction of Age from 3D Facial Surfaces

B.1 Feature Selection through Stability Selection for the various landmarking and phenotyping procedures

After computing the eight different trait matrices used as inputs for the problem of age prediction (see section 4.4.4) , we examined whether a feature selection process would result in helpful remarks regarding the comparison between uniform and adaptive landmark approaches. For this part of our study, we employed a Stability Selection (SL) algorithm. The method builds a large number of LLR models on random subsamples of the initial dataset, and yields the percentage of times that each input trait was considered in the models (had a non-zero coefficient associated to it) [111]. For each dataset, we run SL for 10000 rounds, with the size of the randomized subsamples set to 75% of the initial sample size, and the lasso sparsity parameter $\lambda_{lasso} = 0.015$. The SL percentage scores were mapped to the corresponding landmark positions in order to create the SL maps presented in Fig. B.1. We notice that stability maps of

respective curvature indices were very similar for both landmarking approaches, with somewhat increased localization of high values observed for the adaptive case. This was expected, since adapted landmarks were more congregated around age-related regions.

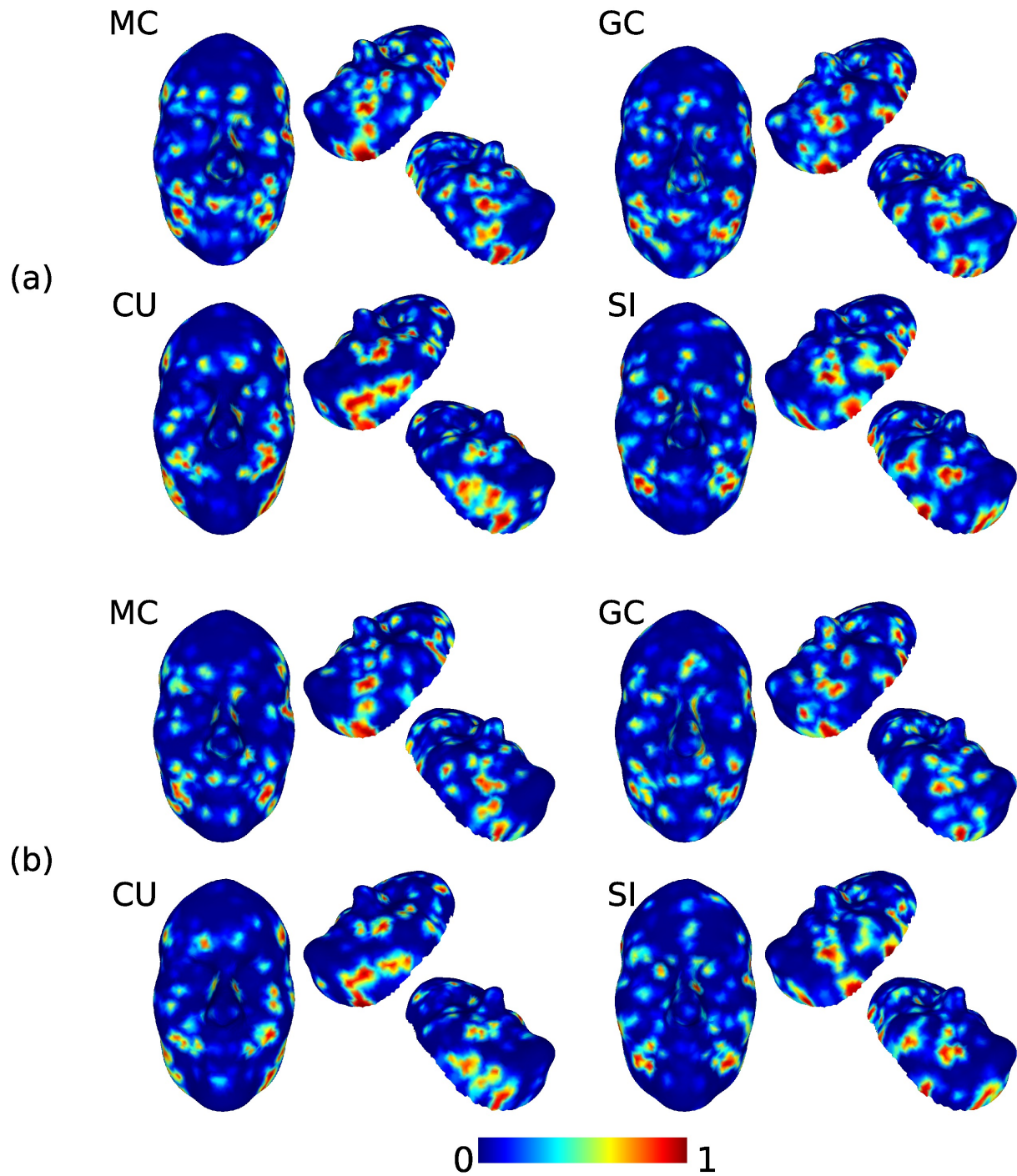


Figure B.1: Age Prediction - Stability Selection facial maps for (a) traits based on uniformly positioned corresponding points and (b) traits based on adaptively positioned corresponding points. Maps are colorized according to the percentage of times that curvature traits had non-zero coefficients in a group of lasso models, built on random subsamples of the data.

Appendix C

Non-parametric Predictive Methodology for Manifold-Valued Responses

C.1 Proof of Proposition 5.1

Proposition C.1. *Let $\{y_i | y_i \in \mathbb{R}^q, i = 1, \dots, N\}$ be observed points on a q dimensional Euclidean space. The sum of squared deviations from the mean $\bar{y} = \frac{1}{N} \sum_{i=1}^N y_i$ can be written in terms of pairwise distances between points as follows:*

$$\sum_{i=1}^N \|y_i - \bar{y}\|^2 = \frac{1}{2N} \sum_{i=1}^N \sum_{j=1}^N d_E^2(y_i, y_j) , \quad (\text{C.1})$$

where $d_E(\cdot, \cdot)$ denotes the Euclidean distance function.

Proof.

$$\begin{aligned}
\sum_{i=1}^N \|y_i - \bar{y}\|^2 &= \sum_{i=1}^N (y_i - \bar{y})^2 \\
&= \sum_{i=1}^N y_i^2 + N\bar{y}^2 - 2\bar{y} \sum_{i=1}^N y_i \\
&= \sum_{i=1}^N y_i^2 + N\bar{y}^2 - 2\bar{y}N\bar{y} \\
&= \sum_{i=1}^N y_i^2 - \frac{1}{N} \left(\sum_{i=1}^N y_i \right)^2 \\
&= \sum_{i=1}^N y_i^2 - \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N y_i y_j \\
&= \frac{2N \sum_{i=1}^N y_i^2 - 2 \sum_{i=1}^N \sum_{j=1}^N y_i y_j}{2N} \\
&= \frac{\sum_{i=1}^N \sum_{j=1}^N (y_i^2 + y_j^2) - \sum_{i=1}^N \sum_{j=1}^N 2y_i y_j}{2N} \\
&= \frac{1}{2N} \sum_{i=1}^N \sum_{j=1}^N (y_i - y_j)^2 \\
&= \frac{1}{2N} \sum_{i=1}^N \sum_{j=1}^N \|y_i - y_j\|^2 \\
&= \frac{1}{2N} \sum_{i=1}^N \sum_{j=1}^N d_E^2(y_i, y_j) .
\end{aligned} \tag{C.2}$$

□

C.2 Out-Of-Sample Prediction

We present here the details of the OOS prediction as proposed by Bengio et al. [15].

Let $\mathcal{Y}$ be a closed subset of $\mathbb{R}^q$, $n \in \mathbb{N}$, μ a strictly positive Borel measure on $\mathcal{Y}$, and $g : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ a positive definite kernel on $\mathcal{Y}$, meaning that for any finite set of points $\{y_i | y_i \in$

$\mathcal{Y}, i = 1, \dots, N\}$ and real numbers $\{a_i | a_i \in \mathbb{R}, i = 1, \dots, N\}$

$$\sum_{i=1}^N \sum_{j=1}^N a_i a_j g(y_i, y_j) \geq 0 . \quad (\text{C.3})$$

Assume also that

$$\int_{w \in \mathcal{Y}} g(y, w) d\mu(y) d\mu(w) < \infty . \quad (\text{C.4})$$

Let an induced integral operator $T_g : L_\mu^2(\mathcal{Y}) \rightarrow L_\mu^2(\mathcal{Y})$ be defined as

$$T_g f(y) = \int_{w \in \mathcal{Y}} g(y, w) f(w) d\mu(w) . \quad (\text{C.5})$$

T_g is a self-adjoint, positive, compact operator with a countable system of non-negative eigenvalues $\{\lambda_l\}_{l=1}^\infty$ satisfying $\sum_{l=1}^\infty \lambda_l^2 < \infty$, with corresponding normalized eigenfunctions $\{\phi_l\}_{l=1}^\infty$. The set of eigenfunctions forms an orthonormal basis of $L_\mu^2(\mathcal{Y})$.

Theorem C.2 (Mercer). *Let $\mathcal{Y}$ be a closed subset of $\mathbb{R}^q$, μ a strictly positive Borel measure on $\mathcal{Y}$, and $g : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ a positive definite kernel on $\mathcal{Y}$. Then*

$$g(y, w) = \sum_{l=1}^\infty \lambda_l \phi_l(y) \phi_l(w) . \quad (\text{C.6})$$

Let now $p(y)$ be a density function, corresponding to a measure in $\mathcal{Y}$. In [183], the authors showed that

$$\hat{g}(y, w) = \sum_{l=1}^m \lambda_l \phi_l(y) \phi_l(w) , \quad (\text{C.7})$$

where $\{\lambda_m | m = 1, \dots, m\}$ are the m largest eigenvalues with corresponding eigenfunctions $\{\phi_m | m = 1, \dots, m\}$, is the best approximation of $g(y, w)$ using only m terms, in the sense of minimizing the expected value of the kernel reconstruction error, given by

$$J(m) = \int_{y \in \mathcal{Y}} \int_{w \in \mathcal{Y}} \left(g(y, w) - \sum_{l=1}^m \lambda_l \phi_l(y) \phi_l(w) \right)^2 p(y) p(w) d(y) d(w) . \quad (\text{C.8})$$

Based on the above results, the following work was presented by [14].

Let $S = \{y_i | y_i \in \mathcal{Y} \subset \mathbb{R}^q, i = 1, \dots, N\}$ be a dataset of N observations sampled i.i.d. from a distribution with unknown density function $p(y)$, with $\hat{p}(y)$ denoting the empirical density from sample S .

Let $\mathbf{K}$ be the $N \times N$ matrix of kernel values $K_{ij} = g(y_i, y_j)$, and $\{z_i | z_i \in \mathbb{R}^m, i = 1, \dots, N\}$ be an embedding of points in S taken by decomposition of the matrix $\mathbf{K}$, with:

$$z_{il} = \sqrt{\lambda_l^K} u_{li}^\lambda, \quad l = 1, \dots, m, \quad i = 1, \dots, N, \quad (\text{C.9})$$

where $z_i = (z_{i1}, \dots, z_{im})$ and λ_l^K the l^{th} largest eigenvalue of $\mathbf{K}$ and u_{li}^λ the i^{th} coordinate of the l^{th} eigenvector of $\mathbf{K}$.

Finally let $T_g^N : L^2(\mathbb{R}^m) \rightarrow L^2(\mathbb{R}^m)$ be an induced operator, defined as follows

$$T_g^N f(y) = \frac{1}{N} \sum_{i=1}^N g(y, y_i) f(y_i) . \quad (\text{C.10})$$

Proposition C.3 ([14]). T_k^N has in its image $m < N$ eigenfunctions of the form

$$\phi_l^N(y) = \frac{N}{\lambda_l^K} \sum_{i=1}^N u_{li}^\lambda g(y, y_i) , \quad (\text{C.11})$$

with corresponding non-zero eigenvalues $\lambda_l^N = \frac{\lambda_l^K}{N}$.

Remark It was also shown in [14], that in the limit as $N \rightarrow \infty$, the eigenfunctions of T_g^N corresponding to non-zero eigenvalues converge to the corresponding eigenfunctions of T_g .

Proposition C.4 ([14]). The value of the eigenfunction ϕ_l^N at point y_i is equal to:

$$\phi_l^N(y_i) = \sqrt{\frac{N}{\lambda_l^K}} z_{il} . \quad (\text{C.12})$$

Remark From proposition C.4 it becomes clear that the coordinates of an embedded point z_i are simply the scaled values of the principal eigenfunctions of T_g^N evaluated at point y_i .

By combining the results from propositions C.3 and C.4, the embedding coordinates of a new point y_{new} can be estimated by the OOS formula (5.10).