DATA REGISTRATION AND FEATURE SELECTION IN
FUNCTIONAL REGRESSION: SOME METHODOLOGICAL AND
COMPUTATIONAL DEVELOPMENTS

A Dissertation in
Statistics
by
Tobia Boschi

© 2022 Tobia Boschi

Submitted in Partial Fulfillment
of the Requirements
for the Degree of
Doctor of Philosophy

May 2022
The dissertation of Tobia Boschi was reviewed and approved by the following:

Francesca Chiaromonte  
Professor, Department of Statistics  
Dorothy Foehr Huck and J. Lloyd Huck Chair in Statistics for the Life Sciences  
Dissertation Co-Adviser  
Chair of Committee

Matthew Reimherr  
Associate Professor, Department of Statistics  
Dissertation Co-Adviser

Murali Haran  
Professor and Head, Department of Statistics

Paul Medvedev  
Associate Professor, Department of Computer Science and Engineering  
Associate Professor, Department of Biochemistry and Molecular Biology

Marzia Angela Cremona  
Assistant Professor, Department of Operations and Decision Systems, Université Laval  
Special Member

Ephraim Hanks  
Associate Professor, Department of Statistics  
Chair of Graduate Studies
Abstract

Functional linear models leverage high-dimensional and complex data and present many open and fascinating theoretical and computational challenges. Advanced optimization techniques which involve smoothing and considering projective subspaces are essential to obtain valid estimates and reduce computational costs. This thesis focuses on two main aspects of functional regression, functional registration and feature selection, and contains three main research projects.

In the first one, we propose a new low-dimensional registration procedure that exploits the relationship between response and predictor in a function-on-function regression. In this context, Functional Covariance Components (FCC) provide a flexible and powerful tool to represent the data in a low-dimensional space, capturing the most meaningful modes of dependency between the two sets of curves.

In the second project, we first develop a new, highly-efficient algorithm to solve Group Elastic Net, which exploits the sparsity structure of the Augmented Lagrangian to reduce the computational burden. Next, taking advantage of the properties of Functional Principal Components, we extend our algorithm to the function-on-scalar feature selection framework, where a functional response is modeled against a huge number of potential scalar predictors.

Finally, we employ advanced functional data techniques along with other statistical tools to study the evolution of the COVID-19 epidemic in Italy - using massive amounts of data that we collect, pre-process, and curate from different public sources on the epidemic, as well as socio-demographic, infrastructural and environmental factors that may affect its unfolding.
# Table of Contents

List of Figures vii

List of Tables xviii

Acknowledgments xx

Chapter 1
   Introduction 1
      1.1 Functional Registration .................................. 1
      1.2 Functional Feature Selection .............................. 2
      1.3 Research applications and collaborations ................... 3

Chapter 2
   Covariance based low-dimensional registration for function-on-function regression 5
      2.1 Introduction ............................................. 5
      2.2 Background ............................................. 6
         2.2.1 Registration with warping functions ................. 7
         2.2.2 Registration to low dimensional spaces ............. 8
      2.3 FCC registration ...................................... 9
         2.3.1 Functional Covariance Components .................... 9
         2.3.2 Low-dimensional registration through FCC ........... 11
      2.4 Algorithms .......................................... 12
      2.5 Simulation study and AneuRisk data ....................... 15
         2.5.1 Simulation settings .................................. 15
         2.5.2 Simulation results .................................... 17
         2.5.3 AneuRisk data ...................................... 19
      2.6 Current developments .................................... 24
      2.7 FCC and Sufficient Dimension Reduction ................... 28

Chapter 3
   A Highly-Efficient Group Elastic Net Algorithm with an Application to Function-On-Scalar Regression 29
      3.1 Introduction .......................................... 29
3.2 Preliminaries ................................................................. 32
  3.2.1 Fenchel conjugate function and proximal operator of $\pi(B)$ 32
  3.2.2 Dual formulation and Augmented Lagrangian ..................... 33
3.3 Methodology .............................................................. 34
  3.3.1 SSNAL method .......................................................... 34
  3.3.2 Extension to function-on-scalar regression ....................... 38
  3.3.3 Solution path implementation ..................................... 39
3.4 Simulation study and INSIGHT data .................................. 40
  3.4.1 Simulation results ..................................................... 40
  3.4.2 INSIGHT study .......................................................... 43

Chapter 4
Functional data analysis characterizes the shapes of the first
COVID-19 epidemic wave in Italy ........................................... 45
4.1 Introduction ................................................................. 45
4.2 Results ....................................................................... 47
  4.2.1 Under-counting deaths .............................................. 47
  4.2.2 Two different epidemics .......................................... 49
  4.2.3 Ranking mortality curves ......................................... 51
  4.2.4 Local mobility and positivity as statistical predictors of mortality 52
  4.2.5 The role of socio-demographic, infrastructural and environmental factors ................................................................. 55
4.3 Discussion ................................................................. 59
4.4 Methods .................................................................... 60
  4.4.1 Data retrieval and pre-processing ................................ 60
    4.4.1.1 Scalar covariates ............................................. 61
  4.4.2 Multivariate analysis tools ....................................... 62
  4.4.3 Functional Data Analysis tools .................................. 63

Chapter 5
Conclusions and future work .............................................. 72
5.1 Functional Registration ................................................ 72
5.2 Functional Feature Selection ......................................... 73
5.3 Research applications and collaborations ............................ 74

Appendix A
Chapter 2 Supplementary Material .................................... 76

Appendix B
Chapter 3 Supplementary Material .................................... 86
  B.1 Proofs ................................................................. 86
    B.1.1 Proof of Proposition 1 ......................................... 86
    B.1.2 Proof of Proposition 2 ......................................... 87
    B.1.3 Proof of Proposition 3 ......................................... 87
List of Figures

2.1 Criteria used to aid in the choice of $K$ (horizontal axis), the number of covariance components used to compute the registration templates. The left-most and center-left panels show the cumulative sum of the first $K$ eigenvalues of $\Sigma_{xy}$ and the Total Phase Variation index in (2.7), rescaled so that 0 corresponds to $\text{TPV}(\text{PRO}_{\Sigma_{xy}}(x, 1), \text{PRO}_{\Sigma_{xy}}(y, 1))$ and 1 corresponds to the TPV of the misaligned curves. The center-right and right-most panels show the L1OPE associated with the curves aligned by $cc_{RS}$ and $cc_{H1}$, respectively – this too is rescaled so that 1 corresponds to the L1OPE of the misaligned curves. Blue and red lines correspond to results for two simulation scenarios, and the green line corresponds to results for the AneuRisk data.

2.2 Simulation results. $aln(x)$ and $aln(y)$ are the $H1$ distances between aligned and original curves, and L1OPE is leave-one-out $H1$ prediction error. On the horizontal axis “msl” represents the misaligned curves, followed by curves registered with different procedures – as indicated. y-axis: Values on the vertical axis have been rescaled between 0 and 1, where 0 corresponds to results obtained with the original curves.

2.3 $y^{pred}$ curves obtained with different regressions. “Original” indicates the regression performed on the original curves, and “misaligned” the one performed on the misaligned curves. In all other panels the regression is performed on curves registered with different procedures – as indicated. The two different colors indicate the two groups in each set of curves.

2.4 Curvature ($x$) and WSS1 ($y$) curves for an individual patient. Misaligned curves are drawn in grey, curves registered by $cc_{H1(2)}$ in light-blue line, and curves registered by $er$ in red.

2.5 misaligned $y$ (top left); $y$ registered by $cc_{H1(1)}$ (top center); misaligned $x$ (bottom left); $x$ registered by $cc_{H1(1)}$ (bottom center); $y^{pred}$ curves by $cc_{H1(1)}$ (top right) and by $cc_{H1(2)}$ (bottom right).
2.6 Preliminary results for the Milano Weather Station Data. We consider 10 curves for each variable; namely temperature (left column), wind speed (center column) and benzene (right column). For each variable, we display the original curves (first row), the projection on the first two principal components (second row), the template functions (third row) and the registered functions (fourth row).

3.1 Curves related to a simulation scenario described in Table 3.1 (second row, \( n = 500 \)). The left and center panels display a sample of 10 response curves and 10 linear model errors terms, respectively. The right panel depicts the true non-zeros coefficients curves (dashed lines) and the \( \text{fgen} \) de-biased estimates (solid lines).

3.2 Plots related to the INSIGHT study. The left panel displays the growth curves. The center and right panels depict values of the 10-fold \( \text{cv} \) Mean Squared Error and the e-bic, respectively, against \( c_\lambda \). These are obtained from \( \text{fgen} \) run with 3 different values of \( \alpha \); 0.8 (green line), 0.6 (red line), 0.4 (blue line). \( \text{fgen} \) estimates are de-biased prior to computing both criteria.

4.1 Mortality curves. (a) DPC (dashed) and ISTAT (solid) differential mortality curves (per 100,000 inhabitants) in four example regions; Lombardia, Veneto, Emilia Romagna and Campania. Curves are smoothed with splines, with degree of smoothing selected by generalized cross-validation (see Methods). ISTAT curves “take off” earlier and in some regions are as much as twice as high at their peak – possibly due to many COVID-19 deaths happening at home and/or not being recorded as such in hospitals, especially in the early stages of the epidemic. (b) MAX mortality curves (per 100,000 inhabitants) in the 20 Italian regions, before (top) and after (bottom) the shifts produced by \( \text{probKMA} \) run with \( K = 2 \). In the bottom panel, time is marked as a day number (as opposed to a date); this represents the region-specific time of the epidemic unfolding, and corresponds to actual time (starting on February 16 and ending on April 30) only for regions with no shifts, e.g., Lombardia). Curves are again smoothed with splines, with degree of smoothing selected by generalized cross-validation. Lombardia, Veneto, Emilia Romagna and Campania, also shown in (a), are highlighted in color. In all panels, vertical lines mark the dates of the national lock-down (March 9) and of the suspension of all nonessential production activities (March 23). In the bottom panel of (b) vertical lines still show these dates without shifts; stars on the curves mark the lock-down after the region specific shifts.
Characterizing two epidemics. (a) MAX mortality curves are shown in the top left panel with 65-day portions identified by probKMA with $K = 2$ in red (Group 1; "exponential" pattern) and blue (Group 2; "flat(tened)" pattern). The curve portions are shown again, this time aligned with each other and separated by group, in the bottom panels. Black lines indicate group averages. The shifts produced by probKMA are shown in the top right panel (motifs, groups and shifts for Group 1 are stable across data sets; shifts for Group 2 are less stable and less interpretable – see Appendix Figure C.3). (b) Shifted Group 1 and Group 2 MAX mortality curves are tested against each other with IWTomics. The heatmap at the top shows $p$-values adjusted at all possible scales (from 1 to 65 days). The middle panel shows in detail the top-most row of the heatmap; i.e. the $p$-values adjusted across the whole 65-day interval. The bottom panel shows again the shifted curves. Gray areas in the middle and bottom panels mark days when the difference between the two groups is significant (adjusted $p$-value < 5%). Starting a little over two weeks from the beginning of their epidemic, curves in the two groups differ significantly at all temporal scales.

Functional boxplot and ranking. (a) Functional boxplot of the MAX data set (top) and MAX mortality curves (bottom) color-coded according to their ranking, as shown in the MAX column of (b). In the boxplot, Toscana is the median (black continuous line); Lombardia, Valle d’Aosta and Liguria are identified as outliers (red dashed lines); and the 50% innermost "box" (grey area) include the curves for Trento/Bolzano, Emilia-Romagna, Marche, Friuli Venezia Giulia, Veneto, Toscana, Molise, Abruzzo, Sardegna, Umbria, and Basilicata. Note that the "box" is skewed upwardly. (b) Rankings of the ISTAT (left), MAX (center) and DPC (right) mortality curves. The median regions are in bold, gray rectangles mark the 50% innermost boxes, and pale red rectangles mark outliers (no region is labeled as an outlier in the ISTAT data set; see Methods). The dots representing each region are color-coded (from intense red, through gray, to intense blue) according to their signed depth values (see Methods). In all three data sets, Lombardia’s curve is the most extreme at the very top of the ranking and, in contrast, Veneto’s curve is deep in the bulk close to the median (Toscana for ISTAT and MAX, Friuli Venezia Giulia for DPC). Segments joining the regions across the three rankings show how the top portion remains rather stable, while the mid- and bottom portions contain several crossings. Regions at the top are those characterized by "exponential" epidemics (Group 1), while regions in the middle and at the bottom are those with "flat(tened)" epidemics (Group2), whose curves can more easily switch in their depth ranks.
4.4 **Associating mortality to local mobility and positivity.** (a) Local mobility curves (Google’s “Groceries & pharmacy”) and positivity curves (regularized ratios of new cases to number of tests performed) in the 20 Italian regions. Curves are smoothed with splines, with degree of smoothing selected by generalized cross-validation, and shifted based on probKMA run on the MAX mortality curves with $K = 2$; time is marked as a day number representing the region-specific time of the epidemic unfolding, and corresponds to actual time (starting on February 16 and ending on April 30) only for regions with no shifts, e.g., Lombardia. Vertical lines show the days corresponding to the nationwide lock-down (March 9) and the suspension of all nonessential production activities (March 23) without shifts, stars on the curves mark the lock-down after the region specific shifts. The example regions of Figure 4.1(a) are highlighted in color. (b) Estimated effect surfaces from the joint function-on-function regression of MAX mortality on local mobility and positivity shown in 3D and as contour plots (March 9, without shift, is again marked on both). Early and mid-period local mobility levels are strong positive predictors of mortality at its peak. Positivity has similar but much weaker predictive signals, likely because the effects are subsumed by mobility. Late local mobility has a negative association with mortality at its peak (mobility resumed faster in regions with milder epidemics), and late positivity a strong positive one (positivity remained elevated in regions with worse epidemics). The regression captures a large share of the variability in mortality curves (in-sample $R^2 = 0.90$, LOO-CV $R^2 = 0.52$), with substantial and comparable contributions of the two predictors (partial $R^2$s = 0.62, 0.53).
4.5 **Interdependencies among scalar covariates and regions.** (a) Heatmap of the 20 (regions) x 12 (covariates) data matrix, with dendrograms from separate hierarchical clustering (correlation distance, complete linkage) of the regions (left) and the covariates (top). Color coding within cells represents values of the standardized covariates (centered and scaled to mean 0 and standard deviation 1). Color coding of some cell borders identifies the biclusters in (b). The dendrograms capture a distinct interdependence structure. For instance, there are marked similarities among Lombardia, Veneto, Emilia Romagna and Piemonte, as well as among some groups of southern regions (Sicilia, Campania, Puglia and Calabria; Basilicata, Abruzzo and Molise). There are also marked associations among groups of covariates. The contagion hubs proxies for hospitals, schools and work places, and number of adults per family doctor, vary closely together. So do the contagion hub proxy for public transport and pollution levels; the percentages of individuals affected by diabetes and allergies; and ICU beds and the percentage of individuals over 65. (b) Restricted heat-maps further illustrating interdependencies through two biclusters of regions and covariates. Color-coding within cells corresponds to that in (a), and each bicluster is identified by a border color and its adjusted H-score (an inverse measure of bicluster strength; see Methods). The first bicluster (adjusted H-score = 0.0902) comprises central and southern regions with "flat(tened)" epidemics (Group 2). The second bicluster (adjusted H-score = 0.0942) comprises northern regions with "exponential" epidemics (Group 1) but also northern and central regions from Group 2.

---

xi
4.6 Associating mortality to socio-demographic, infrastructural and environmental factors. (a) Results from marginal function-on-scalar regressions. Mortality curves are regressed against each of the scalar covariates in Table 4.1. The top plot displays the signs of the effect curves estimated on the MAX data. Time, marked as the 65 days of the region specific epidemic unfoldings, is on the vertical axis (the nationwide lockdown on March 9, without shift, is marked by a horizontal line. Red, blue and green indicate, respectively, positive, negative and non-significant portions (i.e., where 95% confidence bands around the estimated effect curve are entirely above, entirely below, or contain 0; see Methods). The bottom plot displays in-sample $R^2$s for the regressions fitted on MAX, ISTAT and DPC data; these are remarkably consistent. The names in red on the horizontal axes indicate the top 5 covariates selected by SnNAL-EN on all three data sets (see Methods); these are also the ones with the largest $R^2$s. (b) Results from the joint function-on-function regression of MAX mortality on local mobility, positivity, and the first principal component (pc1) of the top 5 covariates, used as a "summary" control. This control does not modify the shapes of the estimated effect surfaces for mobility and positivity (shown on top) – which are very similar to the ones in Figure 4.4,b). The estimated effect curve for pc1 shows a positive and significant association with mortality at its peak (bottom right; 95% confidence band in dashes, gray corresponds to non significant portions, vertical dashed line corresponds to March 9, without shift). The sign of this effect is consistent with marginal findings, based on the loadings of the first principal component (bottom left; positive for adults per family doctor, average beds per hospital, average students per classroom and average employees per firm, and negative for average members per household). With the addition of pc1, the regression reaches an in-sample $R^2 = 0.94$ and a LOO-CV $R^2 = 0.7$. The contributions of local mobility and positivity remain high (partial $R^2 = 0.66$ and 0.61, respectively). That of our "summary" covariate is also substantial (partial $R^2 = 0.39$).

A.1 Simulation 1 original curves (left), misaligned curves (center), registered curves by $cc_{H1R}$ (right) for $y$ (top) and $x$ (bottom). . . . . . . . . . . . 77

A.2 Simulation 2 original curves (left), misaligned curves (center), registered curves by $cc_{H1R}$ (right) for $y$ (top) and $x$ (bottom). . . . . . . . . . . . 78

A.3 Original $\beta$ for simulation 1 (left), simulation 2 (right) . . . . . . . . . . . . 78
A.4 Simulation 1 projections of the misaligned curves on the first (top) and second (bottom) principal components (second and fourth column) and canonical directions (first and third columns) of $y$ (first and second columns) and $x$ (third and fourth columns). ................. 79

A.5 Simulation 2 projections of the misaligned curves on the first (top) and second (bottom) principal components (second and fourth column) and canonical directions (first and third columns) of $y$ (first and second columns) and $x$ (third and fourth columns). ................. 79

A.6 Simulation 1 estimated $\beta$'s by different regressions. Original: the regression is performed on the original curves; misaligned: the regression is performed on the misaligned curves. In the other cases the regression is performed on the curves registered by the indicated algorithm. The white zones are values outside the $z$-range on the right. ................. 80

A.7 Simulation 2 estimated $\beta$'s by different regressions. Original: the regression is performed on the original curves; misaligned: the regression is performed on the misaligned curves. In the other cases the regression is performed on the curves registered by the indicated algorithm. The white zones are values outside the $z$-range on the right. ................. 81

A.8 AneuRisk data, projections of the misaligned curves on the first (top) and second (bottom) principal components (second and fourth column) and canonical directions (first and third columns) of $y$ (first and second columns) and $x$ (third and fourth columns). ................. 82

A.9 AneuRisk data, leave-one-out predicted curves by different regressions. ................. 83

A.10 AneuRisk data, leave one predicted curve by different regressions for just one patient. The top left panel represents the original misaligned curve before any registration. ................. 84

A.11 AneuRisk data, estimated $\beta$'s by different the different algorithms. The white zones are values outside the $z$-range on the right. ................. 85

C.1 **Unshifted mortality curves.** MAX, ISTAT and DPC mortality curves (per 100,000 inhabitants) without shift. Vertical lines show the days corresponding to the national lock down (March 9) and the suspension of all non-essential production activities (March 23). ................. 92
C.2 Mortality curves. (a): DPC mortality curves (per 100,000 inhabitants) in the 20 Italian regions – before (top) and after (bottom) the shifts produced by \textit{probKMA} run with $K = 2$. (b): ISTAT mortality curves (per 100,000 inhabitants) in the 20 Italian regions – before (top) and after (bottom) the shifts produced by \textit{probKMA} run with $K = 2$. In all panels, vertical lines mark the dates of the national lock-down (March 9) and the suspension of all non-essential production activities (March 23). In the bottom panels, vertical lines still show these dates without shifts, stars on the curves mark the lock-down after the region-specific shifts.

C.3 Characterizing two epidemics. Results of \textit{probKMA} and \textit{IWTomics} on (a)-(b) DPC mortality curves and (c)-(d) ISTAT curves. (a) and (c): Mortality curves are shown in the top left panel with portions identified by \textit{probKMA} with $K = 2$ in red (Group 1; “exponential” pattern) and blue (Group 2; “flat(tened)” pattern). The curve portions are shown again, this time aligned with each other and separated by group, in the bottom panels. Black lines indicate group averages. The shifts produced by \textit{probKMA} are shown in the top right panel. (b) and (d): Shifted Group 1 and Group 2 mortality curves are tested against each other with \textit{IWTomics}. The heatmap at the top shows $p$-values adjusted at all possible scales (from 1 to 65 days). The middle panel shows in detail the top-most row of the heatmap; i.e. the $p$-values adjusted across the whole 65-day interval (the 65 different values are reported in Table C.1). The bottom panel shows again the shifted curves. Gray areas in the middle and bottom panels mark days when the difference between the two groups is significant (adjusted $p$-value < 5%).

C.4 Characterizing three epidemics. Results of \textit{probKMA} with $K = 3$ on (a) MAX mortality curves, (b) DPC mortality curves and (c) ISTAT curves. Mortality curves are shown in the top left panel with portions identified by \textit{probKMA} with $K = 3$ in red (Group 1; “exponential” pattern), blue (Group 2; “flat(tened)” pattern) and green (Group 3; “extreme” pattern). The curve portions are shown again, this time aligned with each other and separated by group, in the bottom panels. Black lines indicate the average curves of the group. The shifts produced by \textit{probKMA} are shown in the top right panel.

C.5 Unshifted mobility and positivity curves. Local mobility and positivity curves without shift in the 20 Italian regions. Vertical lines show the days corresponding to the national lock down (March 9) and the suspension of all non-essential production activities (March 23).
C.6 **Shifted curves for ISTAT and DPC.** Mortality (per 100,000 inhabitants), mobility, and positivity curves after the shifts produced by \textit{probKMA} with $K=2$. Vertical lines mark the dates of the national lockdown (March 9) and the suspension of all non-essential production activities (March 23) without shifts. Stars on the curves mark the lock-down after the region-specific shifts.

C.7 **Lags characterization.** Projections on the first Principal Components of mortality curves (MAX, ISTAT, and DPC – upper panels), positivity curves (lower-left panel) and mobility curves (lower-mid panel). Here the curves are not shifted, so the horizontal axis represents calendar dates. Projecting the curves on their first Principal Components produces an additional de-noising, which inherently aligns all the peaks. The thick dark lines are average curves in each panel, and the vertical lines mark the date of their maxima – also expressed in number of days from February 16th. For mobility, we also mark the beginning of the national lockdown (March 9th; 23 days from February 16th), which corresponds to a severe drop in the curves. We notice lags of about 20 days and 1 week between, respectively, the peak and the stark drop in mobility, and the peak in positivity (on average across all regions). The lag between the peak in positivity and the peak in mortality is about 10 days (again on average across all regions). The lower-right panel shows the individual mobility-positivity and positivity-mortality lags for each region in group 1 (those characterized by an “exponential” epidemic pattern). Albeit with some variation across regions, adding up these lags produces an overall delay of approximately 1 month between the peak mobility and that of mortality.

C.8 **MAX mortality residuals.** (a): residuals of the function-on-function regression of MAX mortality on local mobility and positivity. (b): residuals of the function-on-function regression of MAX mortality on local mobility, positivity, and the first principal component of the top 5 covariates. In both panels curves from Group 1 are in red, and curves from Group 2 are in blue. Residuals with positive signs indicate regions for which the true mortality curve is above the estimated mortality curve. Conversely, residuals with negative signs indicate regions for which the true mortality curve is below the estimated mortality curve.
C.9 **Associating mortality to mobility and positivity - ISTAT and DPC.** Results from the joint function-on-function regression of ISTAT and DPC mortality on mobility and positivity. The top row shows the estimated effect surfaces (the March 9 date is marked) with respective partial $R^2$ (for in-sample $R^2$ and LOO-CV $R^2$ see Table C.2). The bottom row shows the regression residuals (for barplots interpretation see Figure C.8).

C.10 **Associating mortality to mobility.** Results from the function-on-function regression of mortality on local mobility. The top row displays the estimated effect surface (the March 9 date is marked) with respective in-sample $R^2$ (for LOO-CV $R^2$ see Table C.2). The bottom row displays the regression residuals (for barplots interpretation see Figure C.8).

C.11 **Associating mortality to positivity.** Results from the function-on-function regression of mortality on positivity. The top row displays the estimated effect surface (the March 9 date is marked) with respective in-sample $R^2$ (for LOO-CV $R^2$ see Table C.2). The bottom row displays the regression residuals (for barplots interpretation see Figure C.8).

C.12 **Correlations between all covariates.** Histogram showing the distribution of pair-wise correlations for all the 68 covariates in Table C.3. 24% of the correlations exceed 0.5 in absolute value. Note: Variance Inflation Factors cannot be computed for the whole set of 68 variables since the sample size here is only $n = 20$ (the Italian regions).

C.13 **Correlation-based dendrogram for all covariates.** Dendrogram of all the covariates in Table C.3, based on the correlation distance $d(x_1, x_2) = 1 - |corr(x_1, x_2)|$ and complete linkage. The 12 selected covariates are shown in red, while covariates dating 2016 or earlier are in grey.

C.14 **Associations between the 12 selected covariates.** Exploratory matrix containing scatterplots with loess regressions, marginal densities, and correlation between pairs of covariates.
C.15 **Marginal function-on-scalar regressions.** Results for marginal function-on-scalar regressions. Mortality curves are regressed against each of the scalar covariates in Table 1. The top-row displays the signs of the effect curves estimated when just one intercept is included in the model. The bottom-row displays the signs of the effect curves estimated when we consider two different intercepts for curves in Group 1 and curves in Group 2. Time is on the vertical axis (the national lockdown on March 9, without shift, is marked by a horizontal line). Red, blue and green indicate, respectively, positive, negative, and non-significant portions (i.e., where 95% confidence bands around the estimated effect curve are entirely above, entirely below, or contain 0).

C.16 **Function-on-function regression of mortality on mobility, positivity, and a control scalar covariate.** Each row shows some results from the joint function-on-function regression mortality on local mobility, positivity, and one of the top 5 covariates selected by $fgen$, used as control. In particular, we display the estimated effect surfaces for mobility and positivity (the March 9 date, without shift, is marked) with their respective partial $R^2$s. The scalar control covariates associated with each row are the following: 1: Adults per family doctor, 2: Ave. beds per hospital (whole), 3: Ave. students per classroom, 4: Ave. employees per firm, 5: Ave. members per household.

C.17 **Associating mortality to mobility, positivity and first principal component - ISTAT and DPC.** Results from the joint function-on-function regression of ISTAT and DPC mortality on mobility, positivity, and the first principal component (pc1) of the top 5 covariates, used as a “summary” control. Each panel shows the estimated effect surfaces for mobility and positivity and the estimated effect curve for pc1 with respective partial $R^2$ (for in-sample $R^2$ and LOO-CV $R^2$ see Table C.2). For interpreting the regression residuals see Figure C.8.
List of Tables

2.1 AneuRisk data. SPE and L1OPE for regressions performed on misaligned data and on data produced by different registration procedures. For FCC and FPC performed with H1 and RS, results are shown for both 1 and 2-dimensional approximations. Values are rescaled to have SPE and L1OPE equal to 100 on the misaligned data. The last row reports CPU times (in seconds) required by the different registration procedures. . . . . 21

3.1 a, b and c report CPU time in seconds for fgen, sklearn and glmnet, respectively. For fgen we also report the number of iterations in parenthesis. r is the number of selected features, l is the range parameter of the Matern process used to generate the coefficients. . . . . . . . . . . . . 41

4.1 Scalar covariates potentially affecting COVID-19 mortality. [1] The percentages of over 65, 70, 75 and 80 are highly correlated at the resolution of regions; we took over 65 as representative. [2] The prevalence of diabetes, hypertension and chronic bronchitis are highly correlated at the resolution of regions; we took diabetes as representative (allergies are not as highly collinear and were retained as separate). [3] Availability of ICU beds is also directly relevant for withstanding the impact of COVID-19 surges. [4] Average members per household is not a direct proxy of inter-generational contacts, but it may capture some of its effects. . . . . . . 56

A.1 Simulation results. Columns: H1 distances between the curves in each row and the original x and y curves; in sample H1 prediction error; leave-one-out H1 prediction error; Euclidean distances between Ŝβ and the true β. Rows: Org, the original curves; msl, the misaligned curves; followed by curves registered with different procedures – as indicated. . . . . . . 76
A.2 Simulation results. Columns: $L^2$ distances between the curves in each row and the original $x$ and $y$ curves; in sample $L^2$ prediction error; leave-one-out $L^2$ prediction error. Rows: Org, the original curves; msl, the misaligned curves; followed by curves registered with different procedures – as indicated.

B.1 $a$, $b$ and $c$ report mean CPU time in seconds for $fgen$, $sklearn$ and $glmnet$, respectively, over 20 replications of the same simulation scenario. In parenthesis we report standard errors. For each scenario we consider three values of $c_\lambda$, which are held fixed over the replications.

B.2 $a$, $b$ and $c$ report CPU time in seconds for $fgen$, $sklearn$ and $glmnet$, respectively. The full $c_\lambda$ grid consists of 100 log-spaced points between 1 and 0.01. We truncate the path search when max active components are selected. $runs$ is the corresponding number of explored $c_\lambda$ values. We fix 1000 seconds as time limit.

C.1 $IWTomics$ adjusted p-value curves. Values of the IWTomics adjusted p-values for each time point and each of the three dataset. Note that the smallest possible p-value is 0.001 since we are employing 1000 permutations in the IWTomics test.

C.2 Functional regression models (in-sample) $R^2$, LOO-CV $R^2$ and partial $R^2$s. For each functional linear model which regresses mortality on the covariates listed in the first column, the table reports the (in-sample) $R^2$, the LOO-CV $R^2$ and the partial $R^2$s.

C.3 Covariates. List of all scalar covariates considered.

C.4 Variance Inflation Factors of covariates. Variance Inflation Factors (VIF) for the 12 scalar covariates used in the main analysis.

C.5 Function-on-scalar feature selection. Top five scalar covariates selected by $fgen$ considering as response the MAX, ISTAT, and DPC mortality curves.
Acknowledgments

First, I would like to thank my advisor Dr. Francesca Chiaromonte. She dragged me into this Ph.D., and she has been an excellent mentor and a wise friend. She guided and protected me during this journey, transmitting her passion for research and teaching me how to believe in myself.

I thank my co-advisor, Dr. Matthew Reimherr, for supporting and giving me new ideas and new energies in a moment when all my research seemed to go wrong.

Then, of course, I have to thank all the friends I met in the past five years, both in the department and outside it. They helped me in the worst moments and shared with me the best ones. Thank you for filling these years with unforgettable experiences and making me feel like State College was a new home.

A special thank goes to Umbi: I was lucky to meet such a kind friend who has been and will continue to be an important piece of my life.

It is with a lot of pride that I thank my parents. Without them, their love, their sacrifices, I would not be the person I am today.

Finally, I thank my girlfriend, Ludovica. You have been the brightest smile in the hardest instants and the first companion of the happiest ones.
Chapter 1

Introduction

Functional Data Analysis (FDA) has established itself as an important and dynamic area of statistics, offering effective new tools applicable to a broad range of scientific domains – including medicine, the life sciences, physics, business, and engineering. Just as in the classical framework, linear regression is one of the most fundamental tools in FDA and presents many open and fascinating theoretical and computational challenges. Indeed, Functional Linear models leverage high dimensional and complex data. In these models, one or more parameters are curves - i.e., infinite-dimensional objects that must be estimated from a finite sample by imposing smoothness conditions and considering appropriate projective subspaces. For these reasons, advanced computing and optimization techniques for solving functional regression are crucial to obtain good estimates and reduce computational costs.

My Ph.D. thesis focuses on two main aspects of functional regression: functional registration and feature selection. The methods I developed are particularly relevant for “Omics” and Biomedical studies, where one has to deal with a large number of predictors and with responses that are suitable for a functional data representation, such as longitudinal measurements or biomedical images.

1.1 Functional Registration

Registration is a critical issue in FDA, and tackles the problem of separating phase (horizontal) and amplitude (vertical) variation in a statistically meaningful way. The work on registration is introduced in Chapter 2.

The core of Chapter 2 is the article Boschi, Chiaromonte, Secchi, and Li (2021) published in STAT. We developed a new registration technique which exploits the information given by Functional Covariance Components (FCC) to register a functional
response and a functional predictor simultaneously. In a function-on-function regression context, FCC provides a flexible and powerful tool to represent the data in a low-dimensional space \cite{Wagner_and_Kneip_2019}, capturing the most meaningful modes of dependency between response and predictor curves. To implement our approach we used a novel parallel algorithm coded in \texttt{R}. We then applied FCC registration to the \textit{AneuRisk} data \cite{Sangalli_etal_2014}. The aim of the \textit{AneuRisk} project is to investigate the interplay between morphological properties of artery walls and hemodynamic factors, and shed light on the possible causes of aneurysmal pathology. In our work we analyzed the relationship between the curvature and the wall shear stress of the internal carotid artery – but enhancements to the FCC procedure could provide additional insight on this biomedical application, e.g., allowing us to classify the patients in two groups based on the position of their aneurysm.

The last section of Chapter 2 briefly presents an ongoing project where we extend the applicability of FCC registration to more than one predictor and curves with different domains. The main challenge of this multivariate FCC (mFCC) registration is the definition of a subspace containing all the information relevant for the relationships between more than two sets of curves \cite{Chiou_etal_2014}. We are applying mFCC to analyze the \textit{Milano Weather Station Data}, which contains pollution and weather-related variables measured every hour for a period of two months in different locations across the city of Milan. Our goal is to improve the results of functional regression analyses aimed at understanding the relationships between pollutants (such as levels of carbon monoxide, benzene, nitrogen, etc.) and weather-related variables (such as temperature, precipitation, humidit, etc).

## 1.2 Functional Feature Selection

Our work on feature selection has been recently published in the proceedings of NeurIPS 2021 \cite{Boschi_Reimherr_and_Chiaromonte_2021} and is presented in Chapter 3.

Functional feature selection combines statistics, machine learning, and advanced optimization techniques to simultaneously perform variable selection, smoothing, and parameter estimation – which is particularly challenging in ultra-high dimensional contexts; that is, when the number of predictors to select from is much larger than the sample size. To develop a new mathematically sound and computationally efficient methodology to solve ultra high-dimensional problems, we started from the Dual Augmented Lagrangian (DAL) – which allows one to exploit the sparsity of the second-order information in the
augmented Lagrangian problem to significantly reduce computational burden without sacrificing accuracy. Tomioka and Sugiyama (2009) and, more recently, Li et al. (2018) developed a DAL algorithm to solve the classic LASSO problem. In Boschi et al. (2020), we proposed a DAL algorithm to solve the classic Elastic Net problem, which considers both the $l_1$ and $l_2$ penalizations. Then, in Boschi, Reimherr, and Chiaromonte (2021), we extended this approach to the Group Elastic Net and the Function-on-Scalar feature selection frameworks ($fgen$).

Several proposals have been recently introduced to tackle Function-on-Scalar feature selection (Chen et al., 2016; Parodi et al., 2018). However, all existing approaches become slow and inefficient when the number of predictors is very large. This setting is common and particularly relevant in many domain applications – for instance, in Genome-Wide Association Studies (GWAS), where the potential predictors are a massive number (tens or hundreds of thousands, sometimes millions) of genetic mutations, such as Single Nucleotide Polymorphisms (SNPs). We exploited the computational efficiency of $fgen$ (implemented in python) to analyze data from the INSIGHT project (Paul et al., 2014) and detect SNPs most related to obesity risk in children. We identified one relevant SNP located in NTM, a well-known gene recently connected to body mass and food addiction (Kichaev et al., 2019).

1.3 Research applications and collaborations

Chapter 4 contains the article Boschi, Di Iorio, Testa, Cremona, and Chiaromonte (2021) published in Scientific Reports. This work is the product of research collaborations which extend beyond Penn State and aim to employ advanced functional data methods to study the evolution of the COVID-19 epidemic in Italy – using massive amounts of data that we collect, pre-process, and curate from different public sources on the epidemic, as well as socio-demographic, infrastructural and environmental factors that may affect its unfolding.

In particular we used functional clustering, registration, regression, and future selection techniques to investigate patterns of COVID-19 mortality across 20 Italian regions and their association with mobility, positivity, and socio-demographic, infrastructural, and environmental covariates. We characterized heterogeneous and staggered epidemics in different areas of Italy and we found that mobility and positivity can predict COVID-19 mortality, also when controlling for relevant covariates. Among the latter, primary care appears to mitigate mortality, and contacts in hospitals, schools and workplaces to
aggravate it.

These research collaborations expand outside the work presented in Chapter 4. We are now working on a new project, contrasting mortality and mobility patterns between the first and second waves of COVID-19 (both pre-vaccine) across the 107 Italian provinces. Using classic and functional feature selection techniques, we are further investigating how the timing of restrictions policies can help control the epidemic. Moreover, in Cintia et al. (2020), using Italian data at different geographic scales, we investigated the relationship between human mobility, which subsumes many facets of the population’s response to the evolving context, and the spread of COVID-19. We found that the time needed to switch mobility off and bring the net reproduction number below the critical threshold of 1 is about one week.
Chapter 2  
Covariance based low-dimensional registration for function-on-function regression

2.1 Introduction

Functional Data Analysis (FDA) comprises sophisticated tools and methods applicable to scientific fields ranging from the geosciences to the social and biomedical sciences. A critical component of FDA is data registration or alignment, which separates phase (horizontal) and amplitude (vertical) variation in a statistically meaningful way [Marron et al. (2015)]. Common alignment approaches utilize warping functions, which monotonically transform the data domain making them as similar as possible to an overall template function. Kneip and Ramsay (2008) and Wagner and Kneip (2019) recently introduced alignment procedures based on reductions to low-dimensional functional spaces, usually obtained through a Functional Principal Components (FPC) basis.

In a function-on-function regression context, we propose an alternative basis for low-dimensional registration obtained through Functional Covariance Components (FCC). This exploits the relationship between response and predictor, preserving information relevant to the regression. Our novel procedure registers response and predictor curves simultaneously based on the most significant modes of covariation between them, and can be seen as the functional parallel of Partial Least Square (PLS) in finite dimension. (Geladi and Kowalski 1986) Unlike Principal Component regression, PLS identifies a reduced space based on the link between response and predictor, not just variation in the latter.
We implement FCC registration with an efficient new parallel algorithm in \( \mathbb{R} \), which minimizes the H1 distance between each curve and its projection on the FCC basis. This distance considers L2 distances in both the original curves and their derivatives – which is crucial in many applications and accounts for shape and smoothness of the curves.

We benchmark our H1 algorithm against an adaptation of the algorithm proposed by Ramsay and Silverman (2005a) and available in the \texttt{fda} \ R \ package (RS). RS minimizes the smallest eigenvalue of the cross product matrix between each curve and a common template, usually the mean curve. Our adaptation uses a different template for each curve; namely, its projection on the FCC basis. We also benchmark FCC registration, implemented with both H1 and RS, against commonly used methods such as mean-template registration (Ramsay and Silverman, 2005a), elastic shape registration (Tucker et al., 2013) and Principal Components registration (Kneip and Ramsay, 2008). Based on simulations and a biomedical application we show that FCC registration, especially with the novel H1, performs on par with other approaches in terms of alignment, provides an effective low-dimensional representation of the data, and improves regression performance.

The remainder of the chapter is organized as follows. In Section 2.2 we give background on the low-dimensional registration problem. In Section 2.3 we introduce our FCC registration. In Section 2.4 we describe its implementation based on RS and on our novel H1. In Section 2.5 we report simulation results and an application to AneuRisk data on 52 patients, where we study the relationship between wall shear stress on the carotid walls and some of its geometrical features. In Section 2.6 we present some current developments related to a multivariate version of FCC. In Section 2.7, we conclude by pointing out an important interpretation of FCC registration in the context of Sufficient Dimension Reduction.

### 2.2 Background

Consider a sample comprising two sets of \( n \) (paired) curves, \( x(t) = x_1(t), \ldots, x_n(t) \) and \( y(t) = y_1(t), \ldots, y_n(t) \). For simplicity, we assume a common domain \([a, b]\) for the curves, though all our developments hold when domains differ for \( x \) and \( y \). We also assume \( x_i, y_i \in H^2([a, b]) \); i.e. that curves are twice differentiable with both first and second derivatives belonging to \( L^2([a, b]) \). For such curves, the L2 squared norm is defined integrating the square over the domain; that is \( \|x_i\|_{L_2}^2 = \int_a^b x_i^2(t)dt \). The H1 squared norm, also known as the Sobolev W12 norm, is defined integrating the sum of two squares over the domain – that of the function and that of its derivative; that
is \( \|x_i\|_{H^1}^2 = \int_a^b \left( x_i^2(t) + \frac{d}{dt}x_i^2(t) \right) dt \). Notably, this allows one to capture jointly the size, i.e. the level, and the smoothness of the function. More details on Sobolev spaces and the \( H^1 \) norm can be found in Adams and Fournier (2003).

In practice, to construct each curve \( x_i \) (or \( y_i \)) starting from any number, say \( S_i \), of raw observations \( x_{i1}^{(R)}, \ldots, x_{iS_i}^{(R)} \), we use a linear combination of \( J \) B-spline functions of order \( m \), \( \phi_j^m \), with \( J - m + 2 \) equidistant knots. We write

\[
  x_i(t) = \sum_{j=1}^J c_{ij} \phi_j^m(t)
\]

where the coefficients \( c_{ij} \) are chosen to minimize \( \sum_{s=1}^{S_i} (x_{is}^{(R)} - x_i(s))^2 + \lambda \int_a^b \left( \frac{d^2}{dt^2} x_i(t) \right)^2 dt \) (the first term is a sum of least squares distances, the integral is a smoothing penalization term, and \( \lambda \) is a smoothing parameter). Note that for any \( m > 3 \), we do indeed obtain curves in \( H^2 \). Different basis systems or entirely different methods (e.g. free-knot regression splines as in Sangalli et al., 2009b) could be used, but a basis system is necessary for our registration procedure. In the following we will use notation similar to Horváth and Kokoszka (2012a). The covariance function between \( x \) and \( y \) is \( \sigma_{xy}(t, s) = \text{E}[(x(s) - \text{E}x(s))(y(t) - \text{E}y(t))] \), and the variance functions of \( x \) and \( y \) correspond to \( \sigma_{xx} \) and \( \sigma_{yy} \). The covariance operator between \( x \) and \( y \) is \( (\Sigma_{xy} v)(t) = \int_a^b \sigma_{xy}(t, s)v(s)ds \) for \( v \in L^2([a, b]) \), and the variance operators for \( x \) and \( y \), \( \Sigma_{xx} \) and \( \Sigma_{yy} \), are defined similarly. On the sample, the covariance function is \( \hat{\sigma}_{xy}(s, t) = n^{-1} \sum_{i=1}^n (x_i(s) - \bar{x}(s))(y_i(t) - \bar{y}(t)) \) and the covariance operator is \( (\hat{\Sigma}_{xy} v)(t) = \int_a^b \hat{\sigma}_{xy}(t, s)v(s)ds \), where \( \bar{x}(t) = n^{-1} \sum_{i=1}^n x_i(t) \) and \( \bar{y}(t) = n^{-1} \sum_{i=1}^n y_i(t) \) are sample means.

### 2.2.1 Registration with warping functions

Functional data exhibit both *amplitude*, which pertains to the size of a curve’s features (e.g. peaks and valleys) ignoring their position in the domain, and *phase* variation, which pertains to the location of the features ignoring their sizes. The distinction is critical: if curves share a common pattern (e.g. a common set of peaks) which is misaligned along the domain, ignoring phase variability can lead to an inability to capture important structure in the data and to inefficiency in modeling. The problem of separating amplitude and phase variation in a statistically meaningful way is called *registration*. Here, we focus on *warping*-based registration. A warping function \( h \) is defined as an element of the space \( \mathcal{H} \subset H^2([a, b]) \) of all continuous, strictly increasing functions such that \( h(a) = a \) and \( h(b) = b \) (Ramsay and Silverman, 2005a). Each curve \( x_i \) is associated to a specific
warping function \( h \) which captures its phase variation through a (possibly non-linear) transformation of the domain. More specifically, \( h \) is chosen so that the registered curve \( x_i(h_i(t)) \) is as similar as possible to a given template. Since monotone transformations do not change shape features, \( x_1(h_1(t)), \ldots, x_n(h_n(t)) \) have the same sequences of peaks and valleys as the original curves and, ideally, exhibit only amplitude variation. Note that the condition \( h(a) = a \) and \( h(b) = b \) excludes simple horizontal translations – imposing that all registered curves have common start and end points. While natural in many applications [Ramsay and Silverman (2002)], this requirement can be modified in specific situations. We represent warping functions as in Kneip and Ramsay (2008): we consider

\[
h_i(t; w_i) = a + (b - a) \frac{\int_a^t e^{w_i(u)} du}{\int_a^b e^{w_i(u)} du},
\]

(2.2)

where \( w_i \in H^2([a, b]) \), and since for any scalar \( c \in \mathbb{R} \) the functions \( w \) and \( w + c \) lead to the same warping \( h \), we standardize warping functions by imposing that \( \int_a^b w_i(t) dt = 0 \) \( \forall i \) and \( n^{-1} \sum_{i=1}^n w_i(u) = 0 \ \forall u \in [a, b] \).
2.3 FCC registration

We now focus on a function-on-function regression context, considering two sets of paired curves \( x(t) = x_1(t), \ldots, x_n(t) \) and \( y(t) = y_1(t), \ldots, y_n(t) \), with \( x_i, y_i \in H^2([a,b]) \forall i \). The function-on-function regression model is defined as (Horváth and Kokoszka (2012a))

\[
y_i(s) = \alpha(s) + \int_a^b x_i(s) \beta(s,t) dt + \epsilon_i(s) \quad i = 1, \ldots, n,
\]

where \( y \) is the functional response, \( x \) the functional predictor, \( \alpha \) the functional intercept in \( H^2([a,b]) \), \( \beta(s,t) \) the regression coefficient surface, and the \( \epsilon \)'s are i.i.d. Gaussian random elements in \( H^2([a,b]) \), independent of the \( x \)'s, with mean function 0 and common variance operator. Again, without loss of generality, we assume \( x \) and \( y \) to have the same domain. We develop a registration procedure that, exploiting the relationship between \( x \) and \( y \), optimizes predictive performance in the function-on-function regression. This regression-driven registration can be compared to the clustering-driven registration introduced in Sangalli et al. (2010): the aim is not only the alignment itself, but also the performance of a specific statistical analysis. To achieve our goal, instead of utilizing the leading modes of variation within one set of curves as in FPC, we utilize the leading modes of covariation between the two sets of response and predictor curves. In the remainder of this section, we describe how the covariance operator can guide a low-dimensional registration procedure.

2.3.1 Functional Covariance Components

To capture the linear associations, we identify the pair of linear combinations of the curves having the largest covariance, then (orthogonally) the pair with the second largest covariance, etc. We use the terms covariance functions or Functional Covariance Components (FCC) to refer to these pairs. These are similar in spirit to the pairs produced by Functional Canonical Correlation Analysis (Leurgans et al. 1993), which maximizes the correlation instead of the covariance. Notably, whether the components are produced considering covariance or correlation, they allow a low-dimensional representation of an infinite-dimensional relationship between two sets of curves. But they are not just an effective dimension reduction tool; investigating these linear combinations is useful for
many statistical problems, including regression (see [He et al. (2010)]) for the functional case and [Borga et al. (1997)] for the classical multivariate case). The first covariance weights (or scores) are defined as

\[
\langle \xi_1, x_i \rangle = \int_a^b \xi_1(t) \left( x_i(t) - \bar{x}(t) \right) dt, \quad i = 1, \ldots, n
\]

\[
\langle \eta_1, y_i \rangle = \int_a^b \eta_1(t) \left( y_i(t) - \bar{y}(t) \right) dt, \quad i = 1, \ldots, n
\]

where the covariance functions \( \xi_1 \) and \( \eta_1 \) maximize the covariance between \( u_1 = (\langle \xi_1, x_1 \rangle, \ldots, \langle \xi_1, x_n \rangle)' \) and \( v_1 = (\langle \eta_1, y_1 \rangle, \ldots, \langle \eta_1, y_n \rangle)' \), under the constraint \( \int_a^b \xi_1^2(t) dt = \int_a^b \eta_1^2(t) dt = 1 \). The process is then iterated: to select the \( j \)-th covariance functions \( \xi_j \) and \( \eta_j \) and form the \( j \)-th vectors \( u_j \) and \( v_j \), one again maximizes covariance considering functions with L2 norm equal to 1 and adding the requirement that each covariance function is orthogonal to prior ones; that is, \( \text{cov}(u_\ell, u_j) = \text{cov}(u_\ell, v_j) = \text{cov}(v_\ell, v_j) = 0 \quad \forall \ell < j \). \( \text{cov}(u_j, v_j) \) declines at each iteration, until subsequent modes of covariation become negligible.

The identification of FCCs can be formulated as an eigen-decomposition or Singular Value Decomposition (SVD); the pairs of covariance functions and weights are obtained from the SVD of the covariance operator \( \Sigma_{xy} = UDV^T \). The \( \xi \)'s and the \( \eta \)'s are provided by the columns of \( U \) and \( V \), respectively; that is, they are the left and right eigenfunctions of \( \Sigma_{xy} \). \( D \) is a diagonal matrix such that \( (D)_{jj} = \text{cov}(u_j, v_j) \). Note that also Functional Canonical Correlation Analysis can be formulated as an SVD using \( R_{xy} = \Sigma_{xx}^{-1/2} \Sigma_{xy} \Sigma_{yy}^{-1/2} \) [He et al. (2004)]. If the goal is to regress \( y \) on \( x \), another natural operator to consider is the linear regression operator \( L_{xy} = \Sigma_{xx}^{-1} \Sigma_{xy} \). However, for our purposes the covariance operator \( \Sigma_{xy} \) has three main advantages with respect to \( R_{xy} \) and \( L_{xy} \).

First, and most important, the bases given by the SVD of \( R_{xy} \) and \( L_{xy} \) do not allow one to reconstruct the original curves: if we scale the covariance using their variances, then also their projections are scaled and one cannot obtain the original \( x \)'s (\( y \)'s) from the projections. Since the projections are the target functions of our registration procedure, it is preferable not to scale them using the variance of \( x \) or \( y \), to ensure that they live in the same space as the original curves. Second, \( R_{xy} \) and \( L_{xy} \) require the inversion of at least one functional operator, which may be singular or almost singular in some applications. By contrast, \( \Sigma_{xy} \) does not require any operator inversion and always allows us to perform a low-dimensional registration, independently of the subsequent regression. Third, even though regressing \( y \) on \( x \) is intrinsically asymmetric, it is useful to develop a registration procedure that aligns the two sets of curves and captures their modes of
dependence in a symmetric way. The SVD of $L_{xy}$ or $R_{xy}$ does not do this, while the SVD of $\Sigma_{xy}$ does; our registration would be the same if $x$ became the response and $y$ the predictor.

### 2.3.2 Low-dimensional registration through FCC

Here we lay out the details of our FCC low-dimensional registration. Let $\xi_1, \ldots, \xi_{K_x}$ and $\eta_1, \ldots, \eta_{K_y}$ be the first $K_x$ and $K_y$ covariance components, i.e. the left and right eigenfunctions of $\Sigma_{xy}$, respectively. We take these as basis systems to define the templates functions. For each curve $x_i(y_i)$, the template $x_0i(y_0i)$ is given by projecting $x_i(y_i)$ on $\xi_1, \ldots, \xi_{K_x}$ ($\eta_1, \ldots, \eta_{K_y}$). In symbols

$$
 x_{0i}(t) = \text{PRO}_{\Sigma_{xy}}(x_i, K_x)(t) = \sum_{j=1}^{K_x} a_{ij} \xi_j(t) + \bar{x}(t), \quad i = 1, \ldots, n
$$

$$
 y_{0i}(t) = \text{PRO}_{\Sigma_{xy}}(y_i, K_y)(t) = \sum_{j=1}^{K_y} b_{ij} \eta_j(t) + \bar{y}(t), \quad i = 1, \ldots, n
$$

(2.5)

where the coefficients $a_{ij}$ and $b_{ij}$ are given by

$$
 a_{ij} = \langle \xi_j, x_i(t) - \bar{x}(t) \rangle = \int_a^b \xi_j(t)(x_i(t) - \bar{x}(t))dt,
$$

$$
 b_{ij} = \langle \eta_j, y_i(t) - \bar{y}(t) \rangle = \int_a^b \eta_j(t)(y_i(t) - \bar{y}(t))dt.
$$

Note that, unlike in (2.3), the goal here is not to find a subspace $L_K$ through which to reconstruct the curves once they have been registered. We use the subspace associated to $\Sigma_{xy}$ to define the registration templates and then find two sets of warpings $h_x$ and $h_y$ minimizing a specific criterion (see Section 2.4). Thus, our procedure does not estimate $L_K$ and the two sets of warpings simultaneously, and it is not affected by identifiability issues. Note also that, even though $x$ and $y$ are registered concurrently, our procedure allows one to fix two different dimensions $K_x$ and $K_y$ for the basis expansion of $x$ and $y$.

The choice of $K_x$ and $K_y$ plays an important role. Theoretically, $x$ and $y$ live in infinite dimensional functional spaces. However, in practice, the rank of $\Sigma_{xy}$ (and thus the number of possible pairs of covariance components) is finite and bounded above by the minimum between the number of basis functions used for expressing $x$ and $y$. In applications, one can choose $K_x$ and $K_y$ empirically, according to the complexity of each set of curves and to the number of components needed to capture their most meaningful modes of variability. Importantly, for the registration to be effective, $K_x$. 

11
and $K_y$ should not be very large. If they are, $\text{PRO}_{\Sigma_{xy}}(x, K_x)$ and $\text{PRO}_{\Sigma_{xy}}(y, K_y)$ capture almost all the variability and resemble very closely the original misaligned curves $x$ and $y$ – so the procedure ineffectively aligns curves to templates very similar to the curves themselves. We propose three quantitative criteria to select dimensions under the simplifying assumption that $K_x = K_y = K$.

First, we measure the cumulative sum of the eigenvalues of $\Sigma_{xy}$ which, for every $K$, captures the share of dependency explained by the first $K$ components. Second, we measure the total phase variation of $\text{PRO}_{\Sigma_{xy}}(x_i, K)$ and $\text{PRO}_{\Sigma_{xy}}(y_i, K)$ generalizing a similarity index in Sangalli et al. (2009a). For two generic curves $g_i$ and $g_j$, let

$$\rho(g_i, g_j) = \frac{\int_a^b \frac{d}{dt}g_i(t) \frac{d}{dt}g_j(t)}{\left(\int_a^b \frac{d}{dt}g_i(t)^2\right)^{0.5} \left(\int_a^b \frac{d}{dt}g_j(t)^2\right)^{0.5}}$$

(2.6)

where $|\rho(g_i, g_j)| \leq 1$ and $\rho(g_i, g_j) = 1$ if and only if $\exists m_1 \in \mathbb{R}^+, m_2 \in \mathbb{R} : g_i = m_1 g_j + m_2$, i.e. the curves exhibit only amplitude variation. For two generic sets of curves $g$ and $q$, we compute the means $\rho_g$ and $\rho_q$ of the pairwise $\rho$ indexes in the two sets, and define the Total Phase Variation (TPV) of $g$ and $q$ as

$$\text{TPV}(g, q) = \frac{1}{2} \left(1 - \frac{1}{2}(\rho_g + \rho_q)\right).$$

(2.7)

Note that $\text{TPV}(g, q) \in [0, 1]$ and that $\text{TPV}(g, q) = 0$ if and only if $\rho_g = \rho_q = 1$, i.e. all curves in $g$ and $q$ have no phase variation. We choose $K$ so that $\text{TPV}(\text{PRO}_{\Sigma_{xy}}(x, K), \text{PRO}_{\Sigma_{xy}}(y, K))$ remains smaller than $\text{TPV}(x, y)$; otherwise the templates would have larger phase variability than the original, misaligned curves – which is not reasonable. Third, we measure regression performance; we register the curves using different $K$ values and evaluate the subsequent regressions in terms of leave-one-out prediction error. Figure 2.1 illustrates the three criteria on simulated data and data from the AneuRisk application described in Section 2.5.

### 2.4 Algorithms

Next, we describe the implementation of FCC registration with two different algorithms; the RS of Ramsay and Silverman (2005a) and our new H1. Both find warpings that maximize similarity between registered curves and templates, but with different definitions of similarity.

Each warping function $h_i$ is defined starting from the function $w_i$ (see (2.2)), which
in turn is expressed as a linear combination of B-splines with \( J = J_w \) (see (2.1))

\[
    h_i(t; c_i) = a + (b - a) \frac{\int_a^b e^{\sum_{j=1}^{J_w} c_{ij} \phi_j(u)} du}{\int_a^b e^{\sum_{j=1}^{J_w} c_{ij} \phi_j(u)} du}
\]  

(2.8)

and for both algorithms the target functions for the \( i \)-th curves are \( x_{0i} = \text{PRO}_{\Sigma_{sx}}(x_i, K_x) \) and \( y_{0i} = \text{PRO}_{\Sigma_{sy}}(y_i, K_y) \), as defined in (2.5).

RS is coded in the function `register.fd` of the `fda` R package. For each \( x_i \) (or \( y_i \), separately) it selects the \( c_{ij} \) coefficients in (2.8) as to minimize \( \mu_2(T(h_i)) + \lambda \int_a^b \left( \frac{d^2}{dt^2} h_i(t) \right)^2 dt \), where \( \mu_2 \) is the smallest eigenvalue of the symmetric operator

\[
    T(h_i) = \begin{pmatrix}
        \int_a^b x_{0i}(t) \, dt & \int_a^b x_{0i}(t)x_i(h_i(t)) \, dt \\
        \int_a^b x_i(h_i(t))x_{0i}(t) \, dt & \int_a^b x_i(h_i(t))^2 \, dt
    \end{pmatrix}
\]

A second derivative roughness penalty is applied to \( h_i \) for regularization purposes and the minimization problem is carried out using a gradient line search method. A strong limitation of RS was pointed out by Vantini (2012); if the L2 norm of \( x_i \) is close to 0, then \( \mu_2 \) is also close to 0. Therefore, in some cases, the minimization algorithm could just reduce the L2 norm of \( x_i \), without forcing the curve to be similar to the template.

Our H1 selects the \( c_{ij} \)'s as to minimize a weighted H1-squared distance to the target

\[
    d_{H1}(x_i, x_{0i})^2 = \alpha \int_a^b (x_i(h_i(t); c_i) - x_{0i})^2 dt + (1 - \alpha) \int_a^b \left( \frac{d}{dt} x_i(h_i(t); c_i) - \frac{d}{dt} x_{0i} \right)^2 dt.
\]  

(2.9)

This is a Sobolev distance (Deza and Deza, 2014) which takes into account the L2 distances between the two sets of curves as well as their derivatives. Bauer et al. (2016) describes how this kind of metrics operate both on the size and the regularity of the underlying curves. Note that (2.9) carries out the registration procedure by finding, for each curve \( x_i \), the warping function \( h_i \) that minimizes the H1 distance between the aligned curve \( x_i(h_i) \) and the target function \( x_{0i} = \text{PRO}_{\Sigma_{sx}}(x_i, K_x) \). Considering also the distance between the derivatives is crucial in many applications, since it controls the smoothness of the registered curves and forces their shapes to be closer to those of the templates.

The weighting in (9) is introduced to account for the fact that curves and derivatives can have markedly different ranges. When analyzing simulated and AneuRisk data in Section 2.5 we set \( \alpha = \text{range}(x) / \text{range} \left( \frac{d}{dt} x \right) \), where ranges (for curves and derivatives) are computed evaluating curves on a grid, sorting the resulting values and taking 5th
and 95th percentiles.

Once the $c_{ij}$’s are computed for $i = 1, \ldots, n$, the functions $w_i, \ldots, w_n$ are standardized so that $\mu_{w_i}(u) = \frac{1}{n} \sum_{i=1}^{n} w_i(u) = 0$ \(\forall u \in [a, b]\) and $\int_{a}^{b} w_i(u) du = 0$ \(\forall i\). For each curve the minimization is implemented in $J_w$ dimensions through the function `newuoa` of the minqa R package. Note that, unlike RS, we do not introduce a roughness penalty. However, the warpings are curves with at most $J_w$ basis. Thus, choosing a small $J_w$ reduces computation time and, as an added bonus, limits the warpings’ complexity acting as a form of regularization. Importantly though the smallest viable $J_w$ is 3; for $J_w = 2$ the warpings are linear and cannot be bound to have the same starting and ending point, or the registration would not affect the curves at all. In simulations (Section 2.5) $J_w = 3$ gives the best results both in terms of alignment and in terms of subsequent regression performance.

Finally, for both RS and H1, each registered $x_i(h_{x_i})$ (or $y_i(h_{y_i})$) is evaluated as follows: (I) $x_i(t)$ and $h_{x_i}(t)$ are evaluated on a fine grid $t_{\text{fine}}$ to obtain $x_{i,fine}$ and $h_{x_i,fine}$; (II) $h_{x_i}^{-1}(t)$ is built with $h_{x_i,fine}$ on the abscissa and $t_{\text{fine}}$ on the ordinate axis, producing $h_{x_i,fine}$; (III) $x_i(h_{x_i}(t))$ is built with $h_{x_i,fine}$ on the abscissa and $x_{i,fine}$ on the ordinate axis.

For initialization, we could set all $c_{ij}$’s to 0 in both algorithms. However, in H1 we increase efficiency using as initial values the $c_{ij}$’s produced by RS. Also, since each curve is minimized separately, we implemented a parallel version of H1. Given the complexity of the function we minimize, computing the exact number of flops required by `newuoa` and H1 is not trivial. However, we provide information on CPU times in Subsection 2.5.3, where we analyze the AneuRisk data benchmarking H1 against other common registration procedures.

We conclude here with a few remarks. First, note that instead of determining the target functions from the SVD of $\Sigma_{xy}$, one could consider the SVD of $\Sigma_{x(h)y(h)}$ and the projections of the aligned curves $x_i(h_i)$ and $y_i(h_i)$. However, $\Sigma_{x(h)y(h)}$ is a function of all $c_{ij}$ coefficients – so one would need to solve a very complicated and computationally intensive minimization in $2n \cdot J_w$ dimensions. Second, note that both RS and H1 can also be used for FPC registration – replacing the SVD of $\Sigma_{xy}$ by the eigen-decomposition of $\Sigma_{xx}$ for $x$, and of $\Sigma_{yy}$ for $y$. These determine the target functions $x_{0i} = \text{PRO}_{\Sigma_{xx}}(x_i, K_x)$ and $y_{0i} = \text{PRO}_{\Sigma_{yy}}(y_i, K_y)$. Third, and perhaps most important for our purposes in this chapter, note that the registration could be improved applying H1 and RS iteratively. In Section 2.5, we consider only one iteration for both algorithms, since repeating the procedure does not lead to sizable improvements of the results on both the simulated data.
Scheme 1 Generation and Processing of Simulated Data

1: Generate the original predictor curves: $x^{org}_i(t), i = 1, \ldots, n$
2: Generate the regression coefficient surface $\beta(t,s)$
3: Generate Gaussian error curves: $\epsilon_i(s), i = 1, \ldots, n$
4: Compute the original response curves:
   $$y^{org}_i(s) = \int_0^b x^{org}_i(t) \beta(t,s) dt + \epsilon_i(s), i = 1, \ldots, n$$
5: Generate the warpings $h_x(t)$ and $h_y(t), i = 1, \ldots, n$
6: Misalign the curves, computing $x^{msl}_i(t) = x^{org}_i(h_x^{-1}(t))$ and $y^{msl}_i(t) = y^{org}(h_y^{-1}(t)), i = 1, \ldots, n$
7: Apply a registration procedure to the misaligned curves, producing $\hat{h}_x$ and $\hat{h}_y, i = 1, \ldots, n$
8: Compute the registered curves $x^{aln}_i = x^{msl}_i(\hat{h}_x)$ and $y^{aln}_i = y^{msl}_i(\hat{h}_y), i = 1, \ldots, n$
9: Fit the regression of $y^{aln}$ on $x^{aln}$, producing $\hat{\beta}$ and the fitted curves $y^{reg}_i, i = 1, \ldots, n$

and the AneuRisk data. Code for the H1 algorithm is provided as a supplementary file. All other codes and scripts, including those for generating simulated data, are available upon request.

2.5 Simulation study and AneuRisk data

2.5.1 Simulation settings

We next present a simulation study to investigate the performance of FCC registration in comparison to other approaches. Specifically, we run both RS and the new H1, considering as target functions the projections of the data on the FCC basis as well as the FPC basis, for a total of 4 combinations. In the following, these are indicated as $tar_{alg}(k)$, where $tar$ is the type of target function, $alg$ is the algorithm used to perform the registration, and $k$ is the number of components used to compute the projection. Thus, we compare $cc_{RS}(k), cc_{H1}(k), pc_{RS}(k),$ and $pc_{H1}(k)$ (in some instances we drop the dimension index; in these cases $k = 1$ by default). In addition, we consider two commonly used and very effective registration procedures: mean registration, indicated as $\mu_{RS}$ (Ramsay and Silverman, 2005a), and elastic registration (Srivastava et al., 2011; Tucker et al., 2013) indicated as $er$. These align curves using as template the mean and the Karcher mean, respectively. Note that, unlike the other approaches, $\mu_{RS}$ and $er$ use a common template for all the curves.

We study two simulation settings as described in Scheme 2, and for each we generate and process the data following Scheme 1. To evaluate the alignment performance of a registration procedure, we consider the total squared H1 distances between original
and aligned curves $\sum_{i=1}^{n} d_{H1}(x_i^{org}, x_i^{aln})^2$ and $\sum_{i=1}^{n} d_{H1}(y_i^{org}, y_i^{aln})^2$. To evaluate the regression performance, we consider three different criteria evaluated on a fine grid: (I) the in-sample H1 prediction error $\text{SPE} = \sum_{i=1}^{n} d_{H1}(y_i^{org}, y_i^{reg})^2$; (II) the leave-one-out H1 prediction error $\text{L1OPE} = \sum_{i=1}^{n} d_{H1}(y_i^{org}, y_i^{pred})^2$, where $y_i^{pred}$ is the predicted curve obtained with the $\hat{\beta}_{(-i)}$ from a fit without the $i$-th curves pair; and (III) the Euclidean distance between $\beta$ and $\hat{\beta}$.

In all simulations, we consider $n = 20$. To generate $x^{org}$, $\beta$, $\epsilon$, $y^{org}$, $h_x$ and $h_y$, we always start from $S = 100$ equidistant raw observations in $[0, 1]$. These are obtained as described in Scheme 2 for $x^{org}$ and the two directions of $\beta$, and from Normal distributions with mean 0 and variance specific to simulations for $\epsilon$ ($\sigma = 1, 1.3$), as to render errors commensurate to the size of the coefficients used in the expansion of the response. The raw observations for $y^{org}$ are then obtained following the model in Scheme 1. Finally, the raw observations for $h_x$ and $h_y$ are obtained as

$$h_i(t) = \frac{e^{zi} - 1}{e^{zi} - 1} \quad i = 1, \ldots, n,$$

where $z_i \sim N(0, 1)$. Based on the raw observations, we then use B-splines with $J = 20$, order 4 and smoothing parameter $\lambda = 10^{-7}$ (see (2.1)) for $x^{org}$, $\epsilon$, $y^{org}$, $h_x$ and $h_y$. We use the same parameters for the B-splines of all target functions employed in the registration procedures, the FCC and FPC directions, and for the B-splines of $x^{msl}$ and $y^{msl}$.
B-splines of both directions of $\beta$ also have $J = 20$ and order 4, but $\lambda$ (which is the same for both directions) is specific to simulations ($\lambda = 10^{-4}, 10^{-7}, 10^{-8}$, as to minimize the L1OPE of the regression performed on the original curves). The B-splines of $w_x$ and $w_y$ have order 4 and $\lambda = 10^{-7}$, but $J_w$ is set to 3 as discussed in Section 2.4.

Note that in the two simulation scenarios, both predictor and response curves are designed to form two different groups. This allows us to better illustrate the advantages of low-dimensional registration procedures in comparison to $\mu_{RS}$ and $er$ – which are not capable of retaining the separation between groups. Also, in all simulations, the responses are constructed without adding an intercept; both the original $x$’s and $y$’s have means very close to 0. For this reason, we do not include the intercept $\alpha$ when applying the linear model (2.4). For both FCC and FPC registration, and in both simulation scenarios, we choose $K_x = K_y = 1$ based on the criteria described in Section 2.3.2 (see Figure 2.1). In both simulations, the first component of $\Sigma_{xy}$ explains more than 90% of the covariation between $x$ and $y$ and the TPV index grows considerably between $K = 1$ and $K = 2$, becoming very similar to that of the misaligned curves. The L1OPE is minimized for $K = 1$ and increases significantly for $K > 1$ for both $cc_{RS}$ and $cc_{H1}$. Furthermore, as we can see in Appendix Figures A.4 and A.5, just one component is sufficient to capture the two groups while markedly reducing the complexity of the target curves. In contrast, $\text{PRO}_{\Sigma_{xy}}(x, 2), \text{PRO}_{\Sigma_{xy}}(y, 2), \text{PRO}_{\Sigma_{xx}}(x, 2)$ and $\text{PRO}_{\Sigma_{yy}}(y, 2)$ are already too similar to the unregistered curves. Importantly we also note that, in Simulation 2, $\text{PRO}_{\Sigma_{xy}}(x, 1)$ differs markedly from $\text{PRO}_{\Sigma_{xx}}(x, 1)$, and $\text{PRO}_{\Sigma_{xy}}(y, 1)$ from $\text{PRO}_{\Sigma_{yy}}(y, 1)$. In this scenario, the FCC and FPC registrations are guided by different templates. We thus expect them to lead to different results.

### 2.5.2 Simulation results

Salient results are summarized in Figures 2.2 and 2.3. Figure 2.2 shows performance in aligning predictor and response curves, and predicting the latter (out of sample) with different registration procedures. These performance values, together with the SPE and the distance between $\beta$ and $\hat{\beta}$, are reported in Appendix Table A.1. Figure 2.3 shows $y^{\text{pred}}$, the leave-one-out predicted response curves, from function-on-function regressions on the original data (which are perfectly aligned), the data after misalignment, and the data after the application of different registration procedures. The true $\beta$’s of the two simulations are shown in Appendix Figure A.3.

In **Simulation 1** the $x$ curves are characterized by two peaks and are easy to align. The surface $\beta$ is smooth, so also the $y$ curves are easy to align. In Figure 2.3 (upper
Figure 2.1: Criteria used to aid in the choice of $K$ (horizontal axis), the number of covariance components used to compute the registration templates. The left-most and center-left panels show the cumulative sum of the first $K$ eigenvalues of $\Sigma_{xy}$ and the Total Phase Variation index in (2.7), rescaled so that 0 corresponds to $TPV(\text{PRO}_{\Sigma_{xy}}(x, 1), \text{PRO}_{\Sigma_{xy}}(y, 1))$ and 1 corresponds to the TPV of the misaligned curves. The center-right and right-most panels show the L1OPE associated with the curves aligned by $cc_{RS}$ and $cc_{H1}$, respectively – this too is rescaled so that 1 corresponds to the L1OPE of the misaligned curves. Blue and red lines correspond to results for two simulation scenarios, and the green line corresponds to results for the AneuRisk data.

panels) we see that the $y^{pred}$ curves produced by $er$ do not split in two groups. The $y^{pred}$ curves obtained from low-dimensional registration with FPC and FCC appear very similar – which is not surprising given the easy nature of the registration problem in this scenario. However, in Figure 2.2 we see that FCC outperforms FPC in terms of both alignment and regression. Moreover, our new H1 algorithm outperforms RS. Regarding estimation of $\beta$ (see Appendix Table A.1), all procedures have similar performance. Appendix Figure A.6 shows how closely $\hat{\beta}$ obtained from data aligned with our procedure resembles the surface estimated from $x^{org}$ and $y^{org}$ (prior to misalignment). In fact, as can be seen in Appendix Figure A.1, the registered curves are very similar to the original ones.

In Simulation 2 the $x$ curves have multiple peaks and are difficult to align. The surface $\beta$ is also rather rough, leading to hard to align $y$ curves. In this scenario FCC, and in particular $cc_{H1}$, gives the best alignment for both the predictor and the response, as we can see in Figure 2.2. Indeed, the registered curves produced by $cc_{H1}$ (shown in Appendix Figure A.2) are very close to the original ones. Notably, $cc_{H1}$ improves the performance by a factor of 10 with respect to $er$ and $\mu_{RS}$ and by a factor of 3 with respect to $pc_{H1}$ – and the latter fails in predicting the order of magnitude of some curves.
Figure 2.2: Simulation results. \( aln(x) \) and \( aln(y) \) are the \( H1 \) distances between aligned and original curves, and \( L1OPE \) is leave-one-out \( H1 \) prediction error. On the horizontal axis “msl” represents the misaligned curves, followed by curves registered with different procedures – as indicated. y-axis: Values on the vertical axis have been rescaled between 0 and 1, where 0 corresponds to results obtained with the original curves.

In terms of regression results, \( H1 \) again dominates \( RS \), and \( cc_{H1} \) outperforms all other registration procedures. This is also evident in Figure 2.3 (middle panels): \( y_{\text{pred}} \) obtained from data aligned with \( cc_{H1} \) and \( cc_{RS} \) are very similar to \( y_{\text{pred}} \) obtained from the original curves (prior to misalignment). Again, \( y_{\text{pred}} \) from \( \mu_{RS} \) and \( er \) do not clearly split in two groups, and the curves registered with \( \mu_{RS} \) and \( er \) resemble the original \( x_{\text{org}} \) and \( y_{\text{org}} \) even less than their misaligned versions \( x_{\text{msl}} \) and \( y_{\text{msl}} \). Appendix Figure A.7 shows the estimated \( \beta \). None of the algorithms captures the peaks that characterize the true \( \beta \) and the \( \beta \) estimated on the original curves. This is due to the difficulty of the registration problem: since aligned curves are not close enough to \( x_{\text{org}} \) and \( y_{\text{org}} \), the relationship between \( x \) and \( y \) is described by a surface that resembles the true \( \beta \).

Finally Appendix Table A.2 reports the same quantities described in Appendix Table A.1, but considering the L2 distance instead of the H1 distance. This does not affect the relative performance of the different methods. In conclusion, in both simulation scenarios the FCC registration, and in particular the one based on our novel \( H1 \) algorithm, improves the alignment of both predictor and response curves. More importantly, FCC improves the quality of the subsequent regression analysis, with a considerable reduction of the \( L1OPE \) with respect to all other techniques considered.

### 2.5.3 AneuRisk data

We apply our new registration procedure to the \textbf{AneuRisk} dataset described in \cite{Sangalli2014}. This data was collected by image reconstruction of three-dimensional
Figure 2.3: $y^{pred}$ curves obtained with different regressions. “Original” indicates the regression performed on the original curves, and “misaligned” the one performed on the misaligned curves. In all other panels the regression is performed on curves registered with different procedures – as indicated. The two different colors indicate the two groups in each set of curves.

cerebral angiographies of 65 patients, with the aim of investigating the interplay between morphological properties of artery walls and hemodynamic factors, and shed light on the possible causes of aneurysmal pathology. Patients in the study belonged to two groups: patients with an aneurysm on the Willis circle, after the final bifurcation of the Internal Carotid Artery (ICA), and patients with an aneurysm on the last tract of the ICA or without an aneurysm. In this chapter we analyze the relationship between the axial derivative of the local average of the ICA wall shear stress (WSS1), our functional response $y$, and the ICA curvature, our functional predictor $x$. $y$ is a hemodynamic factor obtained via computational fluid dynamics in the ICA geometries (Passerini et al., 2012). $x$ is a morphological feature of the ICA and is computed as described in Sangalli et al. (2009a). Since $y$ is available only for 52 out of the 65 patients, the sample size for
Table 2.1: AneuRisk data. SPE and L1OPE for regressions performed on misaligned data and on data produced by different registration procedures. For FCC and FPC performed with H1 and RS, results are shown for both 1 and 2-dimensional approximations. Values are rescaled to have SPE and L1OPE equal to 100 on the misaligned data. The last row reports CPU times (in seconds) required by the different registration procedures.

<table>
<thead>
<tr>
<th></th>
<th>ms1</th>
<th>μ_RS</th>
<th>cr</th>
<th>p_RS(1)</th>
<th>cc_RS(1)</th>
<th>p_RS(2)</th>
<th>cc_RS(2)</th>
<th>p_CH1(1)</th>
<th>cc_CH1(1)</th>
<th>p_CH1(2)</th>
<th>cc_CH1(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPE</td>
<td>100</td>
<td>108</td>
<td>38</td>
<td>135</td>
<td>106</td>
<td>134</td>
<td>80</td>
<td>114</td>
<td>85</td>
<td>93</td>
<td>77</td>
</tr>
<tr>
<td>L1OPE</td>
<td>100</td>
<td>95</td>
<td>33</td>
<td>119</td>
<td>114</td>
<td>123</td>
<td>92</td>
<td>98</td>
<td>87</td>
<td>90</td>
<td>79</td>
</tr>
</tbody>
</table>

CPU time:

- 119
- 96
- 95
- 98
- 95
- 84
- 1438
- 623
- 1984
- 285

For both x and y, we have 450 raw observations for each of the n = 52 subjects. Their domain is transformed to be [0, 1], and we use B-splines with J = 20, order 4 and smoothing parameter $\lambda = 10^{-7}$ to represent both x and y. Again, the same parameters are used for the basis of the FPC and FCC directions, the basis of the projections, and the basis of both the directions of $\beta$. The smoothing parameter for $\beta$, $\lambda_\beta$, varies for each
registration and is selected to minimize the L1OPE. We do not include the intercept \( \alpha \) in (2.4), since a linear model without intercept improves the L1OPE for all the procedures. Just as in the simulation study, the basis of the \( w \)'s are B-splines with \( J_w = 3 \), order 4, and smoothing parameter \( \lambda = 10^{-7} \) (increasing the number of basis elements here slows down the computation and does not appreciably improve performance).

For this application, we measured CPU times (in seconds) for the different registration procedures. Times obtained with a MacBook Pro 2017 with two 3.3 GHz Intel Core i7 processors are reported in Table 2.1. Due to the complexity of the newuoa algorithm, H1 registration takes longer than \( er \) and RS. However, H1 is the only registration algorithm with a parallel implementation; this allows one to substantially reduce its cost using...
multiple processors as needed. Focusing on H1, note also that FPC registration is more expensive than FCC registration. Table 2.1 also contains SPE and L1OPE values achieved by different regressions. Here, SPE and L1OPE of each procedure are computed contrasting predicted curves to the aligned response curves produced by the procedure itself. For this reason, comparisons among procedures should be interpreted with caution. Registering with the RS algorithm, both in FPC and FCC, leads to poor regression performance – sometimes worse than that achievable on the misaligned curves. On the contrary, registering with the H1 algorithm improves regression performance (over the misaligned curves) in all cases. Specifically, $cc_{H1(1)}$ and $cc_{H1(2)}$ give the second best performances both in SPE and in L1OPE.

Notably, er registration appears to outperform all other procedures. However, this is due to the fact that er produces very complex warping functions, which capture too much variability and modify the nature of the curves. To illustrate this behavior, consider the curvature ($x$) and WSS1 ($y$) of one patient reported in Figure 2.4. $cc_{H1(2)}$ moves and changes the amplitude of the curves’ features, but does not affect their number and the general pattern of the curves themselves. On the other hand, er leads to over-smoothed curves for both $x$ and $y$ and does not preserve their features: $x$ has 3 peaks instead of 4, and $y$ has 2 peaks instead of 3. These over-smoothed $x$ and $y$ are more similar to each other, which explains why it is easier to predict $y$ from $x$; regression performance increases, but this entails an unnatural change in the curves’ shape and might lead to the loss of important information on the relationship between response and predictor. Predicted response curves based on all registration procedures are displayed in Appendix Figure A.9, and estimated $\beta$’s in Appendix Figure A.11. Appendix Figure A.10 shows again predicted response curves based on all registration procedures – but focusing on just one patient. The original misaligned response curve of the patient is shown as well. Here one can clearly appreciate how only the curves predicted by FCC procedures retain all the peaks of the original curve, while other procedures tend to combine together the two highest peaks. Figure 2.5 shows the misaligned curves for curvature and WSS1, the curves registered by $cc_{H1(2)}$, and the predicted response curves based on $cc_{H1(2)}$ and $cc_{H1(1)}$. In terms of $x$ registration, $cc_{H1(2)}$ combines all the curves’ peaks in two big groups, one at about 0.3 and one at about 0.6. The registration of $y$, on the contrary, does not seem to reveal notable common landmarks or substantially modify the original curves – which have a more complex nature than the $x$’s. In terms of predicted response curves, $cc_{H1(1)}$ captures the general pattern but also clearly misses some modes of variability in the aligned $y$’s. As expected, curves predicted based on $cc_{H1(2)}$ exhibit more variability (two
components can generate more complex predictions). However, they are still substantially
different from the registered $y$’s.

The results presented here are still rather partial, but they are interesting and
informative, and allow us to appreciate the advantages of the FCC registration. Indeed,
FCC is the only approach which preserves the original features both in registration
and in prediction, while improving prediction quality with respect to all other methods
(except for er). Being able to generate accurate prediction without losing critical feature
information (e.g., the number of peaks) is crucial for this application. Indeed, a good
prediction of the wall shear stress of the internal carotid artery, would allow one to avoid
its direct computation, which is time consuming and computationally intense.

Beyond approximating wall shear stress, the broader purposes in analyzing this data
concern medical diagnosis. For instance, it would be important to accurately classify
patients in two groups based on the position of their aneurysm. Further extensions of our
FCC registration may provide relevant insight; in particular, it would be very useful to
design a generalization of the procedure able to handle multiple functional predictors at
once, and/or curves with different domains. Indeed, considering several predictors could
help explain variation in the first derivative of wall shear stress. Also interestingly, the
ability to simultaneously register the response and multiple predictors would allow us to
align the three spatial coordinates of the ICA centerline before computing the curvature,
instead of starting directly from the curvature computed on coordinates already aligned
by Sangalli et al. (2009a). More details on this extension are given in the next section.

2.6 Current developments

In this section, we briefly present an ongoing project that aims to develop a multivariate
version of our Functional Covariate Components registration (mFCC), which is able
to register curves with different domains and to tackle regressions with more than one
functional predictor. The two main novelties of mFCC with respect to FCC are (i) the use
of warping functions which allow horizontal translations; (ii) the definition of a functional
subspace which captures the relationships between more than two set of curves.

Following Sangalli et al. (2009a), we consider warping functions of the form

$$h(t) = m_0 + m_1 t,$$  \hspace{1cm} (2.11)

where $m_0, m_1 \in \mathbb{R}$. These warpings are linear transformations of the time argument $t$.  

24
Scheme 3 Definition of the template functions in multivariate FCC

1: Consider $S$ sets of curves $x^1, \ldots, x^S$:

2: **Reduction Step.** For each set of curves $s$, find the projection of the curves on their first $k$ principal components:

$$\text{PRO}^1 = \text{PRO}_{xx}(x^1, k), \ldots, \text{PRO}^S = \text{PRO}_{xx}(x^S, k)$$

3: **Regression Step.** For each set of projection curves $s$, perform a functional regression which considers $s$ as response variable and all the other projection sets as predictors:

$$\text{PRO}^s(s) = \alpha(s) + \left( \sum_{i \neq s} \int_a^b \text{PRO}^i(s) \beta(s, t) dt \right) + \epsilon(s) \quad (2.10)$$

4: **Target Functions.** Define the target functions for each set $s$ as the fitted curves $\hat{\text{PRO}}^s$ from the regression in step 2:

$$x_0^1 = \hat{\text{PRO}}^1, \ldots, x_0^S = \hat{\text{PRO}}^S$$

The coefficient $m_0$ produces a horizontal translation of the domain. Thus, the registration procedure is well defined even when the original curves have different domains. The coefficient $m_1$ creates a dilation – allowing one to increase or reduce the size of the features. Note that $m_1$ produces the same transformation throughout the domain. By contrast, the warpings defined in (2.2) are more flexible – allowing a curve to have different levels of dilation in different regions of the domain. However, warpings of the form (2.11) require the estimation of just two parameters. This simplifies the optimization algorithm, reducing computational burden and increasing stability. Moreover, the fact that this set of warpings is closed under composition makes them perfectly suited for iterative procedures.

The main challenge of mFCC is the definition of a subspace which contains information relevant to the relationships between more than two set of curves. We are considering several possible development routes. One route follows Chiou et al. (2014), who introduced a multivariate version of functional PCA allowing one to simultaneously characterize the variation and covariation of several sets of curves. Another route we are considering extends the tensor canonical correlation analysis for multivariate data (Luo et al., 2015) to the functional framework.
We propose the procedure described in Scheme 3 to estimate the templates of mFCC. In the first step, we separately and independently perform dimension reduction for each
variable, i.e. each set of curves. In particular, for each set, we compute the projections on the first $k$ principal components of the curves. In the second step, we use these projections to find a subspace which summarizes the relationship between all sets of curves. The key idea is to regress each set of projections on all others and take the resulting fitted curves as templates curves for our procedure. Our intuition is that these fitted curves span the subspace which best captures the relationship between response and predictors. Indeed, just as in the scalar framework, the functional linear regression estimates are nothing but the projections of the original curves in the space spanned by all predictors, i.e. the minimum distance approximation of the original curves in the subspace which contains all predictors’ information. Note that the procedure described in Scheme 3 is symmetric in the functional variables, as was FCC – where the covariance gives the same roles to response and predictor curves. Also mFCC does not depend on which functional variable we will elect as response in the post-registration regression analysis. Even though the regression framework is intrinsically asymmetric, our proposal aligns multiple sets of curves simultaneously without differentiating their roles ex-ante. It captures their modes of dependence in a symmetric way, allowing any of the variables to be regressed on the others post-registration.

We will investigate in detail the procedure in Scheme 3 and provide a fully characterized mathematical structure to back up our intuition. Notably though, such a registration procedure would be extremely flexible, extending the FCC to functional regression frameworks comprising multiple functional, scalar, and even categorical predictors. To incorporate the information given by scalar and categorical predictors in the procedure, one can simply add these variables in the regression formula (2.10). But we do not perform a dimension reduction on these non-functional variables, since they already live in a finite space. By contrast, the main aim of dimension reduction – which can be seen as a “de-noising” step – is to represent the curves in a finite dimensional subspace and highlight paths and structures relevant to describe their relationships.

We are applying mFCC to study the Milano Weather Station Data made available by the Open BigData challenge. These data contain pollution and weather-related variables measured in different locations in the city of Milan from November 1st 2014 to December 31st 2014. Due to the longitudinal nature of the observations and the variability of the sensors used to collect them, a functional registration technique would adequately reflect the nature of the data. Our goal is to exploit the power of mFCC to improve a functional regression analysis and better understand the relationship between pollutants (such as
levels of carbon monoxide, benzene, nitrogen, etc.) and weather-related variables (such as temperature, precipitation, humidity, etc.).

Figure 2.6 shows some preliminary results for the Milano Weather Station Data. We plot the curves relative to 10 different days for temperature, wind speed and benzene. For each variable, we display the original curves, the projections on the first 2 principal components, the target functions computed as described in Scheme 3 and the registered curves by mFCC. Note how the registered curves (last row) have different domains: mFCC allows horizontal translation during the registration process.

2.7 FCC and Sufficient Dimension Reduction

We conclude this chapter by pointing out an important interpretation of FCC registration in the context of Sufficient Dimension Reduction (SDR). SDR is a set of techniques to handle regression problems with a large number of predictors (see Li, 1991; Cook and Weisberg, 2009; Adragni and Cook, 2009; Ma and Zhu, 2013 for details). Unlike variable selection, which assumes that among all available predictors only a few are truly related to the response, SDR assumes that the response depends only a few linear combinations – potentially loading on many, or even all predictors. The goal of SDR is thus to estimate a minimal, so-called central subspace able to capture all regression information. This framework has recently been extended to the case of functional data (Lee et al., 2013; Li et al., 2017). The link between covariance components and SDR can be evinced from the work of Borga et al. (1997) and Li and Duan (1989) in the finite-dimensional multivariate context. Borga et al. (1997) point out the relationship between Canonical Correlation Analysis and the direction identified by the OLS regression. Li and Duan (1989) prove that, under some assumptions on the distribution of the predictors, OLS can be considered as the very first and simplest SDR technique when the central subspace has dimension 1. In the infinite-dimensional functional framework, the reduction produced by our FCC registration can in fact be seen as an SDR procedure – estimating a subspace of dimension 1 (or larger) which captures the function-on-function regression information. Many interesting developments can be pursued based on this connection, such as finding a registration procedure based on other SDR techniques, e.g. Sliced Inverse Regression (SIR) (Li, 1991) and its several functional versions (fSIR) – see Ferré and Yao (2003), Ferré and Yao (2005), and Wang et al. (2015).
Chapter 3  
A Highly-Efficient Group Elastic Net Algorithm with an Application to Function-On-Scalar Regression

3.1 Introduction

As problems involving very large and potentially structured data become ever more ubiquitous, attention is being devoted to the integration of approaches and techniques from the areas of Feature Selection and Functional Data Analysis (FDA). Indeed, more and more regression applications comprise a large number number of variables – some of which are scalar and some of which are suitable for a functional representation, such as longitudinal measurements or biomedical images (Sørensen et al., 2013; Ullah and Finch, 2013; Cremona et al., 2019). A great deal of recent work has been concerned with feature selection in these applications. Matsui and Konishi (2011); Gertheiss et al. (2013); Fan et al. (2015) study the case where the response is scalar and the features are functional. Chen et al. (2016); Fan and Reimherr (2016); Barber et al. (2017); Parodi et al. (2018); Mirshani and Reimherr (2019) tackle the so called function-on-scalar case, where the response is functional and the features are scalar – focusing on settings in which the number of features is bigger than the number of observations. However, recent developments in optimization have demonstrated that substantial computational gains can still be made when the number of features is massive (e.g. ~1e6). In this work, we present Functional Group Elastic Net (fgen), a novel and highly efficient method to
solve the function-on-scalar feature selection problem in ultra-high-dimensional settings – where the number of features is indeed massive and much larger than the number of observations. The ability to solve these problems with a lower computational burden is increasingly critical. Given the complex, noisy nature of much contemporary data, changing some aspects of their pre-processing or some of the tuning parameters involved in the analysis can lead to completely different results (Krawczyk and Cano 2018; Murdoch et al. 2019). For this reason, repeating an analysis multiple times (e.g., with different choices of data preprocessing pipelines, or to tune certain meta-parameters) is paramount to capture significant signals and ensure the stability of outcomes (Yu and Kumbier, 2020). Substantial reductions in computational burden enable such repetition, allowing scientists and practitioners to conduct truly meaningful and reproducible analyses.

Group Elastic Net incorporates the group structure (Yuan and Lin 2006) and the Elastic Net penalty (Zou and Hastie 2005) into a penalized regression framework. The former allows one to represent each feature (or component) by a group of variables. The latter induces sparsity and regularizes the estimates. We consider the case where all groups have the same size \( k \). The minimization problem is formulated as follows:

\[
\min_B \left( \frac{1}{2} \| XB - Y \|_2^2 + \lambda_1 \sum_{i=1}^p \| B_i \|_2 + (\lambda_2/2) \sum_{i=1}^p \| B_i \|_2^2 \right). \tag{3.1}
\]

Let \( p \) be the number of features, \( n \) the number of statistical units, and \( \| \cdot \|_2 \) the \( l_2 \) norm for matrices, i.e. the Frobenius norm, and vectors. Then, \( X \in \mathbb{R}^{n \times p} \) is the design matrix (that we assume to have standardized columns), \( Y \in \mathbb{R}^{n \times k} \) the response matrix, and \( B \in \mathbb{R}^{p \times k} \) the coefficient matrix. In other words, (3.1) describes a sparse multi-task model where the response and each of the features are represented by a group of \( k \) coefficients (Zhang and Yang 2018). Throughout this chapter, we follow the notation in Johnson et al. (2014) and we use the subscripts \( i \) and \( (i) \) to indicate the \( i \)-th row and the \( i \)-th column of a matrix, respectively. Thus, \( B_i \in \mathbb{R}^k \) are the coefficient values associated with the \( i \)-th group, and \( X_{(i)} \in \mathbb{R}^n \) are the observed values relative to the \( i \)-th feature. Before proceeding, note that (3.1) can be expressed as

\[
\min_B \left( h(XB) + \pi(B) \right), \tag{P}
\]

where \( h(XB) = (1/2) \| XB - Y \|_2^2 \) is the least-squares loss function and \( \pi(B) = \sum_{i=1}^p \pi(B_i) = \lambda_1 \sum_{i=1}^p \| B_i \|_2 + (\lambda_2/2) \sum_{i=1}^p \| B_i \|_2^2 \) is the Group Elastic Net penalty function. The first term in \( \pi \) is not differentiable and creates sparsity at the group level, i.e., if a
component is selected, then all its coefficients are selected and vice-versa. The second term is a Ridge-type penalty which reduces model complexity and tries to control variance inflation due to multicollinearity – feature selection models are indeed known to be less effective and not reliable in scenarios characterized by very high collinearity among features (Katrutsa and Strijov, 2015). \( \lambda_1 \) and \( \lambda_2 \) are penalty parameters \( > 0 \) and control the weight of the two penalties with respect to the least square loss.

To solve (3.1), we develop a new Semi-smooth Newton Augmented Lagrangian (SsNAL) algorithm. We then extend it to the function-on-scalar regression framework by means of Functional Principal Components (FPC) (James et al., 2000; Chiou et al., 2004; Hall and Hosseini-Nasab, 2006). SsNAL exploits the sparsity induced by the augmented Lagrangian second order information to guarantee a super-linear convergence and greatly reduce the computational cost. This methodology, first introduced by Tomioka and Sugiyama (2009) and Tomioka et al. (2011), has been recently used in several applications, e.g., to regular Lasso (Li et al., 2018), constrained Lasso (Deng and So, 2019), and Elastic-Net (Boschi et al., 2020). However, incorporating the group structure significantly increases the dimension of the problem. Indeed, (3.1) is not separable and the optimization must be carried out jointly across the coordinates of the outcome. Therefore, considering a new group penalty while preserving the efficiency of the method requires that we carefully redefine a set of all-new mathematical operators and the theory behind them.

We implemented an efficient version of \texttt{fgen} in \texttt{python} and benchmarked it against the two best Group Elastic Net solvers we found in the literature: the \texttt{python} package \texttt{sklearn} (Pedregosa et al., 2011) and the \texttt{R} package \texttt{glmnet} (Friedman et al., 2010), which is written in \texttt{fortran}. Both of these solvers implement a highly optimized coordinate descent algorithm (Friedman et al., 2010; Breheny and Huang, 2015) and outperform competitors such as FISTA (Beck and Teboulle, 2009; Bonnefoy et al., 2015), ADMM (Deng et al., 2013; Zhu, 2017), and proximal gradient (Chen et al., 2010) by at least one order of magnitude in terms of CPU time. Our simulation results demonstrate that in sparse scenarios \texttt{fgen} is at least 3 times faster than \texttt{glmnet} and more than 10 times faster than \texttt{sklearn}. We also applied \texttt{fgen} to the Intervention Nurses Start Infants Growing on Healthy Trajectories (INSIGHT) study (Paul et al., 2014), which investigates risk factors for childhood obesity. Specifically, we examined the association between hundreds of thousand of Single Nucleotide Polymorphisms (SNPs) and growth curves, which represent a functional outcome.

The remainder of the chapter is organized as follows. In Section 3.2 we describe the
Group Elastic Net problem and introduce some preliminary results. In Section 3.3 we present our new methodology and illustrate how to extend it to a function-on-scalar feature selection problem. In Section 3.4 we investigate the performance of our method on simulated data and apply it to data from INSIGHT. Proofs of theoretical results and additional simulations are included in Appendix B. The fgen code is available at https://github.com/tobiaboschi/fgen

3.2 Preliminaries

In this section we define the Group Elastic Net problem and we introduce some results related to Fenchel conjugate functions and proximal operators, which are essential tools in our developments.

3.2.1 Fenchel conjugate function and proximal operator of \( \pi(B) \)

Fenchel conjugate functions (Fenchel, 1949) allow one to more readily define the dual problem (Boyd and Vandenberghe, 2004) of (P), which is called the primal problem. Let \( \mathcal{X} \subseteq \mathbb{R}^p \) be a convex set and \( f : \mathcal{X} \rightarrow \mathbb{R} \). Then, the conjugate function of \( f \) is \( f^* : \mathcal{X}^* \rightarrow \mathbb{R} \) defined as

\[
\pi^*(Z) = \sup_{x \in \mathcal{X}} \langle Z, x \rangle - f(x),
\]

where \( \mathcal{X}^* = \{ z : \sup_{x \in \mathcal{X}} (\langle z, x \rangle - f(x)) < \infty \} \). \( \langle \cdot, \cdot \rangle \) indicates the inner product, i.e., the dot product. If \( \mathcal{X} \subseteq \mathbb{R}^{p \times k} \), i.e., if \( z \) is a matrix, the definition is still valid but \( \langle \cdot, \cdot \rangle \) is the Frobenius inner product. In our first proposition we provide a closed form solution for the Group Elastic Net penalty conjugate function (see Appendix Section B.1.1 for a proof).

**Proposition 1.** Given \( Z \in \mathbb{R}^{p \times k} \), the conjugate function of \( \pi \) has the form

\[
\pi^*(Z) = \sum_{i=1}^{p} \pi^*(Z_i) = (2\lambda_2)^{-1} \sum_{i=1}^{p} \left( [\|Z_i\|_2 - \lambda_1]_+ \right)^2,
\]

where \( [\cdot]_+ \) is the positive part operator; \( [s]_+ = s \) if \( s > 0 \) and 0 otherwise.

Note that \( \pi^*(Z) \) is a continuous differentiable function. This is a more general result than the one presented in Li et al. (2018), Boschi et al. (2020), because we extend the definition of \( \pi^* \) to the case where \( Z \) is a matrix and not just a vector. In the simple scenario where \( k = 1 \), i.e., when every group consists of just one variable, we obtain again the conjugate function of the standard Elastic Net penalty. Notably, starting from a non-separable objective function, we derive a \( \pi^* \) which does separate. As we will see in
Section 3.3.1, this allows one to induce a new level of sparsity in the Lagrangian problem which is actually key for the massive computational advantage offered by \texttt{fgen}.

**Proximal operators** \cite{Rockafellar1976} are fundamental in many optimization algorithms. Given a lower semi-continuous convex function $f : \mathbb{R}^p \to \mathbb{R}$, the proximal operator of $f$ at $x$ with parameter $\sigma > 0$ is denoted as $\text{prox}_{\sigma f}(x)$ and defined as

$$\text{prox}_{\sigma f}(x) = \arg \min_t \left( f(t) + (2\sigma)^{-1} \|t-x\|_2^2 \right).$$

If $f : \mathbb{R}^{p \times k} \to \mathbb{R}$, i.e. if $x$ is a matrix, then $\text{prox}_{\sigma f} : \mathbb{R}^{p \times k} \to \mathbb{R}^{p \times k}$ and $\|\cdot\|_2$ is the Frobenius norm. \cite{Parikh2014} and \cite{Beck2017} (Chapter 6) provide numerous examples and properties. Combining their results, one can easily find the form of the proximal operator of $\pi(B)$ provided in our second proposition (see Appendix Section B.1.2 for a proof).

**Proposition 2.** The proximal operator of $\pi(B)$ is:

$$\text{prox}_{\sigma \pi}(B) = \left( \text{prox}_{\sigma \pi}(B_1), \ldots, \text{prox}_{\sigma \pi}(B_p) \right)^T,$$

where

$$\text{prox}_{\sigma \pi}(B_i) = (1 + \sigma \lambda_i)^{-1} \left[ 1 - \frac{1}{\|B_i\|_2^2} \sigma \lambda_i \right] B_i. \quad (3.3)$$

Note that $\text{prox}_{\sigma \pi}(B) : \mathbb{R}^{p \times k} \to \mathbb{R}^{p \times k}$. To implement \texttt{fgen}, one also needs the proximal operator of $\pi^*$, which can be obtained through the Moreau decomposition:

$$x = \text{prox}_{\sigma \pi}(x) + \sigma \text{prox}_{\pi^*/\sigma}(x/\sigma), \quad \sigma > 0. \quad (3.4)$$

### 3.2.2 Dual formulation and Augmented Lagrangian

Here we introduce the dual Group Elastic Net problem and its augmented Lagrangian. From \cite{Boyd2004}, a possible dual formulation of \((P)\) is

$$\min_{V,Z} \left( h^*(V) + \pi^*(Z) \right) \quad \text{s.t.} \quad X^TV + Z = 0 \quad (D)$$

where $V \in \mathbb{R}^{n \times k}$ and $Z \in \mathbb{R}^{p \times k}$ are the dual variables matrices. In particular $V_i, Z_i \in \mathbb{R}^k$ are the dual variables associated with the $i$-th group. $h^*$ and $\pi^*$ are the Fenchel conjugate functions of $h$ and $\pi$, respectively. Specifically, $h^*(V) = (1/2) \|V\|_2^2 + \langle Y, V \rangle$ and $\pi^*(Z)$ is given in Proposition 1. We can now define the augmented Lagrangian function and the \textit{Karush-Kuhn-Tucker} (KKT) system associated with \((D)\). The augmented Lagrangian is given by

$$\mathcal{L}_\sigma(V, Z, B) = h^*(V) + \pi^*(Z) - \sum_{i=1}^p \langle B_i, V^TX_{(i)} + Z_i \rangle + (\sigma/2) \sum_{i=1}^p \|V^TX_{(i)} + Z_i\|_2^2, \quad (3.5)$$
where \( \sigma > 0 \). \( B \) is both the primal variable and the Lagrangian multiplier which penalizes the constraints’ violations. The KKT system is given by the following three equations:

\[
\nabla h^*(V) - XB = 0, \quad 0 = \nabla \pi^*(Z) - B = 0, \quad X^TV + Z = 0. \tag{3.6}
\]

Note that \( \nabla h^*(V) = V + Y \). A closed form of \( \nabla \pi^*(Z) \) is not essential for our SsNAL method. The KKT equations will be useful to determine the convergence of our algorithm, since the set \((V^*, Z^*, B^*)\) solves the KKT (3.6) if and only if \((V^*, Z^*)\) and \(B^*\) are the optimal solutions of \((D)\) and \((P)\), respectively (Boyd and Vandenberghe, 2004).

### 3.3 Methodology

In this section we present our new methodology. First, we introduce a SsNAL algorithm to solve the Group Elastic Net problem. Next, we illustrate how to extend it to the function-on-scalar regression framework. Finally, we describe how to implement a solution path over different values of \( \lambda_1 \).

#### 3.3.1 SsNAL method

The SsNAL method is summarized in Algorithm 1. It consists of an Augmented Lagrangian method characterized by an inner subproblem. The subproblem is solved with a Semi-smooth Newton method which exploits the sparsity of the augmented Lagrangian second order information and greatly reduces computational costs. We now provide the details of its implementation and some important theoretical results.

From Rockafellar (1976a), one can find the optimal solution of \((D)\) by solving the Augmented Lagrangian method described in Algorithm 1. The essential part of the algorithm is the subproblem (3.7). As described in Li et al. (2018), an approximate solution \((\bar{V}, \bar{Z})\) for a given \(B\) can be found as

\[
\bar{V} = \arg \min_V \mathcal{L}_\sigma \left( V \mid \bar{Z}, B \right), \quad \bar{Z} = \arg \min_Z \mathcal{L}_\sigma \left( Z \mid \bar{V}, B \right). \tag{3.10}
\]

With a slight abuse of notation, we indicate by \( L_\sigma(V \mid Z, B) \) the function \( L_\sigma(V, Z, B) \) where the parameter \( Z \) and \( B \) are fixed. Similarly for \( L_\sigma(Z \mid V, B) \). Our third proposition provides explicit forms for \( \mathcal{L}_\sigma(V \mid \bar{Z}, B) \) and \( \bar{Z} \) (see Appendix Section B.1.3 for a proof).

**Proposition 3.** Define \( \psi(V) := \mathcal{L}_\sigma \left( V \mid \bar{Z}, B \right) \). Then, for the Group Elastic Net problem
Algorithm 1 Semi-smooth Augmented Lagrangian (SsNAL) method

**Augmented Lagrangian method**

Start from the initial values $V^0, Z^0, B^0, \sigma^0$

**while** not converged **do**

(1) Given $B^k$, find $V^{k+1}$ and $Z^{k+1}$ which approximately solve the inner subproblem

\[
(V^{k+1}, Z^{k+1}) \approx \arg \min_{V, Z} \mathcal{L}_\sigma (V, Z | B^k) \tag{3.7}
\]

(2) Update the Lagrangian multiplier $B$ and the parameter $\sigma$:

\[
B^{k+1} = B^k - \sigma_k (X^T V^{k+1} + Z^{k+1})
\]

\[
\sigma^{k+1} \uparrow \sigma^\infty \leq \infty \tag{3.8}
\]

**end while**

**Semi-smooth Newton method for (3.7)**

To solve (3.7) and find $(V^{k+1}, Z^{k+1})$:

**while** not converged **do**

(1) Find the descent direction $D^j$ solving exactly or by conjugate gradient the linear system

\[
\partial^2 \psi (V^j) \text{vec}(D^j) = -\text{vec} (\nabla \psi (V^j)) \tag{3.9}
\]

(2) **Line search** [Li et al., 2018]: choose $\mu \in (0, 1/2)$ and reduce the step size $s^j$ until

\[
\psi (V^j + s^j D^j) \leq \psi (V^j) + \mu s^j \langle \nabla \psi (V^j), D^j \rangle
\]

(3) Update $V$: $V^{j+1} = V^j + s^j D^j$

(4) Update $Z$: $Z^{j+1} = \text{prox}_{\frac{\pi}{\sigma}} \left( \frac{B^k}{\sigma} - X^T V^{j+1} \right)$

**end while**

we have

\[
(a) \ \psi(V) = h^*(V) + \frac{1 + \sigma \lambda_2}{2 \sigma} \sum_{i=1}^p \| \text{prox}_{\sigma \pi} \left( B_i - \sigma V^T X(i) \right) \|_2^2 - \frac{1}{2 \sigma} \sum_{i=1}^p \| B_i \|_2^2 \tag{3.11}
\]

\[
(b) \ \tilde{Z} = \text{prox}_{\frac{\pi}{\sigma}} \left( \frac{B}{\sigma} - X^T \tilde{V} \right),
\]

where $\text{prox}_{\frac{\pi}{\sigma}} (B/\sigma - X^T \tilde{V}) = \left( \text{prox}_{\frac{\pi}{\sigma}} (B_1/\sigma - \tilde{V}^T X(1)), \ldots, \text{prox}_{\frac{\pi}{\sigma}} (B_p/\sigma - \tilde{V}^T X(p)) \right)^T$.

\[\tilde{Z}\] has a closed form. To find $\tilde{V}$ one has to minimize $\psi$ or, equivalently, find the solution of $\nabla \psi = 0$. Note that $\psi$ is continuous and differentiable, and thus $\nabla \psi$ is well defined.

To solve the subproblem (3.7), we propose the Semi-smooth Newton method in
Algorithm 1. $V$ and $Z$ are updated iteratively – $Z$ according to the rule in Proposition 3, and $V$ by minimizing $\psi$ through one Newton step. The main computational cost is solving the linear system (3.9). This leads to our next crucial result (see Appendix Section B.1.4 for a proof).

Theorem 1. Let $T = B - \sigma X^TV$, $\hat{X} = X \otimes I_k$ (the $nk \times pk$ Kronecker product between $X$ and the $k \times k$ identity matrix), $\hat{2}\psi$ be the generalized Hessian of $\psi$, and $\partial \text{prox}_{\sigma\pi}$ be the Clarke sub-differential of $\text{prox}_{\sigma\pi}$ (Clarke, 1990). Then we have

$$(i) \nabla \psi(V) = V + Y - X \text{prox}_{\sigma\pi}(T) \quad (ii) \hat{2}\psi(V) = I_{nk} + \sigma \hat{X} \partial \text{prox}_{\sigma\pi}(T) \hat{X}^T$$

Moreover, let $Q \in \mathbb{R}^{pk \times pk}$ be the block-diagonal matrix $Q = \begin{bmatrix} P_1 & \cdots & \end{bmatrix}$, where each $P_i$ is a squared $k \times k$ matrix defined as

$$(iii) P_i = \begin{cases} (1 + \sigma \lambda_2)^{-1} \left( 1 - \frac{1}{\|T_i\|_2^{-1}} \sigma \lambda_1 \right) I_k + \frac{1}{\|T_i\|_2^{-3}} \sigma \lambda_1 T_i T_i^T & \|T_i\|_2 > \sigma \lambda_1 \\ 0 & \text{o.w.} \end{cases} \quad (3.13)$$

Then $Q \in \partial \text{prox}_{\sigma\pi}(T)$ and $\hat{2}\psi(V) \text{vec}(D) = (I_{nk} + \sigma \hat{X} Q \hat{X}^T) \text{vec}(D)$ for every $D \in \mathbb{R}^{nk \times k}$ in the domain of $V$ – where $\text{vec}(D) \in \mathbb{R}^{nk}$ is obtained by stacking all the columns of $D$.

Note that, while in Li et al. (2018); Deng and So (2019); Boschi et al. (2020) $\nabla \psi$ and $\hat{2}\psi$ are a vector and a matrix, respectively, here the dimensions of these operators increase due to the group nature of the problem. In particular, $\nabla \psi$ becomes a matrix and $\hat{2}\psi$ a higher order tensor – which we express as an $nk \times nk$ matrix by stacking its dimensions. Moreover, $Q$ is not a simple diagonal matrix as in the previous SsNAL algorithms, but is now characterized by blocks associated to the different groups of variables.

Theorem 1 is critical for preserving the efficiency of $f_{\text{gen}}$, while integrating groups into the problem. First, it states that solving (3.9) is equivalent to solving $(I_{nk} + \sigma \hat{X} Q \hat{X}^T) \text{vec}(D) = - \text{vec} \left( \nabla \psi(V) \right)$. Second, the form of $Q$ still allows one to induce sparsity in the linear system and drastically reduce the computational cost. Indeed, let $\mathcal{J} = \{ j : \|T_j\|_2 \geq \sigma \lambda_1 \}$ and let $r = |\mathcal{J}|$ be the cardinality of $\mathcal{J}$. Then the linear system (3.9) is equivalent to

$$(I_{nk} + \sigma \hat{X}_\mathcal{J} Q_\mathcal{J} \hat{X}_\mathcal{J}^T) \text{vec}(D) = - \text{vec} \left( \nabla \psi(V) \right).$$

(3.14)

Here, $\hat{X}_\mathcal{J} \in \mathbb{R}^{nk \times rk}$ is defined as $\hat{X}_\mathcal{J} = X_\mathcal{J} \otimes I_k$, with $X_\mathcal{J} \in \mathbb{R}^{nxr}$ being the sub-matrix of $X$ restricted to the columns in $\mathcal{J}$. In addition, $Q_\mathcal{J} \in \mathbb{R}^{rk \times rk}$ is the block-diagonal
matrix formed by all the $P_i$ such that $i \in \mathcal{J}$. Using the Cholesky factorization (the generalized Hessian is positive semidefinite) the total cost of solving the linear system reduces from $O(nk^3(n^2 + p^2 + np))$ to $O(nk^3(n^2 + r^2 + nr))$. This includes computing $\hat{X}_\mathcal{J} Q_\mathcal{J} \hat{X}_\mathcal{J}^T$, which is $O(nrk^3(n + r))$, and the Cholesky factorization, which is $O(n^3k^3)$. Because of the sparsity induced by the Group Elastic Net penalty, $r$ is usually much smaller than $p$ – implying a substantial computational gain. Even when $p$ is very large ($\sim 10^6$), one can still solve the linear system efficiently, as long as the dimension $k$ of each group is relatively small ($< 10^2$). Furthermore, if $r < n$, which is often the case when the solution of the Group Elastic Net problem is sparse, one can factorize an $rk \times rk$ matrix using the Sherman-Morrison-Woodbury formula (Van Loan and Golub, 1983):

$$(I_{nk} + \sigma \hat{X}_\mathcal{J} Q_\mathcal{J} \hat{X}_\mathcal{J}^T)^{-1} = I_{nk} - \hat{X}_\mathcal{J} ((\sigma Q_\mathcal{J})^{-1} + \hat{X}_\mathcal{J}^T \hat{X}_\mathcal{J})^{-1} \hat{X}_\mathcal{J}^T. \quad (3.15)$$

The total cost is further reduced from $O(nk^3(n^2 + r^2 + nr))$ to $O(rk^3(n^2 + r^2 + nr + 1))$, including the computation of $Q_\mathcal{J}^{-1}$ which can be done with a cost of $O(rk^3)$ by inverting each one of the $P_i$ blocks independently. Finally, if in the first iterations of the algorithm $n$ and $r$ are both larger than $10^4$, one can solve (3.9) approximately using the conjugate gradient method (Polyak, 1969).

To determine the convergence of the Augmented Lagrangian and the Semi-smooth Newton methods, we check the residuals of the third and first KKT in (3.6), respectively, i.e.

$$\text{res(kkt}_3) = \frac{\sum_{i=1}^p \| V^T X_i + Z_i \|_2}{1 + \sum_{i=1}^p \| V_i \|_2 + \sum_{i=1}^p \| Z_i \|_2}, \quad \text{res(kkt}_1) = \frac{\sum_{i=1}^n \| V_i + Y_i - X_i B \|_2}{1 + \sum_{i=1}^n \| Y_i \|_2}. \quad (3.16)$$

Taking the $l2$-norm of the KKT residuals, normalizing them and using them to assess convergence is a common procedure in the literature (Li et al., 2018; Deng and So, 2019). Both methods have a super-linear convergence rate. Accordingly, the convergence rate of the entire algorithm, which is the sum of the convergence rate of the two sub-problems (Tomlina and Sugiyama, 2009), is still super-linear. Thus, as we show in Section 3.4, $f_{\text{gen}}$ typically converges in very few iterations. The convergence analysis here follows directly from that in Tomlina et al. (2011) and Boschi et al. (2020), where SsNAL convergence is proved for the standard Elastic Net. The formal proof is in Appendix Section B.2 and leverages results in Rockafellar (1976a,b); Luque (1984); Li et al. (2018) and the fact that $\pi^*$ is a continuous differentiable function – which is also true for the Group Elastic Net.
Algorithm 2 Functional Group Elastic Net method

1. Perform FPC of $B$ and find the first $k$ basis components $(\gamma_1, \ldots, \gamma_k)$ with their eigenvalues $(\rho_1, \ldots, \rho_k)$
2. Find the first $k$ FPC scores for each response function $Y_i$: $Y_i = (\langle Y_i, \gamma_1 \rangle_{L^2}, \ldots, \langle Y_i, \gamma_k \rangle_{L^2})$
3. Using $Y$ as response matrix, apply SsNAL to solve (3.1) and find the coefficient scores estimates $B$.
4. Project $B$ into the FPC basis to find the coefficient curve estimates: $B_i = \sum_{j=1}^k B_{(i,j)} \gamma_j$

3.3.2 Extension to function-on-scalar regression

We now extend $fgen$ to the function-on-scalar features selection problem. In function-on-scalar regression, a functional response is regressed on a set of scalar predictors. Assuming the response belongs to the Hilbert Space $L^2([a,b])$, the optimization problem (3.1) becomes

$$\min_B \left( \frac{1}{2} \| XB - Y \|_{L^2}^2 + \lambda_1 \sum_{i=1}^p \| B_i \|_{L^2} + (\lambda_2/2) \sum_{i=1}^p \| B_i \|_{L^2}^2 \right)$$

(3.17)

where $Y$ and $B$ are functional objects with $n$ and $p$ rows, respectively. Each row $Y_i$ is a response function and each row $B_i$ is a coefficient function. The squared $L^2$-norm of a function $f$ is $\| f \|_{L^2}^2 = \int_a^b f^2$, where the inner product between two functions $f$ and $g$ is $\langle f, g \rangle_{L^2} = \int_a^b fg$.

Applying SsNAL directly to (3.17) is not straightforward and would substantially hinder its efficiency. First, the definition of conjugate functions and proximal operators in functional spaces would require a new theoretical background. Second, and perhaps most important from a practical standpoint, computing integrals is much more expensive than computing euclidean norms. For these reasons, in Algorithm 2 we take advantage of Functional Principal Components (FPC) (Horváth and Kokoszka, 2012b; Kokoszka and Reimherr, 2017a) to solve an optimization problem of the same type as (3.1), which is in fact a very close approximation to (3.17). In particular, we build a response matrix $Y$, where each group $i$ is formed by the first $k$ FPC scores of the the function $Y_i$ (Fan and Reimherr, 2016). The level of approximation of $fgen$ thus depends on the number of FPC scores $k$, i.e the dimension of each group. Indeed, given a function $f$ and its FPC basis $\{\gamma_i\}_{i=1}^\infty$, we have

$$\| f \|_{L^2} = \sum_{j=1}^\infty \| \langle f, \gamma_j \rangle_{L^2} \|_2$$

(3.18)

This property – which is true for every orthonormal basis system – plays a crucial role
in the extension of our SsNAL approach to the function-on-scalar regression, since it allows one to approximate the \( L^2 \) function norm with the standard \( l_2 \) matrix norm. Consequently, one can use the FPC scores to construct the response matrix \( Y \) and the coefficients matrix \( B \) in (3.1) starting from the response functions \( Y \) and the coefficient functions \( B \). The number of selected FPC scores determines the dimension \( k \) of each group in the Group Elastic Net problem. In many applications just a few FPC scores allow one to obtain a very close approximation of the original functions. Indeed, among the many orthonormal bases one could envision, FPC has the advantage of being the most parsimonious allowing one to reconstruct the response curves using fewer coefficients than any other orthonormal basis. In scenarios investigated by simulation in Section 3.4, \( k = 5 \) is sufficient to capture more than the 99\% of the \( L^2 \)-norm. This produces an almost perfect approximation of (3.17) while fully preserving \( \text{fgen} \) efficiency.

### 3.3.3 Solution path implementation

To evaluate different values of the penalty parameter \( \lambda_1 \), we implement an efficient solution path search. We compute the solution for a decreasing sequence sequence of \( \lambda_1 \), starting from \( \lambda^{\text{max}} = \max_i \| (X_i)^T Y \| \) which selects 0 active features. When we move to the next \( \lambda_1 \) value, we use the solution obtained at the previous value for initialization (warm start). The two consecutive solutions tend to be close, and \( \text{fgen} \) converges in very few iterations – usually just one. We also allow the user to specify a maximum number of selected features; when this number is reached the path search is stopped, further reducing computation.

To guide the choice of \((\lambda_1, \lambda_2)\) we propose two quantitative criteria: \textit{k-fold Cross Validation (cv)} and an Extended \textit{Bayesian Information Criterion (e-bic)} \cite{Chen2012}, which modifies the standard BIC to also include the number of features \( p \). In symbols, we have

\[
\text{e-bic}(B) = k \log \left( \frac{\text{rss}(B)}{(nk)} \right) + (k\nu) \left( \log(nk) + \log p \right) / n \tag{3.19}
\]

where \( \text{rss}(B) \) is the residual sum of squares associated with the solution \( B \), and \( \nu \) are the Group Elastic Net degrees of freedom. From \cite{Tibshirani2012}, \( \nu = \text{tr} \left( X_J \left( X_J^T X_J + \lambda_2 I_r \right)^{-1} X_J^T \right) \). Note that \( \text{cv} \) can be very computationally expensive because it requires to run \( \text{fgen} \) multiple times for each value of \( \lambda_1 \) and \( \lambda_2 \) under consideration. In contrast, \( \text{e-bic} \) can be computed directly from the original solution. Before evaluating both criteria, we de-bias the \( \text{fgen} \) estimates following the approach
suggested by [Belloni et al. (2014); Zhao et al. (2017)]. First, we run fgen, then, we fit a standard least squares on the selected features. In the next section, following standard practice in the literature – e.g., [Friedman et al. (2010); Pedregosa et al. (2011)] – we rewrite $\lambda_1$ and $\lambda_2$ as $\lambda_1 = c_\lambda \lambda^{\text{max}}$ and $\lambda_2 = (1 - \alpha) c_\lambda \lambda^{\text{max}}$, with $c_\lambda \in (0, 1]$ and $\alpha \in (0, 1)$. $c_\lambda$ determines the reduction with respect of $\lambda^{\text{max}}$, $\alpha$ controls the relative weight of the two penalties.

### 3.4 Simulation study and INSIGHT data

In this section we use synthetic data to illustrate the computational efficiency of fgen, and apply our new method to a Genome Wide Association Study (GWAS) on childhood obesity. In the simulations, we benchmark fgen against the two best Group Elastic Net solvers we found in the literature: the python package sklearn and R package glmnet, which is written in fortran. Other functional-on-scalar feature selection methods, such as the ones proposed by [Barber et al. (2017); Parodi et al. (2018); Mirshani and Reimherr (2019)], have a computational burden more than two order of magnitudes larger than fgen and could not complete instances with $p > 10^4$.

#### 3.4.1 Simulation results

We generate synthetic data as follows. The entries of the design matrix $X \in \mathbb{R}^{n \times p}$ are each drawn independently from a standard normal distribution. The response curves are created as $Y = XB + \epsilon$. $B$ contains $p_0$ non-zero curves. These and the errors $\epsilon$ are generated from a 0 mean Gaussian process with a Matern covariance function (Cressie
Table 3.1: $a$, $b$ and $c$ report CPU time in seconds for \texttt{fgen}, \texttt{sklearn} and \texttt{glmnet}, respectively. For \texttt{fgen} we also report the number of iterations in parenthesis. $r$ is the number of selected features, $l$ is the range parameter of the Matern process used to generate the coefficients.

$$C(t, s) = \omega^2 \left( \Gamma(\nu)2^{\nu - 1} \right)^{-1} \left( (l)^{-1}(2^{\nu})^{1/2} |t - s| \right)^\nu K_\nu \left( (l)^{-1}(2^{\nu})^{1/2} |t - s| \right), \quad (3.20)$$

where $K_\nu$ is a modified Bessel function. In particular, we set the point-wise variance $\omega^2 = 1$ and the range $l = 0.25$ (this determines how fast the curves dependency decays). The smooth parameter $\nu$ is equal to 3.5 for $B$ and to 1.5 for $\epsilon$, i.e. the errors are rougher than the coefficients. Each curve is sampled at 1000 evenly spaced points between 0 and 1. Figure 3.1 shows instances of the response and error curves, $Y$ and $\epsilon$, and the true non-zero coefficient curves in $B$ along with their de-biased estimates produced by \texttt{fgen} (the underlying simulation parameters are those in Table 3.1, second row, $n = 500$). In all scenarios, \texttt{fgen} is run with both the tolerances in (3.16) set to $10^{-6}$ (we set the same tolerance for \texttt{sklearn} and \texttt{glmnet}) and $\mu$ in (11) set to 0.2. We start from $\sigma^0 = p_0/p$.
Figure 3.2: Plots related to the INSIGHT study. The left panel displays the growth curves. The center and right panels depict values of the 10-fold cv Mean Squared Error and the \( e\text{-bic} \), respectively, against \( c_\lambda \). These are obtained from \texttt{fgen} run with 3 different values of \( \alpha \): 0.8 (green line), 0.6 (red line), 0.4 (blue line). \texttt{fgen} estimates are de-biased prior to computing both criteria.

and increase it by a factor of 5 every iteration. If we start from smaller values of \( \sigma \), the algorithm needs more iterations to converge, while if \( \sigma^0 \) is too large, \texttt{fgen} does not converge to the optimal solution. We set \( \lambda_1 = c_\lambda \lambda^\text{max} \) and \( \lambda_2 = (1 - \alpha)c_\lambda \lambda^\text{max} \), where \( c_\lambda \in (0, 1], \alpha \in (0, 1) \), and \( \lambda^\text{max} = \max_i \| (X_i)^T Y \| \). Note that for \texttt{glmnet} and \texttt{sklearn} we need to divide \( \lambda^\text{max} \) by \( n \) since both solvers divide the least squares loss in (3.1) by the number of observations.

Table 3.1 reports CPU times for \texttt{fgen}, \texttt{sklearn} and \texttt{glmnet} under different simulation settings. \texttt{fgen} is the fastest solver in almost every instance. When both \( n \) and \( p \) are large and the solution is sparse, \texttt{fgen} is approximately 6 times faster than \texttt{glmnet} and more than 30 times faster than \texttt{sklearn}. Note that the super-linear convergence rate allows \texttt{fgen} to converge in very few iterations (no more than 4 in all cases). The CPU time increases with \( k \) for all solvers. However, \( k = 5 \) already captures more than the 99% of the \( L^2 \)-norm in all the scenarios considered. If we decrease \( \alpha \) from 0.8 to 0.5, \texttt{fgen} need even fewer iterations to converge, increasing its computational gain with respect to the competitors. Considering rougher coefficients (created with a Matern process with range parameter \( l = 0.1 \)) does not affect the relative performance of the algorithms. The instance with an active set of 486 features is the only one where \texttt{fgen} performs worse than its competitors. As expected, in the presence of non-sparse solutions \texttt{fgen} loses some of its efficiency. However, to tune the penalty parameters in practice, one evaluates a sequence of \( c_\lambda \) values starting from very sparse solutions. In the first steps of the solution path, \texttt{fgen} exploits sparsity and is very efficient. In the following steps, it still converges very quickly thanks to the \textit{warm-start} approach described in Section 3.3.3. In Appendix Table B.2 we compare the solution path computing time. \texttt{fgen} outperforms
the other solvers in every scenario, being approximately 2 times faster than \texttt{glmnet} and from 10 to more than 30 times faster than \texttt{sklearn}. Finally, to gauge uncertainty in our CPU time evaluations, we replicated a subset of the instances explored in Table 3.1 20 independent times. Mean CPU times and standard errors over such replicates are reported in Appendix Table B.1. Results agree with those obtained considering just one replication. Furthermore, one can notice that \texttt{fgen} has also a smaller variability in CPU time when \( n = 5000 \).

Taking into account that \texttt{glmnet} (written in \texttt{fortran}) and \texttt{sklearn} are highly optimized packages, the results above provide strong evidence in support of our method. We also tracked prediction performance for all methods and in all simulation settings considered, but we do not report them here since all three solvers solve the same convex minimization problem and therefore converged to the same solution in all settings.

### 3.4.2 INSIGHT study

Here, we apply \texttt{fgen} to data from the Intervention Nurses Start Infants Growing on Healthy Trajectories (INSIGHT) study (Paul et al., 2014). In particular, we focus on data collected to investigate genetic variants that may affect the risk of childhood obesity. As the prevalence of obesity increases also among children, examining possible causes and risk factors has become an essential public health concern. INSIGHT provides genome-wide Single Nucleotide Polymorphisms (SNPs) information for a cohort of very young children, along with longitudinal information on their growth. Selecting SNPs that may affect growth is thus a GWAS (a Genome-Wide Association Study) – where the outcome is a growth curve. In recent years, many GWASs have identified SNPs strongly associated with obesity phenotypes (Locke et al., 2015). Before proceeding with the analysis, we point out that due to high feature collinearity, low signal-to-noise ratio, and ultra-high dimensionality, GWAS data are very hard to examine and users should be very careful in interpreting results; e.g., selected SNPs may just be proxies for other causal SNPs in their vicinity. While being well aware of all its complexities and potential pitfalls, our main aim in presenting a GWAS analysis is to show the efficiency and the broad applicability of \texttt{fgen}. Our functional outcome captures the evolution of weight/height ratios (Daniels et al., 2015) measured at birth and at 4, 16, 28, and 40 weeks for a total of \( n = 210 \) children. The growth curves – shown in the left panel of Figure 3.2 – are fitted as in Craig et al. (2019) using Principal Analysis by Conditional Estimation (Chen et al., 2017). Building a smooth curve for each child allows one to capture information along the entire time domain and at the same time to de-noise and mitigate the effect of
outlying/anomalous raw measurements. The SNPs collected in INSIGHT are available upon request at dbGaP using the access number phs001498.v1.p1. The growth curves are based on privacy protected data and cannot be made publicly available.

Craig et al. (2019) used flame (Parodi et al., 2018) to solve the function-on-scalar feature selection problem. To do so, they had to reduce the analysis from $p = 342325$ to 10000 SNPs with various preliminary screening steps. The computational efficiency of fgen allows us to inspect all the 342325 SNPs simultaneously. The center and left panels of Figure 3.2 display 10-fold $cv$ and $e\text{-}bic$ for different values of $\alpha$ and $c_{\lambda}$. Both criteria identify just one ($c_{\lambda} = 0.99$) dominant SNP, rs79187646. Without drawing any strong domain conclusion, we remark that the selected SNP appears to be relevant in the literature. Notably, this is the same SNP selected by Boschi et al. (2020), where the same SNP data were associated to BMI at age 3 – a scalar response. Also notably, this dominant SNP was not among the SNPs identified by Craig et al. (2019). Based on the U.S. National Library of Medicine, rs79187646 is located in NTM, a well known gene. According to the NHGRI-EBI GWAS Catalog, many GWASs, including two recent studies (Kichaev et al., 2019; Pulit et al., 2019), have connected NTM to body mass, food addiction, intake of sweet substances and other obesity-related traits.
Chapter 4

Functional data analysis characterizes the shapes of the first COVID-19 epidemic wave in Italy

4.1 Introduction

This chapter describes an impactful research application where we use functional clustering, registration, regression, and future selection techniques, including those presented in previous chapters, to study the evolution of the COVID-19 epidemic in Italy.

At the end of January 2020, two Chinese tourists were hospitalized in Rome and tested positive to SARS-CoV-2. At the beginning of February, a group of Italian citizens was repatriated from Wuhan – among them, one tested positive. As the news media reported these headlines, neither the Italian public nor the Italian authorities appeared to perceive an imminent threat, though retrospective analyses now suggest that the virus may have been circulating in the north of the country as far back as December 2019 – e.g., detection of SARS-CoV-2 in the wastewater of Milan and Turin (La Rosa et al., 2021). The first recorded non-travel related COVID-19 case occurred in Codogno (Lombardia) – where a 38 years old male visited the hospital first on February 17, and then again on February 19 with worsening respiratory symptoms; in this date, he was tested and diagnosed. On February 20, two individuals tested positive in Vo Euganeo (Veneto). Notably, the outbreaks in Lombardia and Veneto took two very different paths, something many observers attributed to the early response and aggressive testing strategy adopted by the regional authorities in Veneto (Mugnai and Bilato, 2020; Lavezzo et al., 2020). After some initial, much debated inconsistencies (e.g., hesitations in implementing
local lock-downs in areas hosting major industrial production hubs, contested decisions to move patients between hospitals and nursing homes and to keep major sports events open to the public in Lombardia), starting in early March, local and central authorities took progressively more stringent measures to limit mobility and social gatherings – culminating with a general nationwide lock-down on March 9 and the suspension of all nonessential production activities on March 23 (starting in early May, activities restarted and mobility and gathering restrictions were gradually loosened).

Lock-down notwithstanding, based on official records, Italy saw a total of \( \approx 35,200 \) COVID-19 deaths as of the beginning of August 2020. While other countries (e.g., the U.S. and Brazil) have reached much higher death counts, Italy’s relative death toll remains rather stark at 58.25 per 100,000 inhabitants. This may be partially attributable to the fact that Italy’s population is very old (nationally, the median age is almost 46 years and the percentage of individuals over 65 almost 22%), and that age itself correlates with conditions such as type II diabetes, hypertension and chronic respiratory ailments, which substantially worsen illness and increase the likelihood of death for individuals affected by the virus. But perhaps the most striking aspect of the COVID-19 epidemic in Italy has been its heterogeneity. Some parts of Lombardia and of other regions in the industrialized north were hit early and especially hard, yet other demographically and socio-economically similar areas fared better. Moreover, most of the central and southern regions of the country experienced a much milder epidemic – notwithstanding waves of relocations from employment-related domiciles in the north back to family homes in the center and south around the time of the nationwide lock-down. Potential contributors to this heterogeneity discussed by both scientists and the media include human density characteristics; centralized, hospital-based vs distributed, primary health care systems; and pollution levels (Wu et al., 2020; Coccia, 2020; Frumento and Sylos Labini, 2020).

A broad and extremely sophisticated literature exists on epidemiological models, which many research groups are utilizing both to aid policy through forecasts and to dissect what happened, in Italy and around the world. We did not utilize these models. Instead, we applied a mix of statistical tools from the field of Functional Data Analysis (Ramsay and Silverman, 2005b; Kokoszka and Reimherr, 2017b), some well-established, and some recently developed by our group. FDA offers very powerful approaches to analyze data sets composed of curves or surfaces, exploiting information in their shapes. These techniques, which have been successfully applied in a variety of scientific domains (Ramsay and Silverman, 2007; Ullah and Finch, 2013; Cremona et al., 2019), can effectively
complement traditional epidemiological analyses and provide useful insights (Carroll et al., 2020). We used them to characterize patterns of COVID-19 deaths occurring around the country and analyze their statistical association with two key predictors; namely, mobility and positivity (the fraction of performed tests returning positive results). We also considered various socio-demographic, infrastructural and environmental covariates. We focused on the period from February 16, right before the first cases were recorded in Codogno and Vo’ Euganeo, to April 30, right before the first lock-down relaxations (restarting of manufacturing and construction activities at the beginning of May). Based on data availability, we performed our analyses at the spatial resolution of regions, which is suboptimal for several reasons. An epidemic is certainly better studied at a much finer resolution (municipalities, urban areas, perhaps the provinces within which Italian regions are further partitioned) and so are its links to predictors and covariates whose signals may dilute when aggregated at the regional level. Moreover, operating with 20 observational units (the Italian regions) limits the size of the statistical models one can reliably fit on the data. The techniques we employed allowed us to pinpoint significant trends working with what we could retrieve from public data sources. Unquestionably though, access to data at higher resolution would allow more nuanced, in-depth analyses and likely produce sharper results.

The remainder of the chapter is organized as follows. In Section 4.2, we present the main findings of our study. In Section 4.3, we discuss the potential impact of our results. In Section 4.4, we describe data and methods used in the analysis.

4.2 Results

Below we describe the salient outcomes of our analyses. After addressing some shortcomings in publicly available COVID-19 deaths records, we characterize two starkly different epidemic patterns and rank regional mortality curves. Next, we relate mortality to mobility and positivity, and to a number of socio-demographic, infrastructural and environmental factors.

4.2.1 Under-counting deaths

Since February 24, the Italian Civil Protection agency (Dipartimento della Protezione Civile; DPC) has released daily counts of recorded COVID-19 deaths at the coarse resolution of regions (only the number of recorded cases are released at the finer resolution
Figure 4.1: Mortality curves. (a) DPC (dashed) and ISTAT (solid) differential mortality curves (per 100,000 inhabitants) in four example regions; Lombardia, Veneto, Emilia Romagna and Campania. Curves are smoothed with splines, with degree of smoothing selected by generalized cross-validation (see Methods). ISTAT curves “take off” earlier and in some regions are as much as twice as high at their peak – possibly due to many COVID-19 deaths happening at home and/or not being recorded as such in hospitals, especially in the early stages of the epidemic. (b) MAX mortality curves (per 100,000 inhabitants) in the 20 Italian regions, before (top) and after (bottom) the shifts produced by probKMA run with $K = 2$. In the bottom panel, time is marked as a day number (as opposed to a date); this represents the region-specific time of the epidemic unfolding, and corresponds to actual time (starting on February 16 and ending on April 30) only for regions with no shifts, e.g., Lombardia). Curves are again smoothed with splines, with degree of smoothing selected by generalized cross-validation. Lombardia, Veneto, Emilia Romagna and Campania, also shown in (a), are highlighted in color. In all panels, vertical lines mark the dates of the national lock-down (March 9) and of the suspension of all nonessential production activities (March 23). In the bottom panel of (b) vertical lines still show these dates without shifts; stars on the curves mark the lock-down after the region specific shifts.

of provinces). In Italy and elsewhere, official death records have often been criticized as undercounts (Ciminelli and Garcia-Mandicó 2020). Alternative data sources do exist, e.g., daily mortality rates – which can be contrasted to those from prior years to gauge differential mortality. In Italy these are provided by the National Statistical Institute (ISTAT) at the resolution of municipalities. We aggregated the data over
municipalities belonging to the same region and subtracted averages over the past 5 years (2015-19, see Methods). Figure 4.1(a) shows smoothed DPC and ISTAT differential mortality curves (per 100,000 inhabitants) for some example regions (Lombardia, Veneto, Emilia Romagna and Campania). The under-counting in the official DPC records was dramatic, especially in badly affected areas and in the initial stages of the epidemic. However, ISTAT differential mortality curves have themselves limitations, especially in less affected areas, where they can fluctuate at small levels and even take negative values idiosyncratically or reflecting other COVID-19 related phenomena (e.g., increases in mortality due to untreated emergencies or reductions in mortality due to fewer accidents during the lock-down). We therefore formed maxima curves (MAX), where the largest between the DPC and the ISTAT datum is taken in each day and for each region, and then smoothed. These are shown in Figure 4.1(b) (DPC and ISTAT smoothed curves for all regions are shown in Appendix Figures C.1 and C.2). We repeated our analyses on all three data sets; given the small number of observational units at our disposal ($n = 20$ regions), this allowed us to borrow strength replicating results across data sets, with their differences and limitations.

### 4.2.2 Two different epidemics

Italy saw the unfolding of two very different epidemics; a relatively mild one in the majority of the country, and a tragic, seemingly out of control one in its most hard-hit regions. These two epidemics can be effectively characterized with probKMA, an FDA technique designed to identify recurrent motifs within a set of curves, and group the curves based of the motifs they comprise (Cremona and Chiaromonte, 2020). Here, the motifs are the temporal patterns of deaths that characterize alternative epidemic unfoldings, which may in fact start at different times in different curves (regions). Thus, the algorithm also produces the shifts required to align regions comprising the same motif to each other. ProbKMA is similar to a $K$-mean algorithm; it requires the user to specify the number of motifs ($K$) at the outset, and to select a distance – which can be defined on the curve levels, their derivatives, or a combination of both (see Methods).

The solution with $K = 2$ and distance defined on curve levels depicts two starkly different epidemics, shown for the MAX curves in Figure 4.2(a). Allowing for shifts, these are represented by 65-day long motifs. Group 1 undergoes a steep ascent (the “exponential” pattern) followed by a slower descent from the peak; it includes many northern regions. Based on the shifts, Lombardia was first, followed by Emilia Romagna, Marche, Liguria, Piemonte, Trento/Bolzano, and last Valle d’Aosta. Lombardia and
Figure 4.2: Characterizing two epidemics. (a) MAX mortality curves are shown in the top left panel with 65-day portions identified by probKMA with $K = 2$ in red (Group 1; 'exponential' pattern) and blue (Group 2; 'flat(tened)' pattern). The curve portions are shown again, this time aligned with each other and separated by group, in the bottom panels. Black lines indicate group averages. The shifts produced by probKMA are shown in the top right panel (motifs, groups and shifts for Group 1 are stable across data sets; shifts for Group 2 are less stable and less interpretable – see Appendix Figure C.3). (b) Shifted Group 1 and Group 2 MAX mortality curves are tested against each other with IWTomics. The heatmap at the top shows $p$-values adjusted at all possible scales (from 1 to 65 days). The middle panel shows in detail the top-most row of the heatmap; i.e. the $p$-values adjusted across the whole 65-day interval. The bottom panel shows again the shifted curves. Gray areas in the middle and bottom panels mark days when the difference between the two groups is significant (adjusted $p$-value < 5%). Starting a little over two weeks from the beginning of their epidemic, curves in the two groups differ significantly at all temporal scales.

Valle d’Aosta presented the most extreme peaks – but Valle d’Aosta’s descent was steeper (with a second late ascent likely due to data recording imprecisions; Valle d’Aosta is a very small region with only $\approx 125,000$ inhabitants). Group 2 follows a “flat(tened)” pattern; it includes all regions in southern and central Italy and, remarkably, Veneto – where the curve was successfully curbed. The shifts produced for this group are less stable and less meaningful in terms of interpretation, as flatter profiles leave more leeway in aligning curves against each other. All results (except for the shifts in Group 2) are rather consistent when using DPC and ISTAT curves (see Appendix Figures C.3(a)
and C.3(c)), and when using distances defined on derivatives instead of curve levels. The solution with $K = 3$ places Lombardia (ISTAT curves) or Lombardia and Valle d’Aosta (MAX and DPC curves) in a cluster of their own (see Appendix Figure C.4). We also validated our results using a modification of funBI (Di Iorio and Vantini, 2019), a functional biclustering technique, and IWTomics (Cremona et al., 2018), a functional testing technique which contrasts two sets of aligned curves pinpointing the locations and scales at which they differ (see Methods). Figure 4.2(b) shows how, starting a little over two weeks from the beginning of their motif (wherever that was in each curve), Group 1 and Group 2 differ significantly at all temporal scales (see also Appendix Figures C.3(b) and C.3(d)).

Why the two epidemics? The pattern of deaths characterizing Group 1 may be due, in large part, to the fact that the virus had circulated silently in the north of Italy for a long period of time before any kind of behavioral changes by the general public, medical protocols, or mitigation policies by local and central authorities were put in place. Mounting evidence suggests that a large share of COVID-19 cases are asymptomatic and yet contagious (Lavezzo et al., 2020); their numbers may have increased until a pent-up reservoir of virus found its way to vulnerable individuals – some researchers also hypothesize Antibody-Dependent-Enhancement of SARS-CoV-2 (Cegolon et al., 2020), and thus a role for re-infections. But a variety of additional factors may have contributed to shaping the two epidemics; we explore some below.

### 4.2.3 Ranking mortality curves

Non-parametric FDA methods can be used to rank curves based on the notion of depth – from the innermost to the most extreme, and to identify outliers (Sun and Genton, 2011; López-Pintado and Romo, 2009). Figure 4.3 shows a functional box plot of the MAX mortality curves and a depth ranking of the curves in the DPC, ISTAT and MAX data sets – shifted based on probKMA run with $K = 2$ and restricted to their aligned 65-day portions. The ranking is directional; we attributed signs to the depth measurements, so that curves far over or under the median curve are at the top or bottom of the ranking, respectively (see Methods). The top portion of the ranking comprises regions with "exponential" epidemics (Group 1) and is rather stable across data sets; Lombardia and Valle d’Aosta are consistently among the most extreme curves (they are also identified as outliers in the MAX and DPC data sets). The mid- and bottom portions of the ranking comprise regions with "flat(tened)" epidemics (Group 2) and are less stable across data sets, as the flatter profiles can more easily switch in their depth...
ranks. However, Toscana (which is the median in the MAX and ISTAT data sets) and Veneto are consistently among the deepest, most central curves. This analysis highlights again the tragic epidemic unfolding in Lombardia, and, by contrast, confirms how Veneto managed to “flatten” its curve back into the bulk.

4.2.4 Local mobility and positivity as statistical predictors of mortality

Next, we focus on two key variables. The first is one of the most discussed policy-actionable variables, mobility, which has been curtailed to various degrees through lock-down measures in most of the countries affected by COVID-19. The second is one of the most discussed sentinel indicators, positivity, i.e. the fraction of performed tests returning positive results. For both these variables daily values for the period February 16 – April 30 can be obtained from data in the public domain at regional resolution.

We considered differential mobility curves provided by Google for the category “Grocery & pharmacy”. These express the fractional reduction with respect to January 2020 levels, and refer to mobility linked to first necessities – such as buying food, medicine, etc. For Italy, they are provided at the resolution of regions. Even though individuals were allowed to leave their homes for these necessities also during the most restrictive phase of the lock-down, the reduction captured by Google’s “Grocery & pharmacy” was substantial. Mobility in weekdays fell by roughly 0.30, i.e. 30%, in the week after the lock-down (March 9), and further decreased in following weeks – reaching the lowest levels (between approximately −0.60 and −0.40 depending on the region) in the week after the suspension of nonessential production activities (March 23). It then slowly increased, getting back in a range between approximately −0.40 and −0.20 at the very end of April (see Appendix Figure C.5). In Lombardia, the peak MAX mortality was between March 20 and 25 – i.e., roughly, simultaneous to the lowest mobility and two weeks after its first substantial drop. Notably, in most Italian regions mobility during lock-down weekends reached −1.00, i.e. −100%. For comparison, in the state of New York, which had among the strongest restriction measures in the U.S., Google’s “Grocery & pharmacy” never fell below −0.40. We refer to Google’s “Grocery & pharmacy” curves as local mobility because they measure how much individuals move around where they live, as opposed to how much individuals move from place to place – e.g., to go from Wuhan to Milan, or from Milan to Palermo, or New York City. Obviously both types of mobility are relevant for the spread of a virus, and definitions depend on scale/resolution,
but the first one is the one we analyzed.

To construct positivity curves, we combined daily public records on number of tests performed and number of new cases, which are also provided by the Italian Civil Protection agency. Taking daily ratios of new cases on tests performed is clearly imperfect, because of (variable and unreported) delays in test results. But regularizing and smoothing these ratios (see Methods) produced a reasonable proxy. Smoothed positivity surpassed 0.1, i.e. 10%, as early as February 20 in some hard hit regions, peaked in a staggered fashion throughout March, and fell below 0.10 for all regions by around April 22 (see Appendix Figure C.5). Lombardia surpassed 0.10 around February 22 and peaked around March 15-18; that is, roughly, about a month and about a week prior to the peak of MAX mortality, respectively. Though we cannot draw exact parallels (our positivity curves are approximate and smoothed), this is consistent with what was observed, e.g., in New York City – where positivity was above 0.10 approximately from March 6-7 to May 12-13 and peaked at about 0.70 around March 28, with deaths peaking between April 5 and 13.

To anchor local mobility and positivity curves to the epidemic unfolding in each region, we shifted them congruently with the mortality curves. Figure 4.4(a) displays shifted curves based on probKMA run on MAX data with $K = 2$ (Appendix Figure C.6 displays shifted curves based on probKMA run on DPC and ISTAT data). The horizontal axis now indicates again days in the region-specific epidemic unfolding, restricted to the 65-day portions where mortality curves align forming the two probKMA motifs.

We then used function-on-function regressions (Ramsay and Silverman, 2005b; Horváth and Kokoszka, 2012b) to model the statistical dependence of mortality on local mobility and positivity; in symbols, we fit the joint model $y(t) = \alpha(t) + \int \beta_{mob}(s,t)x_{mob}(s)ds + \int \beta_{pos}(s,t)x_{pos}(s)ds + \epsilon(t)$, where $y(t)$ is the response curve, i.e. mortality, $\alpha(t)$ is the intercept, $\epsilon(t)$ is the model error, and $x_{mob}(s)$ and $x_{pos}(s)$ are the predictor curves – mobility and positivity, respectively. These predictors are integrated over time, with “effects” represented by surfaces; $\beta_{mob}(t,s)$ is the association of mortality at time $t$ with local mobility at time $s$, and similarly $\beta_{pos}(t,s)$ for positivity (see Methods).

Figure 4.4(b) shows the effect surfaces for local mobility and positivity estimated using the MAX curves as response. $\hat{\beta}_{mob}(t,s)$ suggests that local mobility levels early on and mid-way through the epidemic (e.g., around the March 9 lock-down date for Lombardia) are strong positive predictors of mortality at its peak, with the early predictive signal stronger than the mid-way one. In contrast, the local mobility level late in the epidemic has a negative association with mortality at its peak, likely reflecting a faster resumption of mobility in regions with milder epidemics. $\hat{\beta}_{pos}(t,s)$ suggests that positivity levels early
on and mid-way through the epidemic are also positive predictors of mortality at its peak – though the predictive signals are substantially weaker than those of mobility, likely because they are confounded with the latter. However, the positivity level late in the epidemic has a marked positive association with mortality at its peak. Here the signal is "detangled" from that of mobility, and one finds a sort of retrospective signature; regions which fared worse still had heightened positivity in the late stages of their epidemics. The data at our disposal does not allow an accurate evaluation of the lags that might occur between mobility, positivity and mortality. However, we performed some additional analyses to investigate this. We further denoised the curves projecting them on their first functional principal components (Ramsay and Silverman, 2005b), and measured the distances between the peaks of such projections. On average, there were $\approx 20$ days between the peak of mobility and the peak of positivity, and $\approx 10$ days between the peak of positivity and the peak of mortality (see Appendix Figure C.7). Back to the estimated effect surfaces, we found them to be remarkably similar across the three data sets (MAX, DPC and ISTAT). The joint models all have in-sample $R^2$s above 90% and leave-one-out cross-validated (LOO-CV) $R^2$s above 50% (see Appendix Table C.1), with strong and comparable contributions of local mobility and positivity (e.g., for the MAX curves, the partial $R^2$s are 62% and 53%, respectively). Also, while this is not the case for all regions, residuals are rather consistent across data sets for Veneto, whose mortality is well predicted, and for Lombardia, whose mortality is always and sizably underestimated (see Appendix Figures C.8(a) and C.9).

In order to further assess the roles of local mobility and positivity, we also considered marginal function-on-function regressions for mortality on each, separately; in symbols, $y(t) = \alpha(t) + \int \beta_{\text{mob}}(s,t)x_{\text{mob}}(s)ds + \epsilon(t)$ and $y(t) = \alpha(t) + \int \beta_{\text{pos}}(s,t)x_{\text{pos}}(s)ds + \epsilon(t)$. Effect surface estimates for local mobility are very similar to those in the joint models for all three data sets (see Appendix Figure C.10). Those for positivity confirm a strong association with mortality at its peak, but are less defined in terms of time profile (see Appendix Figure C.11). In summary, we find substantial evidence that local mobility and positivity are associated with COVID-19 mortality, and can predict it with some lag-time. Though the data at our disposal does not allow us to pinpoint lag lengths with accuracy, our analysis does support their roles as policy-actionable and monitoring variables, respectively. We also find that, even when considered jointly, these variables are not enough to fully account for the massive numbers of COVID-19 deaths recorded in Lombardia, the worst hit region in the country.
4.2.5 The role of socio-demographic, infrastructural and environmental factors

We considered 68 scalar (non-longitudinal) covariates retrieved from public sources, proxying for socio-demographic, infrastructural and environmental factors debated by scientists and policy-makers during the epidemic (see Appendix Table C.2). Many of these are suboptimal proxies; they refer to the closest times we could find data for (in some cases 2016 or earlier) and are, too, at the coarse resolution of regions. We performed an initial screen among these covariates to guarantee reasonable data quality (eliminating older and less complete data sets), facilitate interpretations and control collinearity. Appendix Figure C.12 shows a histogram of the pair-wise correlations, about a quarter of which exceeds 0.5 in absolute value, and Appendix Figure C.13 shows a dendrogram where the covariates agglomerate in distinct groups. We thus selected 12 covariates which were relatively recent (2017 or 2018) and well spread across the dendrogram groups. These capture aging of the population; prevalence of pre-existing conditions believed to affect disease severity; quality of distributed primary health care vs. centralized hospital-based health care; the potential of hospitals and nursing homes, but also schools, workplaces, households and public transport to act as contagion hubs; and pollution levels (see Table 4.1; Appendix Figure C.14 provides marginal densities, pair-wise scatter plots and correlations for the 12 selected covariates).

Even this restricted set of 12 covariates presents a distinct interdependence structure (see covariates dendrogram in Figure 4.5(b) and Variance Inflation Factors in Appendix Table C.3). For instance, our contagion hubs proxies for hospitals, schools and work places, and our (inverse) proxy for quality of distributed, primary health care (number of adults per family doctor), tend to vary closely together across regions. Also, our contagion hub proxy for public transport and pollution levels tend to vary together (this is not counter-intuitive, as both increase in more industrialized regions with large metropolitan areas), as do the percentages of individuals affected by diabetes and allergies, and our proxy for quality of centralized, hospital-based health care (ICU beds per 100,000 inhabitants) and the percentage of individuals over 65.

Conversely, some regions show similar profiles across covariates (see regions dendrogram in Figure 4.5(a)). For instance, Lombardia, Veneto, Emilia Romagna and Piemonte have strong similarities, as do groups of southern regions (e.g., Sicilia, Campania, Puglia and Calabria; Basilicata, Abruzzo and Molise). An interesting characterization is produced using the Cheng and Church’s biclustering algorithm (Cheng and Church 2000),
<table>
<thead>
<tr>
<th>Covariate</th>
<th>Description</th>
<th>Year and Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Over 65</td>
<td>Aging of the population [1]</td>
<td>2018, ISTAT</td>
</tr>
<tr>
<td>% Diabetics</td>
<td>Prevalence of relevant pre-existing conditions [2]</td>
<td>2018, ISTAT</td>
</tr>
<tr>
<td>% Allergic</td>
<td>Another potentially relevant pre-existing condition</td>
<td>2018, ISTAT</td>
</tr>
<tr>
<td>Adults per family doctor</td>
<td>Quality of distributed, primary health care</td>
<td>2017, Ministry of Health</td>
</tr>
<tr>
<td>ICU beds per 100K inhabitants</td>
<td>Quality of centralized, hospital-based health care [3]</td>
<td>2018, Ministry of Health</td>
</tr>
<tr>
<td>Ave. beds per hospital (whole)</td>
<td>Ability of hospitals to act as contagion hubs</td>
<td>2018, Ministry of Health</td>
</tr>
<tr>
<td>Ave. beds per nursing home (ward)</td>
<td>Ability of nursing homes to act as contagion hubs</td>
<td>2018, Ministry of Health</td>
</tr>
<tr>
<td>Ave. students per classroom</td>
<td>Ability of schools to act as contagion hubs</td>
<td>2018, Ministry of Education</td>
</tr>
<tr>
<td>Ave. employees per firm</td>
<td>Ability of work places to act as contagion hubs</td>
<td>2017, ISTAT</td>
</tr>
<tr>
<td>Ave. members per household</td>
<td>Ability of households to act as contagion hubs [4]</td>
<td>2017, ASR Lombardia</td>
</tr>
<tr>
<td>Public transport rides per capita</td>
<td>Ability of public transport to act as contagion hub</td>
<td>2017, ISTAT</td>
</tr>
<tr>
<td>PM10</td>
<td>Pollution levels (particulates)</td>
<td>2018, ISTAT</td>
</tr>
</tbody>
</table>

Table 4.1: **Scalar covariates potentially affecting COVID-19 mortality.** [1] The percentages of over 65, 70, 75 and 80 are highly correlated at the resolution of regions; we took over 65 as representative. [2] The prevalence of diabetes, hypertension and chronic bronchitis are highly correlated at the resolution of regions; we took diabetes as representative (allergies are not as highly collinear and were retained as separate). [3] Availability of ICU beds is also directly relevant for withstanding the impact of COVID-19 surges. [4] Average members per household is *not* a direct proxy of inter-generational contacts, but it may capture some of its effects.

which we implement with an adjusted mean squared residue, or H-score ([Di Iorio et al., 2020](#)). A bicluster is a subset of regions which exhibit similar behavior across a subset of covariates. Figure 4.5(b) shows two biclusters with similar adjusted H-score values, obtained through the same run of the algorithm. The first bicluster comprises central and southern regions, all with "flat(tened)" epidemics (Group 2). Its regions have low ratios of adults to family doctors, limited concentrations in hospitals, nursing homes, work places and public transport, and low pollution levels. They also have high percentages of diabetic individuals and limited availability of ICU beds. The second bicluster comprises northern regions with "exponential" epidemics (Group 1), such as Lombardia, Emilia-Romagna and Piemonte, but also northern and central regions with "flat(tened)"
epidemics (Group 2), such as Veneto, Friuli Venezia Giulia and Toscana. Its regions have
high ratios of adults to family doctors, high concentrations in hospitals, work places and
classrooms, and tend to have large percentages of individuals over 65. They also have
low percentages of diabetic individuals and medium or small-sized households.

Next, we used functional regressions with a two-fold aim: pursue a more direct,
systematic assessment of the associations between the scalar covariates and COVID-19
mortality; and use the scalar covariates as controls in models comprising mobility and
positivity to re-assess these key predictors. We stress again that the coarse resolution of
the data poses serious limitations for these analyses, because it may dilute some predictive
signals and because it bounds us to a small sample size. With only \( n = 20 \) observational
units (the regions), fitting functional regression models comprising many terms (e.g.,
several scalar covariates and possibly their interactions; mobility and positivity curves
along with more than one scalar covariate) produces unstable, overfit outcomes. Thus, we
evaluate only the marginal effects of the scalar predictors, and the effects of mobility and
stability with one scalar control at a time. The marginal function-on-scalar regressions
of mortality curves on each of the 12 covariates have in-sample \( R^2 \)'s ranging between
\( \approx 20 \) and 65\%. Here the 'effects' are curves; \( \beta_x(t) \) represents the association of mortality
at time \( t \) with the covariate \( x \). For 8 of the covariates the \( \hat{\beta}_x(t) \)'s show the expected
signs throughout the peak period of the epidemic. In particular, the (inverse) proxy for
quality of distributed, primary health care is the strongest marginal predictor; adults
per family doctor shows a very large positive association with mortality. Also hospital,
school and work place contagion hub proxies show strong positive associations with
mortality. Nursing homes and public transport contagion hub proxies, pollution and
the percentage of individuals over 65 are positive but comparatively weaker marginal
predictors. For 4 of the covariates the \( \hat{\beta}_x(t) \)'s show unexpected signs. The percentages of
diabetics and allergic individuals show negative associations with mortality, likely due
to the fact that their prevalence is high(er) in areas which were spared the brunt of the
epidemic. In fact, estimated effect curves become positive when a differential intercept is
included in the model to account for different overall mortality levels in Group 1 and
Group 2 regions (see Appendix Figure C.15). Also the average number of members per
household shows a negative association with mortality. Its small range of variation across
regions (\( \approx 2.0 - 2.8 \), mean 2.3, s.d. 0.16) may not allow it to properly proxy the effect of
household contagions. At the same time, a strong negative correlation with the percentage
of individuals over 65 may not allow it to properly proxy inter-generational contacts;
regions with more elderly people are in fact those with smaller households. The negative
association of average number of members per household with mortality, which persists even when including a differential intercept for Group 1 and Group 2 in the model (see Appendix Figure C.15), may simply be a 'shadow' of this its negative correlation with the percentage of individuals over 65. Finally, ICU beds per 100,000 inhabitants shows a positive association with mortality which, too, persists when including a differential intercept for Group 1 and Group 2 in the model (see Appendix Figure C.15), and may be in part a 'shadow' of positive correlations with percentage of individuals over 65 and average number of beds per hospital. However, this proxy for quality of centralized, hospital-based health care, so prominent to the public debate during the epidemic, is not a negative predictor of mortality in our analysis.

In conclusion, better proxies and finer resolution may reveal stronger aggravating roles for age, nursing homes, public transport and pollution (Wu et al., 2020; Coccia, 2020) and better dissect the roles of chronic conditions, households and inter-generational contacts, and ICU availability (Dowd et al., 2020; Nepomuceno et al., 2020). But our analysis, notwithstanding limitations in the data, suggests important roles of primary care in mitigating mortality, and of contacts in hospitals, schools and work places in aggravating it.

The results of our marginal function-on-scalar regressions, which are summarized in Figure 4.6(a) for MAX mortality curves, are also consistent across data sets (see Appendix Figure C.15) – which lends them support, at least at the resolution of regions. To further validate their stability we ran fgen (Boschi, Reimherr, and Chiaromonte, 2021) – an Elastic Net-like algorithm that performs feature selection for regressions with many predictors, producing reasonably stable outcomes even with small sample sizes and collinear feature (see Chapter 3). Reassuringly, the output of fgen is consistent with the marginal analysis, and again consistent across data sets (see Appendix Table C.5): the top feature is always adults per family doctor, and the top 5 always include, in addition to it, average beds per hospital, average students per classroom, average employees per firm, and average members per household.

Finally, we ran again the function-on-function regression of mortality on local mobility and positivity, and re-evaluated the effects of these predictors introducing in the model one of the top 5 scalar covariates at a time (see Appendix Figure C.16 for results on DPC, ISTAT and MAX data), as well as their first principal component, which explains ≈ 68% of their variability and can act as a 'summary' control (see Figure 4.6(b) for MAX curves and Appendix Figure C.17 for DPC and ISTAT curves). Remarkably, while the control covariate 'subsumes' some of the predictive power in each model,
the estimated effect surfaces of local mobility and positivity retain the same shapes, and they remain very strong and comparable contributors (e.g., for MAX curves in Figure 4.6(b), the overall in-sample $R^2$ reaches 94%, the LOO-CV $R^2$ is 70%, and the partial $R^2$s are 66, 61 and 39%, respectively, for local mobility, positivity and the first principal component; see also Appendix Table C.1). Thus, with all the limitations of the data at our disposal, controlling for relevant covariates does not modify how the epidemic unfolding is associated to local mobility and positivity over time. Introducing socio-demographic, infrastructural and environmental factors in the modeling also does not change what we observed concerning residuals: mortality in Veneto is well predicted, and mortality in Lombardia remains sizably underestimated (see Appendix Figure C.8(b)) for MAX and Appendix Figure C.17 for DPC and ISTAT).

4.3 Discussion

Notwithstanding the limitations of the data employed in this study, using FDA techniques we were able to characterize heterogeneous and staggered epidemics in different areas of Italy – recapitulating and quantitating what scientists, policy makers and the public saw unfolding during the months of February, March and April 2020. In addition, we were able to document strong associations of COVID-19 mortality with local mobility and positivity, which persist in models that control for other relevant covariates. Investigating local mobility and positivity as, respectively, an actionable effector and a sentinel indicator of epidemic strength and progression, possibly to be used to adapt mitigation and containment efforts in real time, will require more and better data. In particular, accurate data on cases and hospitalizations in addition to deaths, and at a resolution much finer than that of Italian regions. Such data would allow a more systematic evaluation of the lags between the temporal patterns of mobility, contagions, illnesses and casualties – an important avenue for future studies, which could again utilize FDA tools. Such data would also be critical to better capture predictive signals in a number of covariates – which may weaken and/or become confounded when aggregating data over broad, internally heterogeneous areas. But our results, along with those of other recent studies (Cintia et al., 2020), do support a role for mobility as a key modulator of COVID-19 spread and for positivity as a monitoring variable. Moreover, they support a role for distributed, primary health care in mitigating mortality, and for hospitals, schools and work places as contagion hubs that may aggravate the epidemic. If confirmed and fine-tuned on higher resolution data, also these findings could inform decision making –
e.g., on short and medium-term investments to boost distributed health care, or “pod” patients, students or employees. Finally, an extension of the temporal span of the data would also be of great interest to properly characterize different phases of the Italian epidemic – including its evolution after the gradual weakening of lock-down measures in May 2020. We believe that our work demonstrates the potential of FDA techniques for analyzing epidemiological data. Our pipelines and the mix of FDA tools used in this study could be applied to COVID-19 data from other parts of the world.

4.4 Methods

4.4.1 Data retrieval and pre-processing

Functional variables

Daily cumulative COVID-19 death counts per region were retrieved from the Italian Civil Protection agency (Dipartimento della Protezione Civile\(^1\)). \textit{DPC mortality curves} from February 24 to April 30 were computed for each region as the daily increments in COVID-19 death counts, divided by the population of the region as of January 1, 2019 (as recorded by ISTAT\(^2\)). DPC mortality curves were set to zero for the period February 16-23, before the Civil Protection agency started releasing data. Daily death counts from all causes in 7270 Italian municipalities (about 93.5\% of the Italian population) for the years 2015-20 were downloaded from the Italian National Institute of Statistics (ISTAT\(^3\)) on June 4, 2020. Data were aggregated by region, and \textit{ISTAT differential mortality curves} from February 16 to April 30 were computed for each region as the daily difference between 2020 deaths and the average daily deaths in 2015-19, divided by the total population of the municipalities included in the death counts as of January 1, 2019\(^4\). \textit{MAX mortality curves} were created taking, for each region and each day, the maximum between DPC mortality and ISTAT differential mortality. Daily measurements concerning “Grocery & pharmacy” mobility from February 16 to April 30 were downloaded for each region from the Google Mobility Report\(^5\) (\textit{local mobility curves}). These measurements

\(^1\)https://github.com/pcm-dpc/COVID-19/tree/master/dati-regioni
\(^2\)http://asti.istat.it/asti
\(^3\)https://www.istat.it/it/files/2020/03/Dataset-decessi-comunali-giornalieri-e-tracciato-record-4giugno.zip
\(^4\)http://dati.istat.it/Index.aspx Popolazione e famiglie/Popolazione/Popolazione residente al 1° gennaio/Tutti i comuni/2019
\(^5\)https://www.google.com/covid19/mobility/
express percent changes with respect to the corresponding daily mobility levels in the first five weeks of 2020 (January 3 to February 6). *Positivity curves* were constructed using raw data from the Italian Civil Protection agency\(^1\). For each day from February 24 to April 30 and each region, we took the ratio between the number of new positive cases and the number of new tests performed. The ratios were truncated at 0 and 1 to account for irregularities in the row data (e.g., positive cases = –1, or positive cases exceeding tests performed, presumably due to delays in test results). Like DPC mortality, positivity curves were set to zero for the period before the Civil Protection agency started releasing data (February 16-23). For all functional data sets, the two self-governing provinces of Trento and Bolzano were considered together as the Trento/Bolzano region, since not all data were available for both provinces separately. The 20 curves in each functional data set were smoothed using cubic smoothing *B*-splines with knots at each day and roughness penalty on the curve second derivative (Ramsay and Silverman 2005b). For each functional data set the smoothing parameter was selected minimizing the average generalized cross-validation error, GCV (Craven and Wahba 1978), across the 20 curves. All computations were performed using the R package *fda* (Ramsay et al., 2011).

### 4.4.1.1 Scalar covariates

We considered a large number of scalar covariates of potential interest (see Appendix Table C.2), and focused on the 12 listed in Table 4.1 and below. In retrieving and computing various measurements, as was done for the functional variables, the provinces of Trento and Bolzano were aggregated into the Trento/Bolzano region. % Over 65 was retrieved from ISTAT\(^2\) at the regional level for the year 2018. % Diabetics and % Allergics were retrieved from ISTAT\(^6\) at the regional level for the year 2018. *Adults per family doctor* was retrieved from the Ministry of Health\(^7\) at the regional level for the year 2017. To compute *ICU beds per 100,000 inhabitants*, we collected the total number of ICU beds in each region in 2018 from the Ministry of Health\(^8\), multiplied by 100,000 and divided by the population of the region\(^2\) as of January 1, 2019. To compute *Ave. beds per hospital (whole)* we used data from the Ministry of Health\(^9\), which provides the number of beds per ward in each hospital in 2018. We first aggregated them over wards belonging to the same hospital, and then averaged over hospitals in each region. *Ave. beds per nursing home (ward)* was also obtained based on data for the year 2018.

\(^1\)http://dati.istat.it/Index.aspx?QueryId=15448
\(^2\)http://www.salute.gov.it/imgs/C_17_pubblicazioni_1203_ulterioriallegati_ulterioreallegato_8_alleg.pdf
\(^3\)http://www.datit skewed.salute.gov.it/dati/dettaglioDataset.jsp?menu=dattididPag=96
\(^4\)http://www.salute.gov.it/imgs/C_17_bancheDati_6_0_1_file.xls
from the Ministry of Health\textsuperscript{10} – here we considered regional averages at the level of wards, without aggregating over wards inside the same nursing home (the ward-level covariate had a slightly higher association with mortality outcomes). To compute \textit{Ave. students per classroom} we used data from the Ministry of Education\textsuperscript{11}, which provides the number of students in each classroom of each school in the country (public or private, at every level of education), for the year 2018. We averaged them over schools in each region. Data for Trento/Bolzano and Valle d’Aosta were missing, and were imputed through random forest imputation \cite{Stekhoven2012}, with default parameters maxiter=10 (maximum number of iterations to be performed given the stopping criterion is not met beforehand) and ntree=100 (number of trees to grow in each forest). To compute \textit{Ave. employees per firm} we used data from ISTAT\textsuperscript{12}, which provides number of employees per firm at the level of municipalities. We averaged them over firms in each region. Data for Valle d’Aosta were missing, and were again imputed through random forest imputation with default parameters. \textit{Ave. members per household} was retrieved from ASR Lombardia\textsuperscript{13} at the regional level for the year 2017. To compute \textit{Public transport rides per capita} we used data from ISTAT\textsuperscript{14} which provides the number of rides per capita for each Italian province in 2017. We multiplied these by the provinces’ population\textsuperscript{2} as of January 1, 2019, summed up over provinces in the same region, and divided by the region population\textsuperscript{2} as of January 1, 2019. To compute \textit{PM10} we used data from ISTAT\textsuperscript{14}, which provides the average annual concentrations of PM10 (in µg/m³) detected by air quality meters distributed over the Italian territory. We averaged them over meters located in each region.

\section*{4.4.2 Multivariate analysis tools}

We used a number of standard multivariate techniques to analyze first the entire set of scalar covariates and then the 12 we focused on – including the extraction of Principal Components, the calculation of Variance Inflation Factors \cite{Allison1999} to evaluate multicollinearity, and clustering based on hierarchical agglomeration \cite{Hastie2001}.\textsuperscript{10,11,12,13,14}
The latter was used both to agglomerate covariates with similar behavior across regions and to agglomerate regions with similar behavior across covariates. Agglomerative hierarchical clustering groups elements in a set with a bottom-up procedure that results in a dendrogram. Each element starts in its own cluster, and pairs of clusters are merged iteratively with a chosen distance for elements and linkage criterion for clusters.

We employed the correlation distance, defined as \( d(x_1, x_2) = 1 - |\text{corr}(x_1, x_2)| \) for two generic elements \( x_1 \) and \( x_2 \), and the complete linkage, defined as \( D(X_1, X_2) = \max_{x_1 \in X_1, x_2 \in X_2} d(x_1, x_2) \) for two generic clusters \( X_1 \) and \( X_2 \) (thus, the distance between two clusters is defined as the furthest distance between their elements).

We also used biclustering on the 20 (regions) by 12 (covariates) data matrix, to identify subsets of regions exhibiting similar behaviors across subsets of covariates. Following standard literature, we sought sub-matrices of the data whose entries are consistent with the "ideal" additive model \( x_{i,j} = \mu + \alpha_i + \tau_j \), where \( \mu \) is the typical value within the bicluster, and \( \alpha_i \) and \( \tau_j \) are additive adjustments for row \( i \) and column \( j \), but we set all \( \alpha_i \)s to 0 in order to find constant column biclusters, i.e., sub-matrices with constant columns (covariates). We employed the Cheng and Church Biclustering Algorithm (Cheng and Church, 2000), a greedy algorithm which finds the largest sub-matrices whose departure from the additive model is below a user-defined threshold. The departure is computed using the H-score (or mean squared residue score); in symbols, \( H(I, J) = \frac{1}{|I||J|} \sum_{i \in I, j \in J} (x_{i,j} - x_{I,j})^2 \), where \( I \) and \( J \) index the sets of rows and columns composing the bicluster, \( x_{i,j} \) is a generic cell in the bicluster and \( x_{I,j} \) is the mean of column \( j \). We implemented this algorithm with a recently proposed adjustment to the H-score (Di Iorio et al., 2020) that corrects a bias towards smaller biclusters in the original formulation. The adjusted H-score is defined as \( H_{\text{adj}}(I, J) = (\prod_{r=2}^{r-1} x_{r-1}^2 \prod_{q=2}^{q-1} s_{q-1}^2)^{-1} H(I, J) \).

### 4.4.3 Functional Data Analysis tools

**Local clustering of curves and functional motif discovery**

We performed local clustering of smoothed mortality curves (DPC, ISTAT and MAX, separately) using probabilistic K-mean with local alignment, probKMA (Cremona and Chiaromonte, 2020). ProbKMA is a K-mean-like algorithm for functional data that finds \( K \) groups in a set of curves based on a local similarity among portions of the curves themselves. This allows the discovery of functional motifs, i.e. of typical local shapes that recur within and across the curves. In symbols, the algorithm finds \( K \) motifs \( v_1, \ldots, v_K \), membership probabilities \( p_{k,i} \) and shifts \( s_{k,i} \) (i.e. the starting points of the
motif instances) for each cluster-curve pair that minimize the generalized least-squares functional

\[ J(v_1, \ldots, v_K, p_{k,i}, s_{k,i}) = \sum_{i=1}^{N} \sum_{k=1}^{K} p_{k,i}^2 d^2(\tilde{x}_i, v_k), \]

where \( \tilde{x}_i \) is the portion of the curve \( i \) corresponding to the shift \( s_{k,i} \), and \( d \) is the distance used to capture local similarity. For each data set, we considered \( K = 2 \) and \( K = 3 \). ProbKMA is probabilistic; it returns as output a membership probability \( p_{k,i} \) for each cluster-curve pair. However, such an output can be turned into a hard partition by assigning each curve to the group with highest membership probability – which is what we did here. Notably, for \( K = 2 \), membership probabilities showed that Lombardia’s and Valle d’Aosta’s extreme mortality patterns were not well accommodated even in the "exponential" group. The algorithm can employ different definitions of similarity \( d \) and thus capture different aspects of curve shapes. We used Euclidean \( (L^2) \) distance between curve levels for our main analysis – in symbols, \( d = \frac{1}{2} \int_0^T (x(t) - v(t))^2 dt \) for two generic curves \( x \) and \( v \) – though using Euclidean distance between curve derivatives produced similar results (not shown). ProbKMA allows the length of the motifs to be extended endogenously starting from a minimal one fixed in input. However, to identify epidemic patterns we ran it with a fixed motif length of 65 days – hence allowing a maximum shift of 10 days between curves (the mortality curves are 75 days long). The same clusters and very similar shifts were obtained with a fixed motif length of 50 days, which allows a maximum shift of 25 days (results not shown). The shifts produced by probKMA with \( K = 2 \) on the three mortality data sets (DPC, ISTAT and MAX) were employed to align, in addition to the mortality curves themselves, local mobility and positivity curves. All subsequent analyses employing shifted curves (tests contrasting groups of curves, functional boxplots and depth analyses, and functional regression models) were therefore restricted to the 65-day portions where mortality curves aligned following the two probKMA motifs. We also validated the groups produced by probKMA with a modified version of funBI (Di Iorio and Vantini, 2019), an algorithm typically used for finding functional biclusters. We used the modified funBI to identify groups of curves characterized by group-specific fixed length motifs, considering all possible sub-curves of a fixed length and clustering them with a divisive hierarchical algorithm (results not shown).

Testing for differences between groups of curves

We employed an Interval-Wise Testing algorithm developed for omics data, IWTomics (Cremona et al. 2018), to test for differences between the two groups of shifted mortality curves produced by probKMA with \( K = 2 \) (again, separately for DPC, ISTAT and MAX). IWTomics is a non-parametric, permutation-based functional hypothesis test. It contrasts
two sets of curves aligned on a common domain to detect locations where the two sets differ significantly, and scales at which such significant differences are displayed (scales correspond to varying degrees of adjustment for multiple testing on intervals of varying lengths). Here locations are represented by the 65 days where the shifted mortality curves are defined, while scales vary from 1 day to the whole 65 days. The test was performed with the R package IWTomics (Cremona et al., 2018). The package allows the user to select among various possible test statistics; we employed the mean.

**Functional boxplots and depth analyses**

The functional boxplot (Sun and Genton, 2011) is an exploratory tool used to visualize functional data. It is constructed after ordering a set of curves based on a depth measure, such as the modified band depth (López-Pintado and Romo, 2009). The statistics employed to construct a functional boxplots are: the 50% central region envelope, the median curve, and the maximum non-outlying envelope. The 50% central region envelope corresponds to the box in a classical boxplot; it contains the 50% deepest, most centrally located curves. The median, i.e. the deepest curve, is inside this box and represents a robust "center" of the functional data set. The maximum non-outlying envelope is obtained by inflating the 50% central region envelope by 1.5 times its range. All curves extending outside of this envelope are flagged as outliers (the fact that the ISTAT data set in Figure 4.3(b) lacks outlying curves based on this definition is due to the width of its 50% central region envelope). We ranked the curves based on their depth measurements, after attributing a sign to such measurements with an ad hoc procedure. We subtract the median from each curve, and consider the share of the domain on which the difference is positive. If this is larger than 50%, we attribute a positive sign to the curve’s depth – otherwise, we attribute a negative sign. Curves can thus be ranked from the most outlying above the median (labeled as positive), down to those close to the median, down to the most outlying below the median (labeled as negative) – see Figure 4.3(b). While this is not a fully general procedure, it works well on the DPC, ISTAT and MAX mortality curves we considered, which are rather unambiguously above/below the median (the share of the domain where the difference from the median is positive is ≥ 70 or ≤ 30% for all curves in all three data sets). Note also that the median curve of a data set, defined as the deepest, does not necessarily have half of the curves above it and half of the curves below it in the signed ranking we created (e.g., Toscana is the median curve in both ISTAT and MAX data sets, but the number of curves above/below it differs).
Functional regression models

We consider models where a functional response variable is regressed against functional predictors and/or scalar covariates (Ramsay and Silverman, 2005b; Kokoszka and Reimherr, 2017b). All are special cases of the general equation (Horváth and Kokoszka, 2012b)

\[ y_i(t) = \alpha(t) + \sum_{\ell=1}^{L} \int \beta_\ell(s,t) x_{i,\ell}(s) ds + \sum_{j=1}^{J} \beta_j(t) x_{i,j} + \epsilon_i(t) \quad i = 1, \ldots, n. \]

\( n \) is the number of observations, in our case \( n = 20 \) regions. \( y_i(t), i = 1, \ldots n \) are the aligned mortality curves (DPC, ISTAT or MAX, modeled separately), \( \alpha(t) \) is a functional intercept and \( \epsilon_i(t), i = 1, \ldots n \) are i.i.d. Gaussian model errors. \( L \) is the number of functional predictors. \( x_{i,\ell}(s), i = 1, \ldots n, \ell = 1, \ldots L, \) are such predictors, measured on the \( n \) observations. The regression coefficient of each functional predictor, \( \beta_\ell(s,t), \) is a surface. \( J \) is the number of scalar covariates. \( x_{i,j}(s), i = 1, \ldots n, j = 1, \ldots J, \) are such covariates, measured on the \( n \) observations. The regression coefficients of each scalar covariate, \( \beta_j(t), \) is a curve. For the marginal regression of mortality on local mobility and mortality on positivity, we have \( L = 1 \) and \( J = 0. \) For the joint regression of mortality on local mobility and positivity, we have \( L = 2 \) and \( J = 0. \) For the marginal regressions of mortality on individual scalar covariates, we have \( L = 0 \) and \( J = 1. \) In Figure C.15 we fit marginal regressions of this type allowing the estimation of two different intercepts: \( \alpha_1(t) \) for curves in Group 1 and \( \alpha_2(t) \) for curves in Group 2. Finally, for the joint regression of mortality on local mobility, positivity and one scalar control variable, we have \( L = 2 \) and \( J = 1. \) To fit all these functional regressions we used the R package `refund` (Goldsmith et al., 2016), which estimates the functional coefficients as well as their standard errors. We used these standard errors to construct confidence bands around the estimated functional coefficients. To gauge the explanatory power of each model, we computed the in-sample \( R^2 \) as well as the Leave-One-Out Cross-Validation (LOO-CV) \( R^2. \) The former is a functional generalization of the classical coefficient of determination defined as \( SS_{reg}/(SS_{reg} + SS_{res}), \) where \( SS_{reg} \) and \( SS_{res} \) are the regression and the residual sum of squares, respectively. To compute the latter, for each observation \( i, \) one replaces the fitted response curve \( \hat{y}_i(t) \) (from the model fitted on all observations) with the predicted response curve \( \hat{y}_{pred,i}(t) \) obtained for \( i \) from the model fitted withholding \( i \) itself. Finally, for models with multiple terms (predictor and/or covariate), the partial \( R^2 \) of each term is computed as \( (R^2 - R^2_{red})/(1 - R^2_{red}), \) where \( R^2 \) is the coefficient of determination of
the complete model, and $R^2_{red}$ that of the model comprising all terms but the one being evaluated.

**fgen for feature selection**

*fgen* ([Boschi, Reimherr, and Chiaromonte, 2021](#)) is an highly efficient algorithm to perform feature selection in a function-on-scalar regression framework – see Chapter 3. It is a generalization of SsNAL-EN ([Boschi et al., 2020](#)), which performs Elastic Net ([Zou and Hastie, 2005](#)) feature selection in a standard regression framework, i.e. when both response and features are scalars. The Elastic Net is a hybrid between LASSO and Ridge, which penalizes both the $L_1$ and the $L_2$ (Euclidean) norm of the regression coefficients. The $L_1$ penalty induces sparsity selecting only the most predictive among the features. The $L_2$ penalty regularizes coefficient estimates mitigating variance inflation due to collinearity. In particular, we used *fgen* to perform feature selection for the regression of mortality against all 12 scalar covariates in Table 4.1. Notably, we selected the same top 5 features across all three data sets (DPC, ISTAT and MAX) (see Appendix Table C.4) – lending strong support to their association with mortality.
Figure 4.3: **Functional boxplot and ranking.** (a) Functional boxplot of the MAX data set (top) and MAX mortality curves (bottom) color-coded according to their ranking, as shown in the MAX column of (b). In the boxplot, Toscana is the median (black continuous line); Lombardia, Valle d’Aosta and Liguria are identified as outliers (red dashed lines); and the 50% innermost "box" (grey area) include the curves for Trento/Bolzano, Emilia-Romagna, Marche, Friuli Venezia Giulia, Veneto, Toscana, Molise, Abruzzo, Sardegna, Umbria, and Basilicata. Note that the "box" is skewed upwardly. (b) Rankings of the ISTAT (left), MAX (center) and DPC (right) mortality curves. The median regions are in bold, gray rectangles mark the 50% innermost boxes, and pale red rectangles mark outliers (no region is labeled as an outlier in the ISTAT data set; see Methods). The dots representing each region are color-coded (from intense red, through gray, to intense blue) according to their signed depth values (see Methods). In all three data sets, Lombardia’s curve is the most extreme at the very top of the ranking and, in contrast, Veneto’s curve is deep in the bulk close to the median (Toscana for ISTAT and MAX, Friuli Venezia Giulia for DPC). Segments joining the regions across the three rankings show how the top portion remains rather stable, while the mid- and bottom portions contain several crossings. Regions at the top are those characterized by "exponential" epidemics (Group 1), while regions in the middle and at the bottom are those with "flattened" epidemics (Group2), whose curves can more easily switch in their depth ranks.
Figure 4.4: **Associating mortality to local mobility and positivity.** (a) Local mobility curves (Google’s “Groceries & pharmacy”) and positivity curves (regularized ratios of new cases to number of tests performed) in the 20 Italian regions. Curves are smoothed with splines, with degree of smoothing selected by generalized cross-validation, and shifted based on probKMA run on the MAX mortality curves with $K = 2$; time is marked as a day number representing the region-specific time of the epidemic unfolding, and corresponds to actual time (starting on February 16 and ending on April 30) only for regions with no shifts, e.g., Lombardia. Vertical lines show the days corresponding to the nationwide lock-down (March 9) and the suspension of all nonessential production activities (March 23) without shifts, stars on the curves mark the lock-down after the region specific shifts. The example regions of Figure 4.1(a) are highlighted in color. (b) Estimated effect surfaces from the joint function-on-function regression of MAX mortality on local mobility and positivity shown in 3D and as contour plots (March 9, without shift, is again marked on both). Early and mid-period local mobility levels are strong positive predictors of mortality at its peak. Positivity has similar but much weaker predictive signals, likely because the effects are subsumed by mobility. Late local mobility has a negative association with mortality at its peak (mobility resumed faster in regions with milder epidemics), and late positivity a strong positive one (positivity remained elevated in regions with worse epidemics). The regression captures a large share of the variability in mortality curves (in-sample $R^2 = 0.90$, LOO-CV $R^2 = 0.52$), with substantial and comparable contributions of the two predictors (partial $R^2$s = 0.62, 0.53).
Figure 4.5: **Interdependencies among scalar covariates and regions.** (a) Heatmap of the 20 (regions) x 12 (covariates) data matrix, with dendrograms from separate hierarchical clustering (correlation distance, complete linkage) of the regions (left) and the covariates (top). Color coding within cells represents values of the standardized covariates (centered and scaled to mean 0 and standard deviation 1). Color coding of some cell borders identifies the biclusters in (b). The dendrograms capture a distinct interdependence structure. For instance, there are marked similarities among Lombardia, Veneto, Emilia Romagna and Piemonte, as well as among some groups of southern regions (Sicilia, Campania, Puglia and Calabria; Basilicata, Abruzzo and Molise). There are also marked associations among groups of covariates. The contagion hubs proxies for hospitals, schools and work places, and number of adults per family doctor, vary closely together. So do the contagion hub proxy for public transport and pollution levels; the percentages of individuals affected by diabetes and allergies; and ICU beds and the percentage of individuals over 65. (b) Restricted heat-maps further illustrating interdependencies through two biclusters of regions and covariates. Color-coding within cells corresponds to that in (a), and each bicluster is identified by a border color and its adjusted H-score (an inverse measure of bicluster strength; see Methods). The first bicluster (adjusted H-score = 0.0902) comprises central and southern regions with 'flat(tened)' epidemics (Group 2). The second bicluster (adjusted H-score = 0.0942) comprises northern regions with 'exponential' epidemics (Group 1) but also northern and central regions from Group 2.
Figure 4.6: **Associating mortality to socio-demographic, infrastructural and environmental factors.** (a) Results from marginal function-on-scalar regressions. Mortality curves are regressed against each of the scalar covariates in Table 4.1. The top plot displays the signs of the effect curves estimated on the MAX data. Time, marked as the 65 days of the region specific epidemic unfoldings, is on the vertical axis (the nationwide lock down on March 9, without shift, is marked by a horizontal line. Red, blue and green indicate, respectively, positive, negative and non-significant portions (i.e., where 95% confidence bands around the estimated effect curve are entirely above, entirely below, or contain 0; see Methods). The bottom plot displays in-sample $R^2$ for the regressions fitted on MAX, ISTAT and DPC data; these are remarkably consistent. The names in red on the horizontal axes indicate the top 5 covariates selected by SnNAL-EN on all three data sets (see Methods); these are also the ones with the largest $R^2$'s. (b) Results from the joint function-on-function regression of MAX mortality on local mobility, positivity, and the first principal component (pc1) of the top 5 covariates, used as a "summary" control. This control does not modify the shapes of the estimated effect surfaces for mobility and positivity (shown on top) – which are very similar to the ones in Figure 4.4(b). The estimated effect curve for pc1 shows a positive and significant association with mortality at its peak (bottom right; 95% confidence band in dashes, gray corresponds to non significant portions, vertical dashed line corresponds to March 9, without shift). The sign of this effect is consistent with marginal findings, based on the loadings of the first principal component (bottom left; positive for adults per family doctor, average beds per hospital, average students per classroom and average employees per firm, and negative for average members per household). With the addition of pc1, the regression reaches an in-sample $R^2 = 0.94$ and a LOO-CV $R^2 = 0.7$. The contributions of local mobility and positivity remain high (partial $R^2 = 0.66$ and 0.61, respectively). That of our "summary" covariate is also substantial (partial $R^2 = 0.39$).
Chapter 5 | Conclusions and future work

5.1 Functional Registration

In Chapter 2, we introduced FCC registration, a new low-dimensional registration procedure based on the covariance operator between a functional response and a functional predictor. The procedure finds the most important modes of covariation between the two sets of curves, aligns them simultaneously, and at the same time performs a reduction of the data – leveraging a low-dimensional representation built upon the dependence between functional response and functional predictor. We implement FCC registration using two different algorithms: the continuous registration algorithm introduced by Ramsay and Silverman (2005a), and a new algorithm based on H1 distances. Both in our simulation study and in our application to the AneuRisk data, FCC registration improved regression performance in comparison to other registration approaches. This was particularly the case in its H1 implementation. Furthermore, FCC registration led to better alignment of the response $y$. Notably, in most cases analyzed, even for simulated curves that comprised two distinct groups, 1-dimensional subspaces sufficed to capture significant modes of covariation.

Future developments

FCC registration lends itself to numerous further developments. First, as mentioned in Section 2.4, the RS and H1 algorithms could be iterated. Iterative registration procedures have been proposed by Kneip and Ramsay (2008) and Sangalli et al. (2009a). Updating the covariance operator, and therefore the projections of $x$ and $y$ on the selected basis systems at each iteration, may improve the registration procedure: in the first iteration we add to the projections in (2.5) the mean of the misaligned data, which may differ from the mean of ideally aligned data and thus introduce noise into the target functions.
Iterating can mitigate the problem, since at each iteration one would add to the target functions the mean of the data as aligned in the previous iteration, which becomes progressively closer to the “right” mean. We plan to explore iterations in future work, paying special attention to the definition of appropriate stopping criteria.

A second avenue for further development is extending our approach to warpings that can modify the domains of the curves will also broaden applicability – allowing one to handle a larger spectrum of real scenarios where curves are measured on different domains. This could proceed in several directions, e.g. allowing shifts, or pursuing local (as opposed to global) curve registration.

Finally, in Section 2.6, we discuss how to extend FCC registration to regressions with more than one functional predictor. This would greatly expand its applicability, since many contemporary data sets contain measurements on several potentially useful functional predictors.

5.2 Functional Feature Selection

In Chapter 3, we proposed a new Functional Group Elastic Net method (fgen) to solve the function-on-scalar feature selection problem. Our proposal starts with the development of a novel, highly-efficient SsNAL algorithm to solve the Group Elastic Net – which is then extended to the function-on-scalar regression framework using a Functional Principal Components representation. Though we could rely on critical prior results (Tomioka and Sugiyama, 2009; Li et al., 2018; Boschi et al., 2020), in order to integrate the group structure into SsNAL, we had to tackle more complex mathematical operators and redefine the theoretical foundation. Our simulations show a substantial reduction in CPU time with respect to the best existing Group Elastic Net solvers. Finally, we applied fgen to a GWAS study detecting a SNP that may affect obesity risk in children.

Future developments

The current version of fgen is limited to the case where each group has the same size and to the function-on-scalar feature selection problem. In the future, we plan to further extend our work investigating more complex optimization problems (e.g., allowing each group to have a different size).

Given the broad applicability of fgen, further improving estimation accuracy by considering more complex penalizations could have a substantial value added. For instance, in GWAS applications, it may help us expand our understanding of the genetic
architecture and the causes behind complex human diseases. One possibility would be to impose smoothness by directly adding specific constraints for the functional coefficients in the penalization, rather than representing the curves using a basis expansion (Lin and Zhang, 2006).

We are also extending \texttt{fgen} to the function-on-function feature selection setting, where both the response and the predictors are curves. This framework increases the dimension of the optimization problem. It requires defining anew the relevant set of mathematical operators and carefully working out high-order tensors to preserve the method’s efficiency. As we develop this new algorithm, we are planning to use it to analyze functional-MRI data employing time series curves from individual voxels, i.e., small three-dimensional cuboids in a brain image, as functional predictors (Qi and Luo, 2018).

5.3 Research applications and collaborations

In Chapter 4 we employed functional clustering, registration, regression, and future selection techniques to study the evolution of the COVID-19 epidemic in Italy. In particular, we investigated patterns of COVID-19 mortality across 20 Italian regions and their association with mobility, positivity, and socio-demographic, infrastructural and environmental covariates during the first wave of the epidemic. Notwithstanding limitations in accuracy and resolution of the data available from public sources, we pinpointed significant trends exploiting information in curves and shapes with Functional Data Analysis techniques. These depict two starkly different epidemics; an “exponential” one unfolding in Lombardia and the worst hit areas of the north, and a milder, "flat(tened)" one in the rest of the country – including Veneto, where cases appeared concurrently with Lombardia but aggressive testing was implemented early on. We found that mobility and positivity can predict COVID-19 mortality, also when controlling for relevant covariates and identify important factors that may have a central role in containing the spread of the virus – such as mobility restrictions, primary health care, and the control of contacts in hospitals, schools, and workplaces.

The techniques we described could capture additional and potentially sharper signals if applied to richer data. In particular, accurate data on cases and hospitalizations in addition to deaths, and at a resolution much finer than that of Italian regions would allow a more systematic evaluation of the lags and the predictive signals between the temporal patterns of mobility, contagions, illnesses and casualties. For instance, we are
now working on a new project, contrasting mortality and mobility patterns between the first and second waves of COVID-19 (both pre-vaccine) across the 107 Italian provinces to further investigate how the timing of restrictions policies can help control the epidemic.
### Appendix A

#### Chapter 2 Supplementary Material

<table>
<thead>
<tr>
<th></th>
<th>simulation 1</th>
<th></th>
<th>simulation 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$aln(x)$</td>
<td>$aln(y)$</td>
<td>$SPE$</td>
<td>$L1OPE$</td>
</tr>
<tr>
<td>$msl$</td>
<td>0</td>
<td>0</td>
<td>26</td>
<td>30</td>
</tr>
<tr>
<td>$\mu_{RS}$</td>
<td>4111</td>
<td>10456</td>
<td>2703</td>
<td>3088</td>
</tr>
<tr>
<td>$er$</td>
<td>4494</td>
<td>41499</td>
<td>10845</td>
<td>11175</td>
</tr>
<tr>
<td>$pc_{RS}$</td>
<td>6137</td>
<td>38050</td>
<td>23781</td>
<td>23449</td>
</tr>
<tr>
<td>$cc_{RS}$</td>
<td>2115</td>
<td>14505</td>
<td>3358</td>
<td>4958</td>
</tr>
<tr>
<td>$pc_{H1}$</td>
<td>1005</td>
<td>1637</td>
<td>840</td>
<td>964</td>
</tr>
<tr>
<td>$cc_{H1}$</td>
<td>356</td>
<td>559</td>
<td>203</td>
<td>227</td>
</tr>
</tbody>
</table>

Table A.1: Simulation results. Columns: $H1$ distances between the curves in each row and the original $x$ and $y$ curves; in sample $H1$ prediction error; leave-one-out $H1$ prediction error; Euclidean distances between $\hat{\beta}$ and the true $\beta$. Rows: Org, the original curves; msl, the misaligned curves; followed by curves registered with different procedures – as indicated.
Table A.2: Simulation results. Columns: $L^2$ distances between the curves in each row and the original $x$ and $y$ curves; in sample $L^2$ prediction error; leave-one-out $L^2$ prediction error. Rows: Org, the original curves; msl, the misaligned curves; followed by curves registered with different procedures – as indicated.

<table>
<thead>
<tr>
<th></th>
<th>simulation 1</th>
<th></th>
<th>simulation 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$aln(x)$</td>
<td>$aln(y)$</td>
<td>SPE</td>
<td>LIOPE</td>
</tr>
<tr>
<td>$m$</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>$\mu_{RS}$</td>
<td>227</td>
<td>1057</td>
<td>388</td>
<td>444</td>
</tr>
<tr>
<td>$\sigma_{RS}$</td>
<td>239</td>
<td>1864</td>
<td>1395</td>
<td>1486</td>
</tr>
<tr>
<td>$\sigma_{er}$</td>
<td>357</td>
<td>3350</td>
<td>3098</td>
<td>3050</td>
</tr>
<tr>
<td>$p_{RS}$</td>
<td>99</td>
<td>176</td>
<td>419</td>
<td>761</td>
</tr>
<tr>
<td>$cc_{RS}$</td>
<td>50</td>
<td>87</td>
<td>96</td>
<td>118</td>
</tr>
<tr>
<td>$p_{H1}$</td>
<td>78</td>
<td>176</td>
<td>85</td>
<td>129</td>
</tr>
<tr>
<td>$cc_{H1}$</td>
<td>19</td>
<td>29</td>
<td>27</td>
<td>38</td>
</tr>
</tbody>
</table>

Figure A.1: Simulation 1 original curves (left), misaligned curves (center), registered curves by $cc_{H1R}$ (right) for $y$ (top) and $x$ (bottom).
Figure A.2: Simulation 2 original curves (left), misaligned curves (center), registered curves by $cc_{H1R}$ (right) for $y$ (top) and $x$ (bottom).

Figure A.3: Original $\beta$ for simulation 1 (left), simulation 2 (right)
Figure A.4: Simulation 1 projections of the misaligned curves on the first (top) and second (bottom) principal components (second and fourth column) and canonical directions (first and third columns) of $y$ (first and second columns) and $x$ (third and fourth columns).

Figure A.5: Simulation 2 projections of the misaligned curves on the first (top) and second (bottom) principal components (second and fourth column) and canonical directions (first and third columns) of $y$ (first and second columns) and $x$ (third and fourth columns).
Figure A.6: Simulation 1 estimated $\hat{\beta}$’s by different regressions. Original: the regression is performed on the original curves; misaligned: the regression is performed on the misaligned curves. In the other cases the regression is performed on the curves registered by the indicated algorithm. The white zones are values outside the $z$-range on the right.
Figure A.7: Simulation 2 estimated $\hat{\beta}$'s by different regressions. Original: the regression is performed on the original curves; misaligned: the regression is performed on the misaligned curves. In the other cases the regression is performed on the curves registered by the indicated algorithm. The white zones are values outside the $z$-range on the right.
Figure A.8: AneuRisk data, projections of the misaligned curves on the first (top) and second (bottom) principal components (second and fourth column) and canonical directions (first and third columns) of $y$ (first and second columns) and $x$ (third and fourth columns).
Figure A.9: AneuRisk data, leave-one-out predicted curves by different regressions.
Figure A.10: AneuRisk data, leave one predicted curve by different regressions for just one patient. The top left panel represents the original misaligned curve before any registration.
\( \tilde{\beta} \)’s - AneuRisk

Figure A.11: AneuRisk data, estimated \( \beta \)’s by different the different algorithms. The white zones are values outside the \( z \)-range on the right.
Appendix B
Chapter 3 Supplementary Material

B.1 Proofs

In this section we provide detailed proofs of our theoretical results.

B.1.1 Proof of Proposition 1

Note that \( \pi(B) = \sum_{i=1}^{p} \pi(B_i) \) is a separable sum. From Boyd and Vandenberghe (2004) we have

\[
\pi^*(Z) = \sum_{i=1}^{p} \pi^*(Z_i) .
\]  

To compute \( \pi^*(Z_i) \), we use a result from Touchette (2005). Define \( g(b) = (Z_i)^T b - \pi(b) = (Z_i)^T b - \lambda_1 \|b\|_2 - (\lambda_2/2) \|b\|_2^2 \), and let \( b^* = \arg \max_{b \in \mathbb{R}^p} g(b) \). Then

\[
\pi^*(Z_i) = g(b^*) .
\]  

To find \( b^* \) we have to solve \( \nabla g(b) = 0 \), where \( \nabla g(b) = Z_i - \lambda_2 - \lambda_1 \left\{ \begin{array}{ll} \|b\|_2^{-1} b & \text{if } \|b\|_2 \neq 0 \\ \{ b : \|b\|_2 \leq 1 \} & \text{otherwise} \end{array} \right. \). Consider the case where \( \|b\|_2 \neq 0 \) and set \( \nabla g(b) = 0 \). We get

\[
Z_i = (\lambda_2 + \|b\|_2^{-1} \lambda_1) x .
\]  

To solve for \( b \) we must first compute \( \|b\|_2 \). Taking the norm of both sides, we get

\[
\|Z_i\|_2 = (\lambda_2 + \|b\|_2^{-1} \lambda_1) \|b\|_2 . \quad \text{Thus, } \|b\|_2 = \lambda_2^{-1} \left( \|Z_i\|_2 - \lambda_1 \right) . \quad \text{Plugging the last expression in } (B.3) \text{ and solving for } b, \text{ we obtain } b^* = \lambda_2^{-1} (1 - \|Z_i\|_2^{-1} \lambda_1) Z_i, \text{ for } \|b^*\|_2 \neq 0.
\]
We need to take into account that \( \text{dom}(p) = \text{range}(p^*) \) and vice-versa. In particular, \( \|b^*\|_2 \neq 0 \) iff \( \|Z_i\|_2 > \lambda_1 \). Therefore, we have

\[
b^* = \begin{cases} 
\lambda_2^{-1} \left(1 - \|Z_i\|_2^{-1} \lambda_1\right) Z_i & \|Z_i\|_2 > \lambda_1 \\
0 & \text{o.w.}
\end{cases}
\]  

(B.4)

From (B.2) we now need to compute \( g(b^*) \). If \( \|Z_i\|_2 \leq \lambda_1 \), then \( g(b^*) = 0 \). If \( \|Z_i\|_2 > \lambda_1 \), after some algebraic manipulations, we obtain \( g(b^*) = (2\lambda_2)^{-1} (\|Z_i\| - \lambda_1)^2 \). Finally, (B.1) gives us the desired result

\[
\pi^*(Z) = (2\lambda_2)^{-1} \sum_{i=1}^{p} \begin{cases} 
(\|Z_i\|_2 - \lambda_1)^2 & \|Z_i\|_2 > \lambda_1 \\
0 & \text{o.w.}
\end{cases}
\]  

(B.5)

### B.1.2 Proof of Proposition 2

Since \( \pi(B) = \sum_{i=1}^{p} \pi(B_i) \) is a separable sum, from \cite{Beck2017} Remark 6.7, we know

\[
\text{prox}_{\sigma \pi}(B) = \left( \text{prox}_{\sigma \pi}(B_1), \ldots, \text{prox}_{\sigma \pi}(B_p) \right)^T.
\]  

(B.6)

From \cite{Fan2016}, we have \( \text{prox}_{\sigma \lambda_1 \|\cdot\|_2}(B_i) = \left[ 1 - \|B_i\|_2^{-1} \sigma \lambda_1 \right]_+ B_i \). From \cite{Beck2017} 6.2.3, we have \( \text{prox}_{(\sigma \lambda_2/2)\|\cdot\|_2^2}(B_i) = (1 + \sigma \lambda_2)^{-1} B_i \). Moreover, \( \|\cdot\|_2 \) is a proper closed and convex function, thus we can compose \( \text{prox}_{(\sigma \lambda_2/2)\|\cdot\|_2^2} \) and \( \text{prox}_{\sigma \lambda_1 \|\cdot\|_2} \) as described in \cite{Parikh2014}, obtaining the desired form

\[
\text{prox}_{\sigma \pi}(B_i) = (1 + \sigma \lambda_2)^{-1} \left[ 1 - \|B_i\|_2^{-1} \sigma \lambda_1 \right]_+ B_i.
\]  

(B.7)

### B.1.3 Proof of Proposition 3

We first prove (b), i.e. \( \tilde{Z} = \text{prox}_{\pi^*/\sigma}(B/\sigma - X^T \tilde{V}) \). If we compute the derivative of \( L_\sigma(Z | \tilde{V}, B) \) with respect to \( Z_i \) and we set it equal to 0, we obtain

\[
B_i/\sigma - (X^T) \tilde{V} - \tilde{Z}_i = \nabla \pi^*(\tilde{Z}_i)/\sigma
\]  

(B.8)

We now use the sub-gradient proximal operators characterization \cite{Correa1992}:

\[
u = \text{prox}_f(t) \text{ if and only if } t - u \in \partial f(u).
\]  

(B.9)
Considering $t = B_i / \sigma - \bar{V}^T X(i)$, $u = \bar{Z}_i$, and $f = \pi^* / \sigma$, the right hand side of (B.9) is true by (B.8). The left hand side of (B.9) gives us $\bar{Z}_i = \text{prox}_{\pi^*/\sigma}\left( B_i / \sigma - \bar{V}^T X(i) \right)$. To conclude the first part of the proof just note that $\bar{Z} = (\bar{Z}_1, \ldots, \bar{Z}_p)^T$.

For the second part of the proof, we need to find $\psi(V) := L_\sigma(V | \bar{Z}, B)$. First, note that by the Moreau decomposition (3.4) $\bar{Z} = B / \sigma - X^T V - (1/\sigma) \text{prox}_{\pi}(B - \sigma X^T V)$. Plugging this into (3.5), after some algebraic manipulations, we obtain

$$\psi(V) = h^*(V) + \pi^*(\bar{Z}) + \frac{1}{2\sigma} \sum_{i=1}^p \left\| \text{prox}_{\pi}(B_i - \sigma V^T X(i)) \right\|_2^2 - \frac{1}{2\sigma} \sum_{i=1}^p \| B_i \|_2^2 . \quad (B.10)$$

We now have to compute $\pi^*(\bar{Z})$. If we set $T = B - \sigma X^T V$, then $\pi^*(\bar{Z}) = \sum_{i=1}^p \pi^*(\text{prox}_{\pi^*/\sigma}(T_i / \sigma))$. In particular

$$\text{prox}_{\pi^*/\sigma}(T_i / \sigma) = T_i / \sigma - (1/\sigma) \text{prox}_{\pi}(T_i) = \begin{cases} (1 + \sigma \lambda_2)^{-1} \left( \lambda_2 + \| T_i \|_2^{-1} \lambda_1 \right) T_i & \| T_i \|_2 > \sigma \lambda_1 \\ T_i / \sigma & \text{o.w.} \end{cases} . \quad (B.11)$$

Composing (B.11) and (3.2), again after some algebraic manipulations, we get

$$\pi^*(\bar{Z}) = (\lambda_2 / 2) \sum_{i=1}^p \left\| \text{prox}_{\pi}(B_i - \sigma V^T X(i)) \right\|_2^2 . \quad (B.12)$$

Next, to prove (ii), note that $\hat{\partial}^2 \psi(V)$ is the $nk \times nk$ symmetric matrix

$$\begin{bmatrix} \frac{\partial^2 \psi}{\partial V_1 \partial V_1} & \cdots & \frac{\partial^2 \psi}{\partial V_1 \partial V_n} \\ \cdots & \ddots & \cdots \\ \frac{\partial^2 \psi}{\partial V_n \partial V_1} & \cdots & \frac{\partial^2 \psi}{\partial V_n \partial V_n} \end{bmatrix} .$$

88
In particular, each block here is the \( k \times k \) matrix
\[
\frac{\partial \psi}{\partial V_t \partial V_s} = \begin{cases} 
I_k + \sigma \sum_{i=1}^p X_{ti} \partial \prox_{\sigma \pi}(T_i) X_{si} & t = s \vspace{1em} \\
\sigma \sum_{i=1}^p X_{ti} \partial \prox_{\sigma \pi}(T_i) X_{si} & t \neq s \end{cases}.
\] (B.13)

Thus, we have
\[
\hat{\partial}^2 \psi(V) = I_{nk} + \sum_{i=1}^p \begin{bmatrix} X_{1i} \partial \prox_{\sigma \pi}(T_i) X_{1i} & \ldots & X_{1i} \partial \prox_{\sigma \pi}(T_i) X_{ni} \\
\vdots & \ddots & \vdots \\
X_{ni} \partial \prox_{\sigma \pi}(T_i) X_{1i} & \ldots & X_{ni} \partial \prox_{\sigma \pi}(T_i) X_{ni} \end{bmatrix} = \begin{bmatrix} I_{nk} + \hat{X} \partial \prox_{\sigma \pi}(T) \hat{X}^T \end{bmatrix}.
\] (B.14)

We now need to show that \( Q \in \partial \prox_{\sigma \pi}(T) \). Note that \( \partial \prox_{\sigma \pi}(T_i) \) is a \( pk \times pk \) block-diagonal matrix, since \( \frac{\partial \prox_{\sigma \pi}(T_i)}{\partial T_j} = 0 \) for \( i \neq j \). Let us focus on \( T_1 \), and let \( t_1, \ldots, t_k \) be its \( k \) elements. Then, \( (\partial \prox_{\sigma \pi}(T_1))_{ij} = \frac{\partial \prox_{\sigma \pi}(t_1)}{\partial t_j} \), for \( i, j = 1, \ldots, k \). Specifically, for \( \|T_i\|_2 \leq \sigma \lambda_1 \), it is straightforward to see that \( \partial \prox_{\sigma \pi}(T_1) = 0 \). For \( \|T_i\|_2 > \sigma \lambda_1 \), knowing that \( \frac{\partial \|T_i\|_2}{\partial t_i} = \|T_i\|_2^{\frac{1}{2}} - \frac{1}{2} \frac{\sigma \lambda_1}{\|T_i\|_2} t_i \), after some algebraic manipulations we obtain
\[
\frac{\partial \prox_{\sigma \pi}(t_i)}{\partial t_j} = (1 + \sigma \lambda_2)^{-1} \begin{cases} 1 - \sigma \lambda_1 \|T_i\|_2^{-1} + \|T_i\|_2^{-3} \sigma \lambda_1 t_i^2 & i = j \\
\|T_i\|_2^{-3} \sigma \lambda_1 t_i t_j & i \neq j \end{cases}.
\] (B.15)

(B.15) shows us that \( P_1 = \partial \prox_{\sigma \pi}(T_1) \). Without loss of generality, we can do the same way for \( T_2, \ldots, T_p \) and prove (iii). To conclude the proof of the theorem, we note that since \( Q \in \partial \prox_{\sigma \pi}(T) \), then \( I_{nk} + \sigma \hat{X} Q \hat{X}^T \in \hat{\partial}^2 \psi(V) \), and from Hiriart-Urruty et al. (1984) we have \( \partial^2 \psi(V) \text{vec}(D) = (I_{nk} + \sigma \hat{X} Q \hat{X}^T) \text{vec}(D) \), for every \( D \) in the domain of \( V \).

**B.2 Convergence Analysis**

**B.2.1 Inexact Augmented Lagrangian Method**

To state the global convergence of Algorithm 1 and the super-linear convergence of the solution \((V^k, Z^k, B^k)\), we refer to theorem 3.2 and theorem 3.3 in Li et al. (2018) – which are in turned based on the fundamental results presented in Rockafellar (1976a,b), and Luque (1984). Here, we just need verify that the theorems assumptions hold. In particular, we met the assumptions on \( h(\cdot) \), since it is the same for Lasso and Elastic...
Net, and it is always possible to implement the stopping criteria for the local convergence analysis described in Li et al. (2018) – Section 3. The main challenge is to verify that the operators $T_f$ and $T_l$ satisfy the error bound condition, since they are different from the Lasso case.

Given the closed proper convex function $f$ in the objective (3.1), and the convex-concave lagrangian function $l$ in (3.5), we define the maximal monotone operators $T_f$ and $T_l$ as in Rockafellar (1976a):

$$T_f(B) = \partial f(B), \quad T_l(V, Z, B) = \{(V', Z', B')|(V', Z', -X') \in \partial l(V, Z, X)\}. \quad (B.16)$$

We have to show that $T_f$ and $T_l$ are metric subregular (Dontchev and Rockafellar 2009), or equivalently that they satisfy the error bound condition (Robinson 1981), also called growth condition (Luque 1984).

In particular, we say that a multivalue mapping $F : B \Rightarrow V$ satisfies the error bound condition at $v \in V$ with modulus $\kappa > 0$ if $F^{-1}(v) \neq \emptyset$ and there exists $\epsilon > 0$ such that if $b \in B$ with $\text{dist}(v, F(b)) \leq \epsilon$, then

$$\text{dist}(b, F^{-1}(v)) \leq \kappa \text{dist}(v, F(b)). \quad (B.17)$$

The regularity of $T_f$ comes from Zhou and So (2017): since $\nabla h$ is Lipschitz continuous and $\pi$ has a polyhedral epigraph, $T_f$ satisfies the error bound condition. Verifying the bound condition for $T_l$ in the Lasso problem is not straightforward. However, in the Group Elastic Net case, we can use some known results given the special form of $\pi^*$ in (3.2). First, note that $\pi^*$ is a piecewise linear-quadratic function. Thus, we can apply Proposition 12.30 in Rockafellar and Wets (2009) and state that the subgradient mapping $\partial \pi^*$ is piecewise polyhedral. Finally, from Robinson (1981) we know that polyhedral multifunctions satisfy the error bound condition for any point $v \in V$. This proves the regularity of $T_l$ and, therefore, the super-linear convergence of the method.

### B.2.2 Semi-smooth Newton Method

Again, to state the super-linear convergence of the sequence $\{V_j\}$ produced by the Semi-smooth Newton Method in Algorithm 1, we can use theorem 3.6 in Li et al. (2018), which is based on the crucial results of Zhao et al. (2010). All the assumptions are easy to verify. $\nabla h^*$ and $\text{prox}_{\sigma p}$ in (3.3) are semi-smooth functions. By proposition 3.3 in Zhao et al. (2010), $D^j$ defined in (3.9) is a descent direction. Finally, recall that the matrix
\((I_{nk} + \sigma \hat{X}_J Q_J \hat{X}_J^T) \in \hat{\partial^2}\psi(V)\) and it is positive semidefinite.

### B.3 Additional Simulation Results

We ran all simulations on a MacBookPro with 3.3 GHz DualCore Intel Core i7 processor and 16GB ram. We reran all python simulations using openblas and mkl as blas systems, with threads=1,2 and openmp, with threads=1,4. In all scenarios, times match those reported in the paper that are obtained considering openblas with 2 threads and openmp with 4 threads. The following versions of sklearn and glmnet are used: scikit-learn==0.22.2 and glmnet==4.1

Table B.1: \(a, b\) and \(c\) report mean CPU time in seconds for fgen, sklearn and glmnet, respectively, over 20 replications of the same simulation scenari. In parenthesis we report standard errors. For each scenario we consider three values of \(c_\lambda\), which are held fixed over the replications.

<table>
<thead>
<tr>
<th>(\alpha = 0.8, \quad l = 0.25)</th>
<th>(n=1000)</th>
<th>(n=5000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(p; p_0)</td>
<td>(k)</td>
<td>(c_\lambda)</td>
</tr>
<tr>
<td>(2(10^4); 10)</td>
<td>5</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td><strong>0.3</strong> (0.00)</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td><strong>0.4</strong> (0.01)</td>
</tr>
</tbody>
</table>

Table B.2: \(a, b\) and \(c\) report CPU time in seconds for fgen, sklearn and glmnet, respectively. The full \(c_\lambda\) grid consists of 100 log-spaced points between 1 and 0.01. We truncate the path search when \(max\) active components are selected. \(runs\) is the corresponding number of explored \(c_\lambda\) values. We fix 1000 seconds as time limit.

<table>
<thead>
<tr>
<th>(\alpha = 0.8, \quad l = 0.25)</th>
<th>(n=500)</th>
<th>(n=1000)</th>
<th>(n=5000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(p; p_0)</td>
<td>(k)</td>
<td>max</td>
<td>runs</td>
</tr>
<tr>
<td>(10^5; 10^2)</td>
<td>5</td>
<td>7</td>
<td><strong>1.6</strong> 26.2 3.2</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>12</td>
<td><strong>2.8</strong> 45.4 5.2</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>16</td>
<td><strong>5.1</strong> 64.7 6.3</td>
</tr>
</tbody>
</table>
Figure C.1: **Unshifted mortality curves.** MAX, ISTAT and DPC mortality curves (per 100,000 inhabitants) without shift. Vertical lines show the days corresponding to the national lock down (March 9) and the suspension of all non-essential production activities (March 23).
Figure C.2: Mortality curves. (a): DPC mortality curves (per 100,000 inhabitants) in the 20 Italian regions – before (top) and after (bottom) the shifts produced by probKMA run with $K = 2$. (b): ISTAT mortality curves (per 100,000 inhabitants) in the 20 Italian regions – before (top) and after (bottom) the shifts produced by probKMA run with $K = 2$. In all panels, vertical lines mark the dates of the national lock-down (March 9) and the suspension of all non-essential production activities (March 23). In the bottom panels, vertical lines still show these dates without shifts, stars on the curves mark the lock-down after the region-specific shifts.
Figure C.3: **Characterizing two epidemics.** Results of probKMA and IWTomics on (a)-(b) DPC mortality curves and (c)-(d) ISTAT curves. (a) and (c): Mortality curves are shown in the top left panel with portions identified by probKMA with $K = 2$ in red (Group 1; “exponential” pattern) and blue (Group 2; “flat(tened)” pattern). The curve portions are shown again, this time aligned with each other and separated by group, in the bottom panels. Black lines indicate group averages. The shifts produced by probKMA are shown in the top right panel. (b) and (d): Shifted Group 1 and Group 2 mortality curves are tested against each other with IWTomics. The heatmap at the top shows $p$-values adjusted at all possible scales (from 1 to 65 days). The middle panel shows in detail the top-most row of the heatmap; i.e. the $p$-values adjusted across the whole 65-day interval (the 65 different values are reported in Table C.1). The bottom panel shows again the shifted curves. Gray areas in the middle and bottom panels mark days when the difference between the two groups is significant (adjusted $p$-value < 5%).
Table C.1: **IWTomics adjusted p-value curves.** Values of the IWTomics adjusted p-values for each time point and each of the three dataset. Note that the smallest possible p-value is 0.001 since we are employing 1000 permutations in the IWTomics test.

<table>
<thead>
<tr>
<th>MAX</th>
<th>ISTAT</th>
<th>DPC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.648</td>
<td>0.996</td>
</tr>
<tr>
<td>2</td>
<td>0.651</td>
<td>0.998</td>
</tr>
<tr>
<td>3</td>
<td>0.648</td>
<td>0.998</td>
</tr>
<tr>
<td>4</td>
<td>0.634</td>
<td>0.996</td>
</tr>
<tr>
<td>5</td>
<td>0.625</td>
<td>0.988</td>
</tr>
<tr>
<td>6</td>
<td>0.608</td>
<td>0.967</td>
</tr>
<tr>
<td>7</td>
<td>0.581</td>
<td>0.939</td>
</tr>
<tr>
<td>8</td>
<td>0.538</td>
<td>0.878</td>
</tr>
<tr>
<td>9</td>
<td>0.513</td>
<td>0.808</td>
</tr>
<tr>
<td>10</td>
<td>0.462</td>
<td>0.711</td>
</tr>
<tr>
<td>11</td>
<td>0.420</td>
<td>0.583</td>
</tr>
<tr>
<td>12</td>
<td>0.381</td>
<td>0.454</td>
</tr>
<tr>
<td>13</td>
<td>0.335</td>
<td>0.319</td>
</tr>
<tr>
<td>14</td>
<td>0.275</td>
<td>0.211</td>
</tr>
<tr>
<td>15</td>
<td>0.210</td>
<td>0.114</td>
</tr>
<tr>
<td>16</td>
<td>0.143</td>
<td>0.048</td>
</tr>
<tr>
<td>17</td>
<td>0.073</td>
<td>0.019</td>
</tr>
<tr>
<td>18</td>
<td>0.022</td>
<td>0.004</td>
</tr>
<tr>
<td>19</td>
<td>0.001</td>
<td>0.002</td>
</tr>
<tr>
<td>20</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>21</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>22</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>23</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>24</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>25</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>26</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>27</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>28</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>29</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>30</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>31</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>32</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>33</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>34</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>35</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>36</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>37</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>38</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>39</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>40</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>41</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>42</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>43</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>44</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>45</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>46</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>47</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>48</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>49</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>50</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>51</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>52</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>53</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>54</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>55</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>56</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>57</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>58</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>59</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>60</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>61</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>62</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>63</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>64</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>65</td>
<td>0.001</td>
<td>0.001</td>
</tr>
</tbody>
</table>
Figure C.4: Characterizing three epidemics. Results of *probKMA* with *K* = 3 on (a) MAX mortality curves, (b) DPC mortality curves and (c) ISTAT curves. Mortality curves are shown in the top left panel with portions identified by *probKMA* with *K* = 3 in red (Group 1; “exponential” pattern), blue (Group 2; “flat(tened)” pattern) and green (Group 3; “extreme” pattern). The curve portions are shown again, this time aligned with each other and separated by group, in the bottom panels. Black lines indicate the average curves of the group. The shifts produced by *probKMA* are shown in the top right panel.
Figure C.5: Unshifted mobility and positivity curves. Local mobility and positivity curves without shift in the 20 Italian regions. Vertical lines show the days corresponding to the national lock down (March 9) and the suspension of all non-essential production activities (March 23).

Figure C.6: Shifted curves for ISTAT and DPC. Mortality (per 100,000 inhabitants), mobility, and positivity curves after the shifts produced by $probKMA$ with $K=2$. Vertical lines mark the dates of the national lock-down (March 9) and the suspension of all non-essential production activities (March 23) without shifts. Stars on the curves mark the lock-down after the region-specific shifts.
Figure C.7: Lags characterization. Projections on the first Principal Components of mortality curves (MAX, ISTAT, and DPC – upper panels), positivity curves (lower-left panel) and mobility curves (lower-mid panel). Here the curves are not shifted, so the horizontal axis represents calendar dates. Projecting the curves on their first Principal Components produces an additional de-noising, which inherently aligns all the peaks. The thick dark lines are average curves in each panel, and the vertical lines mark the date of their maxima – also expressed in number of days from February 16th. For mobility, we also mark the beginning of the national lockdown (March 9th; 23 days from February 16th), which corresponds to a severe drop in the curves. We notice lags of about 20 days and 1 week between, respectively, the peak and the stark drop in mobility, and the peak in positivity (on average across all regions). The lag between the peak in positivity and the peak in mortality is about 10 days (again on average across all regions). The lower-right panel shows the individual mobility-positivity and positivity-mortality lags for each region in group 1 (those characterized by an “exponential” epidemic pattern). Albeit with some variation across regions, adding up these lags produces an overall delay of approximately 1 month between the peak mobility and that of mortality.
Table C.2: Functional regression models (in-sample) $R^2$, LOO-CV $R^2$ and partial $R^2$'s. For each functional linear model which regresses mortality on the covariates listed in the first column, the table reports the (in-sample) $R^2$, the LOO-CV $R^2$ and the partial $R^2$'s.

<table>
<thead>
<tr>
<th>covariates</th>
<th>MAX</th>
<th>ISTAT</th>
<th>DPC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R^2$</td>
<td>LOO-CV $R^2$</td>
<td>partial $R^2$</td>
</tr>
<tr>
<td>mob</td>
<td>0.79</td>
<td>0.54</td>
<td>-</td>
</tr>
<tr>
<td>pos</td>
<td>0.75</td>
<td>0.47</td>
<td>-</td>
</tr>
<tr>
<td>mob + pos</td>
<td>0.90</td>
<td>0.52</td>
<td>mob: 0.62 pos: 0.53</td>
</tr>
<tr>
<td>mob + pos + pc1</td>
<td>0.94</td>
<td>0.70</td>
<td>mob: 0.66 pos: 0.61 pc1: 0.39</td>
</tr>
</tbody>
</table>

Figure C.8: MAX mortality residuals. (a): residuals of the function-on-function regression of MAX mortality on local mobility and positivity. (b): residuals of the function-on-function regression of MAX mortality on local mobility, positivity, and the first principal component of the top 5 covariates. In both panels curves from Group 1 are in red, and curves from Group 2 are in blue. Residuals with positive signs indicate regions for which the true mortality curve is above the estimated mortality curve. Conversely, residuals with negative signs indicate regions for which the true mortality curve is below the estimated mortality curve.
Figure C.9: **Associating mortality to mobility and positivity - ISTAT and DPC.** Results from the joint function-on-function regression of ISTAT and DPC mortality on mobility and positivity. The top row shows the estimated effect surfaces (the March 9 date is marked) with respective partial $R^2$ (for in-sample $R^2$ and LOO-CV $R^2$ see Table C.2). The bottom row shows the regression residuals (for barplots interpretation see Figure C.8).
Figure C.10: **Associating mortality to mobility.** Results from the function-on-function regression of mortality on local mobility. The top row displays the estimated effect surface (the March 9 date is marked) with respective in-sample $R^2$ (for LOO-CV $R^2$ see Table C.2). The bottom row displays the regression residuals (for barplots interpretation see Figure C.8).
Figure C.11: **Associating mortality to positivity.** Results from the function-on-function regression of mortality on positivity. The top row displays the estimated effect surface (the March 9 date is marked) with respective in-sample $R^2$ (for LOO-CV $R^2$ see Table C.2). The bottom row displays the regression residuals (for barplots interpretation see Figure C.8).

Figure C.12: **Correlations between all covariates.** Histogram showing the distribution of pair-wise correlations for all the 68 covariates in Table C.3. 24% of the correlations exceed 0.5 in absolute value. Note: Variance Inflation Factors cannot be computed for the whole set of 68 variables since the sample size here is only $n = 20$ (the Italian regions).
Table C.3: **Covariates.** List of all scalar covariates considered.

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Year and Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resident population, units</td>
<td>2018, ISTAT</td>
</tr>
<tr>
<td>Land area, hectares</td>
<td>2018, ISTAT</td>
</tr>
<tr>
<td>% population over 65</td>
<td>2018, ISTAT</td>
</tr>
<tr>
<td>% population over 70</td>
<td>2018, ISTAT</td>
</tr>
<tr>
<td>% population over 80</td>
<td>2018, ISTAT</td>
</tr>
<tr>
<td>% population over 85</td>
<td>2018, ISTAT</td>
</tr>
<tr>
<td>% male over 18</td>
<td>2018, ISTAT</td>
</tr>
<tr>
<td>% female over 18</td>
<td>2018, ISTAT</td>
</tr>
<tr>
<td>Employees in large supermarket chains, units</td>
<td>2018, ISTAT</td>
</tr>
<tr>
<td>Department stores, units</td>
<td>2018, ISTAT</td>
</tr>
<tr>
<td>Supermarkets, units</td>
<td>2018, ISTAT</td>
</tr>
<tr>
<td>Hypermarkets, units</td>
<td>2018, ISTAT</td>
</tr>
<tr>
<td>Airports, units</td>
<td>2018, ISTAT</td>
</tr>
<tr>
<td>Landed and departed passengers in airports, units</td>
<td>2018, ISTAT</td>
</tr>
<tr>
<td>Landed and departed airplanes in international flights, units</td>
<td>2018, ISTAT</td>
</tr>
<tr>
<td>Healthcare institutes (private and public), units</td>
<td>2018, Ministry of Health</td>
</tr>
<tr>
<td>Public healthcare institutes, units</td>
<td>2018, Ministry of Health</td>
</tr>
<tr>
<td>Days of stay in public and private healthcare institutes</td>
<td>2015, ISTAT</td>
</tr>
<tr>
<td>Days of stay in public healthcare institutes</td>
<td>2015, ISTAT</td>
</tr>
<tr>
<td>Patients in public and private institutes, units</td>
<td>2015, ISTAT</td>
</tr>
<tr>
<td>Patients in public institutes (except for residual psychiatric institutes), units</td>
<td>2015, ISTAT</td>
</tr>
<tr>
<td>Beds in pneumatology in public and private healthcare institutes</td>
<td>2015, ISTAT</td>
</tr>
<tr>
<td>Mechanical lung ventilators, units</td>
<td>2018, Ministry of Health</td>
</tr>
<tr>
<td>People with at least 1 chronic disease, units</td>
<td>2017, Ministry of Health</td>
</tr>
<tr>
<td>% people with diabetes</td>
<td>2017, Ministry of Health</td>
</tr>
<tr>
<td>% people with hypertension</td>
<td>2017, Ministry of Health</td>
</tr>
<tr>
<td>% people with bronchitis</td>
<td>2017, Ministry of Health</td>
</tr>
<tr>
<td>% people with osteoporosis</td>
<td>2017, Ministry of Health</td>
</tr>
<tr>
<td>% people with arthritis</td>
<td>2017, Ministry of Health</td>
</tr>
<tr>
<td>% people with allergy</td>
<td>2017, Ministry of Health</td>
</tr>
<tr>
<td>% people with ulcer</td>
<td>2017, Ministry of Health</td>
</tr>
<tr>
<td>Old-age index</td>
<td>2017, ISTAT</td>
</tr>
<tr>
<td>Life expectancy at birth (female)</td>
<td>2017, ISTAT</td>
</tr>
<tr>
<td>Life expectancy at birth (male)</td>
<td>2017, ISTAT</td>
</tr>
<tr>
<td>Active buses per 1000 inhabitants</td>
<td>2016, ISTAT</td>
</tr>
<tr>
<td>% children going to school with public transportation</td>
<td>2016, ISTAT</td>
</tr>
<tr>
<td>Public expenditure in healthcare per capita</td>
<td>2016, ISTAT</td>
</tr>
<tr>
<td>Factor risk: alcohol</td>
<td>2016, ISTAT</td>
</tr>
<tr>
<td>Factor risk: smoke</td>
<td>2016, ISTAT</td>
</tr>
<tr>
<td>Average household income</td>
<td>2015, ISTAT</td>
</tr>
<tr>
<td>Mobility index (commuting due to work)</td>
<td>2011, ISTAT</td>
</tr>
<tr>
<td>Self-containment index</td>
<td>2011, ISTAT</td>
</tr>
<tr>
<td>Public mobility index</td>
<td>2011, ISTAT</td>
</tr>
<tr>
<td>PM10</td>
<td>2017, ISTAT</td>
</tr>
<tr>
<td>PM2.5</td>
<td>2017, ISTAT</td>
</tr>
<tr>
<td>ICU beds per 100K inhabitants</td>
<td>2018, Ministry of Health</td>
</tr>
<tr>
<td>Beds in pneumatology</td>
<td>2018, Ministry of Health</td>
</tr>
<tr>
<td>Additional beds in ICU on April, 10th 2020</td>
<td>2020, DPC</td>
</tr>
<tr>
<td>Weighted PM10</td>
<td>2018, ISTAT</td>
</tr>
<tr>
<td>Average students per classroom</td>
<td>2018, Ministry of Education</td>
</tr>
<tr>
<td>Average students per school</td>
<td>2018, Ministry of Education</td>
</tr>
<tr>
<td>Gini index for schools</td>
<td>2018, Ministry of Education</td>
</tr>
<tr>
<td>Average beds per nursing home (ward)</td>
<td>2018, Ministry of Health</td>
</tr>
<tr>
<td>Average beds per nursing home (whole)</td>
<td>2018, Ministry of Health</td>
</tr>
<tr>
<td>Gini index for nursing homes</td>
<td>2018, Ministry of Health</td>
</tr>
<tr>
<td>Average beds per hospital (ward)</td>
<td>2018, Ministry of Health</td>
</tr>
<tr>
<td>Average beds per hospital (whole)</td>
<td>2018, Ministry of Health</td>
</tr>
<tr>
<td>Gini index for hospitals</td>
<td>2018, Ministry of Health</td>
</tr>
<tr>
<td>Average number of employees</td>
<td>2017, ISTAT</td>
</tr>
<tr>
<td>Gini index for firms</td>
<td>2017, ISTAT</td>
</tr>
<tr>
<td>Total number of tests between February 25th and April, 30th 2020</td>
<td>2020, DPC</td>
</tr>
<tr>
<td>Total number of tests between February 25th and March, 23rd 2020</td>
<td>2020, DPC</td>
</tr>
<tr>
<td>Total number of tests between March, 23rd and April, 30th 2020</td>
<td>2020, DPC</td>
</tr>
<tr>
<td>Adults per family doctor</td>
<td>2017, Ministry of Health</td>
</tr>
<tr>
<td>Average members per family</td>
<td>2018, ASR Lombardia</td>
</tr>
<tr>
<td>Public transport rides per capita</td>
<td>2017, ISTAT</td>
</tr>
</tbody>
</table>
Figure C.13: **Correlation-based dendrogram for all covariates.** Dendrogram of all the covariates in Table C.3, based on the correlation distance $d(x_1, x_2) = 1 - |\text{corr}(x_1, x_2)|$ and complete linkage. The 12 selected covariates are shown in red, while covariates dating 2016 or earlier are in grey.
Figure C.14: **Associations between the 12 selected covariates.** Exploratory matrix containing scatterplots with loess regressions, marginal densities, and correlation between pairs of covariates.
Table C.4: **Variance Inflation Factors of covariates.** Variance Inflation Factors (VIF) for the 12 scalar covariates used in the main analysis.

<table>
<thead>
<tr>
<th>Covariate</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>%65+</td>
<td>11.315684</td>
</tr>
<tr>
<td>%Dbts</td>
<td>12.105778</td>
</tr>
<tr>
<td>%Allrgs</td>
<td>3.933976</td>
</tr>
<tr>
<td>Adlts/doct</td>
<td>4.245479</td>
</tr>
<tr>
<td>ICUBds/cpt</td>
<td>2.892715</td>
</tr>
<tr>
<td>AvBds/hspt</td>
<td>5.012343</td>
</tr>
<tr>
<td>AvBds/nrsng</td>
<td>2.428207</td>
</tr>
<tr>
<td>AvStdns/clrm</td>
<td>6.529208</td>
</tr>
<tr>
<td>AvEmpls/firm</td>
<td>7.854636</td>
</tr>
<tr>
<td>AvMbrs/hshld</td>
<td>6.223056</td>
</tr>
<tr>
<td>PubTrsp/cpt</td>
<td>4.293915</td>
</tr>
<tr>
<td>PM10</td>
<td>12.858811</td>
</tr>
</tbody>
</table>

Table C.5: **Function-on-scalar feature selection.** Top five scalar covariates selected by *fgen* considering as response the MAX, ISTAT, and DPC mortality curves.

<table>
<thead>
<tr>
<th></th>
<th>MAX</th>
<th>ISTAT</th>
<th>DPC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Adults per family doctor</td>
<td>Adults per family doctor</td>
<td>Adults per family doctor</td>
</tr>
<tr>
<td>2</td>
<td>Ave. beds per hospital (whole)</td>
<td>Ave. students per classroom</td>
<td>Ave. beds per hospital (whole)</td>
</tr>
<tr>
<td>3</td>
<td>Ave. students per classroom</td>
<td>Ave. beds per hospital (whole)</td>
<td>Ave. students per classroom</td>
</tr>
<tr>
<td>4</td>
<td>Ave. members per household</td>
<td>Ave. employees per firm</td>
<td>Ave. employees per firm</td>
</tr>
<tr>
<td>5</td>
<td>Ave. employees per firm</td>
<td>Ave. members per household</td>
<td>Ave. members per household</td>
</tr>
</tbody>
</table>
Figure C.15: **Marginal function-on-scalar regressions.** Results for marginal function-on-scalar regressions. Mortality curves are regressed against each of the scalar covariates in Table 1. The top-row displays the signs of the effect curves estimated when just one intercept is included in the model. The bottom-row displays the signs of the effect curves estimated when we consider two different intercepts for curves in Group 1 and curves in Group 2. Time is on the vertical axis (the national lockdown on March 9, without shift, is marked by a horizontal line). Red, blue and green indicate, respectively, positive, negative, and non-significant portions (i.e., where 95% confidence bands around the estimated effect curve are entirely above, entirely below, or contain 0).
Figure C.16: **Function-on-function regression of mortality on mobility, positivity, and a control scalar covariate.** Each row shows some results from the joint function-on-function regression mortality on local mobility, positivity, and one of the top 5 covariates selected by $\text{fgen}$, used as control. In particular, we display the estimated effect surfaces for mobility and positivity (the March 9 date, without shift, is marked) with their respective partial $R^2$s. The scalar control covariates associated with each row are the following: 1: Adults per family doctor, 2: Ave. beds per hospital (whole), 3: Ave. students per classroom, 4: Ave. employees per firm, 5: Ave. members per household.
mortality $\sim$ mobility + positivity + reduced PC1

ISTAT

DPC

For interpreting the regression residuals see Figure C.8.

Figure C.17: Associating mortality to mobility, positivity and first principal component - ISTAT and DPC. Results from the joint function-on-function regression of ISTAT and DPC mortality on mobility, positivity, and the first principal component (pc1) of the top 5 covariates, used as a “summary” control. Each panel shows the estimated effect surfaces for mobility and positivity and the estimated effect curve for pc1 with respective partial $R^2$ (for in-sample $R^2$ and LOO-CV $R^2$ see Table C.2). For interpreting the regression residuals see Figure C.8.


Vita
Tobia Boschi

Education

- 2016-2022  PhD  Statistics  Penn State University
- 2014-2016  MSc  Mathematical Engineering  Politecnico di Milano
- 2010-2014  BSc  Mathematical Engineering  Politecnico di Milano

Publications

- Functional data analysis characterizes the shapes of the first COVID-19 epidemic wave in Italy  T. Boschi, J. Di Iorio, L. Testa, M. A. Cremona, F. Chiaromonte (2021)  Scientific Reports
- Covariance Based Low-dimensional Registration for Function-on-function Regression  T. Boschi, F. Chiaromonte, P. Secchi, B. Li (2021)  STAT
- The relationship between human mobility and viral transmissibility during the COVID-19 epidemics in Italy  P. Cintia, L. Pappalardo, S. Rinzivillo, D. Fadda, T. Boschi et others  preprint

Awards

- 2021 Departmental Award for Support of Undergraduate Education