Functional Data Analysis & Variable Selection

Nedret Billor

Auburn University
Department of Mathematics and Statistics

Universidad Nacional de Colombia
Medellin, Colombia
March 14, 2016

Functional Data Analysis

Data Types

- Univariate - Contains numbers as its observations (1D)
  X_n random variable.
- Multivariate - Contains vectors as its observations (pD)
  $\mathbf{X}_n$ random vector.
- Functional - Contains vectors of infinite dimensions as its observations (∞D):
  $\mathbf{X}_n(t), t \in [a, b]$ functions.

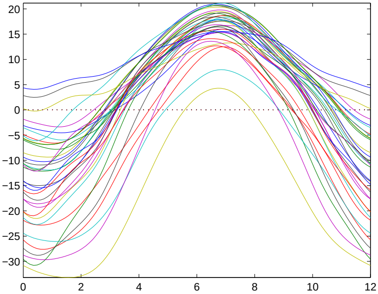


Figure: Functional Data (Ramsay and Silverman, 2005)

Functional Data Analysis (FDA)

FDA : collection of different methods in statistical analysis for analyzing curves or functional data.

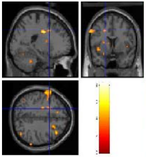
In standard statistical analysis, the focus is:

- on the set of data vectors (univariate, multivariate).

In FDA, the focus is

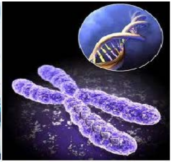
- on the type of data structure such as *curves*, *shapes*, *images*, or *set of functional observations*.

Fields using Functional Data Analysis



Brain Imaging Studies

Chemometrics



Genetics

e-Commerce

Biometrics

Computer Science

Goals

Essentially the same as those of any other branch of statistics:

- to represent the data in ways that aid further analysis,
- to display the data so as to highlight various characteristics,
- to study important sources of pattern and variation among the data,
- to explain variation in an outcome or dependent variable by using input or independent variable information,
- to compare two or more sets of data with respect to certain types of variation....

What are Functional Data about?

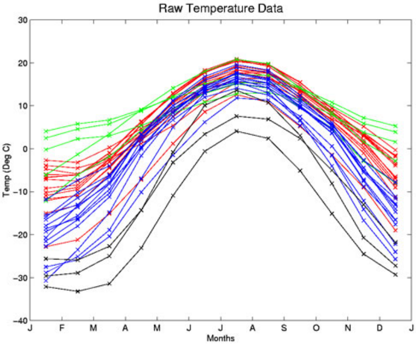
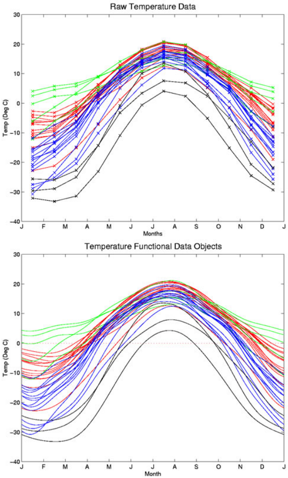


Figure: Canadian Weather Data (Ramsay and Silverman, 2005)

- 'x': Mean temperatures recorded by a weather station for the entire month, collected over 30 years.
- Colors: Geographic climates of the stations.

  Atlantic (red), Continental (blue) Pacific (green), Arctic (black).

What are Functional Data about?

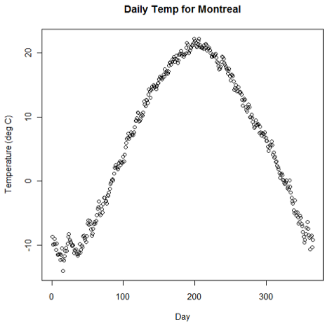


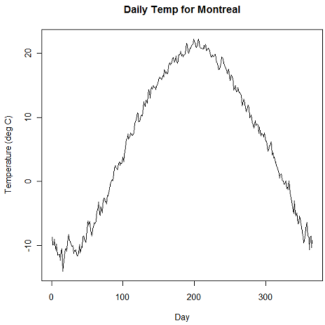
- In FDA we think of the observed data functions as single entities.
- Term *functional* refers to the intrinsic structure of the observed data.
- Functional data are usually observed and recorded discretely as p pairs (t_i, y_i).
- y_i is the 'snapshot' of the function at time t_i.
- The underlying function is assumed to be smooth.

Some Issues to Consider

- Estimation of functional data from noisy discrete observations.
- Estimation of functional data from sparsely sampled observations.
- Numerical representation of infinite-dimensional objects.
- Number of 'predictors', p $>>$ n, the number of observations.

- Estimation of functional data from noisy discrete observations.





How can we represent the temperature pattern of a Canadian city over the entire year?

Discrete to Functional Form

Discrete Form: p pairs of (t_j, y_j).

Functional Form: $y_j = x(t_j) + \varepsilon_j$

- If discrete values are error-less, some interpolation method is used.
- If there is some observational error, some smoothing method is used.

Representing Functional Data by basis functions

- Represent functions:

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t)$$

- Chosen basis system $\phi(t)$ should have features characteristic of the observed data.
- Fourier basis for periodic data, B-spline basis for non-periodic data.
- Interpolation is achieved when $K = p$.
- The degree to which y_j is smoothed is determined by K.

- Two main objectives in function estimation:
  (1) Good fit to data by minimizing $\sum[y_j - x(t_j)]^2$
  (2) Fit should not be too good so that $x(t)$ is locally variable.

- These competing aim correspond to this basic principle,

$$\text{Mean squared error} = \text{Bias}^2 + \text{Sampling variance}$$

where

$$\text{Bias}[\hat{x}(t)] = x(t) - E[\hat{x}(t)]$$
$$\text{Var}[\hat{x}(t)] = E[\{\hat{x}(t) - E[\hat{x}(t)]\}^2]$$

$$\text{Model: } y_j = \sum_{k=1}^{K} c_k \phi_k(t_j) + \varepsilon_j, \; j = 1, \ldots, p$$

- Popular measure to quantify the notion of "roughness" of a function is **curvature**,

$$PEN_2(x) = \int [D^2 x(s)]^2 ds$$

- The penalized residual sum of squares,

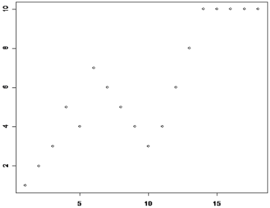
$$PENSSE_m(\mathbf{y}|c) = (\mathbf{y} - \mathbf{\Phi c})^{'} \mathbf{W} (\mathbf{y} - \mathbf{\Phi c}) + \lambda PEN_2(x)$$

where $\mathbf{W}$ is a weight matrix and λ is a smoothing parameter.

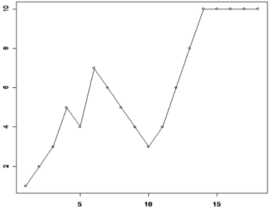
$$\hat{\mathbf{c}} = (\mathbf{\Phi}^{'} \mathbf{W} \mathbf{\Phi} + \lambda \mathbf{R})^{-1} \mathbf{\Phi}^{'} \mathbf{W} \mathbf{y}$$

where $\mathbf{R} = \int D^m \phi(s) D^m \phi^{'}(s) ds$.

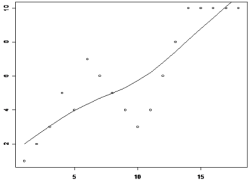
Illustration



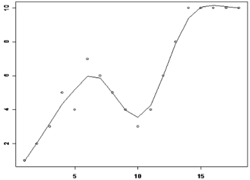
(a) Data Points



(b) Good Fit



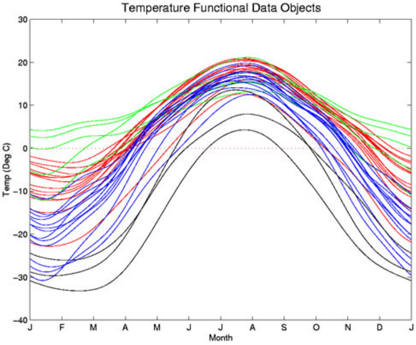
(c) Large λ



(d) Small λ

Statisticians' Questions?

- How can I summarize the patterns?
- Do the summary statistics "mean" and "covariance" have any meaning when I'm dealing with curves?



- Now that we have functional estimated curves of our observed data, we'd like to summarize the estimated temperature curves.

- Climatologists can then use these summaries to talk about typical weather patterns and about variability in these patterns over time and across Canada.

Sample Mean function:

$$\hat{\mu}(t) = \bar{x} = n^{-}1 \sum_{i=1}^{n} x_i(t)$$

Sample variance function:

$$var_x(t) = (n-1)^{-1} \sum_{i=1}^{n} (x_i(t) - \bar{x})^2$$

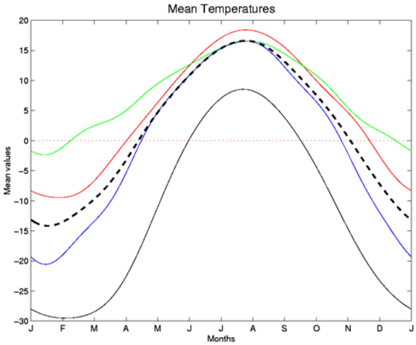
Sample covariance function:

$$\hat{c}(t_1, t_2) = (n-1)^{-1} \sum_{i=1}^{n} (x_i(t_1) - \bar{x}(t_1))(x_i(t_2) - \bar{x}(t_2))$$

Sample correlation function:

$$corr_x(t_1, t_2) = \frac{\hat{c}(t_1, t_2)}{\sqrt{var_x(t_1) var_x(t_2)}}$$

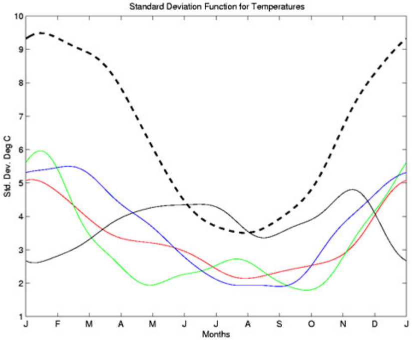
Sample Mean Function



Atlantic stations in red,
Continental in blue, Pacific in green,
and Arctic in black.

- Calculated by averaging the functions pointwise across the replications.
- The mean curves show the distinctive patterns of the four climates.
- Pacific cities are much warmer than the rest of Canada in the winter and spring but have fairly typical summer temperatures.
- The Arctic stations, on the other hand, have temperatures cooler than the average throughout the year.

Sample Standard deviation function



- Simple analogue of the classical standard deviation has similar interpretations.
- SD function suggests:
  the winter months have the greatest variability in recorded temperatures across Canada - approximately 9 degrees Celsius, as compared to the summer months with a standard deviation of about 4 degrees Celsius.

Statisticians' Questions?

Question: What **modes** of variation can we can find in the data?

- How can I determine the primary modes of variation in the data?
- How many typical modes can summarize these thirty-five curves?

Most sets of data display a small number of dominant or substantial modes of variation!

principal components analysis to functional data.

Functional PCA

- $\{x(t), t \in T\}$: a stochastic process where T is some index set which is a bounded interval on $\Re$.
- The principal component scores corresponding to weight γ is generalized to an integral form,

$$Z_i = \int \gamma_j(t) x_i(t) dt.$$

- The weight function $\gamma_j(t)$ is obtained by solving

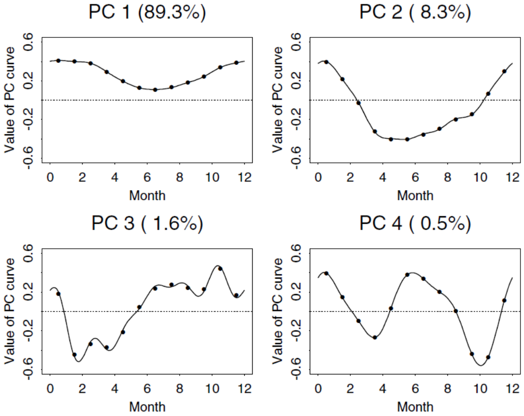
$$\max_{<\gamma_\mathbf{j},\, \gamma_\mathbf{m}>=\mathcal{I}(\mathbf{j}=\mathbf{m}),\, \mathbf{j}\leq\mathbf{m}} N^{-1} \sum \left(\int \gamma_j x_i\right)^2$$

  or equivalent to solving the functional eigenequation

$$\int \psi(s,t)\gamma(t)dt = \lambda\gamma(s) \quad \gamma \in L^2,$$

  where ψ is the covariance function of the $x(t)$.

PCs for Weather Data



First four PC curves estimated from the basis approximation

Lasso-based Variable Selection Methods for Functional Regression Model

What are the Most Important Problems in Statistics?

Classification and **Regression** problems with large numbers of candidate predictor variables occur in a wide variety of scientific fields.
In 1996, Tim Hesterberg asked Brad Efron:

"What are the most important problems in statistics?"

A **single** problem: **Variable selection in regression**.

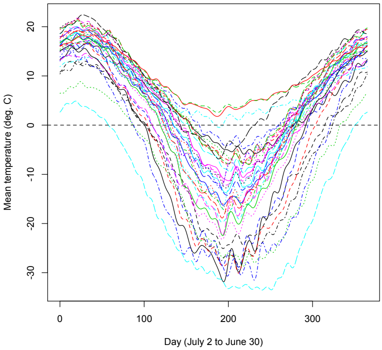
- Hard to argue with this assessment!
- This answer reflects the importance of variable selection in practice since Efron's work has long been strongly grounded in **solving real problems**.

Goals in model selection

- *accurate predictions,*
- *interpretable models—determining which predictors are meaningful,*
- *stability—small changes in the data should not result in large changes in either the subset of predictors used, the associated coefficients, or the predictions, and*
- *avoiding bias in hypothesis tests during or after variable selection*

Functional Regression Model with Functional Predictors and a Scalar Quantitative Response

Simple Functional Regression Model: Canadian Weather Data



Does the total amount of precipitation depend on specific features of the temperature profile of a weather station?

Assumed Model:

$$Y_i = \alpha + \int_{\mathcal{T}} X_i(t)\beta(t)dt + \epsilon_i.$$

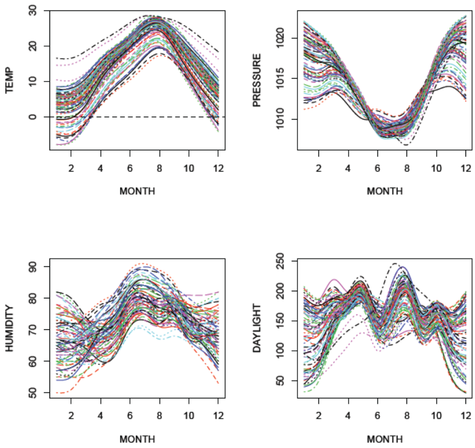
$i = 1, \ldots, 35.$

- Y : scalar response (amount of precipitation).
- $X(t)$: temperature functional predictor.

Simple Functional Regression Model

- Y : scalar response.
- $X(t)$ (functional predictor): squared integrable random function.
- $X_i(t)$ assumed to be $E(X_i(t)) = 0$ and observed without measurement error at a grid of time points.
- $\epsilon_i \sim N(0, \sigma^2)$.
- α: a scalar parameter.
- $\beta(t)$: smooth and squared integrable parameter function.

Multiple Functional Regression Model: Japanese Weather Data



Data from Chronological Scientific Tables 2005 (Matsui and Konishi, 2011)

Assumed Model:

$$Y_i = \alpha + \Sigma_{j=1}^{p} \int_{\mathcal{T}_l} X_{ij}(t)\beta_j(t)dt + \epsilon_i.$$

$i = 1, \ldots, 79.$

Four functional predictors ($p = 4$):

- *Monthly observed average temperatures,*
- *Average atmospheric pressure,*
- *Average humidity,*
- *Time of daylight,*
- *Annual total precipitation (scalar).*

Research Question: Do we need all functional predictors?

Because functional coefficients (β_j) are more complicated objects than scalar coefficients in classical multiple linear regression,

Generally desirable to identify those significant variables in predicting the responses, even if p is small!

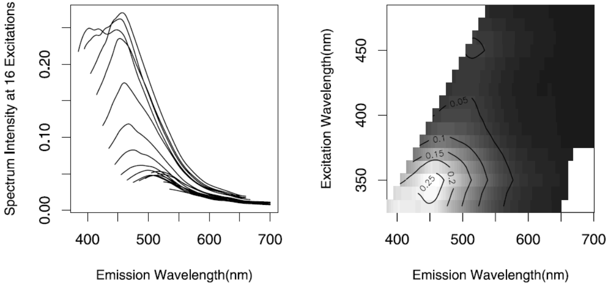
Aim: *Select the functional predictor variables that contribute the most for the prediction of annual total precipitation.*

Some major approaches:

- **Traditional approach**(Ramsay and Silverman, 2005): *represent functional data by an expansion with respect to a certain basis, and subsequent inferences are carried out on the coefficients.*
- **The French school** (Ferraty and Vieu, 2006): *take a nonparametric point of view, extends the traditional nonparametric techniques, most notably the kernel estimate, to the functional case.*
- **Other methods**: *such as put functional regression in the reproducing kernel Hilbert space framework has been developed* (Preda, 2007; Lian, 2007).

Functional Regression Model with Functional / Non-functional Predictors and Binary Response

Objective: *to discriminate the diseased observations from normal based on the high dimensional functional data – the fluorescence spectral measurements*.



Fluorescence spectroscopy data

- Functional predictors: 717 EEM measurements (each contains 16 curves)
- Non-functional predictors : associated with the measurements which may cause systematic difference in spectra, such as tissue type of the measurement site (two levels), or the menopausal status of patients (three levels).

on n=306 patients.

Basic concern: *When there are multiple functions per observation,* **(a) how do we perform a curve selection to select few important curves** & **(b)perform classification based on the selected curves***?*

Classification with functional data: **ALSO** a challenging problem due to the high dimensionality of the observation space.

Variable Selection Methods for Standard Multiple Linear Regression Model

There are

- Traditional Approaches
- Modern Methods (Regularization Methods)

More recently, regularization methods have received much attention for **standard linear regression**:

- LASSO (Least Absolute Shrinkage and Selection Operator),
- SCAD (Smoothly clipped absolute deviation) penalty,
- Adaptive LASSO...

Variable Selection Via Regularization Methods for Functional Regression Model

Variable selection is an important problem in functional regression analysis.

- $\beta(t)$'s are more complicated objects than scalar coefficients in classical multiple linear regression.
- identify those significant $X(t)$ predictors in predicting Y, **even if p is small!**
- Multiple parameters exist for a functional predictor, therefore **group structure based techniques**. These are:
  1. functional group LASSO
  2. functional group SCAD

Functional Regression Model

Suppose that we have n observations

$$\{(X_{ij}(t), Y_i); t \in \Im, i = 1, \ldots, N,\ j = 1, \ldots, p\}.$$

Functional Regression Model:

$$Y_i = \alpha + \Sigma_{j=1}^{p} \int_{\mathcal{T}} X_{ij}(t)\beta_j(t)dt + \epsilon_i, \qquad i = 1, \ldots, N.$$

- Y_i: scalar response.
- $\mathbf{X}_i(t) = (X_{i1}(t), ..., X_{ip}(t))^T$ are functional predictors.

Functional Regression Model

$X_{ij}(t)$ can be discretized on a finite grid & expressed as

$$X_{ij}(t) = \Sigma_{b=1}^{K} a_{ijb}\phi_{jb}(t)$$

- $\phi_{jb}(t)$: basis functions (e.g. Fourier, splines, Gaussian)
- a_{ijb}: basis coefficients.

The coefficient functions β_j expressed as:

$$\beta_j(t) = \Sigma_{b=1}^{K} c_{jb}\phi_{jb}(t)$$

- $\phi_{jb}(t)$: known basis functions,
- c_{jb} : unknown corresponding coefficients.

Functional Regression Model

Our Model becomes:

$$
\begin{aligned}
Y_i &= \alpha + \Sigma_{j=1}^{p} \mathbf{\Phi_{ij}}^T \mathbf{c_j} + \epsilon_i & (1) \\
&= \mathbf{z}_i^T \mathbf{c} + \epsilon_i. & (2)
\end{aligned}
$$

$\mathbf{z}_i = (1, a_{i1}^T \mathbf{J}_{\phi_1}, \ldots, a_{ip}^T \mathbf{J}_{\phi_p})^T$

$\mathbf{c} = (\alpha, \mathbf{c}_1^T, \ldots, \mathbf{c}_p^T)^T$

$\mathbf{J}_b = \int_{\Im} \phi_b^T(t) \phi_b(t) dt$ ($K \times K$ cross-product matrices).

Previous Studies

- Functional Group SCAD: Matsui and Konishi (2011)
- Functional Group SCAD: Lian (2013)
- Functional Group LASSO: Zhu and Cox (2009)
- Functional Group LASSO: Gertheiss et al. (2013)
- Wavelet-Based LASSO: Zhao et al. (2013)

Functional Group LASSO: Zhu and Cox (2009) and Gertheiss et al. (2013)

Both methods:

- for Generalized Linear Model (Classification problem).
- based on Regularization Methods (Functional Group LASSO)

1. **Zhu and Cox** approach: Functional PCs are used to reduce the model to multivariate logistic regression and a grouped Lasso penalty is applied to the reduced model to select useful functional covariates among multiple curves.

2. **Gertheiss et al.**'s approach: penalized likelihood method that simultaneously controls the sparsity of the model and the smoothness of the corresponding coefficient functions by adequate penalization.

Objective function:

$$\Sigma_{i=1}^{n}(Y_i - \alpha - \Sigma_{j=1}^{p} \mathbf{\Phi_{ij}}^T \mathbf{c_j})^2 + P_{\lambda,\varphi}(\beta_j).$$

where the penalty function

$$P_{\lambda,\varphi}(\beta_j) = \lambda(||\beta_j||^2 + \varphi||\beta_j''||^2)^{1/2}.$$

- $||.||^2 = \int(.)^2 dt$ is the L^2 norm.
- β_j'' is the second derivative of β_j.
- λ is the parameter that controls sparseness.
- φ is the smoothing parameter that controls smoothness of the coefficients.

- As $\lambda \uparrow$, $\hat{\beta}$(t) $\rightarrow$ 0 at some value.
- As $\varphi \uparrow$, the departure from linearity is penalized stronger and $\hat{\beta}$(t) becomes closer to a linear function.
- Smaller values for φ result in very wiggly and difficult to interpret estimated coefficient functions.
- For optimal estimates (in terms of accuracy and interpretability), an adequate (λ, φ) combination has to be chosen.
- λ and φ are selected via K-fold cross-validation.

Adaptive group LASSO(Gerthesis et al. (2013)

- $P_{\lambda,\varphi}(\beta_j)$ is modified as:

$$P_{\lambda,\varphi}(\beta_j) = \lambda(\kappa_j||\beta_j||^2 + \varphi\nu_j||\beta_j^{''}||^2)^{1/2}.$$

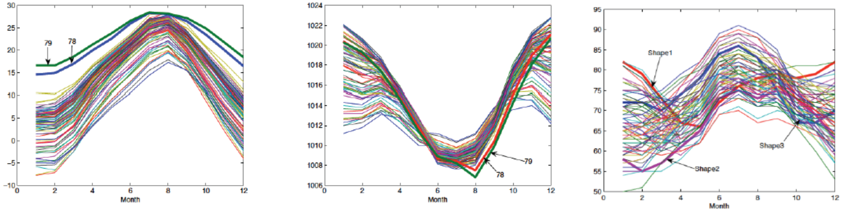
- The weights κ_j and ν_j are chosen in a data-adaptive way to:
  - Reflect some subjectivity about the true parameter functions
  - Allow for different shrinkage and smoothness for the different covariates.

Robust Variable Selection for Functional Regression Models with Functional Predictors and a Scalar Response

Outliers in Japanese Weather Data (Pallavi et al. 2013)

One of the main assumptions in these approaches:

Homogeneity of Data



(a) Temperature (b) Pressure (c) Humidity

Outlier Curve: *That curve that has been generated by a stochastic process with a different distribution than the rest of curves, which are assumed to be identically distributed (Febrero et al,2007).*

Aim: *Select the functional predictor variables that contribute the most for the prediction of annual total precipitation in the presence of outliers.*

Robust Functional Group LASSO: Pannu and Billor, 2015

Two methods based on Gertheiss et al.'s approach: α and c_j can be estimated by minimizing the following:

LAD-groupLASSO

$$\Sigma_{i=1}^{n} |Y_i - \alpha - \Sigma_{j=1}^{p} \mathbf{\Phi_{ij}}^T \mathbf{c_j}| + P_{\lambda,\varphi}(\beta_j).$$

WLAD-groupLASSO

$$\Sigma_{i=1}^{n} w_i |Y_i - \alpha - \Sigma_{j=1}^{p} \mathbf{\Phi_{ij}}^T \mathbf{c_j}| + P_{\lambda,\varphi}(\beta_j).$$

where, $P_{\lambda,\varphi}(\beta_j)$ is the penalty function (Meier et al., 2009) and w_i are the weights for controlling outliers in the functional predictor space.

Adaptive LAD-gLASSO and Adaptive WLAD-gLASSO

Penalty function:

$$P_{\lambda,\varphi}(\beta_j) = \lambda(\kappa_j||\beta_j||^2 + \upsilon_j\varphi||\beta_j^{''}||^2)^{1/2}.$$

- κ_j for smoothness
- υ_j for shrinkage

data adaptive weights.

Functional WLAD- gLASSO

- w_i are obtained using the robust distances of the predictors.
- The outlying observations in the x direction will have large distances and the corresponding weights will be small.
- Therefore, it is expected that the resulting estimator will be robust against the outliers in the response variable and leverage points.

Algorithm

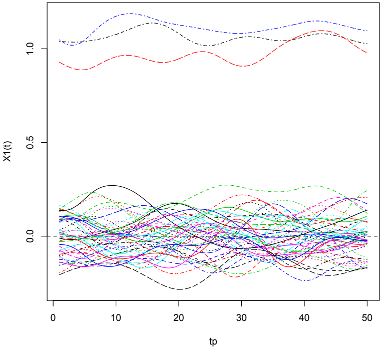
1. Calculate the robust location and scatter estimates, $\tilde{\mu}$ and $\tilde{\Sigma}$ for the location vector and the scatter matrix of the data $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, . . . , $\boldsymbol{x}_n \in \Re^p$.
   - One can use high breakdown point location and scatter estimators such as *MCD* (Minimum Covariance Determinant).
   - The idea behind *MCD* is to find observations whose empirical covariance matrix has the smallest determinant, yielding a pure subset of observations from which to compute standards estimates of location and covariance.
2. Compute the robust distances: $RD(\boldsymbol{x}_i) = (\boldsymbol{x}_i - \tilde{\mu})^T \tilde{\Sigma}^{-1} (\boldsymbol{x}_i - \tilde{\mu})$.
3. Calculate the weights $w_i = min\left\{1, \frac{p}{RD(\boldsymbol{x}_i)}\right\}$ for $i = 1, \ldots, n$.

Numerical Study for Robust Functional Group LASSO

Contamination of Y

- ϵ are generated from the $N(0, 1)$, t_2 and t_7.
- Contamination level: 15%.

Contamination X(t)



Case 1: Asymmetric Contamination

Asymmetric contamination (15%)

Simulation

Performance measures (50 runs):

- SE $= \int(\hat{\beta}_j(t) - \beta_j(t))^2 dt$
- Mean squared Errors of prediction:

$$MSE = 1/n \sum_i (Y_i - \hat{Y}_i)^2.$$

- Mean Absolute Errors of prediction:

$$MAD = 1/n \sum_i \mid Y_i - \hat{Y}_i \mid .$$

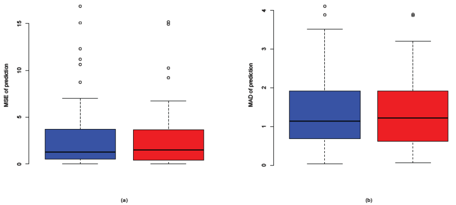
- $p = 10,\ N = 300$
- Time grid: 300 equidistant time points in (0, 300).
- The true model:

$$Y_i = \alpha + \sum_{j=1}^{5} \int_0^{300} \beta_j(t) X_{ij}(t) dt + \epsilon_i.$$

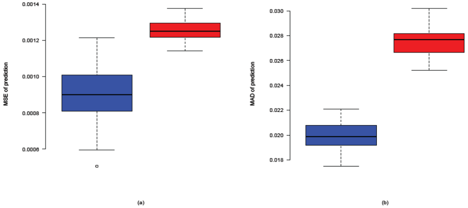
$\epsilon_i \sim N(0, 4)$

- True model depends only on $\beta_1(t)$ - $\beta_5(t)$.

Numerical Study for Robust Functional Group LASSO

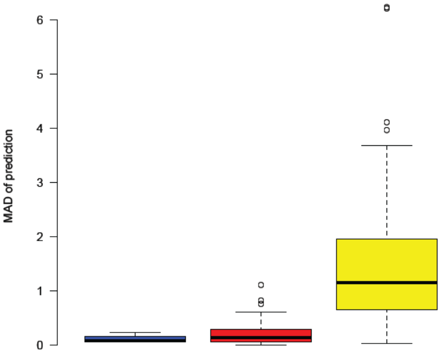


LAD-gLASSO(blue) and gLASSO (red) (No-contamination)



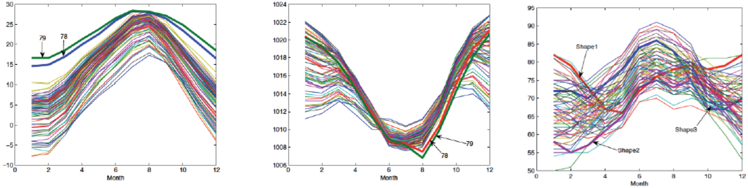
LAD-gLASSO (blue) and gLASSO (red) (15% contamination of Y).

Numerical Study for Robust Functional Group LASSO



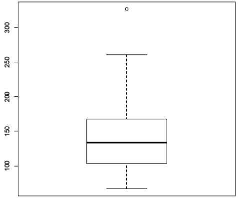
LAD-agLASSO (blue), LAD-gLASSO (red) and classical agLASSO (yellow) at 15% contamination of Y..

Japanese Weather Data

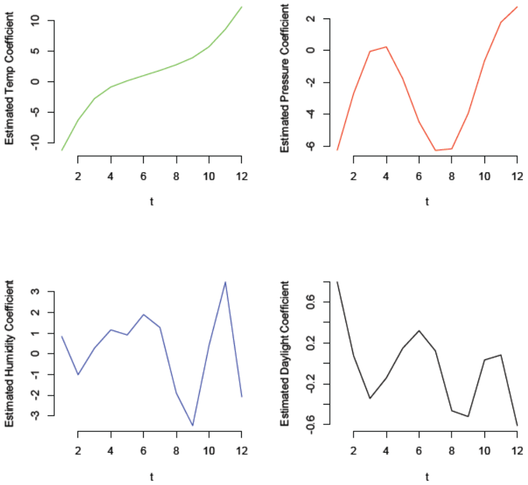


(a) Temperature (b) Pressure (c) Humidity

Outliers in functional predictors

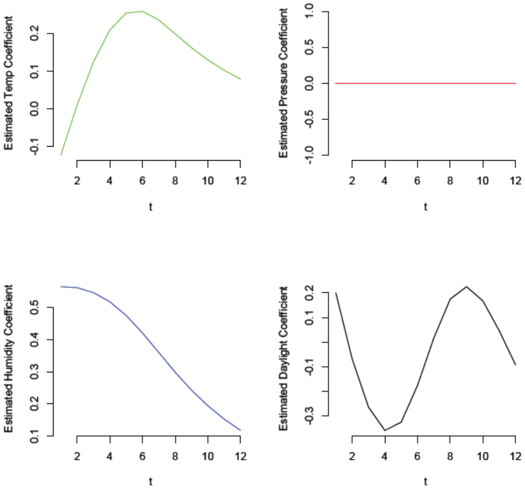


Outliers in scalar response



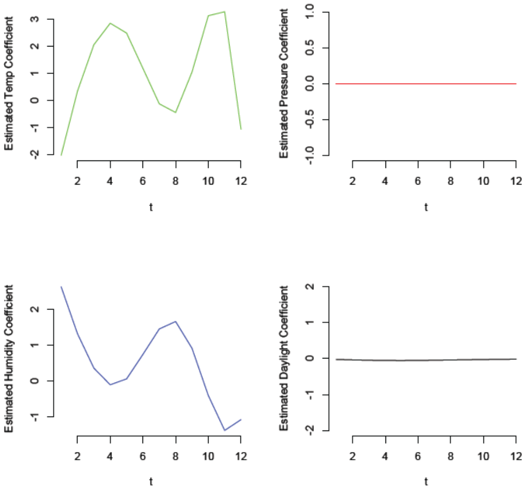
Classical Functional-gLASSO

Japanese Weather Data



Functional WLAD-gLASSO

Japanese Weather Data



Functional Adaptive WLAD-gLASSO

Japanese Weather Data

	TEMP	PRESSURE	HUMIDITY	DAYLIGHT	Avg. Model Size
Functional *WLAD- agLASSO*	1	0.36	0.98	0.40	2.74
Functional *WLAD- gLASSO*	1	0.38	0.96	0.66	3.00
Functional *LAD- gLASSO*	1	0.94	0.98	0.96	3.88

Proportions of runs with the respective functional predictor being selected and average model size.

Average PRESSURE and DAYLIGHT are less frequently selected by functional WLAD-agLASSO!

Selected References

- Gertheiss, J., Maity, A. & Staicu, A-M. (2013). Variable Selection in Generalized Functional Linear Models. Stat. 2: 86-101.
- Pannu, J. and Billor, N. (2015) Robust Group-Lasso for Functional Regression Model, Communication in Statistics: Simulation and Computation. (http://www.tandfonline.com/doi/full/10.1080/03610918.2015.1096375.).
- Lian, H. (2013). Shrinkage estimation and selection for multiple functional regression. Statistica Sinica. 23: 51-74.
- Matsui, H. & Konishi, S. (2011). Variable Selection for Functional Regression Models via the L1 Regularization. Computational Statistics and Data Analysis. 55: 3304- 3310.
- Ramsay, JO. & Silverman, BW. (2005). Functional Data Analysis. Springer.
- Zhao. Y., Ogden, R. T. & Reiss, P. T. (2013). Wavelet-Based LASSO in Functional Linear Regression. Journal of Computational and Graphical Statistics. 21:3, 600-617.
- Zhu, H. & Cox, DD. (2009). A functional generalized linear model with curve selection in cervical pre-cancer diagnosis using fluorescence spectroscopy. IMS Lecture Notes, Monograph Series - Optimality: The Third Erich L. Lehmann Symposium. 57: 173- 189.

Thank You!

Department of Mathematics and Statistics, Auburn University

The Department of Mathematics and Statistics (DMS):

- over 50 professors representing diverse areas
  - pure mathematics
  - applied mathematics
  - statistics
- offer undergraduate programs leading to a Bachelor of Science in Mathematics and Applied Mathematics (with options in Applied Mathematics, Discrete Mathematics, or Actuarial Science)
- graduate programs leading to a Master of Science in Mathematics, Applied Mathematics, Statistics, or Probability and Statistics, and/or the Doctor of Philosophy in Mathematics and the Doctor of Philosophy in Mathematics, concentration in Statistics.
- 130 GTAs

]

List of Research Fields

- Actuarial Mathematics
- Algebra
- Analysis
- Applied Mathematics
- Discrete Mathematics
- Differential Equations
- Geometry
- Linear Algebra
- Numerical Analysis
- Statistics
- Stochastic Analysis
- Topology

Possibilities for

- Research Collaborations
- Graduate Students Recruiting
- Faculty Exchange