Determining optimal treatments based on complex data

R. Todd Ogden

Department of Biostatistics, Columbia University

with Thaddeus Tarpey, Eva Petkova, and Hying Park (NYU), Adam Ciarleglio (George Washington University), Jeff Goldsmith and Joe Schwartz (Columbia University)

February 28, 2019
Motivation

STRESS & ACTIVITY STUDY

Ecological Link of Psychosocial Stress to Exercise: Personalized Pathways Pilot

- 60 intermittently exercising adults, observational data for 6 months (baseline)
- Random assignment for information received, half get:
  - personalized model of stress perception and exercise
  - traditional nomothetic model
- 6 more months observation
- outcome: measures of exercise behavior, health

Research question: Will presentation of a personalized model improve health?
Determining optimal treatments

R. Todd Ogden

Introduction

Functional data

Functional regression

Treatment regimes

A linear functional approach

Simulation

Application

Discussion

References

Daily actigraphy data
Determining optimal treatments

R. Todd Ogden

Introduction
Functional data
Functional regression
Treatment regimes
A linear functional approach
Simulation
Application
Discussion
References

Actigraphy data averaged across days
Example: Child height as a function of age
“Functional data”

Example: Knee angle as children go through a gait cycle
Determining optimal treatments

R. Todd Ogden

Introduction

Functional data

Functional regression

Treatment regimes

A linear functional approach

Simulation

Application

Discussion

References

“Functional data”

Example: Radius and curvature of the carotid artery
“Functional data”

Example: Brain imaging data
Functional data could be curves, spectra, images, time series, etc.

Any data that can be regarded as being a “function” of one or more continuous variables, $X_i(t), t \in \mathcal{T}$.

In practice, each functional observation consists of a finite number of observations made on distinct points: $X_i = (X_i(t_1), X_i(t_2), \ldots, X_i(t_N))$ for $t_1 < t_2 < \ldots < t_N$. 
A functional dataset is typically observed (and stored) just like any multivariate dataset would be (e.g., one row of data per subject).

The key difference is that functional data have a very specific structure. While classical multivariate techniques (MANOVA, PCA, etc.) are permutation-invariant, functional data techniques should not be.

A good functional data technique will exploit the structure of the data.
Common themes with functional data

- Typically, high dimensional (often, ultra high-dimensional)
- Registration
- Smoothing
There is a need for “dimension reduction” — this can be achieved by *regularization*.

- Projecting onto some (smooth) set of basis functions (Fourier basis, polynomials, splines, etc.)
- Penalizing roughness
- Projecting onto some set of basis functions that allow for a sparse representation (wavelets)
Virtually any statistical technique can be extended to handle functional data:

- Box plots
- Discrimination analysis
- Linear models
- Principal components analysis
- MANOVA (becomes “FANOVA”)
- Confidence intervals/bands
- Outlier/influence analysis
- Cluster analysis
- Structural equation modeling / path analysis
- . . .
Ordinary linear regression

Model for each subject:

\[ Y_i = x_i^T \beta + \varepsilon_i, \ i = 1, \ldots, n, \]

where \( \beta \) and each \( x_i \) are \( p \)-vectors.

In matrix form:

\[ Y = X \beta + \varepsilon. \]

The least-squares estimate of \( \beta \) is just

\[ \hat{\beta} = \arg \min_{\beta \in \mathcal{R}^p} \| Y - X \beta \|^2. \]

If \( \text{rank}(X) = p \) then the unique minimizer is \( \hat{\beta} = (X^T X)^{-1} X^T Y. \)

But if \( n < p \), there can be infinitely many \( \beta \in \mathcal{R}^p \) such that \( \| Y - X \beta \|^2 = 0. \) So we generally seek the optimal \( \beta \) within some reasonable subset of \( \mathcal{R}^p. \)
The Hamilton Depression Rating Scale (HAM-D) is a common measure of the severity of MDD symptoms.

\[ Y_i = \nu + \mathbf{x}_i^T \beta + \varepsilon_i \]
Functional regression

In matrix form:

$$Y = \nu 1 + X\beta + \epsilon$$

This can be thought of

- as very high-dimensional regression (hundreds of thousands of voxels per image; ten thousand per “slice”),
- as regression with “functional” data

$$Y_i = \nu + \int x_i(t)\beta(t) \, dt + \epsilon_i$$
Functional linear regression

Given functional predictors $x_1(t), x_2(t), \ldots, x_n(t), \ t \in \mathcal{T}$ and corresponding (scalar) outcome variables $Y_1, Y_2, \ldots, Y_n$, the functional regression model is

$$Y_i = \nu + \int x_i(t) \beta(t) \, dt + \epsilon_i, \ i = 1, \ldots, n.$$  

Examples:

1. Using NIR spectra to predict analytical properties (e.g., water content) (Marx and Eilers, 1999)
2. Using temperature curves to predict annual rainfall (Ramsay and Silverman, 2005)
3. Using measures of white matter integrity across the corpus callosum to predict clinical outcome score (Goldsmith, et al., 2011b)
Interpretation of coefficient function

\[ X_i(s) \rightarrow \beta(s) \rightarrow X_i(s)\beta(s) \rightarrow \int X_i(s)\beta(s) \, ds \]

- Profile x Coefficient (Area Under Curve Shaded)
- Functional Contribution

0
1.6
-2.6
5.2
Individualized treatment decisions

- Continuous response $Y$ (large values are better)
- Treatment (one decision) $A = -1$ or $1$
- Scalar covariates $Z = (1, Z_1, \ldots, Z_p)^\top$ (age, severity, clinical/cognitive measures, etc.)
- Functional observations $X = \{X_1(t), t \in T_1\}, \ldots, \{X_q(t), t \in T_q\}$ (can be 1-D, 2-D, 3-D, \ldots)
- Potential outcomes: $Y^*(-1), Y^*(1)$

We observe $Y = Y^*(1)(A + 1)/2 + Y^*(-1)(1 - A)/2$.

- “Treatment regime”: $g : (Z, X) \rightarrow \{-1, 1\}$
- The “value” of the decision rule $g$ is defined to be the aggregate effect of applying $g$ across the population: $E_{X,Z}E[Y^*(g(Z, X))]$. 
There has been much written about methodology for determining optimal treatment regimes when predictors are scalars.

Data model

\[ E[Y|Z, X, A] = h_{\alpha,\beta}(Z, X) + \frac{A}{2} f_{\gamma,\omega}(Z, X) \]

- \( h_{\alpha,\beta}(Z, X) = \alpha^\top Z + \sum_{\ell=1}^{p} \int \beta_\ell(s)X_\ell(s)ds \)
- \( f_{\gamma,\omega}(Z, X) = \gamma^\top Z + \sum_{\ell=1}^{p} \int \omega_\ell(s)X_\ell(s)ds \)
- \( \beta = \{\beta_1, \ldots, \beta_q\} \) and \( \omega = \{\omega_1, \ldots, \omega_q\} \)


Therefore:

\[ g^{opt}(Z, X) = \text{sign}\{f_{\gamma,\omega}(Z, X)\} \]

\[ = \text{sign}\left\{ \gamma^\top Z + \sum_{\ell=1}^{p} \int \omega_\ell(s)X_\ell(s)ds \right\} \]
Elements of fitting procedure (Ciarleglio, et al., 2015):

- Express functional observations in terms of eigenfunctions of smoothed estimated covariance operator (Goldsmith, et al., 2011a)
- Express the objective function for fitting the data as a loss function in the framework of A-learning (Murphy, 2003)
- Smoothing parameters may be chosen by REML

We would like to consider a procedure for variable selection also.
Modified covariates method (Tian et al., 2014)

Note that: \( E(2YA|Z, X) = f_{\gamma, \omega}(Z, X) \)

Estimate \( \gamma \) and \( \omega \) by minimizing:

\[
\frac{1}{n} \sum_{i=1}^{n} \left( 2Y_i A_i - f_{\gamma, \omega}(Z_i, X_i) \right)^2 \propto \frac{1}{n} \sum_{i=1}^{n} \left( Y_i - f_{\gamma, \omega}(Z_i, X_i) \frac{A_i}{2} \right)^2
\]

So we can estimate \( \gamma \) and \( \omega \) by fitting

\[
Y = f_{\gamma, \omega}(Z, X) \cdot \frac{A}{2} + \varepsilon
\]

\[
= \gamma^T \left\{ Z \cdot \frac{A}{2} \right\} + \sum_{\ell=1}^{p} \int \omega_{\ell}(s) \left\{ X_{\ell}(s) \frac{A}{2} \right\} ds + \varepsilon
\]
Augmentation (Tian et al., 2014)

Since, because of randomization, for any function $a$,
\[ E[A_i a(Z_i, X_i)] = 0, \]

It can be shown that arg min $\frac{1}{n} \sum_{i=1}^{n} \left( Y_i - f_{\gamma, \omega}(Z_i, X_i) \frac{A_i}{2} \right)^2$ converges to the same limit as

\[
\arg \min \frac{1}{n} \sum_{i=1}^{n} \left[ \left( Y_i - f_{\gamma, \omega}(Z_i, X_i) \frac{A_i}{2} \right)^2 + \right.
\frac{A_i}{2} \left\{ \gamma^T Z + \sum_{\ell=1}^{p} \int \omega_{\ell}(s) X_{\ell}(s) ds \right\} E[Y_i | Z_i, X_i] \]

but can often give smaller variance.
Combining estimation with variable selection and roughness penalization

Define \( L_n(\gamma, \omega) = \frac{1}{n} \sum_{i=1}^{n} \left( Y_i - f_{\gamma, \omega}(Z_i, X_i) \frac{A_i}{2} \right)^2 \)

Express functional components in terms of spline basis functions, choose \( \gamma \) and \( \omega \) to minimize (Gertheiss, et al., 2013)

\[
L_n(\gamma, \omega) + \lambda \left\{ \sum_{j=2}^{p+1} J(|\gamma_j|) + \sum_{\ell=1}^{q} P_{\rho_\ell}(\omega_\ell) \right\}
\]

\[
J(|\gamma_j|) = |\gamma_j| \quad P_{\rho_\ell}(\omega_\ell) = \left\{ ||\omega_\ell||^2 + \rho_\ell Q(\omega_\ell) \right\}^{1/2} \quad ||\omega_\ell|| = \int \omega_\ell^2(s)ds
\]

1D: \( Q(\omega_\ell) = \left\| \frac{\partial^2 \omega_\ell}{\partial s^2} \right\|^2 \) \quad 2D: \( Q(\omega_\ell) = \left\| \frac{\partial^2 \omega_\ell}{\partial s^2} \right\|^2 + \left\| \frac{\partial^2 \omega_\ell}{\partial s \partial t} \right\|^2 + \left\| \frac{\partial^2 \omega_\ell}{\partial t^2} \right\|^2 \)

Tuning parameters:
- \( \lambda \) (complexity)
- \( \rho_\ell, \ell = 1, \ldots, q \) (smoothness)
Group lasso

For fixed $\lambda, \rho_1, \ldots, \rho_q$, this objective function can be optimized using procedures for the “group lasso” (Yuan and Lin, 2006).

- Scalar covariates are regarded as groups of size 1.
- Can fit using \texttt{grplasso} in \texttt{R}.
- Tuning parameters may be chosen by cross-validation (although there are $q + 1$ of them . . . ).
Simulated data setup

Baseline Covariates

- **Scalar Predictors:**
  - $Z = (1, Z_1, \ldots, Z_p)^\top$ MVN such that $Z_j \sim N(0, 1)$ and $\text{corr}(Z_j, Z_k) = 0.5 |j-k|$.

- **Functional Predictors:**
  - $q$ functions based on EEG data.

![Graphs of functional predictors](image)

- **Treatment Assignment:** $P(A = 1) = P(A = -1) = 0.5$
Simulation settings

- Correctly/incorrectly specified contrast model
- Moderate/large main/interaction effect
- Small/large numbers of spurious covariates
- Functional covariates with/without error
- Range of sample sizes
Simulation results

- Augmentation improves performance, especially for large main effect and/or when the model is misspecified.
  - **Model Selection**: more true positives (but also more false positives)
  - **Treatment Decision**: considerably better

- These procedures are comparable to or better than “straight lasso”.
  - **Treatment Decision**: comparable especially for large $n$
  - **Model Interpretability**: considerably better
Unfortunately...
Unfortunately... 

Actigraphy study data are not yet available to us!
Unfortunately... 

Actigraphy study data are not yet available to us!

There will be issues of:

- Aggregation
- Registration
- Outliers
- Etc.
Unfortunately... 

Actigraphy study data are not yet available to us!

There will be issues of:

- Aggregation
- Registration
- Outliers
- Etc.

But when we do all this, it’s going to be awesome.
Application: EMBARC data

**Treatment:** Placebo ($n_P = 69$), Sertraline ($n_S = 63$)

** Scalars:**
- HAM-D score
- sex (0 = male, 1 = female)
- age
- years of education

**Functions:** EEG spectra

**Response:** HAM-D score at week 8 (lower scores are preferred)
Random split: 107 in training data; 25 in test data

$B$-spline basis of order 4 with 25 basis functions

5-fold cross validation to choose tuning parameters
### Results: Model fit to training data

<table>
<thead>
<tr>
<th>Variable</th>
<th>$\hat{\gamma}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.001</td>
</tr>
<tr>
<td>HAM-D</td>
<td>-0.009</td>
</tr>
<tr>
<td>Sex</td>
<td>-0.021</td>
</tr>
<tr>
<td>Age</td>
<td>0.012</td>
</tr>
<tr>
<td>Years ed</td>
<td>0.043</td>
</tr>
</tbody>
</table>

### Functional data

- $X_1$ (FZ)
- $X_2$ (FCZ)
- $X_3$ (F4)
- $X_4$ (F3)
- $X_5$ (PZ)
- $X_6$ (POZ)

### Additional comments:

- $\omega^1$ to $\omega^6$ represent different frequency bands.
- The plots show relative amplitude changes over time.
Results: Evaluation in test data

<table>
<thead>
<tr>
<th>Regime (n)</th>
<th>Mean Response</th>
<th>95% Percentile Bootstrap CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Random (25)</td>
<td>12.61</td>
<td>(9.78, 15.56)</td>
</tr>
<tr>
<td>(2) Placebo (12)</td>
<td>12.78</td>
<td>(8.60, 17.45)</td>
</tr>
<tr>
<td>(3) Sertraline (13)</td>
<td>12.46</td>
<td>(8.92, 16.15)</td>
</tr>
<tr>
<td>(MC-AL, MC-CART)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(4) OWL (12)</td>
<td>12.33</td>
<td>(7.83, 17.33)</td>
</tr>
<tr>
<td>(5) MC-A (14)</td>
<td>10.43</td>
<td>(7.19, 13.67)</td>
</tr>
</tbody>
</table>
Current and ongoing work

- Generated Effect Modifiers (GEMs)
  - Search for low-rank projections of high dimensional/functional data that are useful for choosing treatments.

- GEMs with nonparametric link functions
  - Treatment decisions are made on the basis of GEMs but the relationship between GEM and outcome needs not be linear.

- Distance-based methods
  - By defining a distance (e.g., between images) it is possible to make treatment decisions completely nonparametrically.

- Methods for more than two treatments
  - The “modified covariates” method only works for two treatments. This idea can be extended to multiple treatments by constraining the “main effect” to be orthogonal to the “interaction effect”.

Current and ongoing work

- Generated Effect Modifiers (GEMs)
  - Search for low-rank projections of high dimensional/functional data that are useful for choosing treatments.

- GEMs with nonparametric link functions
  - Treatment decisions are made on the basis of GEMs but the relationship between GEM and outcome needs not be linear.

- Distance-based methods
  - By defining a distance (e.g., between images) it is possible to make treatment decisions completely nonparametrically.

- Methods for more than two treatments
  - The “modified covariates” method only works for two treatments. This idea can be extended to multiple treatments by constraining the “main effect” to be orthogonal to the “interaction effect”.

References
References


