Bike-sharing Data Analysis
Functional Data Analysis

Liela Meng
University of California, Davis, Department of Biostatistics

Introduction

The bike-sharing system is a service for individuals to borrow a bike from a ”dock” and return it at another dock (some systems are dockless).

This dataset contains the hourly count of rental bikes between the years 2011 and 2012 in the Capital bike-share system with the corresponding weather (e.g., temperature, wind speed, humidity, etc.) and seasonal information (e.g., holiday, weekend, etc.).

In this analysis, hourly data in January 2011 from the bike-sharing system was analyzed (31 functions in total).

The aim of this project was to visualize the hourly count of total rental bikes, to investigate the dominant modes of variation, and to fit a functional concurrent regression model to estimate the relationship between feeling temperature and the renting counts.

Hypothesis:

- whether rush-hours (8AM, 5PM) have more bike demands;
- Whether the feeling temperature is positively related to the total demands.

Objectives

- Visualize the count of total rental bikes per hour with smooth curves
- Investigate the relationship between feeling temperature and the renting counts.

Methods

One-dimensional local linear kernel smoother:

$$\hat{\beta}_0(x) = \arg min \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \beta(X_i - x)\right)\right)^2 K\left(\frac{x - X_i}{h}\right).$$

Expanechnikov function: $K(x) = \frac{3}{4}(1-x)^2|_{[-1,1]}$.
Bandwidth was subjectively chosen based generalized cross-validation (GCV) method.

Functional principal component analysis (FPCA):

$$X_{iK}(t) = \mu(t) + \sum_{k=1}^{K} A_{ik}\emptyset_k(t), A_{ik} = \int_I (X_i(t) - \mu_t)\phi_k(t)dt.$$

Approximating X_i with K terms, where A_{ik} are the functional principal components (FPCs) and ϕ_k are orthogonal eigenfunctions in descening order.
Using Gaussian as the kernel smooth basis function and setting bandwidth as used in local kernel smoother.

Functional Concurrent Regression Model:

$$Y(t) = \beta_0(t) + \beta_1(t)X(t) + \varepsilon(t), \qquad t = 0,1,\cdots 23.$$

$\beta_0$is non-random function (functional intercept) and β_1 is the non-random coefficient function (functional slope) for feeling temperature.
This varying-coefficient model assumes the value of Y at time t are independent from $X(s)$, $s \neq t$, and $I_X = I_Y = \{0,1,\cdots 23\}$.
In the analysis, the smoothing kernel function is the Gaussian function. Bandwidth for smoothed mean function and smoothed covariance function were subjectively specified based on the automatically estimated smooth method.

Results

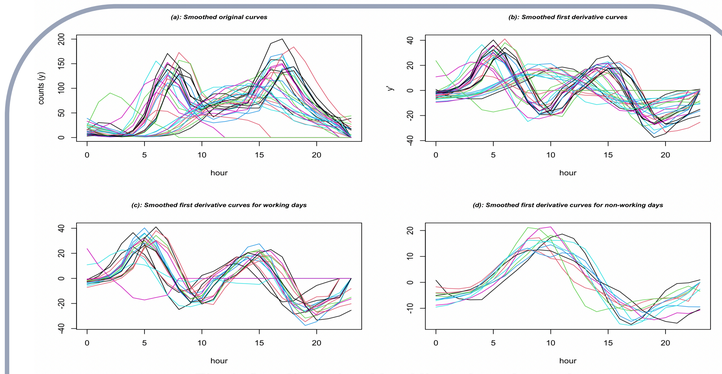

Fig. 1. One dimensional local linear kernel smoother

Since the automatically selected bandwidth value for smoothing the mean function using GCV is 1.15, the subjective bandwidth was chosen around this value. Eventually, bandwidth 2 was specified.

- Figure 1 (a), where smoothing the original count of rented bikes by the hour, we can see two peaks around 8 AM and 5 PM. This trend corresponds to people’s typical commute time.
- Figure 1 (b) is the first derivative of the count, which seems to include two kinds of curves: the first one has peaks at hour 5 and 15, and the other one has one peak at 10 AM. It is natural to attribute the observed difference to whether that day is a working day or not.
- Figure 1 (c, d), working days (c) have curves different from non-working days (d), and those two have the same characteristics we discovered in figure (b).

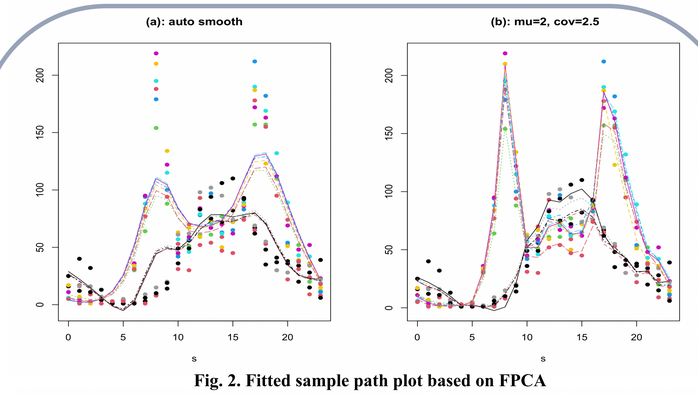

Fig. 2. Fitted sample path plot based on FPCA

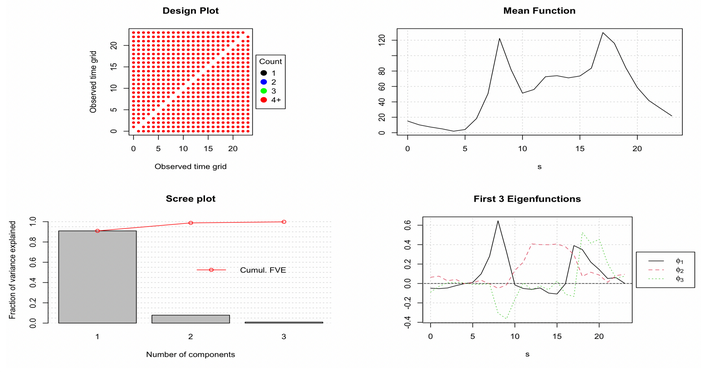

Fig. 3. FPCA results: design plot, mean function, scree plot, first 3 eigenfunctions

Results

Figure 2: we can see the auto fitted plot (2(a)) fails to capture the surge at hour 8 and 16. Thus, I subjectively specified bandwidth for mean and covariance as 2 and 2.5, respectively (2(b)).

Figure 3 shows the FPCA results:

- The mean function still corresponds with what we discovered with the local linear kernel smoother (two peaks at hour 8 and 16).
- PCA1 shows a variation between rush hours and normal hours, while PCA2 shows variation between ”brunch” time and other times.

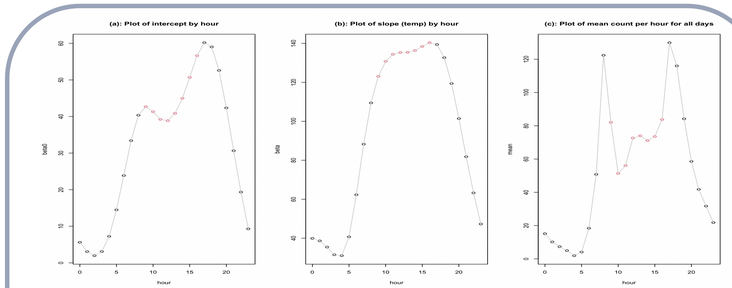

Fig. 4. Plot of estimated intercept and slope per hour

Figure 4 (b): overall, the feeling temperature is positively related to the total counts.
Figure 4 (a,b): from hour 9 to 16, estimations of β0, β show different pattern, which might indicate that in this time frame, the feeling temperature might play a more important role than in other times as the estimated coefficients are increasing at a relatively high level (see red dots).

Discussion

After using the local linear kernel method, we can see that smoothed total count curves have ***two peaks around 8 AM and 5 PM (rush-hours***) for working days and one peak around 10 AM for weekends, and those two features seem to correspond to PCA1 and PCA2.

The functional concurrent regression model was utilized to investigate the relationship between feeling temperature and the counts, and a ***positive relationship was observed*** across all times. In the time frame 9 AM to 4 PM, though we can see the estimated coefficients for temperature are increasing at a relatively high level, we can not explicitly explain how it is associated with the total count. Another thing to notice is that, the assumption ($Y(t) \perp X(s),\ s \neq t$) of the varying-coefficient model might not hold in reality.

Thus, I believe further research should focus more on including more influential factors and try the functional linear model rather than the varying-coefficient model.

Acknowledgements

I would like to thank Professor Hans-Georg Mueller and Han Chen for their help in writing this report. I would like to extend my thanks to Professor JiGuo Cao for offering open course and R codes online for free.

Data was downloaded from UCI Mashing Learning Repository website:
https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset

References

[1] Fanaee-T, H. & Gama, J. Event labeling combining ensemble detectors and background knowledge. Progress in Artificial Intelligence, 1–15 (2013).
[2] Wang, J.-L., Chiou, J.-M. & Mueller, H.-G. Review of Functional Data Analysis 2015. arXiv: 1507.05135 [stat.ME].
[3] Cao, J. Functional Data Analysis Course https : / / github . com / caojiguo / FDAcourse2019.