UC DAVIS, DEPARTMENT OF STATISTICS

Functional Principal Component Analysis on medfly data

Amy Kim

March 14, 2016

1. INTRODUCTION

1.1. SUMMARY

Main interest is how the egg-laying pattern is related to lifetime of Medfly. I conduct Functional Principle Component Analysis (FPCA) on modeling medfly data, which are associated with the time-dynamics of the egg-laying trajectory as the start of modeling. FPCA achieves to reduce 22 dimensions to 10 dimensions when Fraction-of-Variance-Explained threshold (FVE threshold) is 0.9999. In FVE threshold 0.9, it reduces to 3 dimensions. The data modeling based on the FPCA method is precise as the data modeling based on the whole dimension of data. Principal Component Analysis (PCA) does not give any dimensionality reduction on this modeling. The FPCA Poisson model implies the number of egg-laying has negative relationship to lifetime.

1.2. FUNCTIONAL PRINCIPAL COMPONENT ANALYSIS

Functional Principal Component Analysis(FPCA) is a dimentionality reduction method for functional data analysis. By FPCA, I choose Funcional Principal Components, and fit lifetime on the K numbers of FPCs. K will be choosen by FVE 0.9999 and 0.90. Following is brief mathmatical background.

MATHMATICAL BACKGROUND For a random function $X(t) \in L^2([0,T])$, infinite dimension, which indicates also $E(\int X^2(t)dt) < \infty$

Let

$$
\begin{aligned}
E(X(t)) &= \mu(t) \\
Cov(X(t), X(s)) &= G(t,s) \\
\int\int g(s)G(s,t)g(t)dsdt \geq 0, &\quad \forall g \in L^2
\end{aligned}
$$

be the mean function and covaraicance function are continuous in s and t. Covariance function is symmetric, and $G \geq 0$

$$G(t,s) = \Sigma_{j=1}^{\infty}\lambda_i\phi_j(t)\phi_j(s)$$

where λ_j is jth eigen value and $\phi_j(s)$ is jth eigenfucntions.

By Karhunen-Loeve Theorem:

$$
\begin{aligned}
X(t) &= \mu(t) + \Sigma_{j=1}^{\infty}\xi_j\phi_j(t) \\
\text{where} \qquad \xi_i &=< X(t) - \mu(t), \phi_j(t) > \\
&= \int (X(t) - \mu(t)\phi_j(t)dt
\end{aligned}
$$

Here, ξ_j are random component with

$$E(\xi_i) = 0, E(\xi_i\xi_j) = 0, Var(\xi_j) = \lambda_j \quad \text{for} \quad j \neq k$$

We call ξ_j are the Functional Principal Components (FPCs) of X
We choose K

$$\Sigma_{j=1}^{\infty}\lambda_j \approx \Sigma_{j=1}^{K}\lambda_j, \quad \lambda_1 \geq \lambda_2 \geq \cdots \geq 0$$

1.3. Data

The medfly data set consists of number of eggs laid daily for each of 1000 medflies[1] for 25 days and Lifetimes. I suspect a relationship of laying egg pattern over time to lifetime. I subset data by choosing observations that the number of laying eggs for 25 days is more than 100 eggs as well as lifetime is longer than 25 days. Also, exploring data, one shows extreme hat value, I exclude the observation[2]. The number of observation is 661.

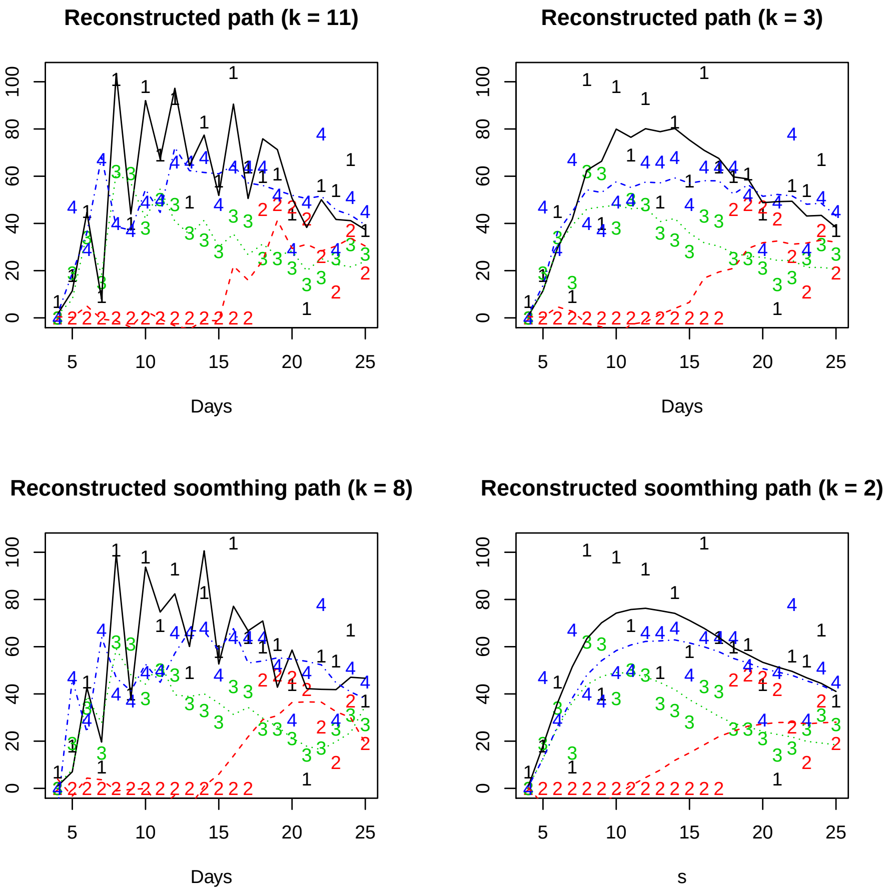
2. Methods

2.1. Dimension Reduction

Dimension Reduction is the process of reducing the number random variables. The Medfly data has egg counts per each 25 days. Each day is my predictors to lifetime. Observed that Day

[1] Data were obtained in Dr. Carey's laboratory

[2] id number is 353



Reconstructed path (k = 11)

Reconstructed path (k = 3)

Reconstructed soomthing path (k = 8)

Reconstructed soomthing path (k = 2)

Figure 2.1: FPCA Path Plot

1 and 3 have no egg counts. Id number 353 medfly has extreme leverage point[3], so it makes that no egg counts in Day 2 as well. I process dimension reduction from 22 dimensions (Day 4 to Day 25)

2.1.1. Functional Principal Component

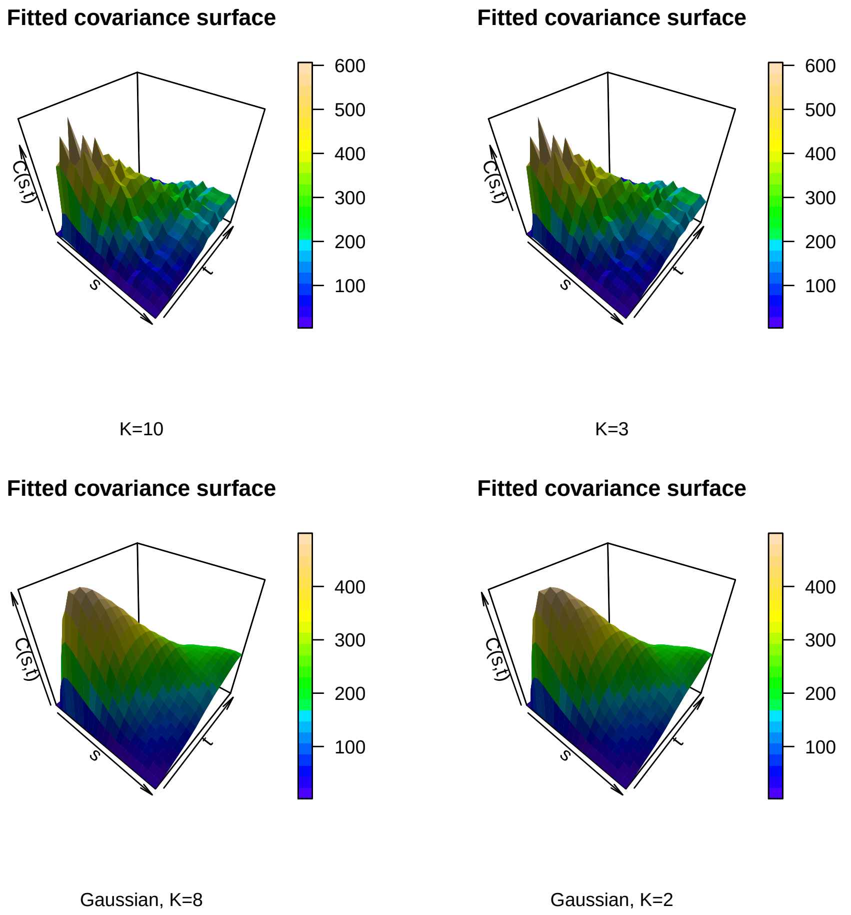
I have $X(t) =$ egg counts, $\quad t = 4, \cdots, 25$ on the data. I assign Fraction-of-Variance-Explained threshold(FVE threshold) is 0.9999. Then, FPCA reduce dimensions from 23 to 10 (K = 10)[4].

Path plot, Fig 2.1, shows how well each observations explained by chosen FPCs. The four observations[5] are randomly chosen, and the path plot shows FPCA has select 10 variables

[3] The leverage returns 1 in Naive Poisson GLM

[4] All calculation is done by R - package:fpapace

[5] ids: 4, 24, 69, 280



Fitted covariance surface

Fitted covariance surface

K=10

K=3

Fitted covariance surface

Fitted covariance surface

Gaussian, K=8

Gaussian, K=2

Figure 2.2: FPCA Covariance Plot

which explain very close to original data[6]. When we want to set our K selection criteria is that FEV equals to .90, FPCA reduces 3 dimensions(K = 3). In the path plot(bottom left), K = 3 is smoother, but still explain well about our original data. Additionally, I estimated the mean and covariance by smoothing[7], then FPCA reduces 8 dimensions with achieveing FEV is 0.9999 and 2 dimensions with FEV 0.90. The estimated covariance plot, Fig 2.2, shows how I estimated covariance functions.

2.1.2. Compare to Principal Component Analysis

I also conduct the Principal Component Analysis to reduce dimension in order to see how FPCA effective. Principal Component Analysis permute the order of data which is not feasiable in the data set, since PCA will remove the time depedencey that should not be reordered.

[6] Original data are 1,2,3, and 4 on each plots

[7] Gaussian

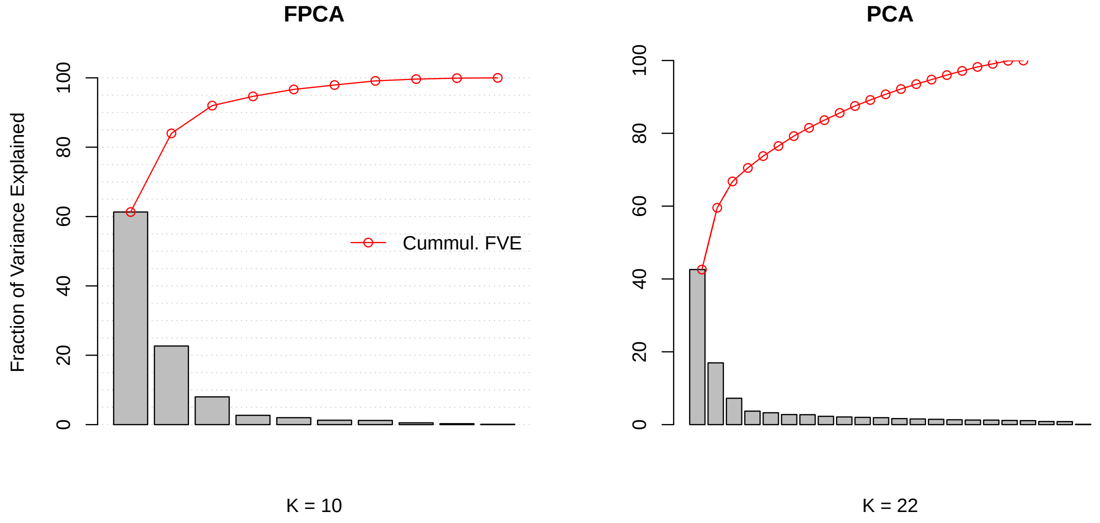


Figure 2.3: Scree Plots

FPCA carries time information in functional data and does not reorder them.

PCA returns we need all 22 variables to achieve 99% explained variance, which does not make any improvement. The screeplot Fig 2.3 To acheive 90% explained variance, PCA recommends 13 variables while FPCA does 3 variables.

2.2. Modeling - Poisson

I'd like to see the relatioship of lifetime and egg laying pattern. The lifetime is recoded in days, so I conduct poisson regression lifetime(count) on the response variables with log link. Here is three approaches of modeling. I fit 3 models which I call Naive approach, FPCA approach, and PCA approach. In Naive approach, I take all 22 variables(daily egg counts for 22 days) to model lifetimes. I choose the 0.90 FVE threshold, then FPCA approach I fit lifetime on the 3 variables(3 functional principal components) while PCA approach fitting on the 13 variables (13 principal components).

Goodness-of-fit According to Residual Plots Fig 2.4, there are no evidence of lack of goodness of fit. FPCA poisson's residuals are spread well and the smoothing line is at the zero horizontal line.

Error Comparison All three model's Quasi R squared and cross-validation prediction error are very closed. Remind of FPCA chosen 3 variables and PCA choose 13 variables, then FPCA poisson model perform efficiently precise.

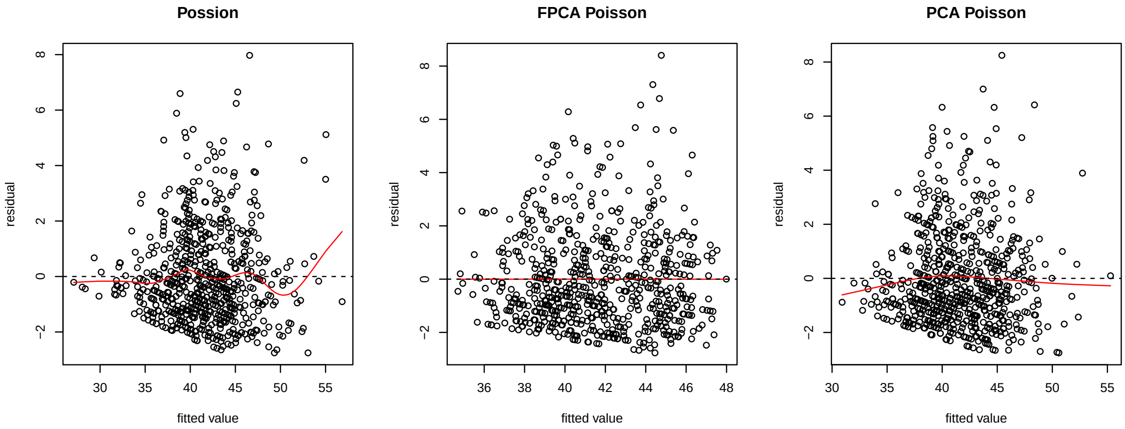


Figure 2.4: Residual Plots

	Poisson	FPCA	PCA
Quasi R2	0.99952	0.99946	0.99949
CV error	134.71570	153.28270	158.36160

Table 2.1: Goodness-of-fit

3. Results

The final model is the FPCA poisson model with first three functional principal components. In FPCA, the first FPC is the direction in feature space along which projections have the largest variance, so usally follows mean function. The second FPC is the direction which maximizes variance among all directions orthogonal to the first. The third FPC is the variance-maximizing direction orthogonal to the second components. From the model summary, Table 3.1, I can conclude the number of egg-laying has negative relationship to lifetime of medfly because the lifetime has negative relationship to first FPC, and positive relationship to second FPC.

Analysis on FPCA poisson model is more meaningful than Naive Poisson Model. From Naive Poisson Summary, Table A.1, D5, D8, D9, and other 7 days are siginificant effect at level 0.05, which indicates relationships between certain days and lifetime. However, it's not applicable to analysis effects of egg-laying on the lifetime.

4. Discussion

I choose the FPCA poisson model for medfly data with first three FPC. Notice the 3 FPC is not siginificant. Each observation in the medfly data has each own egg-laying pattern. across the time. It might be a good expansion to add analysis each observation's variance - random

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.72529	0.00605	616.12686	0.00000
fpca1.90$xiEst1	-0.00066	0.00009	-7.65667	0.00000
fpca1.90$xiEst2	0.00114	0.00014	8.30951	0.00000
fpca1.90$xiEst3	0.00030	0.00021	1.42548	0.15402

Table 3.1: FPCA Poisson Model (FVE 0.90)

effect, and see I can drop the third FPC, then still can hold FVE 0.90.

I also wonder if I add whole data which has 101 days rather than 25 days, how FPCA will perform the dimensionality reduction.

The FPCA possion[8] is called a functional regression. If adding more smoothing techniques to estimate each observations's egg pattern before FPCA. What I have tried is Gaussian with estimated covarince by GCV, and notice smoothing lose some individual charateristics, Bias and Variance Tradoff. Thus, smooth estimating requires sophisticated technique.

A. Appendix

A.1. Reference

[1] Carey, J.R., Liedo, P., Muller, H.G., Wang, J.L., Chiou, J.M. (1998). Relationship of age patterns of fecundity to mortality, longevity, and lifetime reproduction in a large cohort of Mediterranean fruit fly females. J. of Gerontology –Biological Sciences 53, 245-251.

[2] Hadjipantelis, Dai, Ji, Muller & Wang (2016) FunctionalPCA in R *A software primer using fdapace*

A.2. Dimension Reduction

A.3. modeling

[8] As I denote

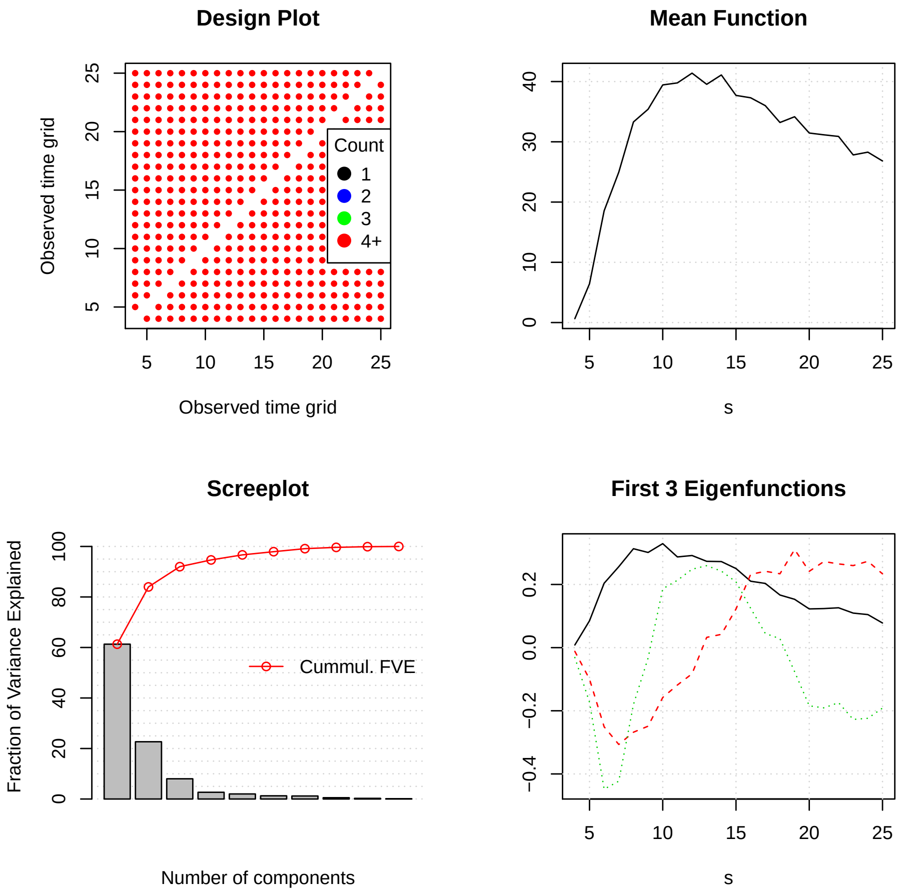


Figure A.1: FPCA Diagonostic Plot

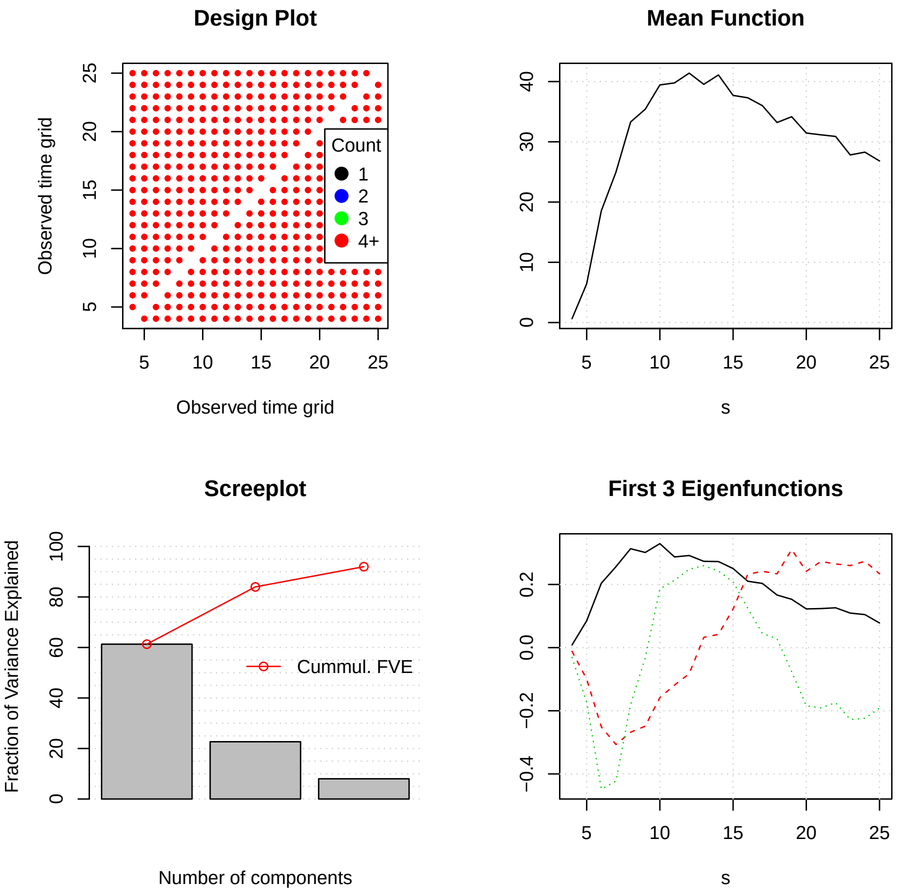


Figure A.2: FPCA Diagonostic Plot - FVE 0.9

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.7486	0.0155	241.15	0.0000
D4	-0.0051	0.0021	-2.37	0.0178
D5	-0.0006	0.0005	-1.12	0.2640
D6	0.0000	0.0003	0.14	0.8886
D7	-0.0004	0.0003	-1.11	0.2664
D8	-0.0011	0.0003	-3.73	0.0002
D9	-0.0012	0.0003	-3.74	0.0002
D10	0.0000	0.0004	0.03	0.9791
D11	-0.0004	0.0004	-1.15	0.2518
D12	0.0012	0.0004	3.34	0.0008
D13	0.0005	0.0004	1.33	0.1847
D14	0.0002	0.0004	0.53	0.5973
D15	-0.0018	0.0004	-4.28	0.0000
D16	0.0017	0.0004	4.14	0.0000
D17	-0.0002	0.0004	-0.46	0.6451
D18	-0.0012	0.0004	-2.69	0.0072
D19	0.0001	0.0004	0.18	0.8560
D20	-0.0001	0.0005	-0.14	0.8881
D21	0.0004	0.0005	0.91	0.3603
D22	-0.0030	0.0005	-5.93	0.0000
D23	-0.0003	0.0006	-0.48	0.6284
D24	0.0021	0.0005	4.20	0.0000
D25	0.0030	0.0005	6.31	0.0000

Table A.1: Result Table from Poisson Model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.7243	0.0061	615.39	0.0000
fpca1$xiEst1	-0.0006	0.0001	-7.41	0.0000
fpca1$xiEst2	0.0012	0.0001	8.71	0.0000
fpca1$xiEst3	0.0002	0.0002	1.11	0.2655
fpca1$xiEst4	0.0004	0.0003	1.43	0.1539
fpca1$xiEst5	0.0005	0.0003	1.64	0.1018
fpca1$xiEst6	-0.0002	0.0003	-0.56	0.5754
fpca1$xiEst7	-0.0020	0.0003	-5.76	0.0000
fpca1$xiEst8	0.0007	0.0004	1.83	0.0668
fpca1$xiEst9	-0.0012	0.0004	-3.22	0.0013
fpca1$xiEst10	0.0001	0.0004	0.19	0.8501

Table A.2: FPCA Poisson Model (0.9999)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.7238	0.0061	615.00	0.0000
pc1$x[, 1:13]PC1	-0.0007	0.0001	-7.71	0.0000
pc1$x[, 1:13]PC2	0.0012	0.0001	8.63	0.0000
pc1$x[, 1:13]PC3	0.0003	0.0002	1.25	0.2102
pc1$x[, 1:13]PC4	0.0004	0.0003	1.46	0.1455
pc1$x[, 1:13]PC5	-0.0006	0.0003	-1.95	0.0507
pc1$x[, 1:13]PC6	-0.0002	0.0003	-0.47	0.6373
pc1$x[, 1:13]PC7	0.0021	0.0003	6.04	0.0000
pc1$x[, 1:13]PC8	-0.0007	0.0004	-1.80	0.0714
pc1$x[, 1:13]PC9	-0.0012	0.0004	-3.18	0.0015
pc1$x[, 1:13]PC10	0.0001	0.0004	0.20	0.8382
pc1$x[, 1:13]PC11	0.0000	0.0004	0.05	0.9641
pc1$x[, 1:13]PC12	0.0015	0.0004	3.55	0.0004
pc1$x[, 1:13]PC13	0.0013	0.0004	2.91	0.0036

Table A.3: PCA Poisson Model (0.90)