Functional Principal Component Analysis on Medfly Data

Amy Kim (atykim@ucdavis.edu)
UC Davis, Department of Statistics

<Summary>

- Modeling Medfly Lifetime on Egg counts for 25 each days (Poisson Model)
- Use Functional Principal Component Analysis (FPCA) to dimensionality reduction before modeling
- FPCA Reduce 22 dimension to 10 dimension with Fraction-of-Variance-Explained(FVE) threshold 0.9999
- Principal Component Analysis(PCA) does not give any dimensionality reduction on the data with same threshold.
- FPCA Poisson modeling on Medfly lifetime is effective and precise as Poisson modeling using non-reduced dimension data

<Introduction>

Dimension Reduction is the statistical process of reducing the number random variables. One of famous dimension reduction method is Principal Component Analysis (PCA). However, PCA is not always feasible. PCA permutes the order of data , so it will remove the time dependency that should not be reordered.

Functional Principal Component Analysis (FPCA) is a dimensionality reduction method for functional data analysis. FPCA carries time information in functional data and does not reorder them.

I conduct Functional Principle Component Analysis (FPCA) on modeling medfly data, which are associated with the time-dynamics of the egg-laying trajectory.

<Data>

The medfly data set consists of number of eggs laid daily for each of 1000 medflies for 25 days and its lifetimes. I suspect a relationship of laying egg pattern over time to lifetime. Before starting analysis, 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. Thus, the data has 22 dimension (Day 4 to Day 25).

<Note>

[1] Data from Dr. Carey's laboratory
[2] R package FDAPACE is used

<Method>

Mathematical Background FPCA

$$X(t) \in L^2([0,T]), \quad X(t) = \mu(t) + \Sigma_{j=1}^{\infty}\xi_j\phi_j(t)$$

$$\xi_i = < X(t) - \mu(t), \phi_j(t) >= \int (X(t) - \mu(t)\phi_j(t)dt$$

ξ_j Functional Principal Component (FPC) of X

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

Choose K such that $\Sigma_{j=1}^{\infty}\lambda_j \approx \Sigma_{j=1}^{K}\lambda_j$

FPCA on Medfly Data

X(t) = egg counts, t = 4, … , 25 on the data. FPCA reduces dimensions from 23 to 10 (K = 10) with FVE threshold : 0.9999. *Fig 1*.

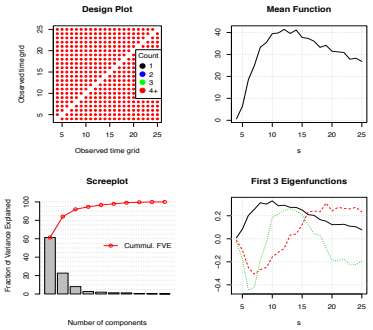

FIG1. Diagnostic Plots of FPCA TVE threshold 0.9999

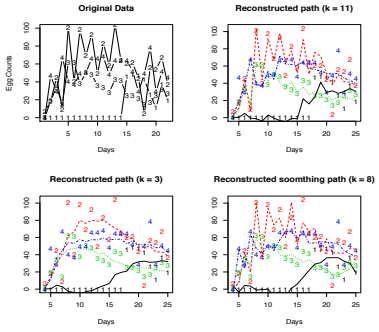

FIG2 Path Plots of 4 observations

Comparing to PCA

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

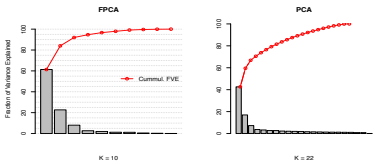

FIG3. Scree Plots of FPCA and PCA

Path plot, *Fig2* , shows how well each observations explained by chosen FPCs. The four observations are randomly chosen, and the path plot shows FPCA has select 10 variables which explain very close to original data. When we want to set our K selection criteria is that FVE 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, then FPCA reduces 8 dimensions with achieving FVE is 0.9999.

Poisson Modeling on Lifetime

Main interest is the relationship of lifetime and egg laying pattern. The lifetime is number of days they lives (Count), so I conduct poisson regression lifetime on each days with log link. Here is three approaches of modeling.

1. Poisson model (22 variables)
   Lifetime ~ each 22 days egg count:
2. FPCA Poisson model (3 variables)
   Lifetime ~ 3 FPCs (ξ_i)
3. PCA Poisson model: (13 variables)
   Lifetime ~ 13 PCs

Both FPCA and PCA FVE threshold 0.90

<Results>

FPCA Poisson model is preferable to other two model.

In Residual Plots, *Fig4* 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.

All three model's Quasi R squared and cross-validation prediction error, *Table1,* are very closed. Remind of FPCA chosen 3 variables and PCA chosen 13 variables, then FPCA Poisson model perform efficiently precise

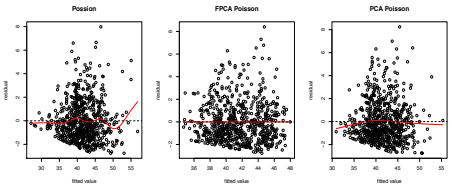

FIG4. Pearson Residual Plots

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

Table1. Goodness of Fit

The FPCA Poisson model has first 3 FPCs. The 1st FPC is the direction in feature space along which projections have the largest variance, so usually follows mean function. The 2nd FPC is the direction which maximizes variance among all directions orthogonal to the 1st. The 3rd FPC is the variance-maximizing direction orthogonal to the second components. The final model implies the egg-laying has negative relationship to lifetime.

	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

Table2. FPCA Poisson Model Summary Table

<Future Work>

- Try Medfly data which has 101 days records, and how the FPCA perform the dimensionality reduction and functional regression
- Adding smoothing techniques to estimate each observations' egg patterns before/or during FPCA.