Principal Components Regression

$$y_i = \beta_0 + \sum \beta_j x_{ij} + \epsilon_i$$

Usual least squares may be inappropriate if

- x is high dimensional, especially in comparison to sample size
- x is highly correlated
- main focus is on future prediction

In these cases, we would like to find some way to reduce the amount of covariate informaion

- Variable selection procedures (unstable)
- Principle components regression

Principal Components Regression

$$y_i = \beta_0' + \sum_{j=1}^{p'} \beta_j' \alpha_{ij} + \epsilon_i$$

Advantages:

- α_{ij} are uncorrelated – stability of estimates
- dimension reduction
- stable variable selection

Choosing p'

- Use enough PCs to caputure 90% of variation
- Maximize adjusted R^2
- Do variable selection (ie, not necessarily leading PCs)

Principal Components Regression

$$y_i = \beta_0 + \sum \beta_j x_{ij} + \epsilon_i$$

Instead, represent

$$\mathbf{x}_i = \sum_{j=1}^{p} \alpha_{ij} \xi_j$$

for ξ_j the principle components of $\mathbf{x}$.

Then model

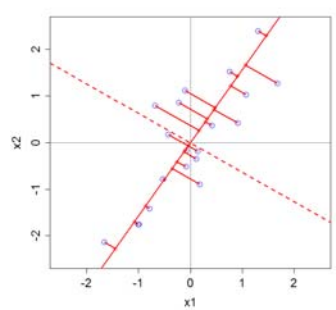
$$y_i = \beta_0' + \sum_{j=1}^{p'} \beta_j' \alpha_{ij} + \epsilon_i$$

for some $p' < p$.

Principal Components Regression

$$y_i = \beta_0' + \sum_{j=1}^{p'} \beta_j' \alpha_{ij} + \epsilon_i$$

This is a bet that most variation in y is in direction of large variation in x.



Bias: variation in y due to PCs not included in the model

If $n > p$, we can consider all p Principle Components. Otherwise, we can't escape this bet.

Models for "Miss Congeniality"

Predict rating from ratings of other movies.

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
Independence Day	-0.249	0.266	-0.103		-0.106	-0.113
Patriot	-0.259		0.308	0.484	-0.186	0.272
Day After Tomorrow	-0.317	0.135	0.119	-0.658	-0.343	0.201
Pirates Caribbean	-0.149		-0.242		-0.338	0.553
Pretty Woman	-0.232	-0.427	-0.303			-0.363
Forrest Gump	-0.113			0.286	-0.299	-0.170
The Green Mile	-0.162			0.303	-0.277	
Con Air	-0.301	0.318	-0.223		0.441	0.105
Twister	-0.302	0.138	-0.165	-0.192	-0.265	-0.489
Sweet Home Alabama	-0.295	-0.599	-0.181	-0.172	0.234	0.350
Pearl Harbor	-0.366	-0.131	0.743		0.258	-0.113
Armageddon	-0.333	0.215			0.195	
The Rock	-0.244	0.307	-0.221	0.192	0.295	
What Women Want	-0.286	-0.297	-0.111	0.183	-0.202	-0.101

BTRY 6150: Applied Functional Data Analysis: Functional Principal Components Regression

Functional PCR

For functional data analysis, we can also employ principal components regression.

$$x_i(t) = \sum f_{ij}\xi_j(t)$$

And we can model

$$y_i = \beta_0 + \sum \beta_j f_{ij} + \epsilon_i$$

or

$$\mathbf{y} = F\beta + \epsilon$$

- reduces to a standard linear regression problem on orthogonal covariates
- avoids the need for cross-validation (assuming number of PCs is fixed)

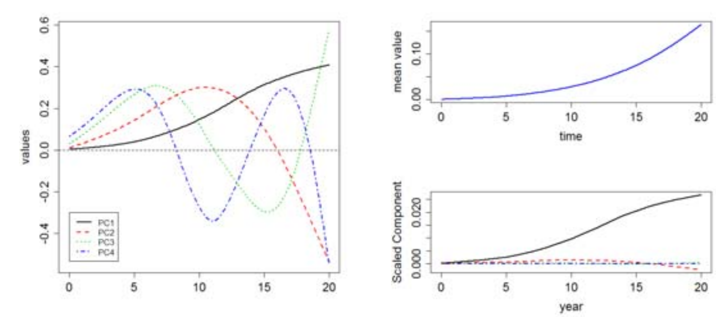
By far the most theoretically studied method.

BTRY 6150: Applied Functional Data Analysis: Functional Principal Components Regression

```
Coefficients:
                       Estimate Std. Error  t value Pr(>|t|)
(Intercept)           3.418e-17  8.105e-03 4.22e-15 1.000000
netpca2$scoresComp.1 -2.484e-01  4.413e-03  -56.277  < 2e-16 **
netpca2$scoresComp.2 -1.914e-01  8.580e-03  -22.310  < 2e-16 **
netpca2$scoresComp.3 -1.169e-01  9.183e-03  -12.734  < 2e-16 **
netpca2$scoresComp.4 -5.635e-02  9.314e-03   -6.050 1.50e-09 **
netpca2$scoresComp.5  2.117e-02  1.017e-02    2.081 0.037426 *
netpca2$scoresComp.6  2.050e-02  1.042e-02    1.967 0.049198 *
netpca2$scoresComp.7 -2.494e-02  1.062e-02   -2.348 0.018906 *
netpca2$scoresComp.8 -6.175e-02  1.102e-02   -5.601 2.19e-08 **
netpca2$scoresComp.9  5.444e-03  1.131e-02    0.481 0.630175
netpca2$scoresComp.10 -2.448e-02 1.154e-02   -2.121 0.033928 *
netpca2$scoresComp.11  2.787e-03 1.176e-02    0.237 0.812658
netpca2$scoresComp.12  2.826e-02 1.285e-02    2.200 0.027825 *
netpca2$scoresComp.13  5.054e-02 1.315e-02    3.844 0.000122 **
netpca2$scoresComp.14 -6.644e-02 1.336e-02   -4.972 6.74e-07 **
```

BTRY 6150: Applied Functional Data Analysis: Functional Principal Components Regression

Principal Components of Market Data



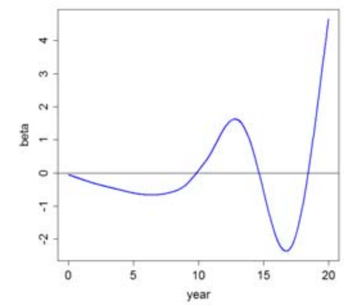

```
> HiFi.pca$varprop
 [1] 9.275017e-01 6.184101e-02 9.802453e-03 5.535989e-04 2.8
```

Principal Components Regression of Market Data

```
> pcamod = lm(y~HiFi.pca$scores[,1:5])
> summary(pcamod)
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              0.35925    0.01110  32.365 1.46e-14 ***
HiFi.pca$scores[, 1:5]1  0.25340    0.04328   5.854 4.19e-05 ***
HiFi.pca$scores[, 1:5]2 -0.74783    0.16763  -4.461 0.000538 ***
HiFi.pca$scores[, 1:5]3  0.93695    0.42103   2.225 0.043003 *
HiFi.pca$scores[, 1:5]4 -4.20080    1.77167  -2.371 0.032624 *
HiFi.pca$scores[, 1:5]5  3.17800    2.45745   1.293 0.216870
---
Signif. codes:  0 Ś***Š 0.001 Ś**Š 0.01 Ś*Š 0.05 Ś.Š 0.1 Š Š 1

Residual standard error: 0.04964 on 14 degrees of freedom
Multiple R-squared: 0.8259,     Adjusted R-squared: 0.7638
F-statistic: 13.28 on 5 and 14 DF,  p-value: 6.733e-05
```


HiFi Data

```
pcabeta = pcamod$coef[2]*HiFi.pca$harmonics[1]+
    pcamod$coef[3]*HiFi.pca$harmonics[2]+
    pcamod$coef[4]*HiFi.pca$harmonics[3]+
    pcamod$coef[5]*HiFi.pca$harmonics[4]+
    pcamod$coef[6]*HiFi.pca$harmonics[5]

plot(pcabeta)
```



Functional Regression Interpretation

$$y_i = \beta_0 + \sum \beta_j f_{ij} + \epsilon_i$$

Recall that $f_{ij} = \int x_i(t)\xi_j(t)dt$ so that

$$y_i = \beta_0 + \sum \int \beta_j \xi_j(t) x_i(t) dt + \epsilon_i$$

So we can interpret

$$\beta(t) = \sum \beta_j \xi_j(t)$$

and use the model

$$y_i = \beta_0 + \int \beta(t) x_i(t) dt + \epsilon_i$$

Confidence Intervals

$$\hat{\beta}(t) = \sum \hat{\beta}_j \xi_j(t)$$

so

$$\text{Var}\left[\hat{\beta}(t)\right] = [\xi_1(t) \ \cdots \ \xi_k(t)]\,\text{Var}\left[\hat{\beta}\right]\begin{bmatrix}\xi_1(t)\\ \vdots \\ \xi_k(t)\end{bmatrix}$$

and recall that

$$\text{Var}[\beta] = \sigma^2 \left(F^T F\right)^{-1}$$

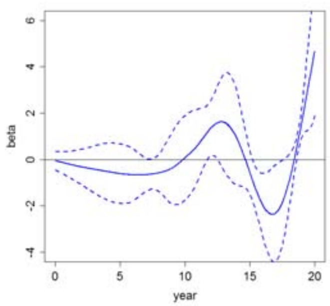
When the PCs are not smoothed, $F^T F$ is diagonal and

$$\text{Var}\left[\hat{\beta}(t)\right] = \sum mbox Var\left[\hat{\beta}_j\right]\xi_i^2(t)$$

When smoothed PCs are used, $F^T F$ may no longer be orthogonal.

Confidence Intervals for HiFis

```
pcacoefvar = summary(pcamod)$coefficients[,2]^2

pcabetavar = pcacoefvar[2]*HiFi.pca$harmonics[1]^2+
    pcacoefvar[3]*HiFi.pca$harmonics[2]^2+
    pcacoefvar[4]*HiFi.pca$harmonics[3]^2+
    pcacoefvar[5]*HiFi.pca$harmonics[4]^2+
    pcacoefvar[6]*HiFi.pca$harmonics[5]^2
```




```
lines(pcabeta+2*sqrt(pcabetavar))
lines(pcabeta-2*sqrt(pcabetavar))
```

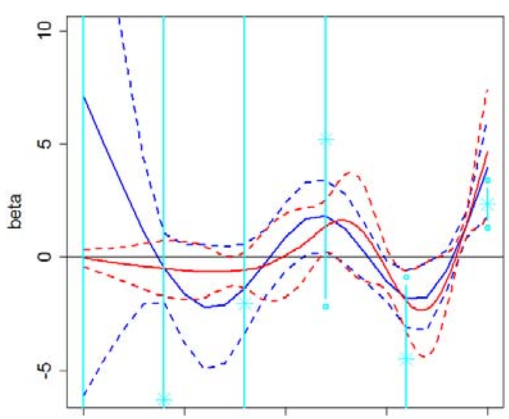

Summary

- Principal components regression = dimension reduction technique
- functional Principal components regression works exactly the same way
- re-interpretation as a basis expansion for $\beta(t)$
- standard errors for $\beta(t)$ calculated from linear regression covariance
- Selecting number of PCs = choice of basis
- Alternative is smoothing; performance depends on how well truth matches your assumptions.

Comparisons

Adjusted R^2:

Pointwise	PCA	Smooth
0.7575	0.7368	0.7705



Problems of Inference

Focus so far on *exploratory* data analysis

- Estimates of relationships or smooths
- Pointwise confidence intervals

Alternative way of assessing meaning:

What would the estimate look like if there were no relationship?

especially because we are using non-parametric methods

This is exactly a hypothesis test

$$H_0 : \beta(t) = 0$$

Permutation Tests

$$H_0 : \beta(t) = 0$$

Problem: smoothing, integrals, basis expansions are not accounted for by traditional tests.

Problem: I don't want to think too hard to create a test

Solution: permutation tests.

More Formally

We need

1. Data $\{y_i, x_i\}_{i=1}^n$
2. A test statistic $T(\mathbf{y}, \mathbf{x})$ measuring the strength of the relationship between y_i and x_i.

Do i in $1, .., B$:

1. Create new $\mathbf{y}'$ by *randomly permuting* the values in y.
2. Record $T_i = T(\mathbf{y}', \mathbf{x})$.

If $T_{obs} = T(\mathbf{y}, \mathbf{x})$ is greater than $1 - p\%$ of the T_i, reject H_0.

Permutation Tests

Given data $\{y_i, x_i\}_{i=1}^n$ we want a non-parametric test of

H_0: y is independent of x

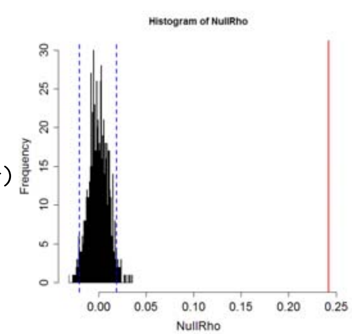
Idea: if H_0 is true, the *pairing* becomes important.

So: try randomly pairing y_j with x_k and see what your model is like

Examples

```
> x = netflix[,1]
> y = netflix[,2]

> Rho = cor(x,y)

> NullRho = rep(0,1000)

> for(i in 1:1000){
>     s = sample(10000)
>     NullRho[i] = cor(x[s],y)
> }

> hist(NullRho))
> abline(v=Rho)
```

```
> C1 = quantile(NullRho,0.025)
> C2 = quantile(NullRho,0.975)
> abline(v=c(C1,C2))
```



Functional Linear Regression and Permutation F-Tests

We have data $\{y_i, x_i(t)\}$ with a model

$$y_i = \alpha + \int \beta(t) x_i(t) dt + \epsilon_i$$

and $\hat{\beta}(t)$ estimated by penalized least squares

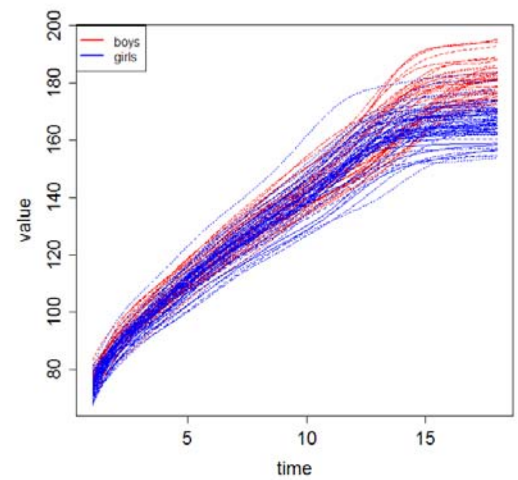
Choose a the usual F statistic as a measure of association:

$$F = \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

Note that scaling doesn't matter.

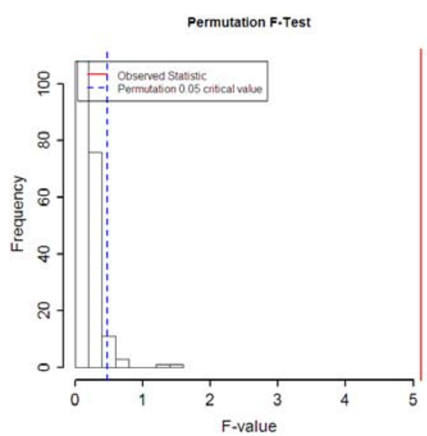
Functional Responses

Are the following populations of functions different?



Significance of HiFi Penetration Prediction

```
> Fpermres = Fperm.fd(y,xfdlist,betalist)
```



Permutation t-Tests

$$t(x_1, x_2) = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{1}{n_1}\mathrm{Var}(x_1) + \frac{1}{n_2}\mathrm{Var}(x_2)}}$$

gives relative difference between two groups of observations.

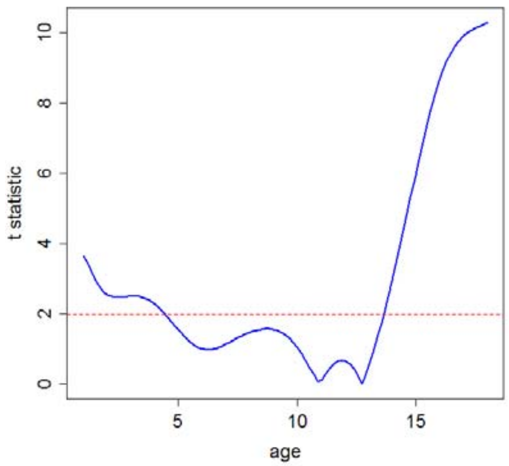
Functional equivalent is a point-wise statistic

$$t(x_1(t), x_2(t)) = \frac{\bar{x}_1(t) - \bar{x}_2(t)}{\sqrt{\frac{1}{n_1}\mathrm{Var}(x_1(t)) + \frac{1}{n_2}\mathrm{Var}(x_2(t))}}$$

Interpretable in its own right.

Example

Is there a difference between the heights of boys and girls?



Max t

Take $T(x_1(t), x_2(t)) = \max t(x_1(t), x_2(t))$.

- provides a single number to test
- can run a permutation test based on $T(x_1(t), x_2(t))$
- powerful against local violations of the null
- may not be so powerful against consistent, small differences between groups of curves
- Other statistics such as $\int t(x_1(t), x_2(t))$ could also be used
- We can also look at the pointwise critical values of the permutation test.

Problems

Usual critical value = 1.98 does not account for

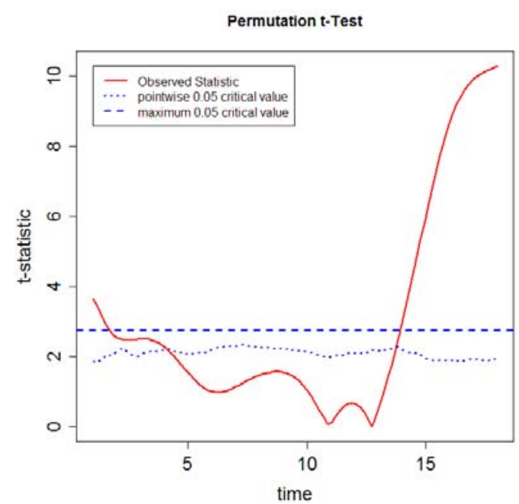
- multiple testing
- correlation between tests

Additionally, making assumptions about distributions of *curves* is difficult.

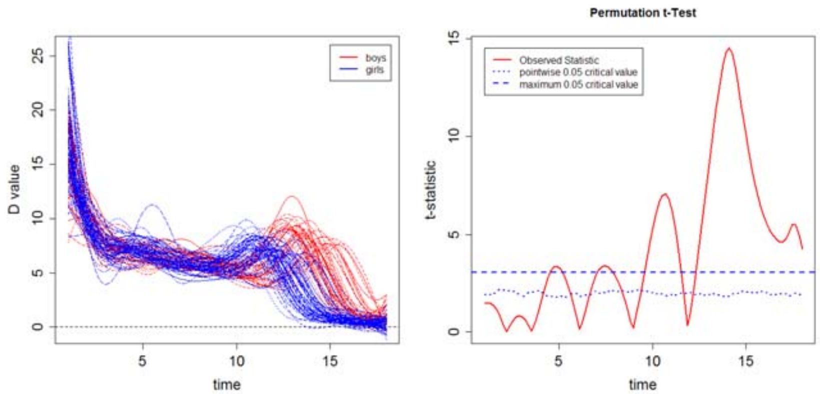
$\Rightarrow$ need permutation tests, but what statistic?

Boys v Girls

```
> res = tperm.fd(hgtffd,hgtmfd)
```



What about Rates of Growth?



Summary

- Permutation tests avoid need for parametric assumptions
- Requires null assumption of independence of y and covariates (cannot test aspects of distribution, eg mean or variance)
- Application in functional linear regression: permutation F tests
- Functional t-tests for differences between groups
  - functional t-statistic - pointwise interpretability between groups
  - base permutation tests on maximum, mean, pointwise values of $t(t)$