**Projection Pursuit**: Project $p$-dimensional data into lower dimensional spaces to reveal interesting features of the data.

**Basic steps:**

1. Choose the dimension of the projection ($k=2$)

2. Choose a criterion to optimize (Projection index)

3. Evaluate the criterion for a number of projections, possibly chosen at random

4. Starting with the best projection from step (3), optimize the criterion to find a local optimum

5. Graphically display the optimal projection

6. Repeat steps (3) through (5) starting at different initial projections
Optimization criterion:

Not affected by "sphering" the data.

Suppose $X_j = \left[x_{j1}, \ldots, x_{jn_j}\right]'$ are i.i.d. with $E(X_j) = \mu$ and $V(X_j) = \Sigma$.

Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0$ be the eigenvalues of $\Sigma$.

Define $A = \left[\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p\right]'$ to be the eigenvectors of $\Sigma$. Use $\bar{X} = \frac{1}{n} \sum_{j=1}^{p} (X_j' - \bar{X})X_j'X_j - \bar{X}'X_j$ to estimate $\Sigma$.

So that the principal component scores are i.i.d. with $E(\bar{z}_j) = 0$, $V(\bar{z}_j) = I$.

"Sphered data" $\bar{X} = DA'(X_j - \mu)$.
Importance of step 3:

- Generally the optimality criterion has many "local" modes.
- Grand Tour


\[ \mathbf{y}_j = \begin{bmatrix} y_{1j} \\ \vdots \\ y_{kj} \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1p} \\ \vdots & \vdots & & \vdots \\ b_{k1} & b_{k2} & \cdots & b_{kp} \end{bmatrix} \begin{bmatrix} z_{1j} \\ \vdots \\ z_{pj} \end{bmatrix} \]

where \[ b_i = \begin{bmatrix} b_{i1} \\ \vdots \\ b_{ip} \end{bmatrix} \quad \text{and} \quad b_k = \begin{bmatrix} b_{k1} \\ \vdots \\ b_{kp} \end{bmatrix} \]

satisfy \[ b_r^t b_s = 0 \quad r \neq s \]
\[ b_r^t b_r = 1 \]

- Note that \( \mathbf{y}_j \) is sphered.
- Choose \( b_1, \ldots, b_k \) to optimize the criterion.
(iii) Let the grand tour pick a random projection.

(iv) Then apply step 4.

Repeat (iii) and (iv) a number of times and select the "best" result.

(step 4) Numerical optimization

- Evaluate the criterion for scores \( Y_1, ..., Y_n \) at projections close to the current projection.
- Follow path of steepest ascent.
- Computer time increases as dimensionality \( p \) increases.

Some optimality criteria:

**Entropy index:**

\[
f(y) \quad \text{density function evaluated at } \ y = \begin{bmatrix} y_1 \\ y_k \end{bmatrix} = B \Xi_{p \times 1}
\]

Pick \( B \) to maximize

\[
E \left\{ \left[ f(y) \right]^d \right\} = \int \left[ f(y) \right]^{d+1} dy
\]

- Use kernel methods to compute a "smoothed" estimate \( \hat{f}(y) \) from the "data" \( Y_1, ..., Y_n \).
- Some other estimate of \( f(y) \).
Kernel estimation of a density

\[ \hat{f}(y) = \frac{1}{n \omega} \sum_{i=1}^{n} \phi \left( \frac{y - y_i}{\omega} \right) \]

where \( \omega > 0 \) is the "band width" or "smoothing" parameter

\[ \phi(\cdot) \] non-negative function decreasing monotonically to zero as \( t_i \to \infty \), \( t_i \to -\infty \)

for any \( i = 1, \ldots, k \)

\[ t_i' = (t_i, \ldots, t_k) \]

e.g. use \( \phi(\cdot) = \frac{1}{\sqrt{2\pi}} \left[ \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \right] \)

\( \alpha = 1 \) Freeman-Tukey Index

Choose \( B = \begin{bmatrix} b_{11} & \cdots & b_{1p} \\ \vdots & \ddots & \vdots \\ b_{k1} & \cdots & b_{kp} \end{bmatrix} \)

to maximize

\[ E\{f(y)\} = \int [f(y)]^2 \, dy \]

- reveals clusters? non-normality?
- minimized by a parabolic density. For \( k = 1 \)
  
  \( f(y) = \max \left\{ 0, \frac{3(5 - y^2)}{20 \sqrt{5}} \right\} \)

\[ -\sqrt{5} \quad 0 \quad \sqrt{5} \]
(α→0) Shannon–Weaver Index

\[ E \left\{ \log(f(y)) \right\} = \int_{\mathbb{R}^n} \log(f(y)) f(y) \, dy \]

... minimized when

\[ f(y) = \frac{k}{17} \left( \prod_{i=1}^{k} \sqrt{2\pi} e^{-\frac{1}{2}y_i^2} \right) \]

... the "uninteresting" standard normal distribution

Fisher information:

\[ E \left\{ \left( \frac{f'(y)}{f(y)} \right)^2 \right\} = E \left\{ \frac{d \log(f(y))}{dy} \right\} \]

\[ = \int_{\mathbb{R}^n} \left( \frac{f'(y)}{f(y)} \right)^2 f(y) \, dy \]

Distance from "normality":

Hellinger distance:

\[ \int_{\mathbb{R}^n} \left[ \sqrt{f(y)} - \sqrt{\phi(y)} \right]^2 \, dy \]

Legendre index:

\[ \int_{\mathbb{R}^n} \left[ f(y) - \phi(y) \right]^2 \frac{1}{\phi(y)} \, dy \]

Hall index: (Hermite index)

\[ \int_{\mathbb{R}^n} \left[ f(y) - \phi(y) \right]^2 \, dy \]
Natural Hermite index:

\[ I^N = \int_{\mathbb{R}^k} \left[ f(y) - \Phi(y) \right]^2 \phi(y) \, dy \]

approximate with expansion using Hermite polynomials with estimated coefficients

\[ \sum_{i=1}^{M} \hat{\alpha}_i \, p_i(y_j) \]

\[ \hat{\alpha}_i = \frac{1}{n} \sum_{k=1}^{n} p_i(y_{jk}) \phi(y_{jk}) \]

Other indices in XGobi:

Central mass

\[ \hat{E}(\Phi(y)) = \frac{1}{n} \sum_{j=1}^{n} \phi(y_j) \]

maximized

Central holes

\[ - \hat{E}(\Phi(y)) = -\frac{1}{n} \sum_{j=1}^{n} \phi(y_j) \]
$\mu^2(A, B \pm c) = (E_A + E_B \pm E_c)^2 - (P_A + P_B \pm P_c)^2$

$\mu^2(A, \pm c) = (E_A \pm E_c)^2 - (P_A \pm P_c)^2$

$E_A$ is energy of particle $A$ (10$^9$ ev)
$P_A$ is momentum of particle $A$

$(\rho)^2$ represents an inner product $\rho/\rho$

Assignment to $\pi_1^+$ and $\pi_2^+$ was done randomly.
Figure 5. 2-D Particle Physiscs Data. (a) First two principal components. (b) Projection similar to that found by Friedman and Tukey (1974).
Summary:

(1) Triangular shaped region
   - largely determined by
     \[ x_3 = \mu^2(p, \pi^-) \]
     \[ x_5 = \mu^2(p, \pi^+) \]
   - constraint on the total squared invariant mass

(2) Six one-dimensional arms extending from the vertices
   - two from each vertex
   - a "line" on the projection could be a higher dimensional object
     - brush/color the points
     - look at many projection (grand tour)
• Three aberrant cases
  - examine these cases
  • "bad" data
  • unexplained phenomena
The grand tour may fail to reveal certain types of patterns or "structure"

Can a "grand tour" find a pattern where no pattern exists?
Conditioning or "slicing"  

1. Choose only those cases with  
   \[ a_i < x_{ij} < b_i \]  
   \[ a_j < x_{ij} < b_j \]  

2. Brushing  

References:  


Regression on principal components

Suppose a model must be constructed to predict a response \((Y)\) from \(p\) explanatory variables \(X_1, X_2, \ldots, X_p\)

Consider the multiple regression model

\[
Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon
\]

If there are strong colinearities among the explanatory variables, the estimated coefficients may be "unreliable" (some linear combinations of the estimated coefficients have large variances)

Why?

Consider a special case with \(p=2\) explanatory variables and no intercept.

\[
y_j = \beta_1 x_{1j} + \beta_2 x_{2j} + \epsilon_j
\]

\[
\begin{bmatrix}
Y_1 \\
\vdots \\
Y_n
\end{bmatrix}
= 
\begin{bmatrix}
x_{11} & x_{12} \\
\vdots & \vdots \\
x_{1n} & x_{2n}
\end{bmatrix}
\begin{bmatrix}
\beta_1 \\
\beta_2
\end{bmatrix}
+ 
\begin{bmatrix}
\epsilon_1 \\
\vdots \\
\epsilon_n
\end{bmatrix}
\]

or

\[
y = X \beta + \epsilon
\]

Then if \(\epsilon \sim \mathcal{N}_n(0, \sigma^2 I)\)

\[
\begin{bmatrix}
b_1 \\
b_2
\end{bmatrix}
= (X'X)^{-1}X'y
\sim \mathcal{N}_2\left(\begin{bmatrix}
\beta_1 \\
\beta_2
\end{bmatrix}, \sigma^2(X'X)^{-1}\right)
\]
If \( X_1 \)'s and \( X_3 \)'s are highly colinear
\[
\begin{array}{c}
X_2 \\
\vdots \\
X_1
\end{array}
\]
then \((X'X)\) has a small positive eigenvalue. Consequently, \((X'X)^{-1}\) has a large positive eigenvalue.

Remedies:

1. **Reduce the number of explanatory variables.** In this case eliminate either \( X_1 \) or \( X_2 \)?

   This may produce a very good prediction formula, but it will not describe the joint effect of \( X_1 \) and \( X_2 \) on \( Y \) as \( X_1 \) and \( X_2 \) increase together.

2. **Compute a new predictor variable which is an “optimal” linear combination of \( X_1 \) and \( X_2 \), say the first principal component and regress \( Y \) on this new variable \( Y_i = \beta_1 x_i + e_i \).**
Example: Hald cement data

**Oroper + Smith, Applied Regression Analysis**, pp 327-332
also appendix B.

\[ X_1 = \text{amount of tricalcium aluminate} \]
\[ 3\text{CaO} \cdot \text{Al}_2\text{O}_3 \]

\[ X_2 = \text{amount of tricalcium silicate} \]
\[ 3\text{CaO} \cdot \text{Si}_2\text{O}_5 \]

\[ X_3 = \text{amount of tetracalcium alumino ferrite} \]
\[ 4\text{CaO} \cdot \text{Al}_2\text{O}_3 \cdot \text{Fe}_2\text{O}_5 \]

\[ X_4 = \text{amount of dicalcium silicate} \]
\[ 2\text{CaO} \cdot \text{Si}_2\text{O}_5 \]

(These variables are measured as percentages of the weight of the clinkers from which the cement was made.)

<table>
<thead>
<tr>
<th></th>
<th>( Y )</th>
<th>( X_1 )</th>
<th>( X_2 )</th>
<th>( X_3 )</th>
<th>( X_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>78.5</td>
<td>7</td>
<td>26</td>
<td>6</td>
<td>60</td>
</tr>
<tr>
<td>2</td>
<td>79.3</td>
<td>1</td>
<td>29</td>
<td>15</td>
<td>52</td>
</tr>
<tr>
<td>3</td>
<td>104.3</td>
<td>11</td>
<td>56</td>
<td>8</td>
<td>20</td>
</tr>
<tr>
<td>4</td>
<td>87.6</td>
<td>11</td>
<td>31</td>
<td>8</td>
<td>47</td>
</tr>
<tr>
<td>5</td>
<td>95.9</td>
<td>7</td>
<td>52</td>
<td>6</td>
<td>33</td>
</tr>
<tr>
<td>6</td>
<td>109.3</td>
<td>11</td>
<td>55</td>
<td>9</td>
<td>22</td>
</tr>
<tr>
<td>7</td>
<td>102.7</td>
<td>3</td>
<td>71</td>
<td>17</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>72.5</td>
<td>7</td>
<td>31</td>
<td>22</td>
<td>44</td>
</tr>
<tr>
<td>9</td>
<td>93.1</td>
<td>2</td>
<td>54</td>
<td>18</td>
<td>22</td>
</tr>
<tr>
<td>10</td>
<td>115.9</td>
<td>21</td>
<td>47</td>
<td>4</td>
<td>36</td>
</tr>
<tr>
<td>11</td>
<td>83.8</td>
<td>1</td>
<td>40</td>
<td>23</td>
<td>34</td>
</tr>
<tr>
<td>12</td>
<td>113.2</td>
<td>11</td>
<td>66</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td>13</td>
<td>109.4</td>
<td>10</td>
<td>68</td>
<td>8</td>
<td>12</td>
</tr>
</tbody>
</table>
Regression on principal components

\[ X_1 = x_0 + x_1 w_{1j} + x_2 w_{2j} + e_j \]

\[ \omega_{1j} = -0.68 X_{1j} - 6.79 X_{2j} + 0.029 X_{3j} + 7.31 X_{4j} \]

\[ \omega_{2j} = 0.64 X_{1j} + 0.020 X_{2j} - 7.35 X_{3j} + 1.08 X_{4j} \]

Use least squares estimation

\[ s_j = 95.145 - 55.33 w_{1j} + 91.94 w_{2j} \]

\[ R^2 = 0.95 \]

Covariance matrix:

\[
\begin{bmatrix}
X_1 & X_2 & X_3 & X_4 \\
X_1 & 34.60 & 20.92 & -31.05 & -24.17 \\
X_3 & -31.05 & -13.89 & 41.03 & 3.17 \\
\end{bmatrix}
\]

Eigenvalues:

\[
510.8, 67.5, 12.4, 0.23
\]

-0.686
-0.948
-0.811
-0.731
Regression on Principal Components

\[ \hat{Y} = 89.2427 + 0.63714 X_1 + 0.3938 X_2 \]

(1.099) (0.093)

(0.099) - 0.7104 X_3 - 0.3046 X_4

(0.033)

(0.099)

Least squares regression on the original variables:

\( R^2 = .9824 \)

Complete model

\[ \hat{Y} = 63.3186 + 1.5431 X_1 + 0.5500 X_2 \]

(0.074) (0.092)

(0.123) 0.155 X_3

(0.074)

(0.092)

\[ \hat{Y} = 53.56009 + 1.4678 X_1 + 0.6617 X_2 \]

(2.382) (1.012)

(0.099)

Write the model in terms of the original variables:

\[ \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 (x_i - \bar{x}_i) + \hat{\beta}_2 (x_i - \bar{x}_i) + \hat{\beta}_3 (x_i - \bar{x}_i) X_{ij} + \hat{\beta}_4 (x_i - \bar{x}_i) \]

\[ = \hat{\beta}_0 + \sum_{i=1}^{p} \hat{\beta}_i \bar{x}_{ij} + \sum_{i=1}^{p} \hat{\beta}_i (x_i - \bar{x}_i) X_{ij} + \sum_{i=1}^{p} \hat{\beta}_i (x_i - \bar{x}_i) \]
Regression on principal components:

Attractive features:

1. Provides some understanding of relationships among explanatory variables
   - $X_1 + X_2 + X_3 + X_4 = \text{constant}$
   - tradeoff between $X_1$ and $X_3$
   - tradeoff between $X_2$ and $X_4$

2. Reduce variance of estimated regression coefficients

3. Better interpretation of how changes in explanatory variable affect the mean response
   
   evolved heat is reduced when
   $X_3$ is substituted for $X$
   $X_4$ is substituted for $X$
Possible drawbacks:

1. Restriction to a subspace of the original explanatory variables
   Predictions at points outside this subspace can be terrible

2. Determination of principal components does not consider the response variable
   - Components with large variances may have little correlation with \( Y \)
   - Components corresponding to near singularities may have substantial predictive ability (the third P.C. is significant in the regression)

Regression on 3 principal components
\[ R^2 = 0.981 \]
\[ \hat{Y}_j = 95.4154 - 0.5532 w_{1j} + 0.9193 w_{2j} \\ (0.659) \quad (0.030) \quad (0.083) \]
\[ - 0.7113 w_{3j} \\ (0.194) \]
\[ = 111.48 + 1.049 X_{1j} + 0.00683 X_{2j} \\ (6.12) \quad (-1.02) \quad (0.108) \]
\[ - 0.4234 X_{3j} - 0.6378 X_{5j} \\ (1.00) \quad (0.094) \]

Regression on 2 principal components
\( (R^2 = 0.954) \)
\[ \hat{Y}_j = 95.4154 - 0.5532 w_{1j} + 0.9193 w_{2j} \\ (0.984) \quad (0.045) \quad (0.125) \]
3. Biased estimates
   - regression parameters
   - mean response
   - predictions

4. Interpretation of P.R. Scores

Other biased regression methods

  Latent root analysis

  Bayesian regression

  Ridge regression

References:

Draper + Smith, *Applied Regression Analysis*
2nd edition, Wiley

Gunst + Mason, *Regression Analysis and Its Applications*, Marcel Dekker

Myers, *Classical and Modern Regression with Applications*, 2nd ed. PWS-KENT