An Introduction to Independent Components Analysis (ICA)

Anish R. Shah, CFA
Northfield Information Services
Anish@northinfo.com

London
Nov 9, 2010
Overview of Talk

• Review principal components

• Introduce independent components
Part I: Principal Components Analysis (PCA)

• At each step, PCA finds the direction that explains the most remaining variation
1\text{st} \text{ principal component of a cylinder}
PCA on the Florida Keys
Principal components of 2 years of annual mutual fund returns

Returns of top (by 10 yr return) US diversified mutual funds
1 outlier excluded
Principal Component Analysis (PCA)

• Setup:
  – Centered observations $y_1 \ldots y_N$
    Each observation is a vector of length $T$
    
    $y_1$ = a stock, as 5 years of monthly returns ($T = 60$)
    $y_1$ = a hectare of the Florida Keys, as latitude & longitude ($T = 2$)
    $y_1$ = a black & white digital photo, as pixel intensities ($T = 10^6$)

• Goal:
  – Fit a linear model that minimizes the squared error between the model and observations
  – Imagine regressing many stocks against a single independent variable
    PCA finds the independent variable that explains the most on average
Two ways to center observations

1- Subtract the average (over t) of each observation, e.g. stock covariance model

2- Subtract the average (over n) of each coordinate, e.g. shape of Florida Keys

Choice of method depends on the application
PCA’s error criterion

• Consider the familiar linear model
  \[ y_i(t) = \beta_i \cdot x(t) + \text{error} \]

• Regression
  – \( SE_i = \text{squared error for observation } i = \sum_{t=1..T} [y_i(t) - \beta_i \cdot x(t)]^2 \)
  – Given \( x \) and \( y \), regression sets \( \beta \) to minimize squared error
    \[ \beta_i^* = x^T y_i / x^T x \]

• PCA
  – \( TSE = \text{total error over all observations} = \sum_{i=1..N} SE_i \)
  – PCA finds the vector \( x \) (and the \( \beta_i \)'s) that minimize TSE
  – To get the next component, repeat on the residuals \( \{y_i - \beta_i \cdot x\} \)
PC’s come from eigenvectors of the matrix of 2\textsuperscript{nd} moments

- \( Y = (T \times N) \) matrix. \( N \) observations, each of length \( T \)

- \( C = Y^T Y = (N \times N) \) matrix of 2\textsuperscript{nd} moments across \( t \)
  - \( C_{i,j} = \sum_{t=1}^{T} y_i(t) y_j(t) \)
  - e.g. the co-movement between securities \( i \) & \( j \) averaged across time

- \( \underline{C} = Y Y^T = (T \times T) \) matrix of 2\textsuperscript{nd} moments across \( n \)
  - \( \underline{C}_{i,j} = \sum_{k=1}^{N} y_k(i) y_k(j) \)
  - e.g. the co-movement between \textit{periods} \( i \) & \( j \) averaged across securities

- \( C \) and \( \underline{C} \) have the same non-zero eigenvalues and yield the same PC’s
  i.e. covariance of 5000 stocks over 60 months can be analyzed as a \( 60 \times 60 \) matrix instead of a \( 5000 \times 5000 \)
PC’s come from eigenvectors of the matrix of 2\textsuperscript{nd} moments (cont.)

- $Y = (T \times N)$ N observations, each of length T
  
  $C = Y^T Y$ (size $N \times N$) \quad $\bar{C} = Y Y^T$ (size $T \times T$)

- $\lambda_i$ = the $i$’th largest eigenvalue of $C$ or $\bar{C}$
  
  $v_i = (\text{length } N)$ normalized eigenvector of $C$ corresponding to $\lambda_i$
  
  $\bar{v}_i = (\text{length } T)$ normalized eigenvector of $\bar{C}$ corresponding to $\lambda_i$
  
  $C = \sum_i \lambda_i v_i v_i^T$ \quad $\bar{C} = \sum_i \lambda_i \bar{v}_i \bar{v}_i^T$

- $v_i = Y^T \bar{v}_i / (\lambda_i)^{\frac{1}{2}}$ \quad $\bar{v}_i = Y v_i / (\lambda_i)^{\frac{1}{2}}$

- $p_i = (\text{length } T)$ $i$\textsuperscript{th} normalized principal component = $v_i = Y v_i / (\lambda_i)^{\frac{1}{2}}$
  
  $e_i = (\text{length } N)$ exposures to the component = $Y^T v_i = (\lambda_i)^{\frac{1}{2}} v_i$

- Average squared error explained by the component = $\lambda_i / (T \times N)$
PCA (cont.)

• More volatile observations have greater impact in determining components
  – To counteract, reweight \( y_i(t) \leftarrow w_i y_i(t) \)
    • e.g. \( w_i = 1 / \sqrt{y_i^T y_i} \) maximizes average correlation
    • \( w_i = \sqrt{\text{mkt cap}_i} \) weights squared error by cap
    • Exposures for the original observations are the reweighted observation’s exposures divided by the weight

• 2 views of PCA
  – a low dimensional representation of something high dimensional, e.g. stock return covariance
  – a way to separate features from noise, e.g. extracting the structural part of stock returns

• PCA yields factors uncorrelated with one another
Uncorrelated doesn’t mean independent

\[ f(x) = x \text{ and } g(x) = x^2 \]

f and g uncorrelated, but \[ g = f^2 \]
An application: face recognition

• Train the model
  – Start with a large # of pictures of faces
  – Calculate the PC’s of this set – called “eigenfaces”
  – For each person, take a reference photo and calculate its loadings on the PC’s

• Model in operation
  – Person looks into camera
  – Compare the image’s eigenface loadings to the reference photo’s
  – If close enough, accept as a match

Image Source: AT&T Laboratories Cambridge
Part II: Independent Components

- Goal: extract the signals driving a process
  
  - Cocktail party problem – separate the sound of several talkers into individual voices
  
  - Stock market returns – extract signals that investors use to price securities, fit predictive model for profit
    
    
An example: Observe 4 linear combinations of 4 signals
Principal Components
Independent Components
ICA – Independent Component Analysis

- **Similar Setup**
  - Assume there exist independent signals
    \[ S = [s_1(t), \ldots, s_N(t)] \]
  - Observe only linear combinations of them, \( Y(t) = A S(t) \)
  - Both \( A \) and \( S \) are unknown!
  - \( A \) is called the mixing matrix

- **Goal**
  - Recover the original signals \( S(t) \) from \( Y(t) \)
  - ie. find a linear transformation \( L \), ideally \( A^{-1} \), such that
    \[ LY(t) = S(t) \] (up to scale and permutation)
ICA – Basic idea

• First get rid of correlation – “whitening”
  – Apply a linear transformation $N$ to decorrelate and normalize the signals: $(NY)^TNY = I$. Let $Z = NY$
  – The whitening transformation isn’t unique – any rotation of whitened signals is white:
    • $W$ rotation $\rightarrow W^TW = I \rightarrow (WZ)^T(WZ) = Z^T(W^TW)Z = Z^TZ = I$
  – Principal components are one source of whitened signals

• Then address higher order dependence
  – Find a rotation $W$ that makes the whitened signals independent, i.e. the columns of $WZ$ independent
  – The optimization problem is $\text{minimize}_W \text{dep}(WZ)$
    • where $\text{dep}(M)$ is a measure of the dependency between the columns of $M$
    • s.t. $W^TW = I$ (W is a rotation)
Notions of independence:

1. Nonlinear decorrelation

• Signals are already decorrelated by whitening
  – \( E[u \cdot v] = E[u] \cdot E[v] \)

• Know that for independent signals \( u \) & \( v \),
  – \( E[g(u) \cdot h(v)] = E[g(u)] \cdot E[h(v)] \) for all functions \( g \), \( h \)

• Are there functions \( \hat{g} \) & \( \hat{h} \) whose nonlinearity captures most of the higher order dependence?
  – Ans: How well a particular function works depends on the shape of the data distribution’s tails

• \( \text{dep}(M) = \) magnitude of the difference between \( E[\hat{g} \cdot \hat{h}] \) and \( E[\hat{g}] \cdot E[\hat{h}] \) when \( \hat{g} \) & \( \hat{h} \) are applied to the columns of \( M \)
Notions of independence:
2. Non-Gaussianity

- Model is $Y = AS$
  - where the columns of $S$ are independent

- Central limit theorem says adding things together makes them more Gaussian

- Unmixed signals should be less Gaussian
A measure of non-Gaussianity: Kurtosis

- **Kurtosis**
  - $4^{th}$ centered moment / squared variance
  - A measure of the mass in the distribution’s tails
  - Highly influenced by outliers
  - Takes values from 0 to $\infty$, Gaussian is 3

- **Excess Kurtosis = Kurtosis – 3**
  - Takes values from $-3$ to $\infty$, Gaussian is 0
  - Maximize the absolute value to find non-Gaussian
  - $\text{dep}(M) = -1 \times |\text{excess kurtosis of columns of } M|$
Background: Information theory (Shannon 1948)

• Entropy
  – \( H(X) = - \sum p(x) \log p(x) \)
  – \# of bits needed to encode X

• Differential Entropy (for continuous random variables)
  – \( h(X) = - \int p(x) \log p(x) \, dx \)
  – For a given variance, Gaussians maximize differential entropy

• Kullback-Leibler Divergence (Relative Entropy)
  – \( D(p \| q) = \int p(x) \log [p(x)/q(x)] \, dx \geq 0 \)

• Mutual Information
  – \( I(X,Y) = D[p(x,y) \| p(x) \, p(y)] \)
    \( = \text{avg } \# \text{ of bits } X \text{ tells about } Y = \text{avg } \# \text{ of bits } Y \text{ tells about } X \)
  – \( I(X,Y) = 0 \iff X \& Y \text{ independent} \)
Information theory puzzle

- Entropy is the idea behind compression and coding

- 12 coins: 11 same, 1 heavier or lighter

Using 3 weighings of a balance, can you identify the odd coin and whether it’s heavier or lighter?

- Each use of the balance returns 1 of 3 values: =, <, >

24 equally likely configurations (12 coin positions x 2 states)

Entropy = \(-\sum p(x) \log p(x) = -\sum_{i=1..24} \frac{1}{24} \log_3 \left(\frac{1}{24}\right) = 2.89\)

It would be impossible if entropy > # of weighings
A measure of non-Gaussianity: Negentropy

- Negentropy
  - Shortfall in entropy relative to a Gaussian with the same variance
  - Useful because scale invariant

- To evaluate, need probability distribution
  - Estimate densities by expansions around a Gaussian density
  - Cumulant (moment) based (Edgeworth, Gram-Charlier)
    - sensitive to outliers
  - By non-polynomial functions, e.g. $x \exp(-x^2/2)$, $\tanh(x)$
    - more robust, but choice of functions depends on the tails

- Maximize an approximation of negentropy
  - $\text{dep}(M) = -1 \times |\text{negentropy of columns of } M|$
Mutual information

• Minimize the mutual information among the signals
  – $\text{dep}(M) = \text{mutual information of columns of } M$

• After manipulating and constraining the signals to be uncorrelated, minimizing mutual information is maximizing negentropy
Another characterization: Maximum likelihood

• Recall model $Y = As$
  
  – let $B = A^{-1}$
  
  – $p(Y) = |\det(B)| \prod_i p_i(b_i^T Y)$

• Find demixing matrix $B$ that maximizes the likelihood of the observations $Y$

• First need an (inexact) model of the $p$’s, similar to density approximation for negentropy
Connections to human wiring

• ICA can be characterized as sparse coding
  – How can signals be represented compactly (each signal loading on a few of the factors) while retaining as much information as possible?
  – A neuron codes only a few messages and rarely fires

• Edges are the independent components in pictures of nature
  – Our visual system is built to detect edges
Summary

• Incredibly clever and powerful tool for extracting information

• Fundamental – can motivate results from many different starting points
References

• Hyvärinen, A, J Karhunen, and E Oja. Independent Component Analysis. 2001


• Bishop, Christopher. Pattern Recognition and Machine Learning. 2007