CPSC 540: Machine Learning

Probabilistic PCA, Factor Analysis, Independent Component Analysis

Mark Schmidt

University of British Columbia

Winter 2018

Outline

1. Probabilistic PCA

2. Factor Analysis

3. Independent Component Analysis

PPCA: $x \overset{V}{\leftarrow} z$ $\quad P(x,z) = P(x|z)P(z)$.

隐变量 => EM Algorithm.

=> $\max \mathbb{E}_{P(z|x)}[\log P(x,z)]$ $\quad$ 认为 $P(z) = N(0,1)$

Gaussian 旋转 平移 误差

认为 x. z 之前有一个线性关系: $x = \underline{V}z + \underline{\mu} + \underline{\varepsilon}$ $\quad$ V: 旋转矩阵.

$\varepsilon \sim N(0, \sigma^2 I_D)$ $\quad$ I_D 唯单位矩阵.

$\mu_x = Vz + \mu, \ \sigma_x = \sigma^2 I \Rightarrow \underline{P(x|z) = N(Vz+\mu, \sigma^2 I)}$

p(x) 是 p(x|z) 边缘分布. p(x|z) 是高斯 => p(x) 也是高斯分布.

$p(z|x) = \frac{p(x,z)}{p(x)}$ 也是高斯 → 隐变量 $z \sim N(0,1)$

$$\mathbb{E}(x) = \mathbb{E}[Vz+\mu+\varepsilon] = V\overset{0}{\underline{\mathbb{E}(z)}} + \overset{\text{确定的}}{\underline{\mathbb{E}(\mu)}} + \overset{0}{\underline{\mathbb{E}(\varepsilon)}} = \mu$$

$$Cov(x) = \mathbb{E}[(x-\mathbb{E}[x])(x-\mathbb{E}[x])^T] = \mathbb{E}[(Vz+\varepsilon)(Vz+\varepsilon)^T] = \mathbb{E}[Vzz^TV^T + Vz\varepsilon^T + \varepsilon z^TV^T + \varepsilon\varepsilon^T]$$

$$= V\overset{Cov(z)=I}{\underline{\mathbb{E}[zz^T]}}V^T + V\overset{Cov(z,\varepsilon)=0}{\underline{\mathbb{E}(z\varepsilon^T)}} + \underline{\mathbb{E}(\varepsilon z^T)}V^T + \overset{Cov(\varepsilon)=\sigma^2 I}{\underline{\mathbb{E}(\varepsilon\varepsilon^T)}}$$

$$= VV^T + \sigma^2 I$$

一般这里要用EM解
缓这个假设太特殊

$\Rightarrow \underline{p(x) = N(\mu, VV^T + \sigma^2 I)}$ $\quad \Rightarrow \begin{cases} p(z) = N(0,1) \\ p(x|z) = N(Vz+\mu, \sigma^2 I) \\ p(x) = N(\mu, VV^T+\sigma^2 I) \end{cases}$

行列式

多元高斯分布: $p(x) = (2\pi)^{-\frac{D}{2}} \underline{|VV^T+\sigma^2 I|}^{-\frac{1}{2}} \exp\left(-(x-\mu)^T(VV^T+\sigma^2 I)^T(x-\mu)\right)$

MLE: $l(\theta) = \sum_{i=1}^{n} \log P_i(x) = -\frac{Dn}{2} 2\pi - \frac{n}{2}\log \underline{|VV^T+\sigma^2 I|} - \sum_{i=1}^{n} (x-\mu)^T \underline{(VV^T+\sigma^2 I)^{-1}}(x-\mu)$

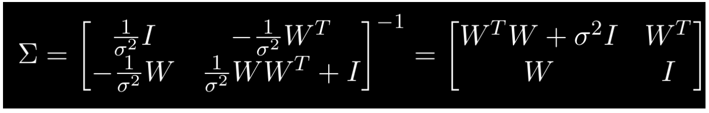
$|VV^T+\sigma^2 I| = |\frac{1}{\sigma^2}VV^T + I| \cdot |\sigma^2 I|$ $\quad$ 使用公式化简 → $\frac{1}{\sigma^2} I - \frac{1}{\sigma^2(1+\sigma^2)} VV^T$

Wondburry Matrix Identity: $(I+UV)^{-1} = I - U(I-VU)^{-1}V$

$\Rightarrow l(\theta) \Rightarrow l(V) = \sum_{i=1}^{n} x_i^T V V^T x_i + const$, s.t. $VV^T = I$

$= Tr(V^T x x^T V) + const.$

$p(\begin{bmatrix} x \\ z \end{bmatrix}) = N(\begin{bmatrix} \mu \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_{xx} & \Sigma_{xz} \\ \Sigma_{zx} & \Sigma_{zz} \end{bmatrix}) \Rightarrow$



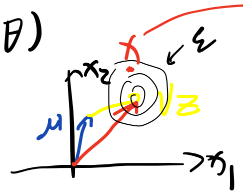
$$\Sigma = \begin{bmatrix} \frac{1}{\sigma^2} I & -\frac{1}{\sigma^2} W^T \\ -\frac{1}{\sigma^2} W & \frac{1}{\sigma^2} W W^T + I \end{bmatrix}^{-1} = \begin{bmatrix} W^T W + \sigma^2 I & W^T \\ W & I \end{bmatrix}$$

EM 算法:

E-step: $p(x) = (2\pi)^{-\frac{D}{2}} |VV^T+\sigma^2 I|^{-\frac{1}{2}} \exp\left(-(x-\mu)^T(VV^T+\sigma^2 I)^T(x-\mu)\right)$

M-step: $\max Q(\theta_{new}|\theta)$

$x = Vz + \mu + \varepsilon$:



是一个 $Vz+\mu+\varepsilon$ 的分布下的采样.

$x = Vz+\mu+\varepsilon$, $\varepsilon \sim N(0, \sigma^2 I)$: PPCA.

不同

而 FA (Factor Analysis) : $x = Vz+\mu+\varepsilon$, $\varepsilon \sim N(0, D)$. D：对角矩阵，对角线上全是σ^2

=> 同理： $p(x) = N(\mu, VV^T + D)$

FA必须用EM才能求解.

独立

	PCA	FA	ICA
Data Rotation	✓	✗	
Feature scaling	✗	✓	

PCA本就是找特征方差最大的方向 故Feature Scaling 就会导致出问题

Expectation Maximization with Many Discrete Variables

- EM iterations take the form

$$\Theta^{t+1} = \underset{\Theta}{\text{argmax}} \left\{ \sum_{H} \alpha_H \log p(O, H \mid \Theta) \right\},$$

  and with multiple MAR variables $\{H_1, H_2, \dots, H_m\}$ this means

$$\Theta^{t+1} = \underset{\Theta}{\text{argmax}} \left\{ \sum_{H_1} \sum_{H_2} \cdots \sum_{H_m} \alpha_H \log p(O, H \mid \Theta) \right\},$$

- In mixture models, EM sums over all k^n possible cluster assignments.
- In binary semi-supervised learning, EM sums over all 2^t assignments to $\tilde{y}$.

- But conditional independence allows efficient calculation in the above cases.
  - The H are independent given $\{O, \Theta\}$ which simplifies sums (see EM notes).
  - We'll cover general case when we discuss probabilistic graphical models.

Today: Continuous-Latent Variables

- If H is continuous, the sums are replaceed by integrals,

$$\log p(O \mid \Theta) = \log\left(\int_H p(O, H \mid \Theta) dH\right) \qquad \text{(log-likelihood)}$$

$$\Theta^{t+1} = \underset{\Theta}{\text{argmax}}\left\{\int_H \alpha_H \log p(O, H \mid \Theta) dH\right\} \qquad \text{(EM update)},$$

  where if have 5 hidden varialbes $\int_H$ means $\int_{H_1}\int_{H_2}\int_{H_3}\int_{H_4}\int_{H_5}$.

- Even with conditional independence these might be hard.

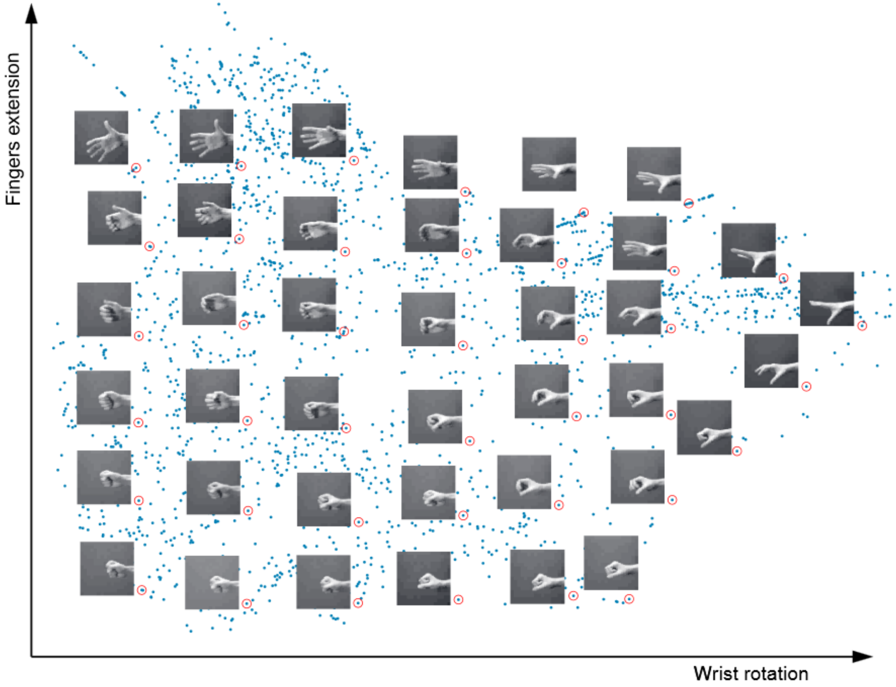
- Gaussian assumptions allow efficient calculation of these integrals.
  - We'll cover general case when we get discuss Bayesian statistics.

Today: Continuous-Latent Variables

- In mixture models, we have a discrete latent variable z^i:
  - In mixture of Gaussians, if you know the cluster z^i then $p(x^i \mid z^i)$ is a Gaussian.
- In latent-factor models, we have continuous latent variables z^i:
  - In probabilistic PCA, if you know the latent-factors z^i then $p(x^i \mid z^i)$ is a Gaussian.
- But what would a continuous z^i be useful for?
- Do we really need to start solving integrals?

Today: Continuous-Latent Variables

- Data may live in a low-dimensional manifold:

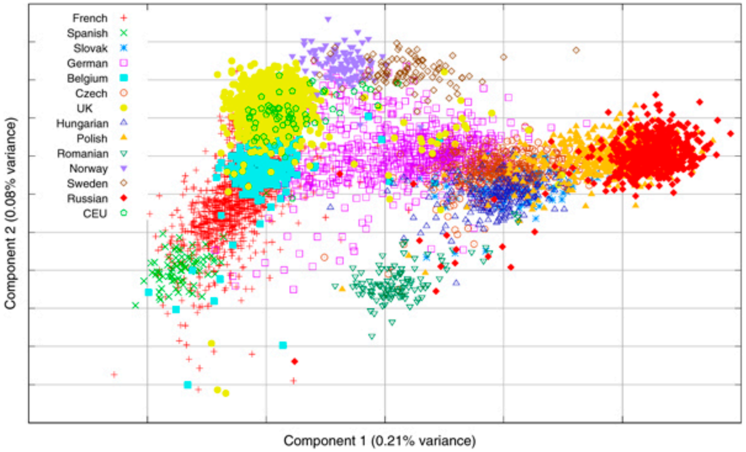


http://isomap.stanford.edu/handfig.html

- Mixtures are inefficient at representing the 2D manifold.

Principal Component Analysis (PCA)

- PCA replaces X with a lower-dimensional approximation Z.
  - Matrix Z has n rows, but typically far fewer columns.
- PCA is used for:
  - Dimensionality reduction: replace X with a lower-dimensional Z.
  - Outlier detection: if PCA gives poor approximation of x^i, could be outlier.
  - Basis for linear models: use Z as features in regression model.
  - Data visualization: display z^i in a scatterplot.
  - Factor discovering: discover important hidden "factors" underlying data.



http://infoproc.blogspot.ca/2008/11/european-genetic-substructure.html

PCA Notation

- PCA approximates the original matrix by factor-loadings Z and latent-factors W,

$$X \approx ZW.$$

  where $Z \in \mathbb{R}^{n\times k}$, $W \in \mathbb{R}^{k\times d}$, and we assume columns of X have mean 0.
- We're trying to split redundancy in X into its important "parts".
- We typically take $k << d$ so this requires far fewer parameters:

$$\underbrace{\begin{bmatrix} \quad\quad\quad\quad\quad\quad \\ \\ \\ \\ \\ \end{bmatrix}}_{X \in \mathbb{R}^{n\times d}} \approx \underbrace{\begin{bmatrix} \quad \\ \\ \\ \\ \\ \end{bmatrix}}_{Z \in \mathbb{R}^{n\times k}} \underbrace{\begin{bmatrix} \quad\quad\quad\quad\quad\quad \\ \\ \end{bmatrix}}_{W \in \mathbb{R}^{k\times d}}$$

- Also computationally convenient:
  - Xv costs $O(nd)$ but $Z(Wv)$ only costs $O(nk + dk)$.

PCA Notation

- Using $X \approx ZW$, PCA approximates each examples x^i as
$$x^i \approx W^T z^i.$$
- Usually we only need to estimate W:
  - If using least squares, then given W we can find z^i from x^i using
$$z^i = \underset{z}{\text{argmin}} \|x^i - W^T z\|^2 = (WW^T)^{-1}Wx^i.$$
- We often assume that W^T is orthogonal:
  - This means that $WW^T = I$.
  - In this case we have $z^i = Wx^i$.
- In standard formulations, solution only unique up to rotation:
  - Usually, we fit the rows of W sequentially for uniqueness.

Two Classic Views on PCA

- PCA approximates the original matrix by latent-variables Z and latent-factors W,

$$X \approx ZW.$$

  where $Z \in \mathbb{R}^{n \times k}$, $W \in \mathbb{R}^{k \times d}$.
- Two classical interpretations/derivations of PCA (equivalent for orthogonal W^T):
  1. Choose latent-factors W to minimize error ("synthesis view"):

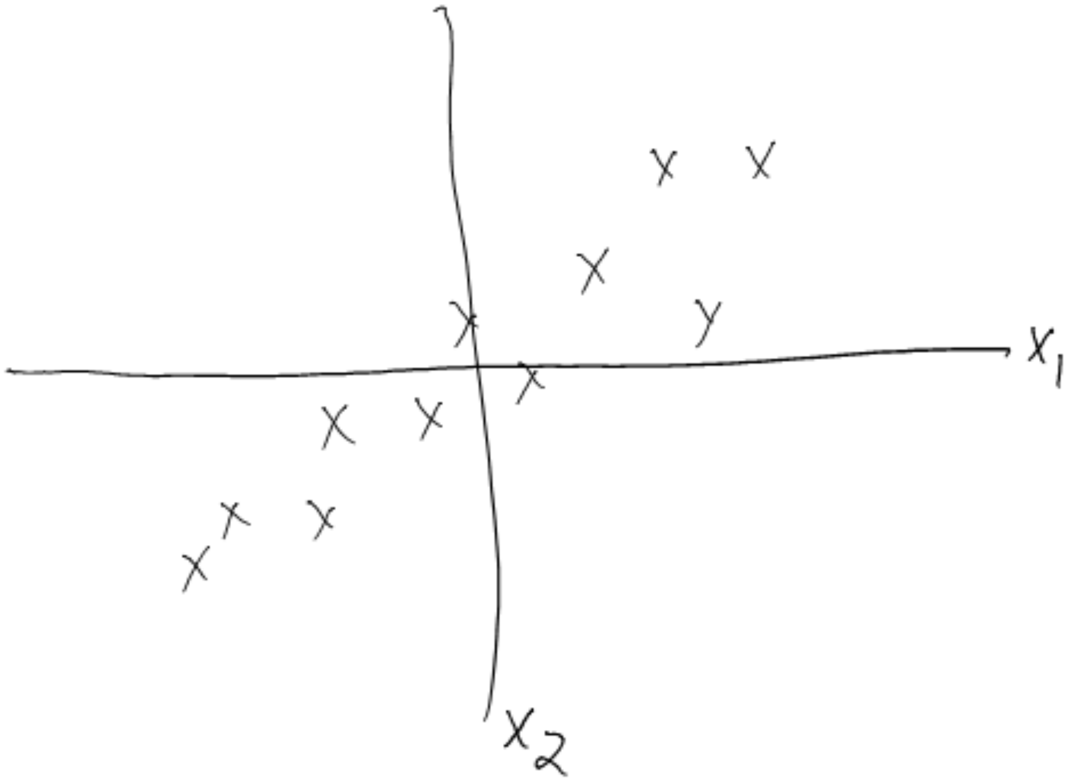
$$\underset{Z \in \mathbb{R}^{n \times k}, W \in \mathbb{R}^{k \times d}}{\text{argmin}} \|X - ZW\|_F^2 = \sum_{i=1}^n \sum_{j=1}^d (x_j^i - (w_j)^T z^i)^2.$$

  2. Choose latent-factors W^T to maximize variance ("analysis view"):

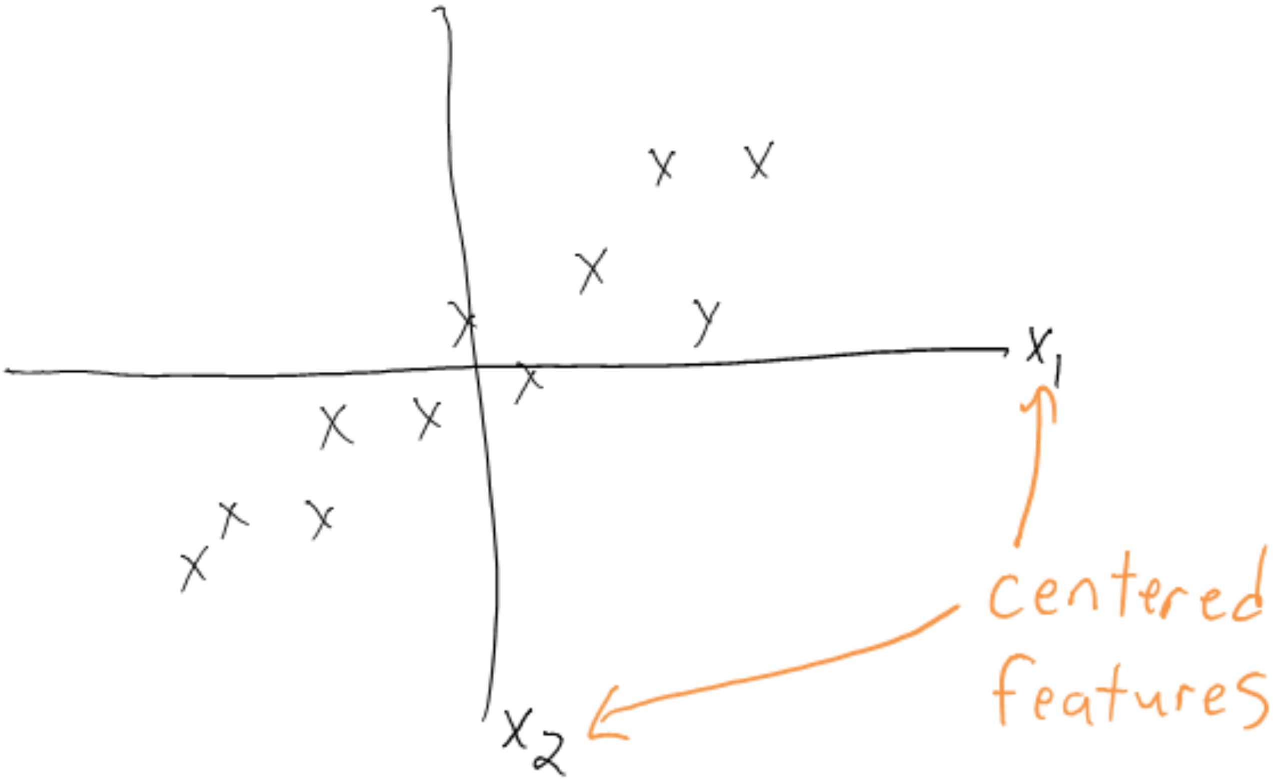
$$\begin{aligned}\underset{W \in \mathbb{R}^{k \times d}}{\text{argmax}} &= \sum_{i=1}^n \|z^i - \mu_z\|^2 = \sum_{i=1}^n \|Wx^i\|^2 \quad (z^i = Wx^i \text{ and } \mu_z = 0)\\ &= \sum_{i=1}^n \text{Tr}((x^i)^T W^T W x^i) = \text{Tr}(W^T W \sum_{i=1}^n x^i (x^i)^T) = \text{Tr}(W^T W X^T X),\end{aligned}$$

  and we note that $X^T X$ is n times sample covariance S because data is centered.

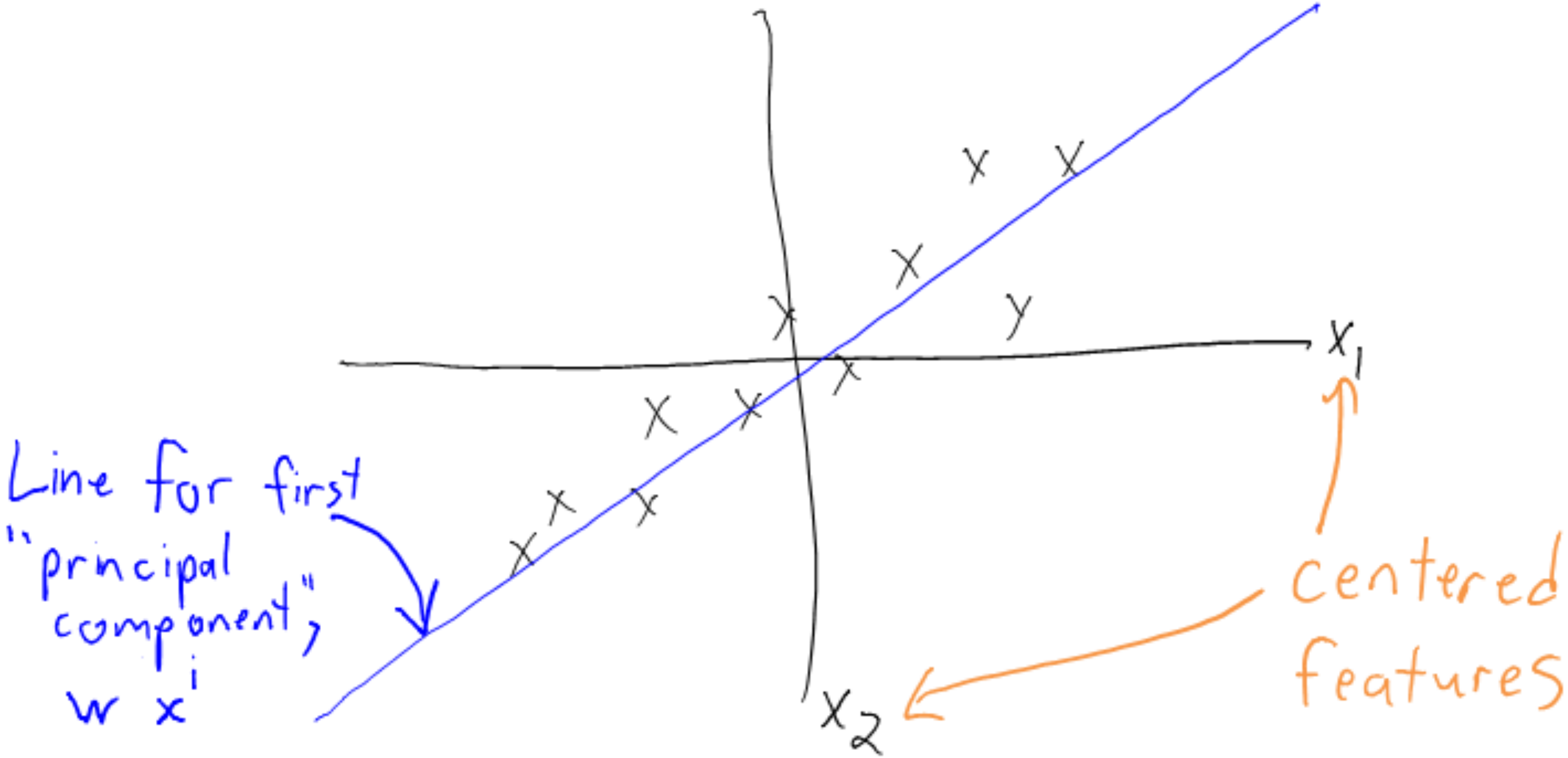
Two Classic Views on PCA



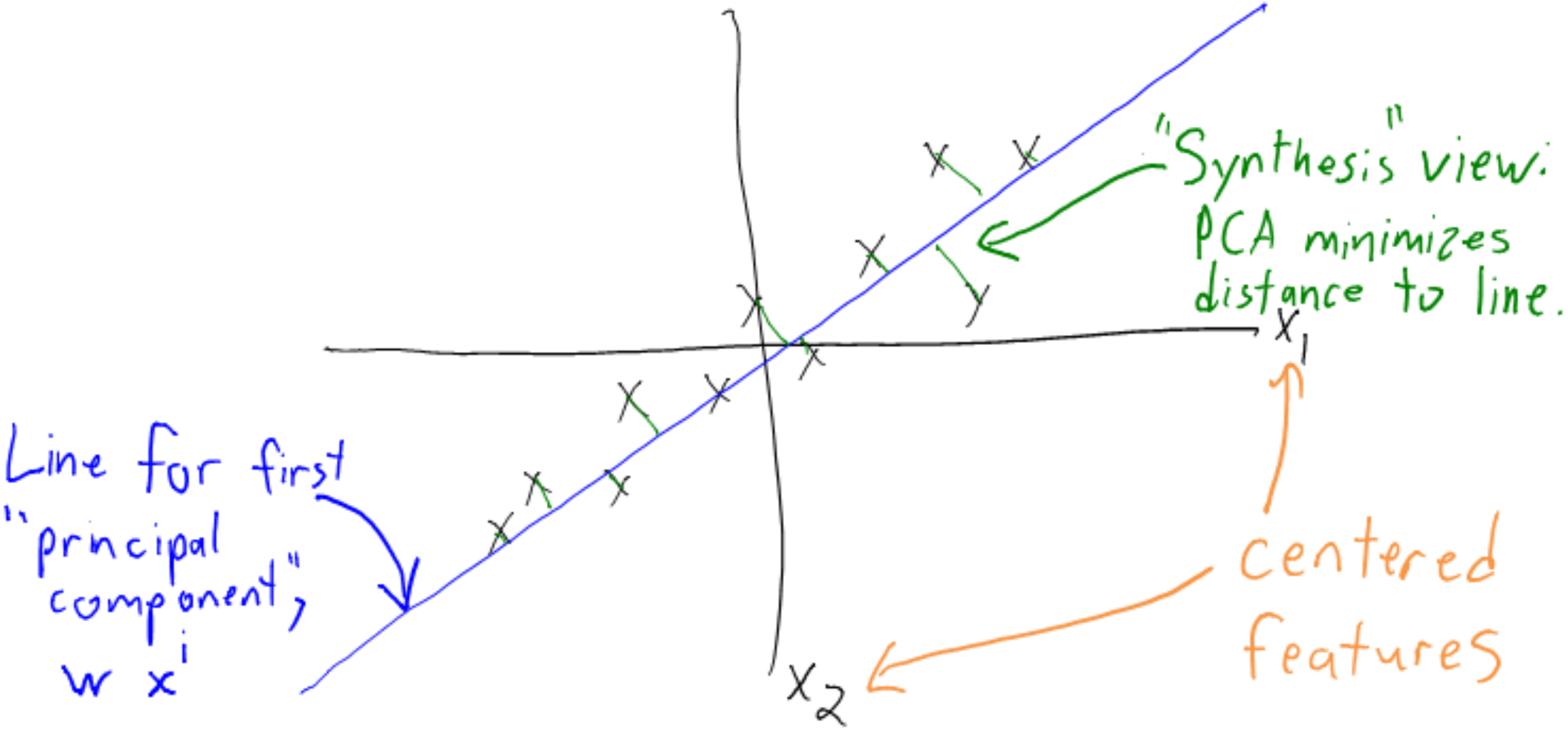
Two Classic Views on PCA



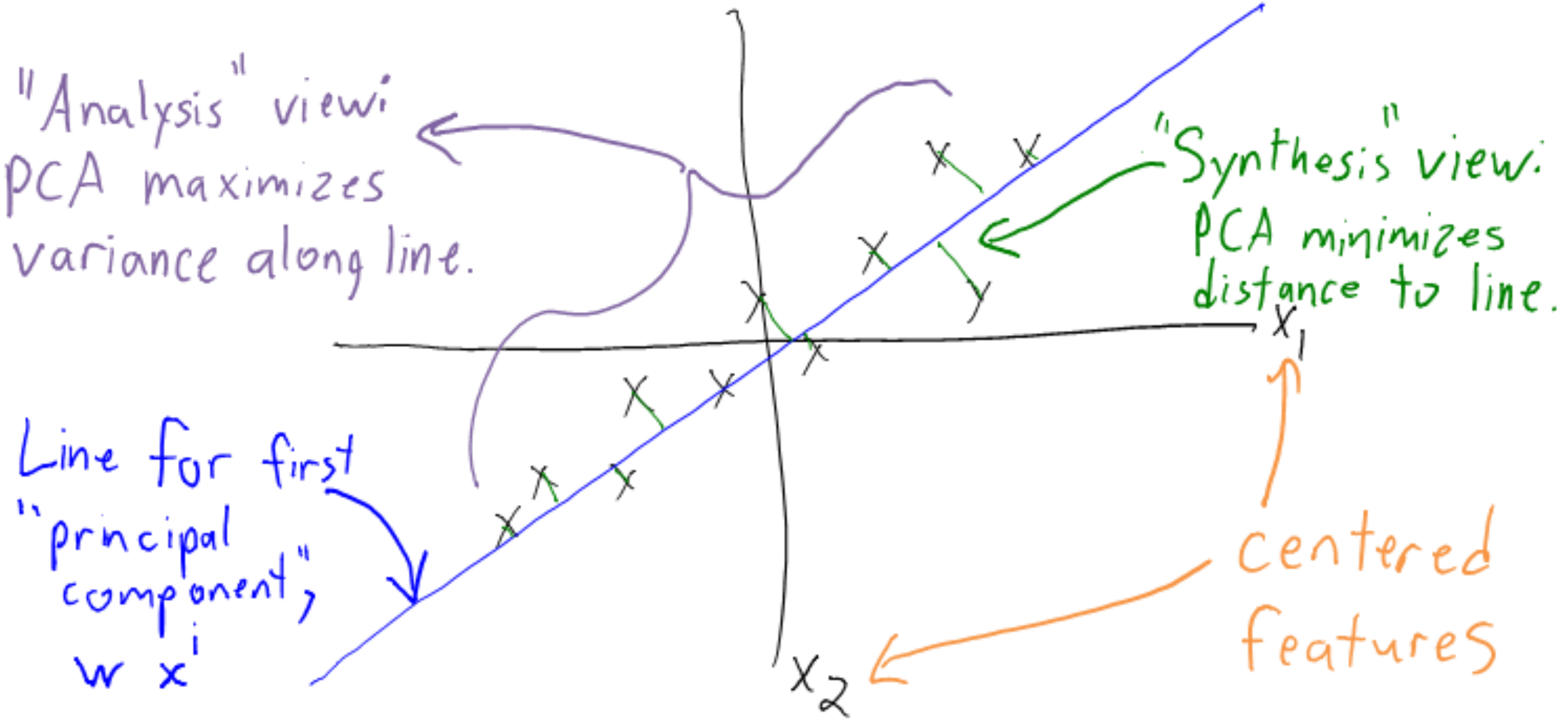
Two Classic Views on PCA



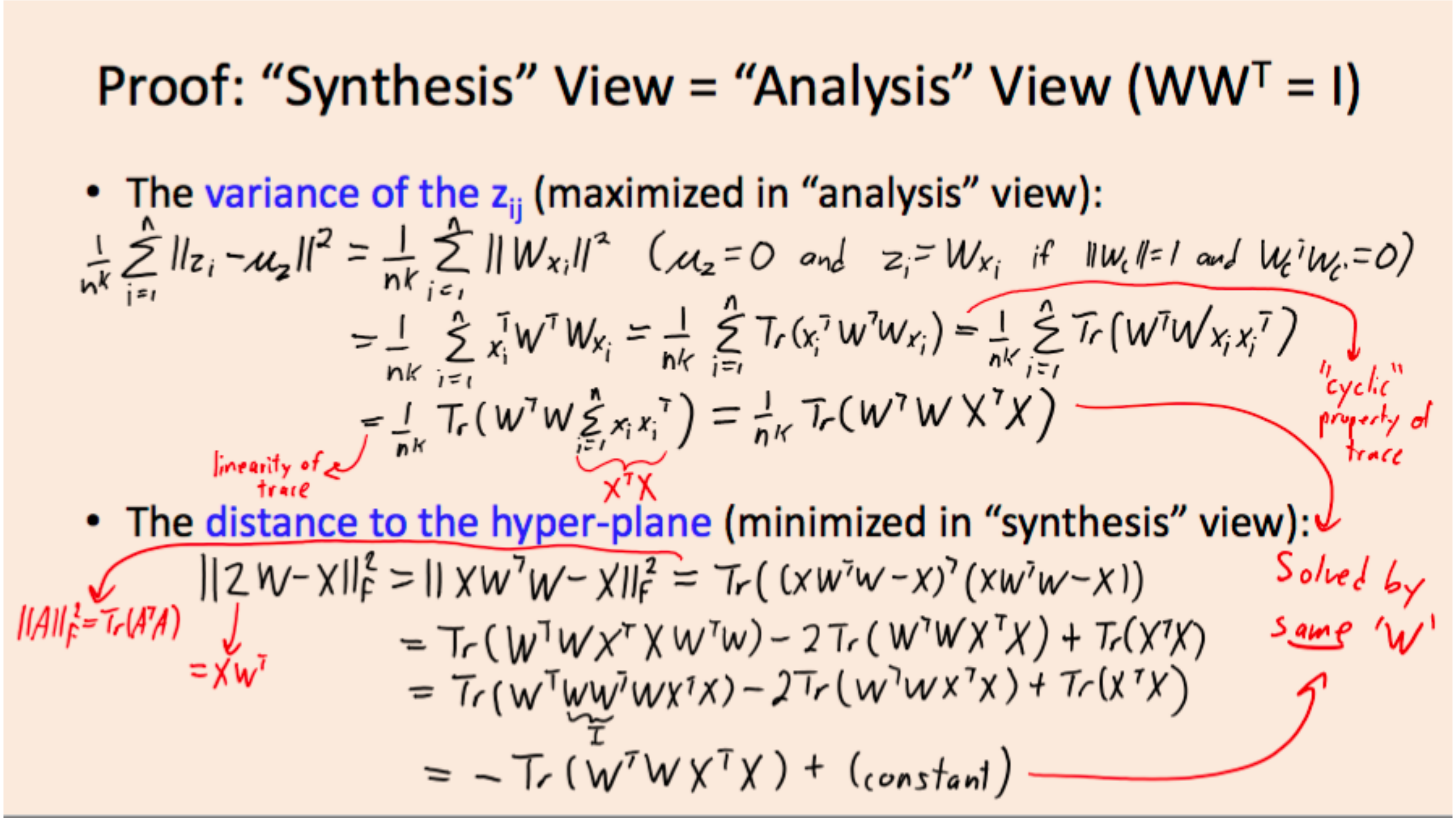
Two Classic Views on PCA



Two Classic Views on PCA


"Analysis" view:
PCA maximizes variance along line.
"Synthesis" view:
PCA minimizes distance to line.
Line for first "principal component";
w x^i
x_1
x_2
centered features

Two Classic Views on PCA



Proof: "Synthesis" View = "Analysis" View ($WW^T = I$)

- The variance of the z_{ij} (maximized in "analysis" view):

$$\frac{1}{nk}\sum_{i=1}^{n}\|z_i-\mu_z\|^2 = \frac{1}{nk}\sum_{i=1}^{n}\|Wx_i\|^2 \quad (\mu_z = 0 \text{ and } z_i = Wx_i \text{ if } \|w_c\|=1 \text{ and } w_c^T w_{c'} = 0)$$

$$= \frac{1}{nk}\sum_{i=1}^{n} x_i^T W^T W x_i = \frac{1}{nk}\sum_{i=1}^{n} Tr(x_i^T W^T W x_i) = \frac{1}{nk}\sum_{i=1}^{n} Tr(W^T W x_i x_i^T)$$

"cyclic" property of trace

$$= \frac{1}{nk} Tr(W^T W \underbrace{\sum_{i=1}^{n} x_i x_i^T}_{X^TX}) = \frac{1}{nk} Tr(W^T W X^T X)$$

linearity of trace

- The distance to the hyper-plane (minimized in "synthesis" view):

$$\|\underset{=XW^T}{Z}W - X\|_F^2 = \|XW^TW - X\|_F^2 = Tr((XW^TW-X)^T(XW^TW-X))$$

$\|A\|_F^2 = Tr(A^TA)$

$$= Tr(W^TWX^TXW^TW) - 2Tr(W^TWX^TX) + Tr(X^TX)$$

$$= Tr(W^T\underbrace{WW^T}_{I}WX^TX) - 2Tr(W^TWX^TX) + Tr(X^TX)$$

$$= -Tr(W^TWX^TX) + (\text{constant})$$

Solved by same 'W'

Probabilistic PCA

- With zero-mean ("centered") data, in PCA we assume that

$$x^i \approx W^T z^i.$$

- In probabilistic PCA we assume that

$$x^i \sim \mathcal{N}(W^T z^i, \sigma^2 I), \quad z^i \sim \mathcal{N}(0, I).$$

(we can actually use any Gaussian density for z)

- We can treat z^i as nuisance parameters integrate over them in likelihood,

$$p(x^i \mid W) = \int_{z^i} p(x^i, z^i \mid W) dz^i.$$

- Looks ugly, but this is marginal of Gaussian so it's Gaussian.
  - Regular PCA is obtained as the limit of σ^2 going to 0.

Manipulating Gaussians

- From the assumptions of the previous slide we have (leaving out ^{i} superscripts):

$$p(x \mid z, W) \propto \exp\left(-\frac{(x - W^T z)^T (x - W^T z)}{2\sigma^2}\right), \quad p(z) \propto \exp\left(-\frac{z^T z}{2}\right).$$

- Multiplying and expanding we get

$$\begin{aligned} p(x, z \mid W) &= p(x \mid z, W) p(z \mid W) \\ &= p(x \mid z, W) p(z) && (z \perp W) \\ &\propto \exp\left(-\frac{(x - W^T z)^T (x - W^T z)}{2\sigma^2} - \frac{z^T z}{2}\right) \\ &= \exp\left(-\frac{x^T x - x^T W^T z - z^T W x + z^T W W^T z}{2\sigma^2} + \frac{z^T z}{2}\right) \end{aligned}$$

Manipulating Gaussians

- So the “complete” likelihood satsifies

$$p(x, z \mid W) \propto \exp\left(-\frac{x^T x - x^T W^T z - z^T W x + z^T W W^T z}{2\sigma^2} + \frac{z^T z}{2}\right)$$

$$= \exp\left(-\frac{1}{2}\left(x^T\left(\frac{1}{\sigma^2}I\right)x + x^T\left(\frac{1}{\sigma^2}W^T\right)z + z^T\left(\frac{1}{\sigma^2}W\right)x + z^T\left(\frac{1}{\sigma^2}WW^T + I\right)z\right)\right),$$

- We can re-write the exponent as a quadratic form,

$$p(x, z \mid W) \propto \exp\left(-\frac{1}{2}\begin{bmatrix} x^T & z^T \end{bmatrix}\begin{bmatrix} \frac{1}{\sigma^2}I & -\frac{1}{\sigma^2}W^T \\ -\frac{1}{\sigma^2}W & \frac{1}{\sigma^2}WW^T + I \end{bmatrix}\begin{bmatrix} x \\ z \end{bmatrix}\right),$$

- This has the form of a Gaussian distribution,

$$p(v \mid W) \propto \exp\left(-\frac{1}{2}(v-\mu)^T\Sigma^{-1}(v-\mu)\right),$$

with $v = \begin{bmatrix} x \\ z \end{bmatrix}$, $\mu = 0$, and $\Sigma^{-1} = \begin{bmatrix} \frac{1}{\sigma^2}I & -\frac{1}{\sigma^2}W^T \\ -\frac{1}{\sigma^2}W & \frac{1}{\sigma^2}WW^T + I \end{bmatrix}$.

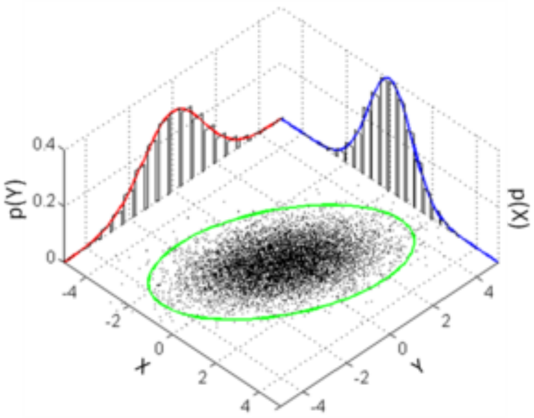
Manipulating Gaussians

- Remember that if we write multivariate Gaussian in partitioned form,

$$\begin{bmatrix} x \\ z \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_x \\ \mu_z \end{bmatrix}, \begin{bmatrix} \Sigma_{xx} & \Sigma_{xz} \\ \Sigma_{zx} & \Sigma_{zz} \end{bmatrix}\right),$$

then the marginal distribution $p(x)$ (integrating over z) is given by

$$x \sim \mathcal{N}(\mu_x, \Sigma_{xx}).$$



https://en.wikipedia.org/wiki/Multivariate_normal_distribution

Manipulating Gaussians

- Remember that if we write multivariate Gaussian in partitioned form,

$$\begin{bmatrix} x \\ z \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_x \\ \mu_z \end{bmatrix}, \begin{bmatrix} \Sigma_{xx} & \Sigma_{xz} \\ \Sigma_{zx} & \Sigma_{zz} \end{bmatrix}\right),$$

then the marginal distribution $p(x)$ (integrating over z) is given by

$$x \sim \mathcal{N}(\mu_x, \Sigma_{xx}).$$

- For probabilistic PCA we assume $\mu_x = 0$, but we partitioned Σ^{-1} instead of Σ.
- To get Σ we can use a partitioned matrix inversion formula,

$$\Sigma = \begin{bmatrix} \frac{1}{\sigma^2}I & -\frac{1}{\sigma^2}W^T \\ -\frac{1}{\sigma^2}W & \frac{1}{\sigma^2}WW^T + I \end{bmatrix}^{-1} = \begin{bmatrix} W^TW + \sigma^2 I & W^T \\ W & I \end{bmatrix},$$

which gives that solution to integrating over z is

$$x \mid W \sim \mathcal{N}(0, W^TW + \sigma^2 I).$$

Notes on Probabilistic PCA

- NLL of observed data has the form

$$-\log p(x \mid W) = \frac{n}{2}\text{Tr}(S\Theta) - \frac{n}{2}\log|\Theta| + \text{const.},$$

  where $\Theta = (W^TW + \sigma^2 I)^{-1}$ and S is the sample covariance.
- Not convex, but non-global stationary points are saddle points.
- Equivalence with regular PCA:
  - Consider W^T orthogonal so $WW^T = I$ (usual assumption).
  - Using matrix determinant lemma we have

$$|W^TW + \sigma^2 I| = |I + \frac{1}{\sigma^2}\underbrace{WW^T}_{I}| \cdot |\sigma^2 I| = \text{const.}$$

  - Using matrix inversion lemma we have

$$(W^TW + \sigma^2 I)^{-1} = \frac{1}{\sigma^2}I - \frac{1}{\sigma^2(\sigma^2+1)}W^TW,$$

    so minimizing NLL maximizes $\text{Tr}(W^TWS)$ as in PCA.

Generalizations of Probabilistic PCA

- Why do we need a probabilistic interpretation of PCA?
  - Good excuse to play with Gaussian identities and matrix formulas?
- We now understand that PCA fits a Gaussian with restricted covariance:
  - Hope is that $W^TW + \sigma I$ is a good approximation of full covariance X^TX.
  - We can do fancy things like mixtures of PCA models.

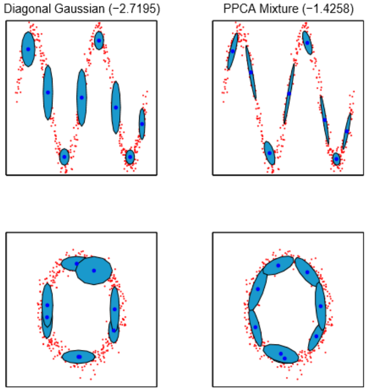


Figure 8: Comparison of an 8-component diagonal variance Gaussian mixture model with a mixture of PPCA model. The upper two plots give a view perpendicular to the major

http://www.miketipping.com/papers/met-mppca.pdf

- We could consider different $x^i \mid z^i$ distribution (but integrals are ugly).
  - E.g., Laplace of student if you want it to be robust.
  - E.g., logistic or softmax if you have discrete x^i_j.

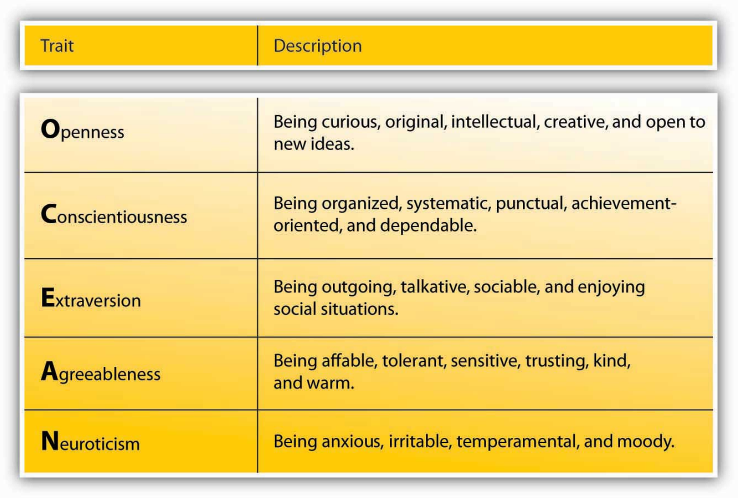
Outline

1. Probabilistic PCA
2. Factor Analysis
3. Independent Component Analysis

Factor Analysis

- Factor analysis (FA) is a method for discovering latent-factors.
    - A standard tool and widely-used across science and engineering.
- Historical applications are measures of intelligence and personality traits.
    - Some controversy, like trying to find factors of intelligence due to race.

(without normalizing for socioeconomic factors)



Trait	Description
Openness	Being curious, original, intellectual, creative, and open to new ideas.
Conscientiousness	Being organized, systematic, punctual, achievement-oriented, and dependable.
Extraversion	Being outgoing, talkative, sociable, and enjoying social situations.
Agreeableness	Being affable, tolerant, sensitive, trusting, kind, and warm.
Neuroticism	Being anxious, irritable, temperamental, and moody.

https://new.edu/resources/big-5-personality-traits

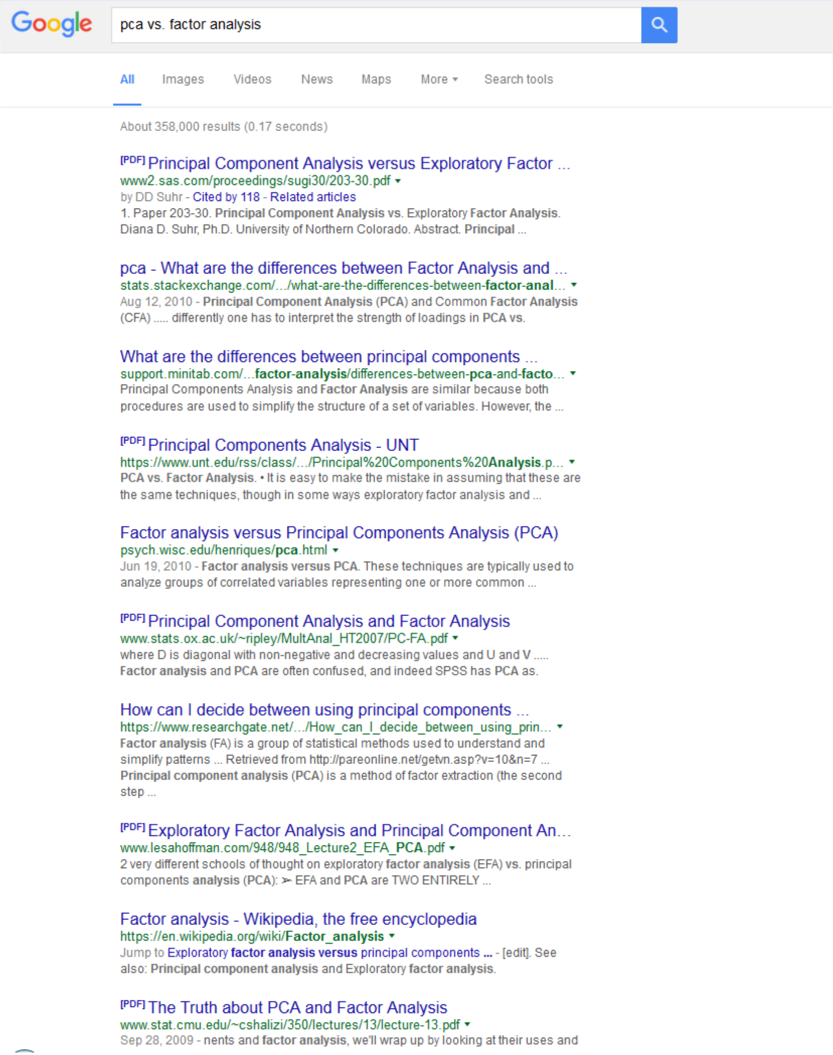
- “Big Five” aspects of personality (vs. non-evidence-based Myers-Briggs):
    - https://fivethirtyeight.com/features/most-personality-quizzes-are-junk-science-i-found-one-that-isnt

Factor Analysis

- FA approximates the original matrix by latent-variables Z and latent-factors W,

$$X \approx ZW.$$

- Which should sound familiar...
- Are PCA and FA the same?
  - Both are more than 100 years old.
  - People are still fighting about whether they are the same:
    - Doesn't help that some software packages run PCA when you call FA.



About 358,000 results (0.17 seconds)

[PDF] Principal Component Analysis versus Exploratory Factor ...
www2.sas.com/proceedings/sugi30/203-30.pdf ▾
by DD Suhr - Cited by 118 - Related articles
1. Paper 203-30. **Principal Component Analysis** vs. Exploratory **Factor Analysis**. Diana D. Suhr, Ph.D. University of Northern Colorado. Abstract. **Principal** ...

pca - What are the differences between Factor Analysis and ...
stats.stackexchange.com/.../what-are-the-differences-between-**factor-anal**... ▾
Aug 12, 2010 - **Principal Component Analysis** (PCA) and Common **Factor Analysis** (CFA) ..... differently one has to interpret the strength of loadings in PCA **vs.**

What are the differences between principal components ...
support.minitab.com/...**factor-analysis**/differences-between-**pca-and-facto**... ▾
Principal Components Analysis and **Factor Analysis** are similar because both procedures are used to simplify the structure of a set of variables. However, the ...

[PDF] Principal Components Analysis - UNT
https://www.unt.edu/rss/class/.../Principal%20Components%20**Analysis**.p... ▾
PCA vs. Factor Analysis. • It is easy to make the mistake in assuming that these are the same techniques, though in some ways exploratory factor analysis and ...

Factor analysis versus Principal Components Analysis (PCA)
psych.wisc.edu/henriques/**pca**.html ▾
Jun 19, 2010 - **Factor analysis versus PCA**. These techniques are typically used to analyze groups of correlated variables representing one or more common ...

[PDF] Principal Component Analysis and Factor Analysis
www.stats.ox.ac.uk/~ripley/MultAnal_HT2007/PC-FA.pdf ▾
where D is diagonal with non-negative and decreasing values and U and V .....
Factor analysis and PCA are often confused, and indeed SPSS has PCA as.

How can I decide between using principal components ...
https://www.researchgate.net/.../How_can_I_decide_between_using_prin... ▾
Factor analysis (FA) is a group of statistical methods used to understand and simplify patterns ... Retrieved from http://pareonline.net/getvn.asp?v=10&n=7 ...
Principal component analysis (PCA) is a method of factor extraction (the second step ...

[PDF] Exploratory Factor Analysis and Principal Component An...
www.lesahoffman.com/948/948_Lecture2_EFA_**PCA**.pdf ▾
2 very different schools of thought on exploratory **factor analysis** (EFA) vs. principal components **analysis** (PCA): ➢ EFA and PCA are TWO ENTIRELY ...

Factor analysis - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/**Factor_analysis** ▾
Jump to Exploratory **factor analysis versus** principal components **...** - [edit]. See also: **Principal component analysis** and Exploratory **factor analysis**.

[PDF] The Truth about PCA and Factor Analysis
www.stat.cmu.edu/~cshalizi/350/lectures/13/lecture-13.pdf ▾
Sep 28, 2009 - nents and **factor analysis**, we'll wrap up by looking at their uses and

PCA vs. Factor Analysis

- In probabilistic PCA we assume

$$x^i \mid z^i \sim \mathcal{N}(W^T z^i, \sigma^2 I), \quad z^i \sim \mathcal{N}(0, I),$$

  and we obtain PCA as $\sigma \to 0$.
- In FA we assume

$$x^i \mid z^i \sim \mathcal{N}(W^T z^i, D), \quad z^i \sim \mathcal{N}(0, I),$$

  where D is a diagonal matrix.
- The difference is that you can have a noise variance for each dimension.
- Repeating the previous exercise we get that

$$x^i \sim \mathcal{N}(0, W^T W + D).$$

- So FA has extra degrees of freedom in variance of individual variables.

PCA vs. Factor Analysis

- We can write non-centered versions of both models:
  - Probabilistic PCA:

$$x^i \mid z^i \sim \mathcal{N}(W^T z^i + \mu, \sigma^2 I), \quad z^i \sim \mathcal{N}(0, I),$$

  - Factor analysis:

$$x^i \mid z^i \sim \mathcal{N}(W^T z^i + \mu, D), \quad z^i \sim \mathcal{N}(0, I),$$

    where D is a diagonal matrix.

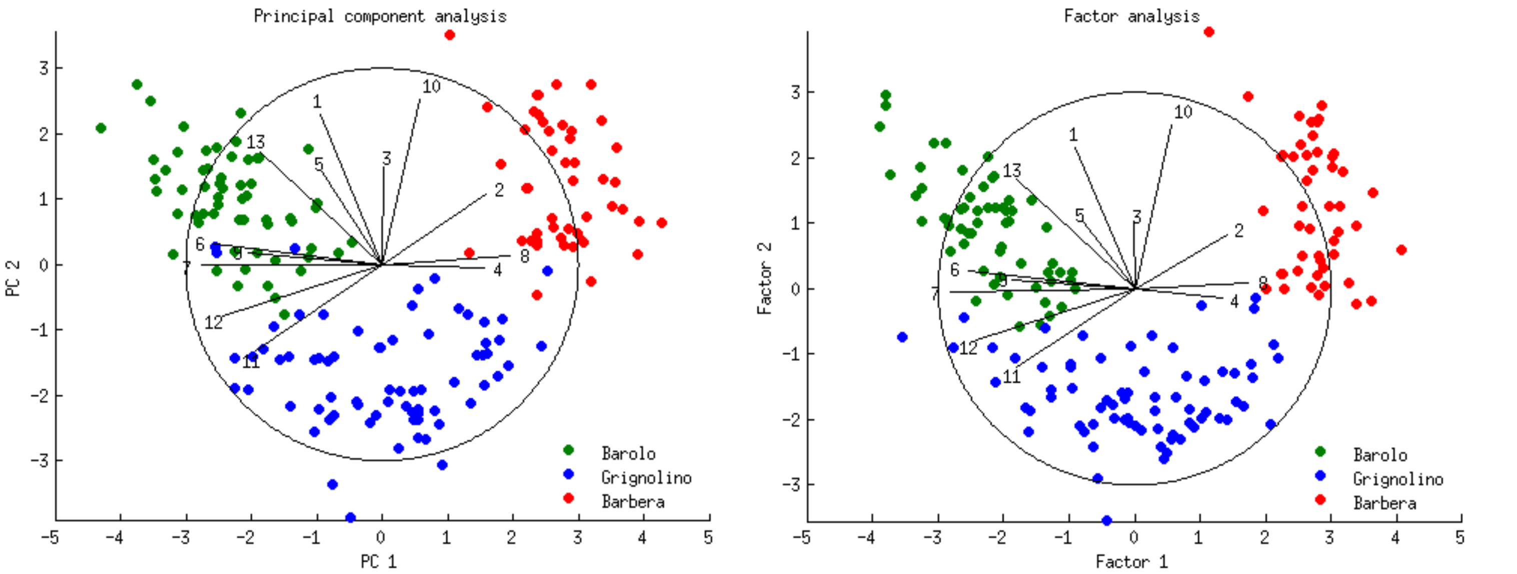
- A different perspective is that these models assume

$$x^i = W^T z^i + \epsilon,$$

  where PPCA has $\epsilon \sim \mathcal{N}(\mu, \sigma^2 I)$ and FA has $\epsilon \sim \mathcal{N}(\mu, D)$.

PCA vs. Factor Analysis

In practice they usually give pretty similar results:



http://stats.stackexchange.com/questions/1576/what-are-the-differences-between-factor-analysis-and-principal-component-analysi

Remember in 340 that difference with PCA and ISOMAP/t-SNE was huge.

Factor Analysis Discussion

- Similar to PCA, FA is invariant to rotation of W,

$$W^TW = W^T\underbrace{Q^TQ}_{I}W = (WQ)^T(WQ),$$

  for orthogonal Q.
  - So as with PCA you can't interpret multiple factors as being unique.
- Differences with PCA:
  - Not affected by scaling individual features.
    - FA doesn't chase large-noise features that are uncorrelated with other features.
  - But unlike PCA, it's affected by rotation of the data.
  - No nice "SVD" approach for FA, you can get different local optima.

Orthogonality and Sequential Fitting

- The PCA and FA solutions are not unique.
- Common heuristic:
  1. Enforce that rows of W have a norm of 1.
  2. Enforce that rows of W are orthogonal.
  3. Fit the rows of W sequentially.
- This leads to a unique solution up to sign changes.
- But there are other ways to resolve non-uniqueness (Murphy's Section 12.1.3):
  - Force W to be lower-triangular.
  - Choose an informative rotation.
  - Use a non-Gaussian prior ("independent component analysis").

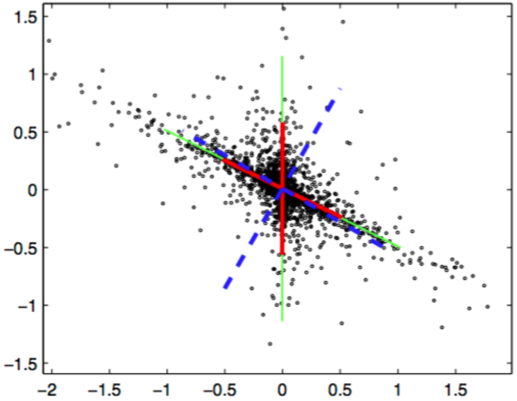
Outline

1. Probabilistic PCA
2. Factor Analysis
3. Independent Component Analysis

Motivation for Independent Component Analysis (ICA)

- Factor analysis has found an enormous number of applications.
  - People really want to find the “factors” that make up their data.

- But factor analysis can’t even identify factor directions.



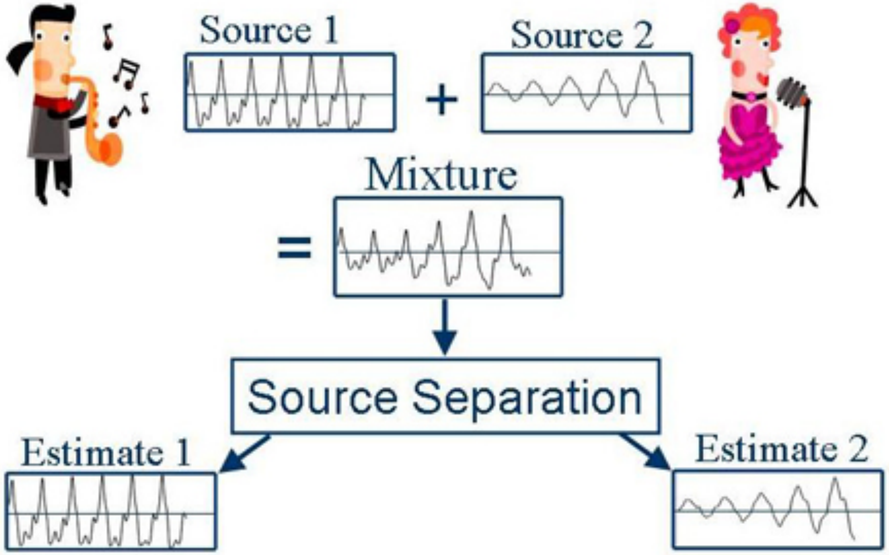
http://www.inf.ed.ac.uk/teaching/courses/pmr/lectures/ica.pdf

Motivation for Independent Component Analysis (ICA)

- Factor analysis has found an enormous number of applications.
  - People really want to find the "factors" that make up their data.
- But factor analysis can't even identify factor directions.
  - We can rotate W and obtain the same model.
- Independent component analysis (ICA) is a more recent approach
  - Around 30 years old instead of > 100.
  - Under certain assumptions it can identify factors.
- The canonical application of ICA is blind source separation.

Blind Source Separation

- Input to blind source separation:
  - Multiple microphones recording multiple sources.

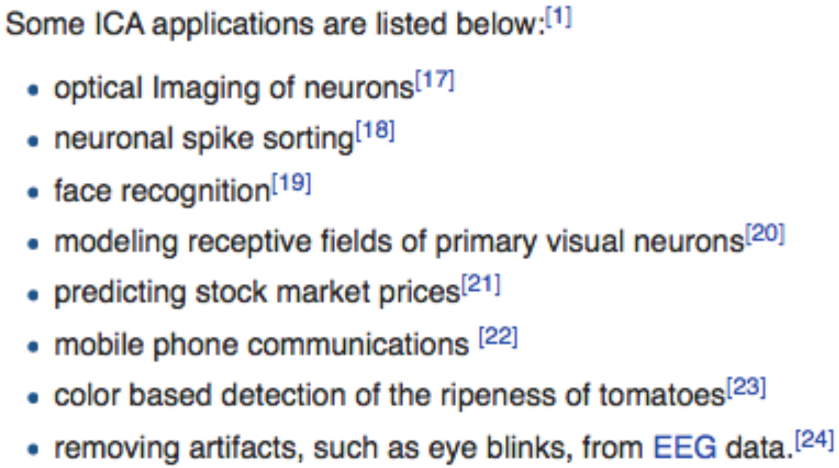


http://music.eecs.northwestern.edu/research.php

- Each microphone gets different mixture of the sources.
  - Goal is to reconstruct sources (factors) from the measurements.

Independent Component Analysis Applications

- ICA is replacing PCA/FA in many applications.



Some ICA applications are listed below:[1]

- optical Imaging of neurons[17]
- neuronal spike sorting[18]
- face recognition[19]
- modeling receptive fields of primary visual neurons[20]
- predicting stock market prices[21]
- mobile phone communications [22]
- color based detection of the ripeness of tomatoes[23]
- removing artifacts, such as eye blinks, from EEG data.[24]

- It's the only algorithm we didn't cover in 340 from the list of "The 10 Algorithms Machine Learning Engineers Need to Know".
- Recent work shows that ICA can often resolve direction of causality.

Limitations of Matrix Factorization

- As in PCA/FA, ICA is a matrix factorization method,

$$X \approx ZW.$$

- Let's assume that $X = ZW$ for a "true" W with $k = d$.
  - Different from PCA where we assume $k << d$.

- There are only 3 issues stopping us from finding "true" W.

3 Sources of Matrix Factorization Non-Uniquness

- Label switching: get same model if we permute rows of W.
  - We can exchange row 1 and 2 of W (and same columns of Z).
  - Not a problem because we don't care about order of factors.
- Scaling: get same model if you scale a row.
  - If we multiply row 1 of W by α, could multiply column 1 of Z by $1/\alpha$.
  - Can't identify scale/sign, but might hope to identify direction.
- Rotataion: we the get same model if we rotate W. pre-multiply W by orthogonal Q.
  - Rotation correspond to orthogonal matrices Q, such matrices have $Q^TQ = I$.
  - If we rotate W with Q, then we have $(QW)^T(QW) = W^TQ^TQW = W^TW$.
- If we could address rotation, we could identify the directions.

Another Unique Gaussian Property

- Consider a prior that assumes the z_c^i are independent,

$$p(z^i) = \prod_{c=1}^{k} p_c(z_c^i).$$

  - E.g., in PPCA and FA we use $\mathcal{N}(0,1)$ for each z_c^i.

- If $p(z^i)$ is rotation-invariant, $p(Qz^i) = p(z^i)$, then it must be Gaussian.

- The (non-intuitive) magic behind ICA:
  - If the priors are all non-Gaussian, it isn't rotationally symmetric.

- Implication: we can identify factors W if at most 1 factor is Gaussian.
  - Up to permutation/sign/scaling (other rotations change distribution).

PCA vs. ICA

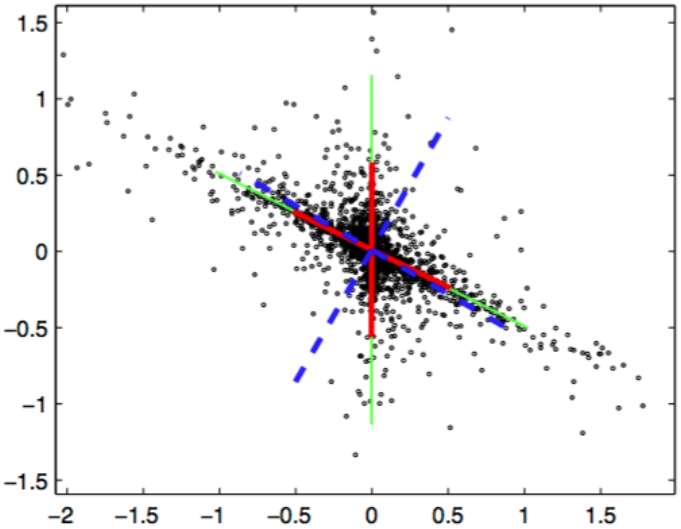


Figure : Latent data is sampled from the prior $p(x_i) \propto \exp(-5\sqrt{|x_i|})$ with the mixing matrix $\mathbf{A}$ shown in green to create the observed two dimensional vectors $\mathbf{y} = \mathbf{A}\mathbf{x}$. The red lines are the mixing matrix estimated by `ica.m` based on the observations. For comparison, PCA produces the blue (dashed) components. Note that the components have been scaled to improve visualisation. As expected, PCA finds the orthogonal directions of maximal variation. ICA however, correctly estimates the directions in which the components were independently generated.

http://www.inf.ed.ac.uk/teaching/courses/pmr/lectures/ica.pdf

Independent Component Analysis

- In ICA we use the approximation,

$$X \approx ZW$$

where we want z_j^i to be non-Gaussian and independent across j.

  - Usually, we "center" and "whiten" the data before applying ICA.

- A common strategy is maximum likelihood ICA assuming a heavy-tailed z_j^i like

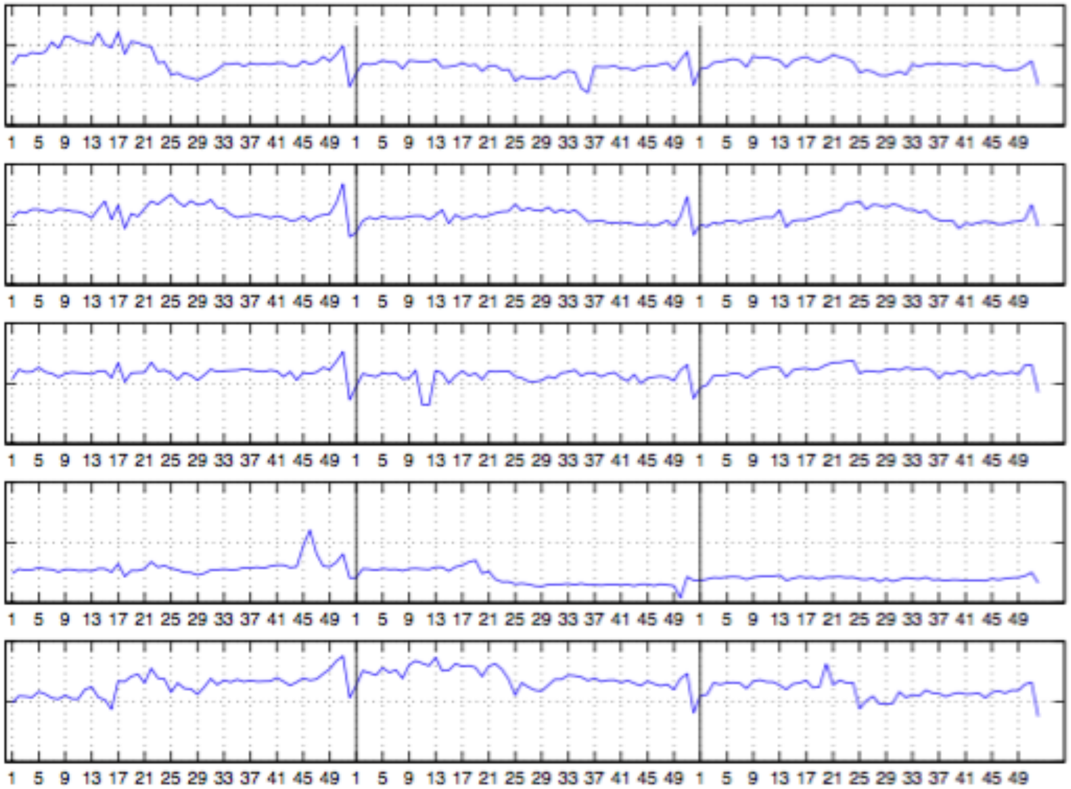
$$p(z_j^i) = \frac{1}{\pi(\exp(z_j^i) + \exp(-z_j^i))}.$$

- Another common strategy fits data while maximizing measure of non-Gaussianity:
  - Maximize kurtosis, which is minimizes by Gaussians.
  - Miniimize entropy, which is maximized with Gaussians.

- The fastICA method is a popular Newton-like method maximizing kurtosis.

ICA on Retail Purchase Data

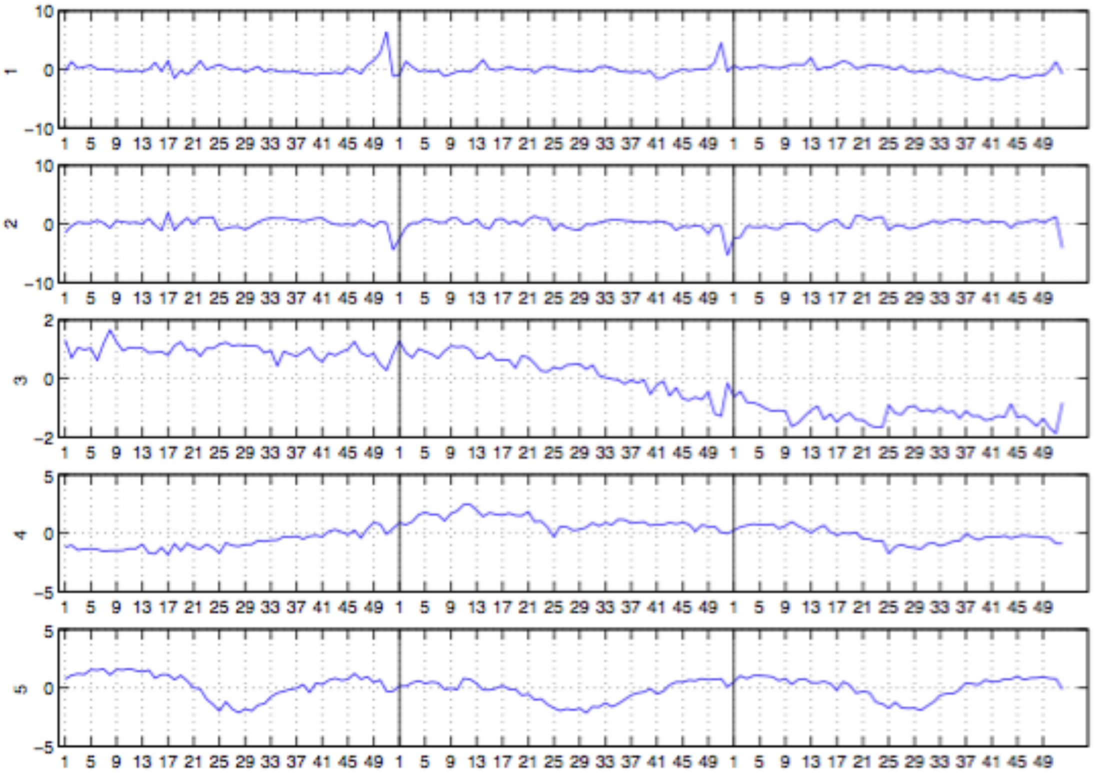
- Cash flow from 5 different stores over 3 years:



http://www.stat.ucla.edu/~yuille/courses/Stat161-261-Spring14/Hyv000-icatut.pdf

ICA on Retail Purchase Data

- Factors found using ICA.
  - 1-2 reflect "holiday season", 3-4 are year-to-year, and 5 is summer dip in sales.



http://www.stat.ucla.edu/~yuille/courses/Stat161-261-Spring14/HyvO00-icatut.pdf

Summary

- PCA is a classic method for dimensionality reduction.
- Probabilistic PCA is a continuous latent-variable probabilistic generalization.
- Factor analysis extends probabilistic PCA with different noise in each dimension.
- Independent component analysis: allows identifying non-Gaussian latent factors.