CPSC 540: Machine Learning

Probabilistic PCA and Factor Analysis

Mark Schmidt

University of British Columbia

Winter 2016

Admin

- Assignment 2:
  - Today is the last day to hand it in.
- Assignment 3:
  - Due February 23, start early.
  - Some additional hints will be added.
- Reading week:
  - No classes or tutorials next week.
  - I'm talking at Robson Square 6:30pm Wednesday February 17.
- February 25:
  - Default is to not have class this day.
  - Instead go to Rich Sutton's talk in DMP 110 at 3:30:
    - "Reinforcement Learning And The Future of Artificial Intelligence".

Last Time: Expectation Maximization

- We considered learning with observed variables O and hidden variables H.
- In this case the "observed-data" log-liklihooed has a nasty form,

$$\log p(O|\Theta) = \log\left(\sum_H p(O,H|\Theta)\right).$$

Last Time: Expectation Maximization

- We considered learning with observed variables O and hidden variables H.
- In this case the "observed-data" log-liklihooed has a nasty form,

$$\log p(O|\Theta) = \log\left(\sum_H p(O,H|\Theta)\right).$$

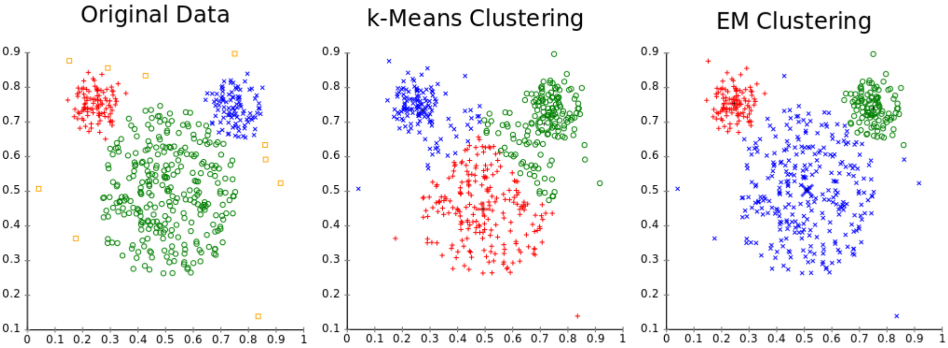
- EM applies when "complete-data" log-likelihood, $\log p(O,H|\Theta)$, has a nice form.
- EM iterations take the form

$$\Theta^{t+1} = \underset{\Theta}{\text{argmin}}\left\{-\sum_H \alpha_H \log p(O,H|\Theta)\right\},$$

where $\alpha_H = p(H|O,\Theta^t)$.

K-Means vs. Mixture of Gaussians

- Applying EM to mixture of Gaussians is similar to k-means clustering:
  - But EM/MoG does probabilistic (or "soft") cluster assignment:
    - Points can have partial membership in multiple clusters.
  - And EM/MoG allows different covariance for each cluster (k-means has $\Sigma_c = I$).
    - Clusters do not need to be convex.



https://en.wikipedia.org/wiki/K-means_clustering

Last Time: Expectation Maximization

- EM iterations take the form

$$\Theta^{t+1} = \underset{\Theta}{\text{argmin}} \left\{ -\sum_H \alpha_H \log p(O, H|\Theta) \right\},$$

  where $\alpha_H = p(H|O, \Theta^t)$.

- Guaranteed to increase likelihood of observed data.

Last Time: Expectation Maximization

- EM iterations take the form

$$\Theta^{t+1} = \operatorname*{argmin}_{\Theta} \left\{ -\sum_{H} \alpha_H \log p(O, H|\Theta) \right\},$$

  where $\alpha_H = p(H|O, \Theta^t)$.
- Guaranteed to increase likelihood of observed data.
- This sums over all possible values of H, which seems intractable.
  - In binary semi-supervised learning (SSL), requires sum over 2^t values possible $\tilde{y}$.

Last Time: Expectation Maximization

- EM iterations take the form

$$\Theta^{t+1} = \underset{\Theta}{\text{argmin}} \left\{ -\sum_H \alpha_H \log p(O, H | \Theta) \right\},$$

  where $\alpha_H = p(H | O, \Theta^t)$.
- Guaranteed to increase likelihood of observed data.
- This sums over all possible values of H, which seems intractable.
  - In binary semi-supervised learning (SSL), requires sum over 2^t values possible $\tilde{y}$.
- Fortunately, conditional independence often allows efficient calculation.
  - See EM note posted (soon) on webpage for details (mixtures and SSL).
  - We'll cover general case when we discus probabilistic graphical models.

Today: Continuous-Latent Variables

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

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

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

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.

Today: Continuous-Latent Variables

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

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

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

  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.
- Integrals like these can be computed under a conjugacy property.
- Today we focus on the Gaussian case.
  - We'll cover general case when we get to Bayesian statistics.

Today: Continuous-Latent Variables

- In mixture models, we have a discrete latent variable z:
  - In mixture of Gaussians, if you know the cluster z then $p(x|z)$ is a Gaussian.

Today: Continuous-Latent Variables

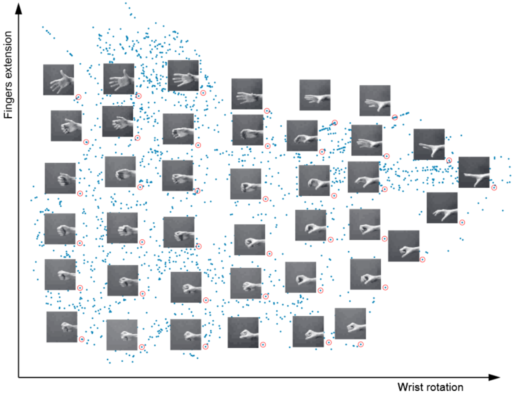
- In mixture models, we have a discrete latent variable z:
  - In mixture of Gaussians, if you know the cluster z then $p(x|z)$ is a Gaussian.
- In latent-factor models, we have continuous latent variables z:
  - In probabilistic PCA, if you know the latent-factors z then $p(x|z)$ is a Gaussian.

Today: Continuous-Latent Variables

- In mixture models, we have a discrete latent variable z:
    - In mixture of Gaussians, if you know the cluster z then $p(x|z)$ is a Gaussian.
- In latent-factor models, we have continuous latent variables z:
    - In probabilistic PCA, if you know the latent-factors z then $p(x|z)$ is a Gaussian.
- But what would a continuous z 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.

Principal Component Analysis (PCA)

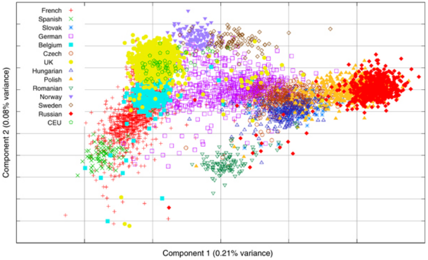
- PCA replaces X with a lower-dimensional approximation Z.
  - Matrix Z has n rows, but typically far fewer columns.
- PCA is used for:
  - Dimensionality reduction: replace X with a lower-dimensional Z.

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.

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,

$$X \approx ZW^T.$$

  where $Z \in \mathbb{R}^{n \times k}$, $W \in \mathbb{R}^{d \times k}$, and we assume columns of X have mean 0.
- We're trying to split redundancy in X into its important "parts".

PCA Notation

- PCA approximates the original matrix by factor-loadings Z and latent-factors W,
$$X \approx ZW^T.$$
where $Z \in \mathbb{R}^{n\times k}$, $W \in \mathbb{R}^{d\times k}$, 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:

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

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

PCA Notation

- Using $X \approx ZW^T$, PCA approximates each examples x^i as

$$x^i \approx Wz^i.$$

PCA Notation

- Using $X \approx ZW^T$, PCA approximates each examples x^i as
$$x^i \approx Wz^i.$$
- Usually we only need to estimate W:
  - If using least squares, then given W we can find z^i from x^i using
$$z^i = \underset{z}{\text{argmin}} \|x^i - Wz\|^2 = (W^TW)^{-1}W^Tx^i.$$

PCA Notation

- Using $X \approx ZW^T$, PCA approximates each examples x^i as

$$x^i \approx Wz^i.$$

- Usually we only need to estimate W:
  - If using least squares, then given W we can find z^i from x^i using

$$z^i = \underset{z}{\text{argmin}} \, \|x^i - Wz\|^2 = (W^T W)^{-1} W^T x^i.$$

- We often assume that W is orthogonal:
  - This means that $W^T W = I$.
  - In this case we have $z^i = W^T x^i$.
- In standard formulations, solution only unique up to rotation:
  - Usually, we fit the columns of W sequentially for uniqueness.

Two Classic Views on PCA

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

$$X \approx ZW^T.$$

  where $Z \in \mathbb{R}^{n \times k}$, $W \in \mathbb{R}^{d \times k}$.
- Two classical interpretations/derivations of PCA:

Two Classic Views on PCA

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

$$X \approx ZW^T.$$

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

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

Two Classic Views on PCA

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

$$X \approx ZW^T.$$

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

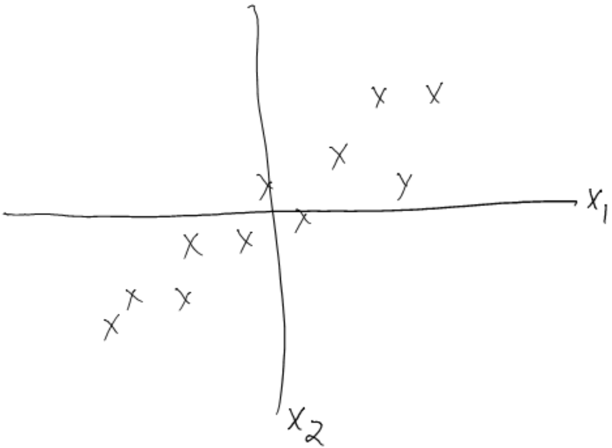
$$\underset{W \in \mathbb{R}^{d \times k}, Z \in \mathbb{R}^{n \times k}}{\text{argmin}} \|X - ZW^T\|_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{d} (x_j^i - (w^j)^T z^i)^2.$$

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

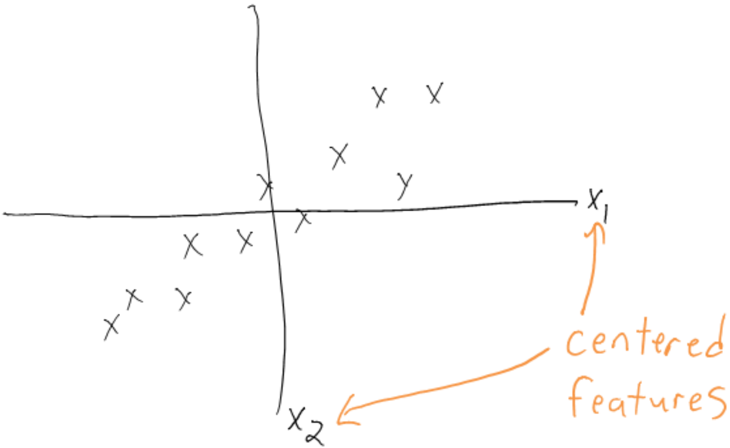
$$\underset{W \in \mathbb{R}^{d \times k}}{\text{argmax}} = \sum_{i=1}^{n} \|z^i - \mu_z\|^2 = \sum_{i=1}^{n} \|W^T x^i - W^T \mu\|^2 \qquad (z^i = W^T x^i)$$

$$= \sum_{i=1}^{n} \|W^T x^i\|^2 = \text{Tr}(W^T X^T X W) = \text{Tr}(W^T \Sigma W) \qquad (\text{Assuming } \mu = 0)$$

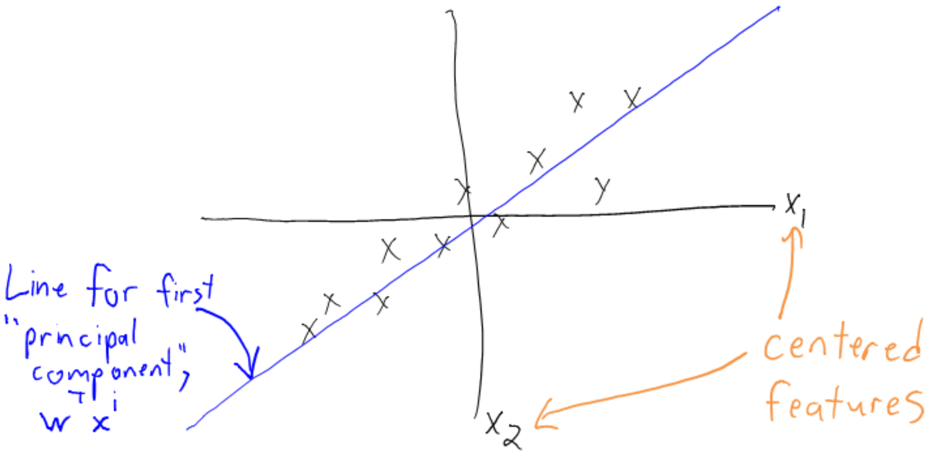
PCA in One Dimension



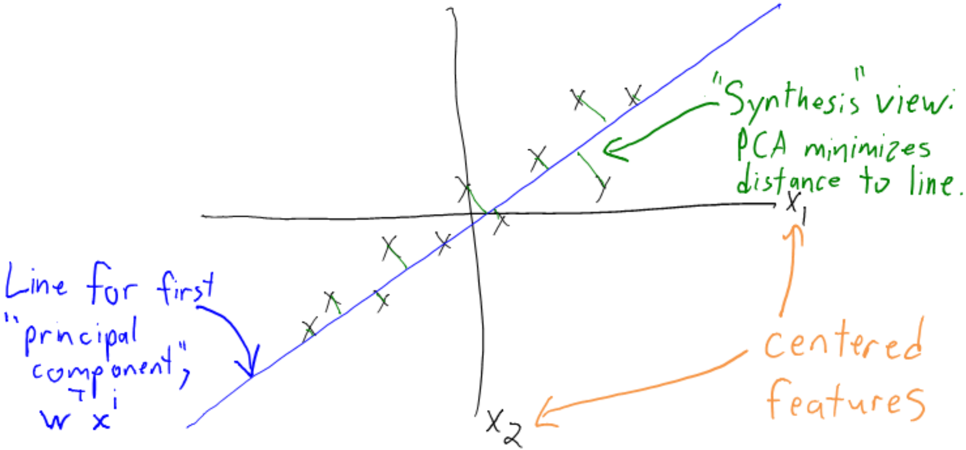
PCA in One Dimension



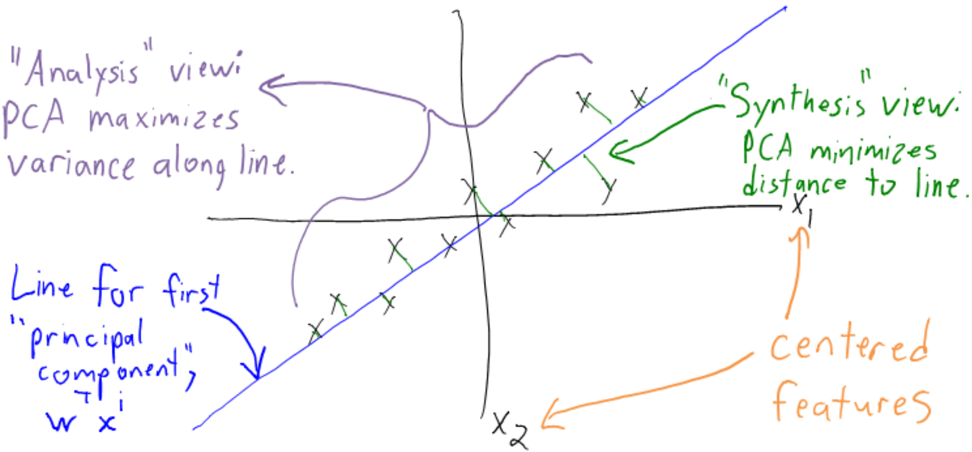
PCA in One Dimension



PCA in One Dimension



PCA in One Dimension



Probabilistic PCA

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

$$x \approx Wz.$$

- In probabilistic PCA we assume that

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

- Note that any Gaussian density for z yields equivalent model.

Probabilistic PCA

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

$$x \approx Wz.$$

- In probabilistic PCA we assume that

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

- Note that any Gaussian density for z yields equivalent model.
- Since z is hidden, our observed likelihood integrates over z,

$$p(x|\Theta) = \int_z p(x, z|\Theta) dz = \int_z p(z|\Theta) p(x|z, \Theta) dz$$

- This looks ugly, but can be computed due to magical Gaussian properties...

Manipulating Gaussians

- From the assumptions of the previous slide we have

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

Manipulating Gaussians

- From the assumptions of the previous slide we have

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

- Multiplying and expanding we get

$$\begin{aligned}
p(x,z|W) &= p(x|z,W)p(z|W) \\
&= p(x|z,W)p(z) \quad \text{(assuming } z \perp W\text{)} \\
&\propto \exp\left(-\frac{(x-Wz)^T(x-Wz)}{2\sigma^2} - \frac{z^Tz}{2}\right) \\
&= \exp\left(-\frac{x^Tx - x^TWz - z^TW^Tx + z^TW^TWz}{2\sigma^2} + \frac{z^Tz}{2}\right) \\
&= \exp\left(-\frac{1}{2}\left(x^T\left(\frac{1}{\sigma^2}I\right)x + x^T\left(\frac{1}{\sigma^2}W\right)z + z^T\left(\frac{1}{\sigma^2}W^T\right)x + z^T\left(\frac{1}{\sigma^2}W^TW + I\right)z\right)\right).
\end{aligned}$$

Manipulating Gaussians

- Thus the joint probability satisfies

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

Manipulating Gaussians

- Thus the joint probability satisfies

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

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

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

- This has the form of a Gaussian distribution,

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

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

Manipulating Gaussians

- Thus we have

$$\begin{bmatrix} z \\ x \end{bmatrix} \sim \mathcal{N}(0, \Lambda^{-1}), \quad \text{with } \Lambda = \begin{bmatrix} I + \frac{1}{\sigma^2} W^T W & -\frac{1}{\sigma^2} W^T \\ -\frac{1}{\sigma^2} W & \frac{1}{\sigma^2} I \end{bmatrix}.$$

- A special case of general result that product of Gaussians is Gaussian.
  - See Bishop/Murphy textbooks for general formula.

Manipulating Gaussians

- Thus we have

$$\begin{bmatrix} z \\ x \end{bmatrix} \sim \mathcal{N}(0, \Lambda^{-1}), \quad \text{with } \Lambda = \begin{bmatrix} I + \frac{1}{\sigma^2} W^T W & -\frac{1}{\sigma^2} W^T \\ -\frac{1}{\sigma^2} W & \frac{1}{\sigma^2} I \end{bmatrix}.$$

- A special case of general result that product of Gaussians is Gaussian.
  - See Bishop/Murphy textbooks for general formula.
- We are interested in the marginal after integrating over z,

$$p(x|W) = \int_z p(x, z|W) dz,$$

but another special property of Gaussians is that marginals are Gaussian.

Manipulating Gaussians

- If we can write our multivariate Gaussian in terms of mean and covariance

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

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

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

Manipulating Gaussians

- If we can write our multivariate Gaussian in terms of mean and covariance

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

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

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

- Using a matrix inversion lemma lets us convert from Λ to Σ, giving

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

Manipulating Gaussians

- If we can write our multivariate Gaussian in terms of mean and covariance

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

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

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

- Using a matrix inversion lemma lets us convert from Λ to Σ, giving

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

- Combining the above we obtain

$$p(x|W) = \int_z p(x, z|W)dz = \frac{1}{\sqrt{2\pi}^{\frac{d}{2}}|WW^T + \sigma^2 I|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}x^T(WW^T + \sigma^2 I)x\right).$$

Notes on Probabilistic PCA

- Regular PCA is obtained as limit of $\sigma \to 0$.

Notes on Probabilistic PCA

- Regular PCA is obtained as limit of $\sigma \to 0$.
- Negative log-likelihood has the form

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

  where $C = WW^T + \sigma^2 I$ and $S = X^T X$.
- Not convex, but all stable stationary points are global minima.

Notes on Probabilistic PCA

- Regular PCA is obtained as limit of $\sigma \to 0$.
- Negative log-likelihood has the form

$$-\log p(x|W) = \frac{n}{2}\mathsf{Tr}(SC) + \frac{n}{2}\log|C| + \text{const.},$$

where $C = WW^T + \sigma^2 I$ and $S = X^T X$.

- Not convex, but all stable stationary points are global minima.
- Can reduce cost from $O(d^3)$ to $O(k^3)$ with matrix inversion/determinant lemmas:
  - Allows us to work with $W^T W$ instead of WW^T.
- We can get $p(z|x, W)$ using that conditional of Gaussians is Gaussian.

Notes on Probabilistic PCA

- Regular PCA is obtained as limit of $\sigma \to 0$.
- Negative log-likelihood has the form

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

  where $C = WW^T + \sigma^2 I$ and $S = X^T X$.
- Not convex, but all stable stationary points are global minima.
- Can reduce cost from $O(d^3)$ to $O(k^3)$ with matrix inversion/determinant lemmas:
  - Allows us to work with $W^T W$ instead of WW^T.
- We can get $p(z|x, W)$ using that conditional of Gaussians is Gaussian.
- We can also consider different distribution for $x^i|z^i$:
  - E.g., Laplace of student if you want it to be robust.
  - E.g., logistic or softmax if you have discrete x^i_j.

Generalizations of Probabilistic PCA

- Why bother with all this math?
    - Good excuse to play with Gaussian identities and matrix formulas?

Generalizations of Probabilistic PCA

- Why bother with all this math?
  - Good excuse to play with Gaussian identities and matrix formulas?
- We now understand that PCA fits a Gaussian with restricted covariance:
  - Hope is that $WW^T + \sigma I$ is a good approximation of full covariance.
  - Lets us understand connection between PCA and factor analysis.

Generalizations of Probabilistic PCA

- Why bother with all this math?
  - Good excuse to play with Gaussian identities and matrix formulas?
- We now understand that PCA fits a Gaussian with restricted covariance:
  - Hope is that $WW^T + \sigma I$ is a good approximation of full covariance.
  - Lets us understand connection between PCA and factor analysis.
- We can do fancy things like mixtures of PCA models.

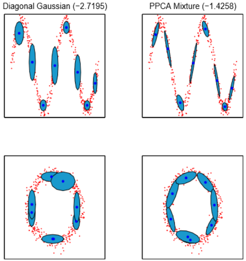


Figure 8: Comparison of an 8-component diagonal variance Gaussian mixture model with a mixture of PPCA model. The upper two plots give a view perpendicular to the major

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

(pause)

Factor Analysis

- Factor analysis (FA) is a method for discovering latent-factors.
- 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

- But a standard tool and very widely-used across science and engineering.

Factor Analysis

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

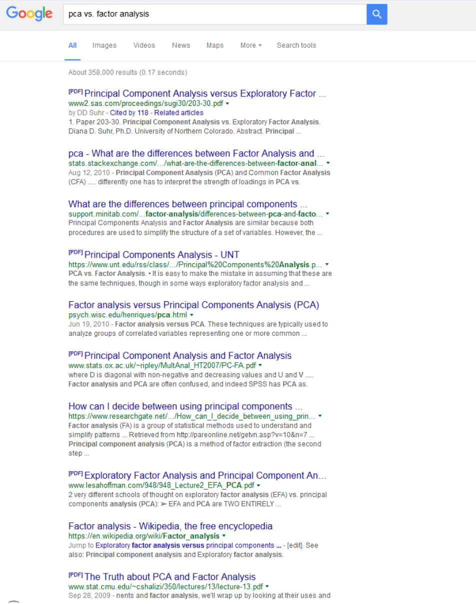
$$X \approx ZW^T.$$

Factor Analysis

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

$$X \approx ZW^T.$$

- 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.



Google pca vs. factor analysis

All Images Videos News Maps More Search tools

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

$$x|z \sim \mathcal{N}(Wz, \sigma^2 I), \quad z \sim \mathcal{N}(0, I),$$

and we obtain PCA as $\sigma \to 0$.

PCA vs. Factor Analysis

- In probabilistic PCA we assume

$$x|z \sim \mathcal{N}(Wz, \sigma^2 I), \quad z \sim \mathcal{N}(0, I),$$

  and we obtain PCA as $\sigma \to 0$.

- In FA we assume

$$x|z \sim \mathcal{N}(Wz, D), \quad z \sim \mathcal{N}(0, I),$$

  where D is a diagonal matrix.

PCA vs. Factor Analysis

- In probabilistic PCA we assume

$$x|z \sim \mathcal{N}(Wz, \sigma^2 I), \quad z \sim \mathcal{N}(0, I),$$

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

$$x|z \sim \mathcal{N}(Wz, D), \quad z \sim \mathcal{N}(0, I),$$

  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

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

PCA vs. Factor Analysis

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

$$x|z \sim \mathcal{N}(Wz + \mu, \sigma^2 I), \quad z \sim \mathcal{N}(0, I),$$

  - Factor analysis:

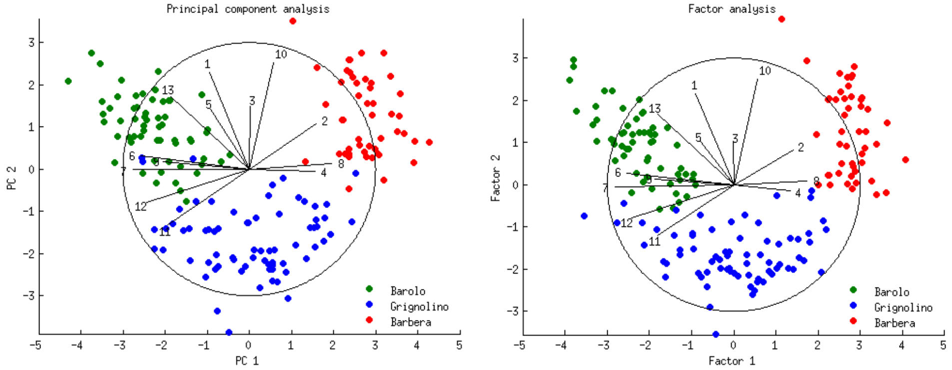
$$x|z \sim \mathcal{N}(Wz + \mu, D), \quad z \sim \mathcal{N}(0, I),$$

    where D is a diagonal matrix.

PCA vs. Factor Analysis

- We can write non-centered versions of both models:
  - Probabilistic PCA:
  $$x|z \sim \mathcal{N}(Wz + \mu, \sigma^2 I), \quad z \sim \mathcal{N}(0, I),$$
  - Factor analysis:
  $$x|z \sim \mathcal{N}(Wz + \mu, D), \quad z \sim \mathcal{N}(0, I),$$
  where D is a diagonal matrix.
- A different perspective is that these models assume
$$x = Wz + \mu + \epsilon,$$
where PPCA has $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ and FA has $\epsilon \sim \mathcal{N}(0, D)$.
- So in FA W still models covariance, but you have extra freedom in variance of individual variables.

PCA vs. Factor Analysis



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

Biplot: lines represent projection of elementary vectors like $e_1 = \begin{bmatrix} 1 & 0 & 0 & \cdots \end{bmatrix}$.

Factor Analysis Discussion

- No closed-form solution for FA, and can find different local optima.
- Unlike PCA, FA doesn't change if you scale variables.
    - FA doesn't chase large-noise features that are uncorrelated with other features.

Factor Analysis Discussion

- No closed-form solution for FA, and can find different local optima.
- Unlike PCA, FA doesn't change if you scale variables.
  - FA doesn't chase large-noise features that are uncorrelated with other features.
- Unlike PCA, FA changes if you rotate data.
  - But if $k > 1$ you can rotate factors,

$$WQ(WQ)^T = WQQ^TW^T = WW^T.$$

  - So you can't interpret multiple factors as being unique.

(pause)

Independent Component Analysis

- More recent development is independent component analysis (ICA).
- Key idea:
  - Can't identify Gaussian factors due to rotation issue.

Independent Component Analysis

- More recent development is independent component analysis (ICA).
- Key idea:
  - Can't identify Gaussian factors due to rotation issue.
  - Replace Gaussian $p(z)$ with product of non-Gaussian distributions,

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

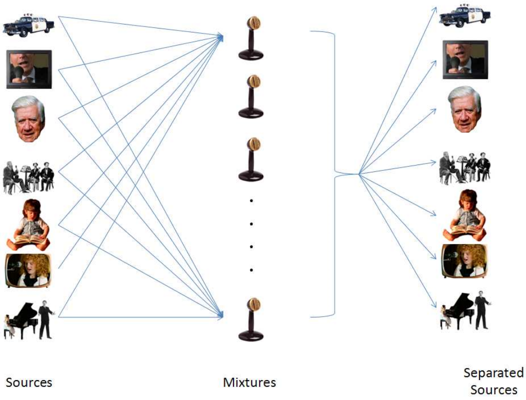
- A common choice is the heavy-tailed

$$p(z_c) = \frac{1}{\pi(\exp(z_c) + \exp(-z_c))}$$

- Now a huge topic: many variations and nonlinear generalizations exist.

ICA for Blind Source Separation

- Classic application is blind source separation.



https://onionesquereality.wordpress.com/tag/over-complete-independent-component-analysis

- Under certain conditions, you can recover true independent factors.

Robust PCA

- A fourth view of PCA is that it solve the rank-constrained approximation problem

$$\underset{W}{\operatorname{argmin}} \|X - W\|_F^2, \quad \text{with } \operatorname{rank}(W) \leq k.$$

Robust PCA

- A fourth view of PCA is that it solve the rank-constrained approximation problem

$$\underset{W}{\operatorname{argmin}} \|X - W\|_F^2, \quad \text{with } \operatorname{rank}(W) \leq k.$$

- We could approximation non-convex rank-constraint using nuclear-norm

$$\underset{W}{\operatorname{argmin}} \|X - W\|_F^2 + \lambda \|W\|_*.$$

Robust PCA

- A fourth view of PCA is that it solve the rank-constrained approximation problem

$$\mathop{\mathrm{argmin}}_{W} \|X - W\|_F^2, \quad \text{with } \mathrm{rank}(W) \le k.$$

- We could approximation non-convex rank-constraint using nuclear-norm

$$\mathop{\mathrm{argmin}}_{W} \|X - W\|_F^2 + \lambda \|W\|_*.$$

- Robust PCA corresponds to using the L1-norm,

$$\mathop{\mathrm{argmin}}_{W} \|X - W\|_1 + \lambda \|W\|_*.$$

- Typically solved by introducing variable $S = X - W$.

Mixtures Models for Classification

- Classic generative model for supervised learning uses

$$p(y^i|x^i) \propto p(x^i|y^i)p(y^i),$$

and typically $p(x^i|y^i)$ is assumed by Gaussian (LDA) or independent (naive Bayes).

Mixtures Models for Classification

- Classic generative model for supervised learning uses

$$p(y^i|x^i) \propto p(x^i|y^i)p(y^i),$$

and typically $p(x^i|y^i)$ is assumed by Gaussian (LDA) or independent (naive Bayes).

- But we could allow more flexibility by using a mixture model,

$$p(x^i|y^i) = \sum_{c=1}^{k} p(z^i = c|y^i)p(x^i|z^i = c, y^i).$$

Mixtures Models for Classification

- Classic generative model for supervised learning uses

$$p(y^i|x^i) \propto p(x^i|y^i)p(y^i),$$

  and typically $p(x^i|y^i)$ is assumed by Gaussian (LDA) or independent (naive Bayes).
- But we could allow more flexibility by using a mixture model,

$$p(x^i|y^i) = \sum_{c=1}^{k} p(z^i = c|y^i)p(x^i|z^i = c, y^i).$$

- Instead of a generative model, we could also take a mixture of regression models,

$$p(y^i|x^i) = \sum_{c=1}^{k} p(z^i = c|x^i)p(y^i|z^i = c, x^i).$$

- Called a "mixture of experts" model:
  - Each regression model is an "expert" for certain values of x^i.

Summary

- PCA is a classic method for dimensionality reduction.

Summary

- PCA is a classic method for dimensionality reduction.
- Probabilistic PCA is a continuous latent-variable probabilistic generalization.

Summary

- PCA is a classic method for dimensionality reduction.
- Probabilistic PCA is a continuous latent-variable probabilistic generalization.
- Product and marginal of Gaussians have nice closed-form expressions.

Summary

- PCA is a classic method for dimensionality reduction.
- Probabilistic PCA is a continuous latent-variable probabilistic generalization.
- Product and marginal of Gaussians have nice closed-form expressions.
- Factor analysis extends probabilistic PCA with different noise in each dimension.

Summary

- PCA is a classic method for dimensionality reduction.
- Probabilistic PCA is a continuous latent-variable probabilistic generalization.
- Product and marginal of Gaussians have nice closed-form expressions.
- Factor analysis extends probabilistic PCA with different noise in each dimension.
- Other latent-factor models like ICA, robust PCA, and mixture of experts.

- Next time: probabilistic graphical models.