Principal Component Analysis (PCA)

COS 424/524, SML 302: Fundamentals of Machine Learning

Professor Engelhardt

COS424/524, SML302

Lecture 16

Exploratory data analysis using dimension reduction

In the last few lectures, we learned about one form of *unsupervised learning* with *latent variable models*: *clustering* data using *mixture models*.

We learned how to use expectation-maximization (EM) to fit latent variable models to observations.

Today, we begin *dimension reduction* using *latent variable models*.

Basic idea behind dimension reduction

Goal: Compute a reduced representation of data i from p-dimensional to K-dimensional, where $K \ll p$.

$$\langle x_1, \ldots, x_p \rangle_i \rightarrow \langle z_1, \ldots, z_K \rangle_i$$

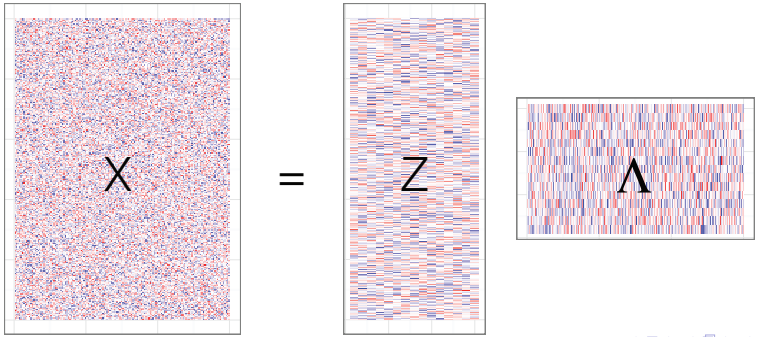
Objective: reconstruct the p-dimensional data from the K-dimensional data as accurately as possible.

Linear dimension reduction

Linear dimension reduction

Linear dimension reduction linearly projects each observation X_i to a **component** or **factor** in K space using weights Λ in feature space:

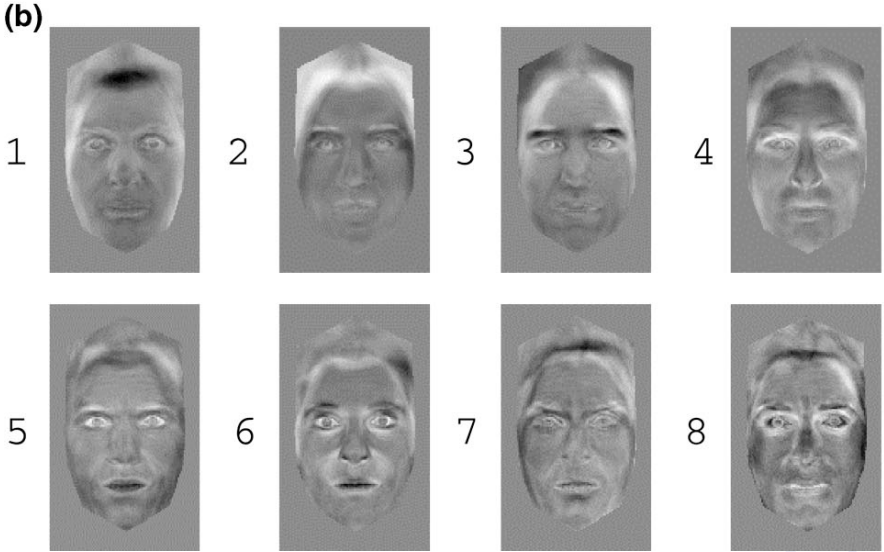
$$X = Z\Lambda$$



Example: Eigenfaces

- The input are pictures of peoples' faces, each a set of pixels
- Each face is represented as a weighted combination of "eigenfaces"
- The latent *components* capture recurring properties in faces; can be used to cluster faces and compare similarity.

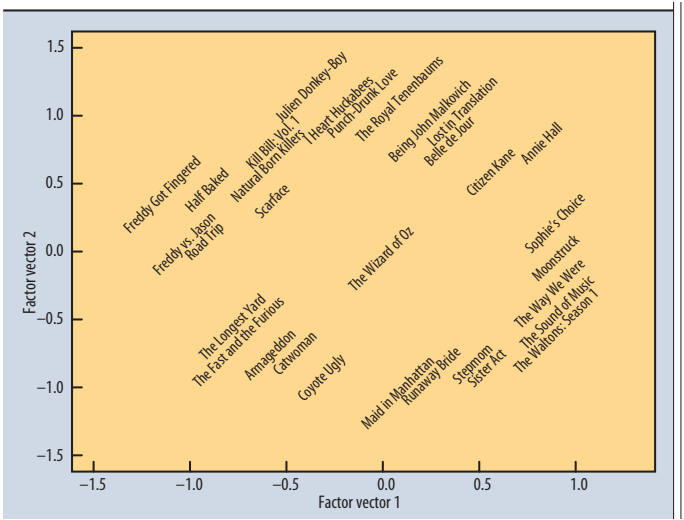
Example: Eigenfaces



Example: Netflix recommendations

- The observed data are users' movie ratings with missing values.
- Users represented as weighted combination of *components* (i.e., movie rater "types"); each component assigns strength to each movie.
- Plot the movies along their first two recovered latent dimensions

Example: Netflix recommendations



Example: In-painting or imputation

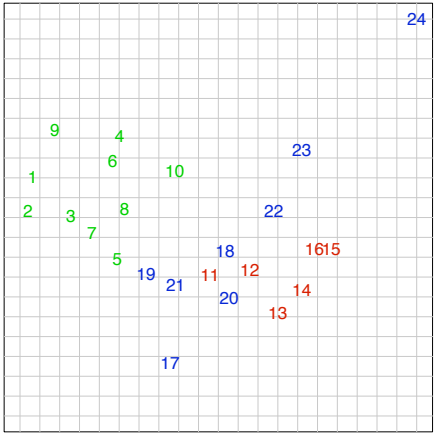
- The data are photographs with text imposed on them.
- Goal: "Erase" the text and fill in missing pieces of the photograph
- The pixels at a point are modeled as a weighted combination of factors; the weights for nearby pixels are assumed to be related

Example: Inpainting (imputation)

Example: Poetry

- A senior thesis about analyzing poetry (David Kaplan)
- Various poetry features were measured and encoded
- Each poem modeled as a weighted combination of latent components
- Plot the first two components with respect to 24 poems.

Example: Poetry



#	TITLE	AUTHOR
1	An Octopus (1935)	Moore
2	No Swan So Fine (1935)	Moore
3	The Steeple Jack (1935)	Moore
4	Poetry (1935)	Moore
5	A Grave (1935)	Moore
6	Marriage (1935)	Moore
7	The Pangolin (1941)	Moore
8	The Paper Nautilus (1941)	Moore
9	His Shield (1951)	Moore
10	Baseball and Writing (1961)	Moore
11	After Apple-Picking	Frost
12	The Wood-Pile	Frost
13	A Servant To Servants	Frost
14	Mending Wall	Frost
15	Home Burial	Frost
16	The Death of the Hired Man	Frost
17	The Day Lady Died	O'Hara
18	Ave Maria	O'Hara
19	A Step Away From Them	O'Hara
20	Music	O'Hara
21	Steps	O'Hara
22	Poem (Lana Turner has collapsed)	O'Hara
23	St. Paul and All That	O'Hara
24	Song (Is it dirty)	O'Hara

Factor 2: Two Frank O'Hara poems

Song (It is dirty)

Is it dirty
does it look dirty
that's what you think of in the city
does it just seem dirty
that's what you think of in the city
you don't refuse to breathe do you
someone comes along with a very bad character
he seems attractive. is he really. yes. very
he's attractive as his character is bad. is it. yes
that's what you think of in the city
run your finger along your no-moss mind
that's not a thought that's soot
and you take a lot of dirt off someone
is the character less bad. no. it improves constantly

you don't refuse to breathe do you

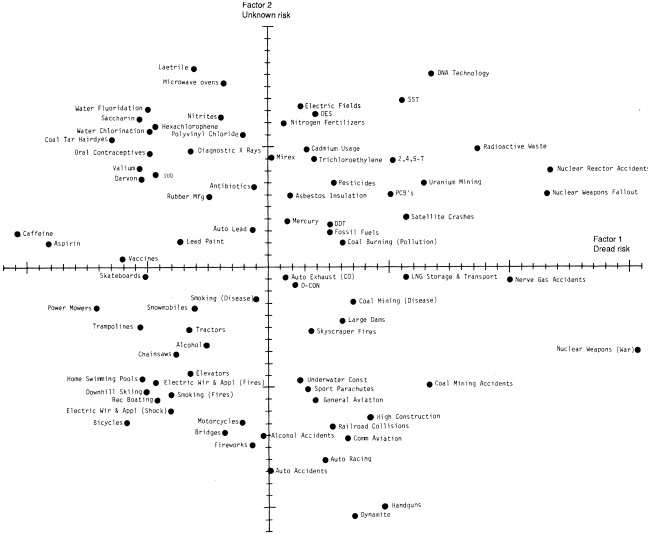
The Day Lady Died

It is 12:20 in New York a Friday
three days after Bastille day, yes
it is 1959 and I go get a shoeshine
because I will get off the 4:19 in Easthampton
at 7:15 and then go straight to dinner
and I don?t know the people who will feed me
I walk up the muggy street beginning to sun
and have a hamburger and a malted and buy
an ugly NEW WORLD WRITING to see what the poets
in Ghana are doing these days
I go on to the bank
and Miss Stillwagon (first name Linda I once heard)
doesn't even look up my balance for once in her life
and in the GOLDEN GRIFFIN I get a little Verlaine
for Patsy with drawings by Bonnard although I do
think of Hesiod, trans. Richmond Lattimore or
Brendan Behan's new play or Le Balcon or Les Nègres
of Genet, but I don't, I stick with Verlaine
after practically going to sleep with quandariness
and for Mike I just stroll into the PARK LANE
Liquor Store and ask for a bottle of Strega and
then I go back where I came from to 6th Avenue
and the tobacconist in the Ziegfeld Theatre and
casually ask for a carton of Gauloises and a carton
of Picayunes, and a NEW YORK POST with her face on it
and I am sweating a lot by now and thinking of
leaning on the john door in the 5 SPOT
while she whispered a song along the keyboard
to Mal Waldron and everyone and I stopped breathing

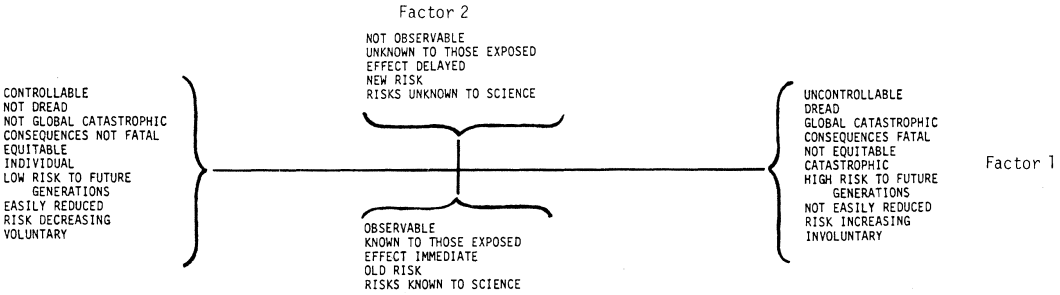
Example: Risk factors *[Slovic 1987]*

- Survey of people's perception of risk
- Answers encoded with two components
- The dimensions were rotated, interpreted
- Visualize relationship between risks and people's perception of risk

Example: Risk factors



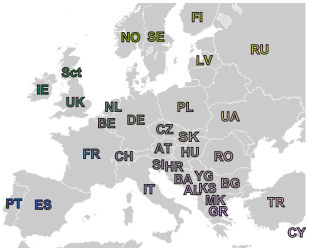
Example: Risk factors

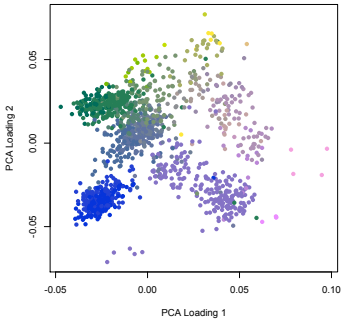


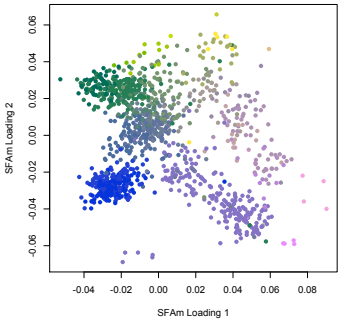
How do the components get labels and interpretations?

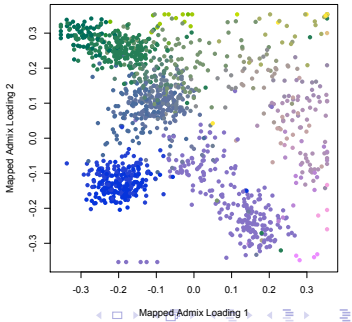
Example: Pop structure *[Novembre & Stephens 2008]*

- Measured $\sim$ 1000 individuals from across Europe at 200,000 genomic locations.
- Linear projection of this enormous matrix down to two dimensions.
- Projected data recapitulate map of Europe.



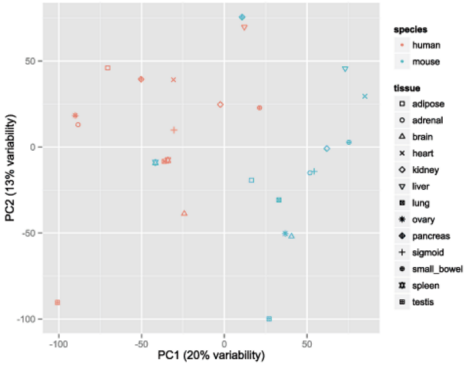


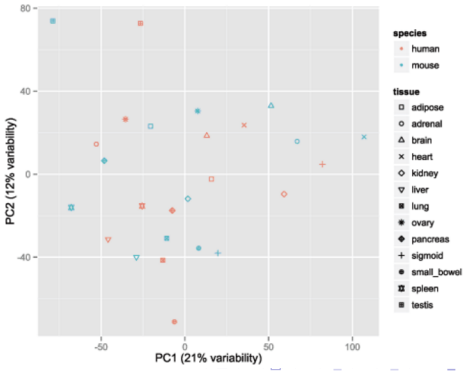




Example: Gene expression across tissues and species

- Measured $\sim$ 13 sample from two species (human, mouse)
- Linear projection of this matrix down to two dimensions
- What has more similar gene expression, tissues or species? *[Gilad & Mizrahi-Man 2008]*





Why perform dimension reduction?

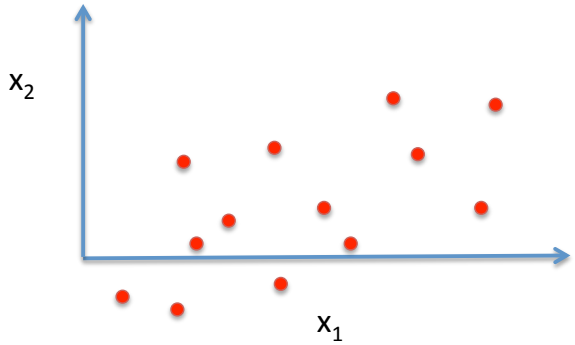
There are many reasons for performing dimension reduction

- *Data compression:* Store data using reduced representation
- *Explore data*: we can visualize high dimensional data in 2D
- *Generalize data*: understand low dimensional structure by studying reduced representation.
- *Imputing data*: fill in missing values of matrix X.
- *Predict data*: given a new sample with no observations, try to predict based on observed patterns in data.

Principal component analysis (PCA)

PCA is the workhorse of dimension reduction methods. Invented multiple times in the last century *[Pearson 1901, Hotelling 1933]*.

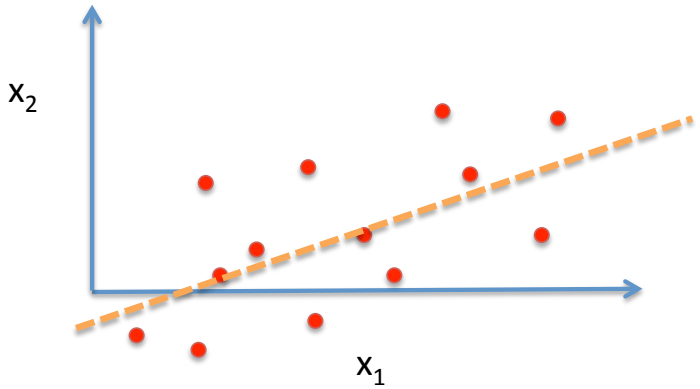
Idea: linearly project observations onto low dimensional subspace in the original feature space.



Let's reduce the dimension of 2D data to 1D.

Principal component analysis

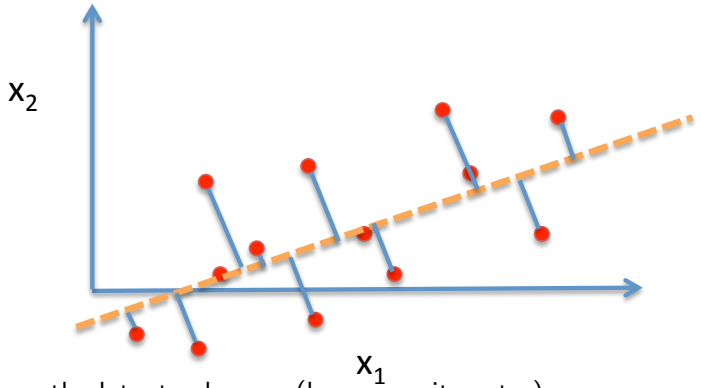
Idea: linearly project observations onto low dimensional subspace in the original feature space.



What is the parameter Λ? What are the latent variables Z?

Principal component analysis

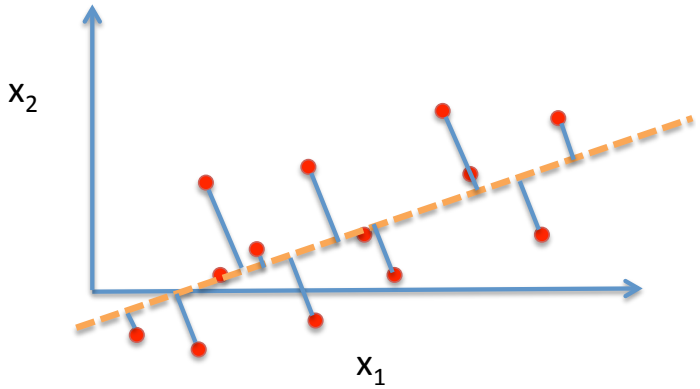
Idea: linearly project observations onto low dimensional subspace in the original feature space.



- Parameters Λ are the latent subspace (here: a unit vector)
- Latent variables Z (*principal components*) are the projections of original points onto line (here: a 1D scalar value)

Principal component analysis

Idea: linearly project observations onto low dimensional subspace in the original feature space.



Note: this is not linear regression. Why not?

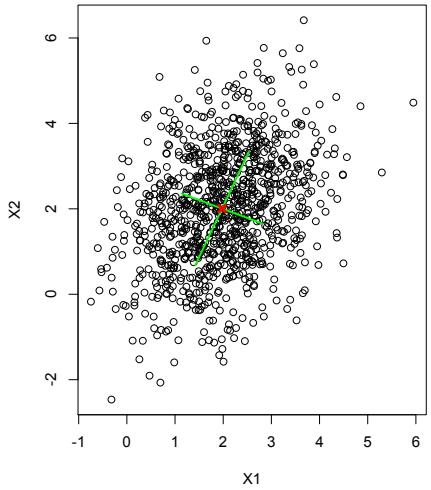
PCA: define the subspace

There are three ways to interpret how PCA defines latent subspace Z

1. **Maximize the variance** of the projection along the K orthogonal components in Z
2. Minimize the **reconstruction error**: $||X - Z\Lambda||$
3. MLE of a parameter in a **latent variable probabilistic model**

PCA interpretations: maximizing variance

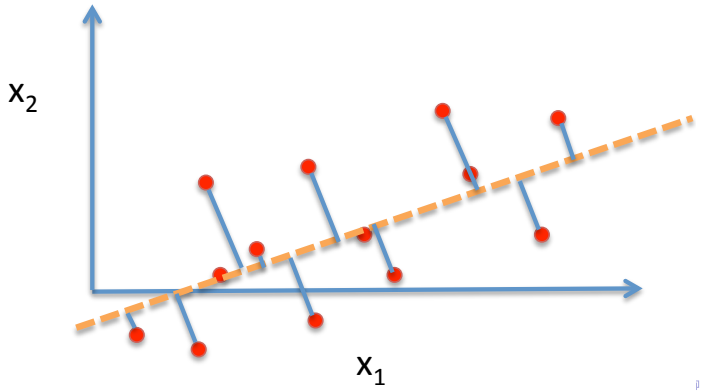
Maximize the variance of the projection along each of the K components *[Hotelling 1933]*.



PCA interpretations: minimizing reconstruction error

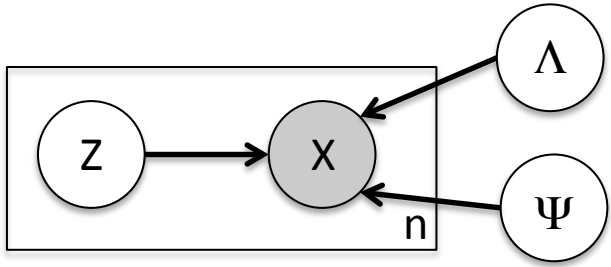
Minimize the **reconstruction error**, i.e., the distance between the original data and its "estimate" in the low dimensional subspace *[Pearson 1901]*.

$$arg \min_{Z,\Lambda} ||X - Z\Lambda||$$



PCA interpretations: MLE

PCA is the maximum likelihood estimate of a parameter in a latent variable probabilistic model *[Tipping 1999, Collins 2002, Roweis 1998]*.



PCA interpretations

We will discuss the first and third perspectives:

- the first because, mathematically, it is important to know how to compute principal components;
- the third because it illuminates how PCA generalizes, although it does not lead to the best algorithms.

Both the optimization problem and the probabilistic model require an understanding of the multivariate Gaussian distribution.

The multivariate Gaussian

A multivariate Gaussian is a joint Gaussian distribution for p dimensional vectors X_i. The parameters are:

- μ: The mean, a $p \times 1$ vector.
- Σ: The covariance matrix, a $p \times p$ positive definite symmetric matrix. (Positive definiteness means that $x^\top \Sigma x > 0$ for all x.)

Each element of Σ is a covariance between the p features,

$$\sigma^2_{j,\ell} = \mathrm{E}[X_j X_\ell] - \mathrm{E}[X_j]\mathrm{E}[X_\ell].$$

The variances of each feature lie along the diagonal.

Multivariate Gaussian density

The density of a multivariate Gaussian is

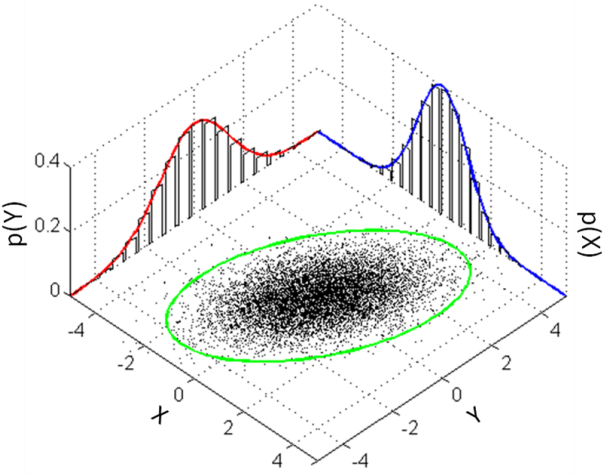
$$p(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right\}.$$

- The first term ensures that the PDF integrates to one.
- The second term is a quadratic form.
- The Gaussian distribution is in the exponential family.

  Exercise: show this.

Multivariate Gaussian

The function $f(x) = (1/2)(x - \mu)^\top \Sigma^{-1}(x - \mu)$ defines contours of equal probability.



MLE of a multivariate Gaussian

The data are $\{X_1, \ldots, X_n\}$, where X_i is a p vector observation.

The MLEs of the MVN parameters are

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

and

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mu})(X_i - \hat{\mu})^\top.$$

These generalize the 1-dimensional case.

Exercise: derive these MLEs.

Marginals and conditionals of the multivariate Gaussian

Let's consider X_1 and X_2 to be a split of the MVN data in dimension $p \geq 2$, where X_1 and X_2 are vectors.

Suppose that X_1 and X_2 are jointly Gaussian:

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \quad \text{and hence } \Psi = \Sigma^{-1} = \begin{bmatrix} \Psi_{11} & \Psi_{12} \\ \Psi_{21} & \Psi_{22} \end{bmatrix}$$

And so we can write:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right]$$

Marginal and conditionals of the multivariate Gaussian

Using this model, we derive the marginal and conditional distributions:

Marginal distribution

$$p(X_1) = \mathcal{N}(X_1|\mu_1, \Sigma_{11});$$

Conditional distribution

$$p(X_1|X_2) = \mathcal{N}(X_1|\mu_{1|2}, \Sigma_{1|2}),$$

where

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2),$$

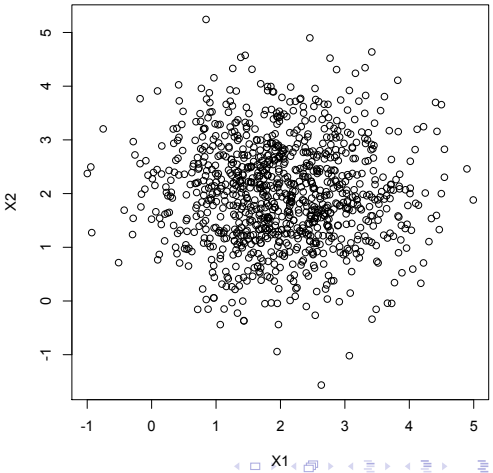
$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

These formulas are derived using the Schur complement of a matrix and the matrix inversion lemma. Exercise: show this.

Multivariate Gaussian example: independent dimensions

Let's generate 1000 points with $p = 2$ from a multivariate Gaussian, where the dimensions are independent and of equal variance:

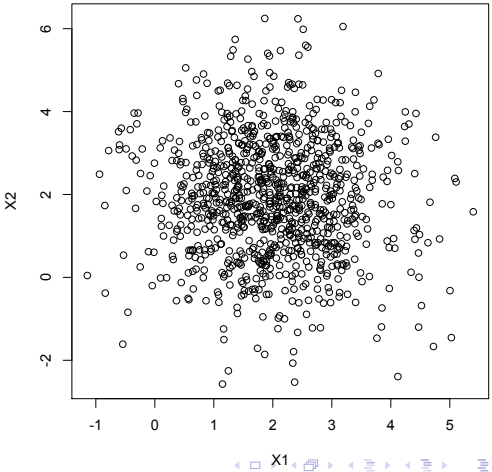
$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 2 \\ 2 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right]$$



Multivariate Gaussian example: independent dimensions

Let's generate 1000 points with $p = 2$ from a multivariate Gaussian, where the dimensions are independent and have different variances:

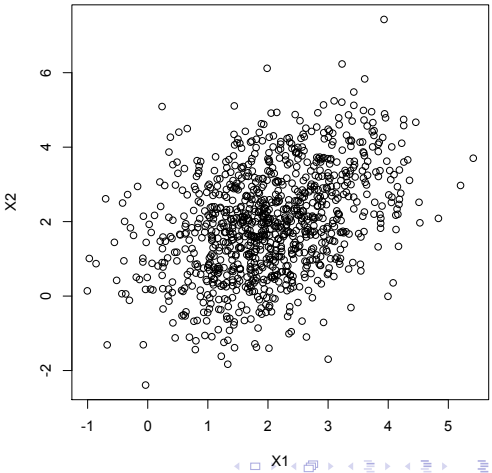
$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}\right]$$



Multivariate Gaussian example: non-independent dimensions

Let's generate 1000 points with $p = 2$ from a multivariate Gaussian, where the dimensions are not independent:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix}\right]$$



Principal components of a set of samples

Principal components performs an orthogonal transformation on a set of correlated features to produce a set of linearly uncorrelated feature vectors.

These orthogonal vectors are called *principal components* (PCs)

- There are at most $\min(p, n)$ PCs for n samples with p features.
- PCs are ordered by proportion of variance that they explain in X.
- First PC captures direction of greatest variance in observations;
- latter PCs capture direction of greatest variance in residuals; orthogonal to all previous PCs.

Principal components & orthogonality

- *Orthogonal vectors*: a pair of vectors are orthogonal if their inner product is zero.
- In two dimensions, orthogonal vectors are at right angles to each other
- Orthogonal vectors are linearly uncorrelated, and appear statistically independent.

Principal components: computation

- PCs are the eigenvectors of the empirical covariance matrix (i.e., for mean-centered samples, $X^T X$)
- Assumption is that underlying data are multivariate Gaussian
- Then PCs are eigenvectors of the covariance matrix of X
- Transformations of X that affect covariance matrix will affect PCs.

Why are PCs eigenvectors of covariance of X?

Principal components: computation

- First, define $Z \in \Re^{K \times n}$ as the set of PCs of matrix X.
- Assumption is that underlying data are multivariate Gaussian
- Compute first PC $z_{i,1} = w_1 x_i$. For orthonormal weights $w_1 \in \Re^p$:

$$\begin{aligned} w_1 &= \arg \max_{||w||=1} \left\{ \sum_{i=1}^{n} (w_1 x_i)^2 \right\} \\ &= \arg \max_{||w||=1} \left\{ w^T X^T X w \right\} \\ &= \arg \max_{w} \left\{ \frac{w^T X^T X w}{w^T w} \right\} \end{aligned}$$

- w_1 corresponds to first eigenvector of empirical covariance of X.

Principal components: computation

- Then, we compute the kth PC $z_{i,k} = w_k x_i$: for weight $w_k \in \Re^p$.
- First, compute $\hat{X}_k = X - \sum_{\ell=1}^{k-1} X w_\ell w_\ell^T$. Then,

$$w_k = \arg\max_w \left\{ \frac{w^T \hat{X}_k^T \hat{X}_k w}{w^T w} \right\}$$

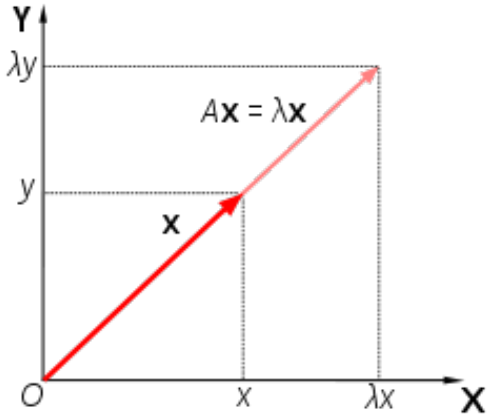
- w_k corresponds to kth eigenvector of empirical covariance of X.

Eigenvectors and eigenvalues: definition

Recall: The function $f(x) = (1/2)(x - \mu)^\top \Sigma^{-1}(x - \mu)$ defines contours of equal probability for a multivariate Gaussian.

Let $A = X^T X$ be a covariance matrix, x (here) be an eigenvector of A, and λ be the corresponding eigenvalue.

Then define x and λ through $Ax = \lambda x$: eigenvectors x of A are the unique vectors for which projection of A onto the eigenvector has the same value as a λ-scaled vector.



[Figure source: wikipedia]

Multivariate Gaussian example: independent dimensions

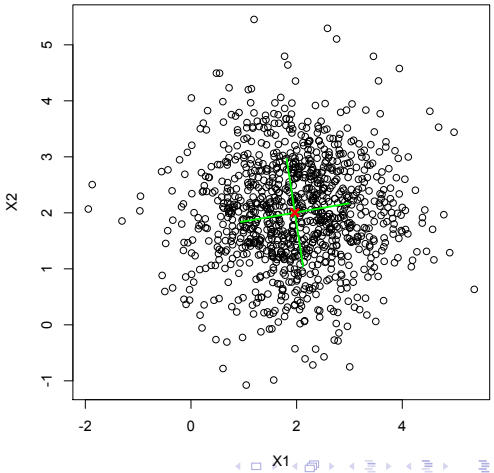
Let's generate 1000 points with $p = 2$ from a multivariate Gaussian, where the dimensions are independent and of equal variance:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right]$$

We can compute PCs and plot the weight vectors.

The length of the kth PC is plotted as the square root of the kth eigenvalue.

Would these be the same vectors for another sample from this MVN?

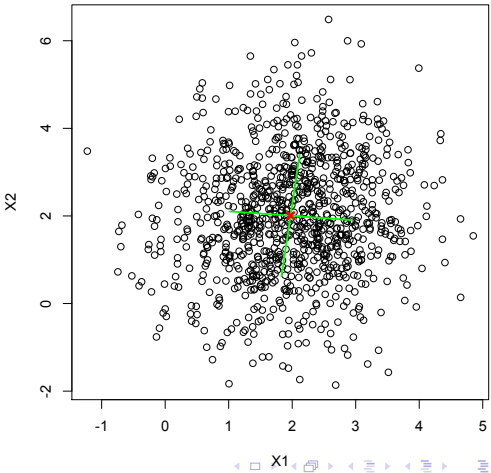


PCs example: independent dimensions

Let's generate 1000 points with $p = 2$ from a multivariate Gaussian, where the dimensions are independent and have different variances:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}\right]$$

We can compute PCs and plot the weight vectors.

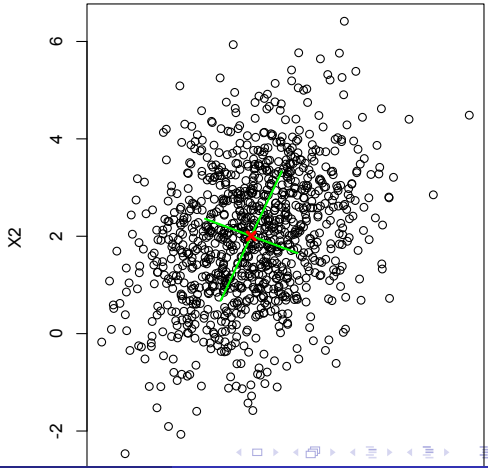


PCs example: non-independent dimensions

Let's generate 1000 points with $p = 2$ from a multivariate Gaussian, where the dimensions are not independent:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix}\right]$$

We can compute PCs and plot the weight vectors.

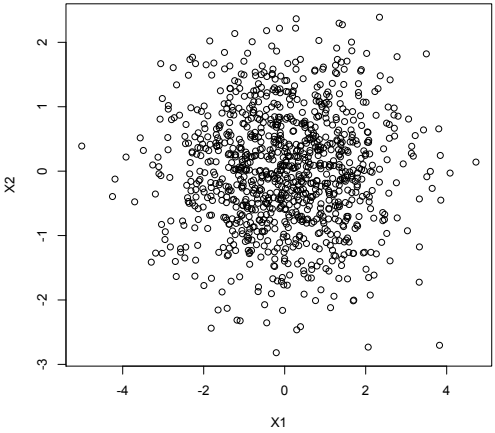


PCs example: non-independent dimensions

We can then plot the PCs from the 1000 samples from:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix}\right]$$

Note that the two dimensions are statistically uncorrelated.

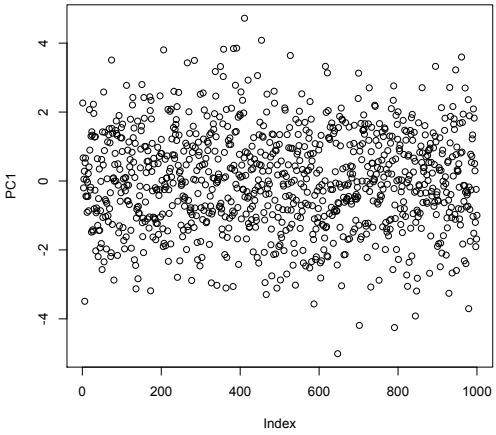


PCs example: dimension reduction

We can project the 1000 samples from:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix}\right]$$

to the first PC only.



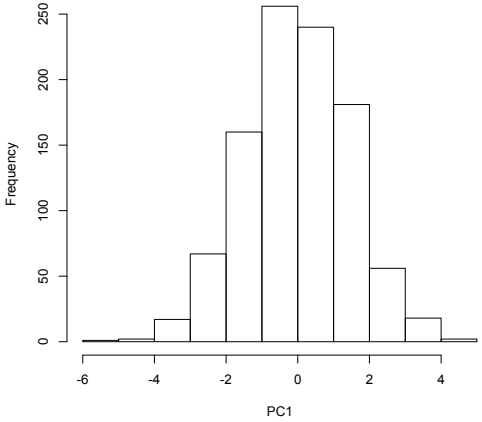
PCs example: dimension reduction

We can project the 1000 samples from:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix}\right]$$

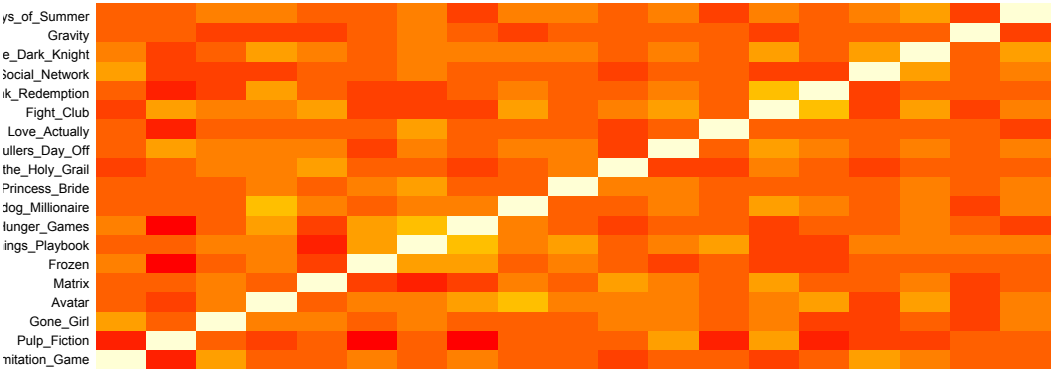
to the first PC only.

Here, I've plotted the histogram of this first PC.



PCs example: Movie ratings

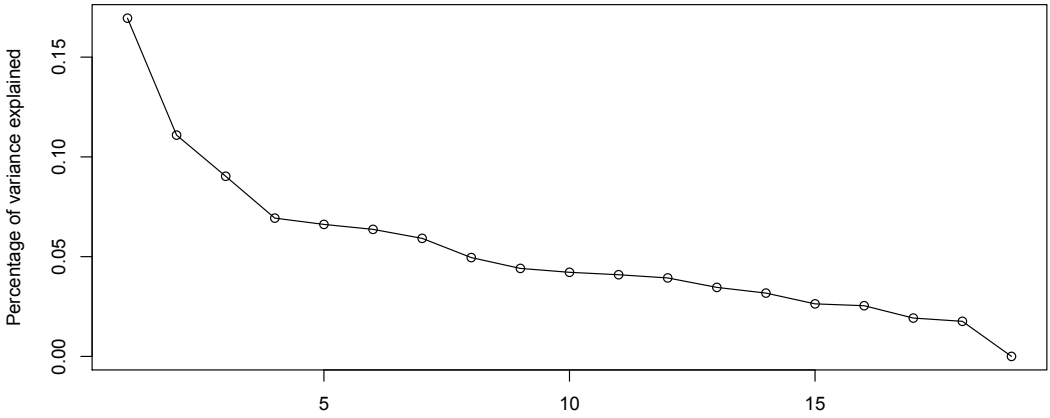
Let's look at the movie ratings data we collected at the start of the class.



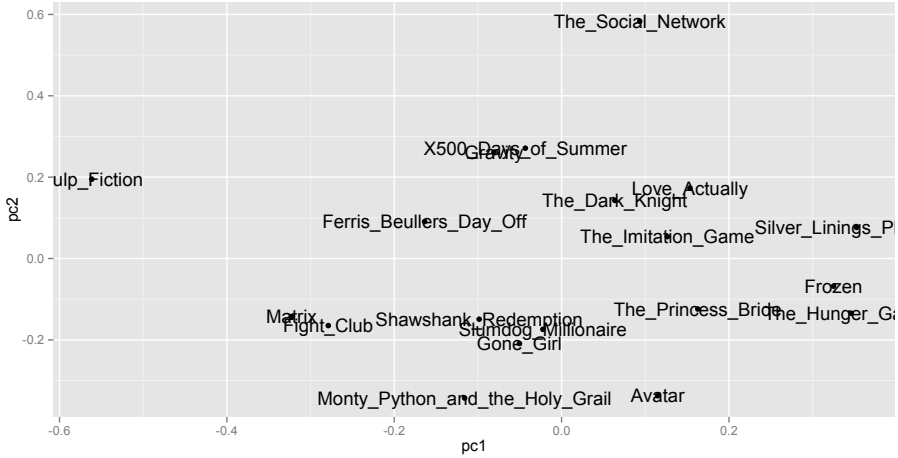
Here, we compute the covariance after mean-centering the movie ratings for each movie.

PCs example: Movie ratings

We can plot the percentage variance explained by each component (normalized eigenvalues, for eigenvalues all greater than or equal to 0)

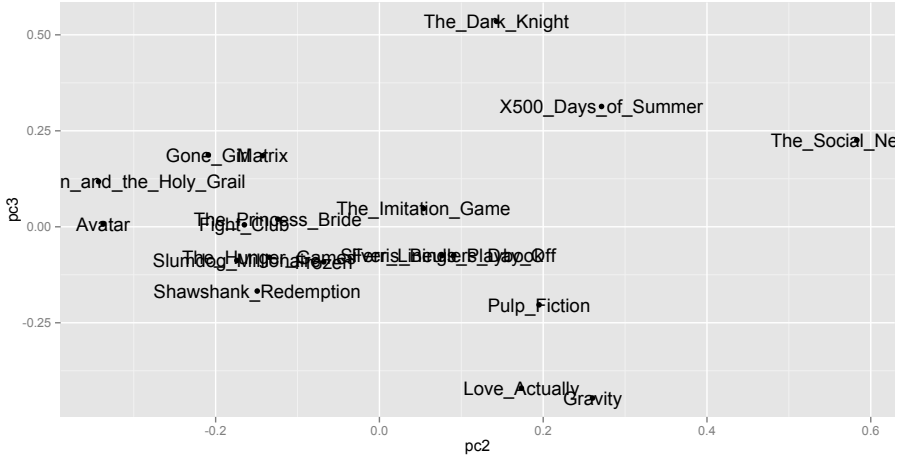


PCs example: Movie ratings, PC1 vs PC2



- PC1: *Pulp Fiction* versus mainstream
- PC2: ??

PCs example: Movie ratings, PC2 vs PC3



- PC2: ??
- PC3: ??

A note on dimensions of principal component basis

Let's consider PCs with respect to the n samples and p features

- The number of dimensions of the underlying basis of the PCs is at most $K \leq \min(n, p)$.
- Often, data lie on a much lower dimension $K \ll \min(n, p)$
- When $p > n$, empirical covariance of X not positive definite; some zero or negative eigenvalues
- As with other unsupervised learning methods, exercise caution when selecting K

Computing PCs using singular value decomposition

Practically, we can compute PCs using singular value decomposition.

$$X = U\Sigma V^T$$

In this equation:

- U is a $n \times n$ matrix with orthonormal columns (left singular values)
- Σ is $n \times p$ diagonal matrix (singular values, square roots of eigenvals)
- V is a $p \times p$ matrix with orthonormal columns (right singular values)

To compute K PCs:

- First K columns of $U\Sigma$ are PCs Z
- First K columns of V are weights (eigenvectors)

Probabilistic PCA

Let's look at the probabilistic interpretation of PCA.

As we've seen, mean squared error can be reinterpreted with Gaussian distributions.

In probabilistic PCA,

$$\begin{aligned} \vec{z}_i &\sim \mathcal{N}_K(0, I) \\ \vec{x}_i \mid \vec{z}_i &\sim \mathcal{N}_p(\vec{\mu} + \Lambda^T \vec{z}_i, I\psi) \end{aligned}$$

In this equation:

- μ is a p-vector
- Λ is a $K \times p$ matrix
- ψ is a scalar (residual variance)

Principal components analysis model

We can assume $\mu = 0$ (i.e., mean center the data). Then

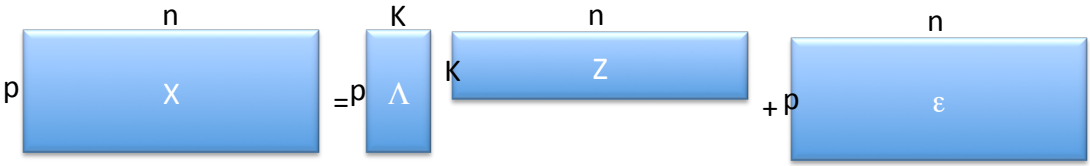
$$\begin{aligned} \vec{z}_i &\sim \mathcal{N}_K(0, I) \\ \vec{x}_i \mid \vec{z}_i &\sim \mathcal{N}_p(\Lambda^T \vec{z}_i, I\vec{\psi}) \end{aligned}$$

How can we compare this model to the optimization formulation:

- No explicit constraint of orthogonality on Λ
- PCA is a MLE solution to this model (with orthogonality constraints)

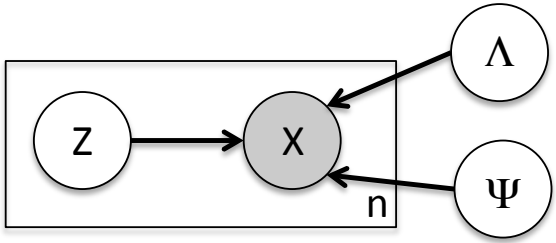
Matrix factorization

Why PCA can be thought of as matrix factorization:



Graphical model of PCA

Graphical model representation:



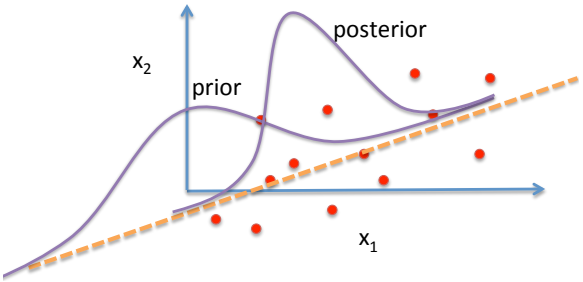
This looks a lot like a mixture model.

In PPCA, z_i is a "latent variable representation" of x_i

$$\mathrm{E}[\vec{X}_i \mid \vec{z}_i] = z_{i,1}\vec{\lambda}_1 + z_{i,2}\vec{\lambda}_2 + \ldots + z_{i,K}\vec{\lambda}_K$$

where $\vec{\lambda}_k$ is kth row of Λ.

Intuition for probabilistic PCA



- The low-dimensional space is the latent variable z
- Before observing x, z has a Gaussian prior distribution
- After observing x, the conditional distribution of $z \mid x, \Lambda$ is Gaussian
- Posterior expectation of z is a low dimensional linear projection of x

Assumptions of PCA

PCA generally assumes:

- features are mean centered
- observations are marginally Gaussian
- residual variance the same across features
- latent structure is low dimensional
- latent structure is a linear subspace

Optimization approach to PCA

- Dimensions of low dimensional space are orthogonal

Probabilistic approach to PCA

- Unidentifiable with respect to labels, scale, rotation

Extensions to PCA and related methods

- Factor analysis: next class
- Bayesian PCA: Regularize with appropriate Bayesian priors
- Independent component analysis (ICA): non-Gaussian z
- Canonical component analysis (CCA): multiple observations
- Latent Dirichlet Allocation – in this course
- Non-negative matrix factorization (NMF)
- Kernelized PCA: project observations to higher dimension
- Linear discriminant analysis (Fisher): PCA but includes class labels
- Sparse PCA: add sparsity in the weight matrix
- Nonlinear PCA: nonlinear projection to latent dimensions
- Many more...

Additional Resources

- MLAPA: Chapter 12
- *Elements of Statistical Learning*: Chapter 14
- Probabilistic PCA: *[Roweis 1998]*
- Probabilistic PCA: *[Tipping 1999]*
- EM and latent factor models: *[Ghahramani & Hinton 1998]*
- Metacademy: *principal component analysis*