All you want to know about GPs: Linear Dimensionality Reduction

Raquel Urtasun and Neil Lawrence

TTI Chicago, University of Sheffield

June 16, 2012

Notation

p	data dimensionality	
q	latent dimensionality	
n	number of data points	
$\mathbf{Y}$	*design matrix* containing our data	$n \times p$
$\mathbf{X}$	matrix of latent variables	$n \times q$

Row vector from matrix $\mathbf{A}$ given by $\mathbf{a}_{i,:}$ column vector $\mathbf{a}_{:,j}$ and element given by $a_{i,j}$.

Online Resources

All source code and slides are available online

- Tutorial homepage is
  - http://ttic.uchicago.edu/~rurtasun/tutorials/GP_tutorial.html.
  - Code available at http://staffwww.dcs.shef.ac.uk/people/N.Lawrence/.

High Dimensional Data

- High dimensional data dominates many application domains.
- Examples include:

  a customer in a data base, where the features might include their purchase history, where they live, their sex, and age;

High Dimensional Data

- High dimensional data dominates many application domains.
- Examples include:

  a customer in a data base, where the features might include their purchase history, where they live, their sex, and age;

  a digitized photograph, where the features include the pixel intensities, time, date, and location of the photograph;

High Dimensional Data

- High dimensional data dominates many application domains.
- Examples include:

a customer in a data base, where the features might include their purchase history, where they live, their sex, and age;

a digitized photograph, where the features include the pixel intensities, time, date, and location of the photograph;

human motion capture data for the movie and games industries, where features consist of a time series of angles at each joint;

High Dimensional Data

- High dimensional data dominates many application domains.
- Examples include:

a customer in a data base, where the features might include their purchase history, where they live, their sex, and age;

a digitized photograph, where the features include the pixel intensities, time, date, and location of the photograph;

human motion capture data for the movie and games industries, where features consist of a time series of angles at each joint;

human speech, where the features consist of the energy at different frequencies (or across the cepstrum) as a time series;

High Dimensional Data

- High dimensional data dominates many application domains.
- Examples include:

a customer in a data base, where the features might include their purchase history, where they live, their sex, and age;

a digitized photograph, where the features include the pixel intensities, time, date, and location of the photograph;

human motion capture data for the movie and games industries, where features consist of a time series of angles at each joint;

human speech, where the features consist of the energy at different frequencies (or across the cepstrum) as a time series;

a webpage or other document, features could consist of frequencies of given words in a set of documents and linkage information between documents;

High Dimensional Data

- High dimensional data dominates many application domains.
- Examples include:

a customer in a data base, where the features might include their purchase history, where they live, their sex, and age;

a digitized photograph, where the features include the pixel intensities, time, date, and location of the photograph;

human motion capture data for the movie and games industries, where features consist of a time series of angles at each joint;

human speech, where the features consist of the energy at different frequencies (or across the cepstrum) as a time series;

a webpage or other document, features could consist of frequencies of given words in a set of documents and linkage information between documents;

gene expression data, features consist of the level of expression of thousands of genes.

High Dimensional Data

- High dimensional data dominates many application domains.
- Examples include:

  a customer in a data base, where the features might include their purchase history, where they live, their sex, and age;

  a digitized photograph, where the features include the pixel intensities, time, date, and location of the photograph;

  human motion capture data for the movie and games industries, where features consist of a time series of angles at each joint;

  human speech, where the features consist of the energy at different frequencies (or across the cepstrum) as a time series;

  a webpage or other document, features could consist of frequencies of given words in a set of documents and linkage information between documents;

  gene expression data, features consist of the level of expression of thousands of genes.

Mixtures of Gaussians

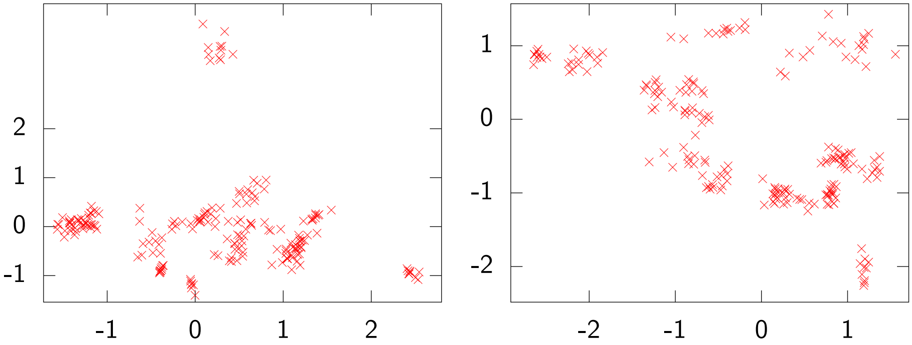


Figure: Two dimensional data sets.

Mixtures of Gaussians

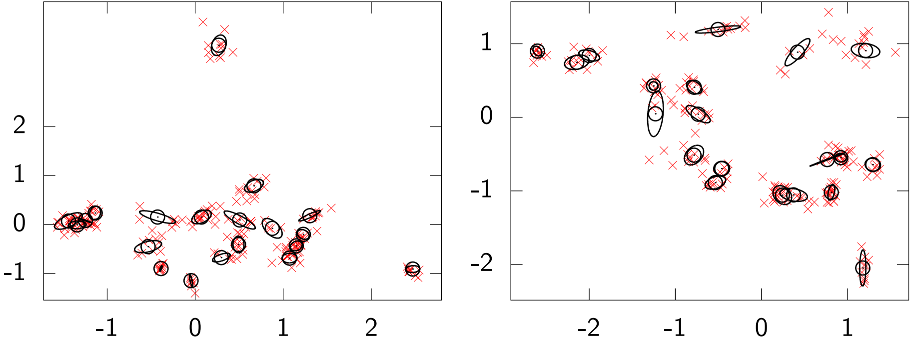


Figure: Complex structure not a problem for mixtures of Gaussians.

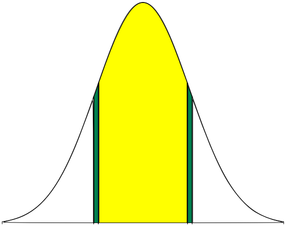
Thinking in High Dimensions

- Two dimensional plots of Gaussians can be misleading.
- Our low dimensional intuitions can fail dramatically.
- Two major issues:
    1. In high dimensions all the data moves to a 'shell'. There is nothing near the mean!
    2. Distances between points become constant.
    3. These affects apply to many densities.
- Let's consider a Gaussian "egg".

Thinking in High Dimensions

- Two dimensional plots of Gaussians can be misleading.
- Our low dimensional intuitions can fail dramatically.
- Two major issues:
  1. In high dimensions all the data moves to a 'shell'. There is nothing near the mean!
  2. Distances between points become constant.
  3. These affects apply to many densities.
- Let's consider a Gaussian "egg".

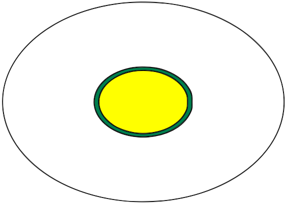
The Gaussian Egg



Volumes: 65.8%, 4.8% 29.4%

Figure: One dimensional Gaussian density.

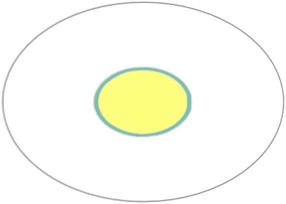
The Gaussian Egg



Volumes: 59.4%, 7.4% 33.2%

Figure: Two dimensional Gaussian density.

The Gaussian Egg



Volumes: 56.1%, 9.2%, 34.7%

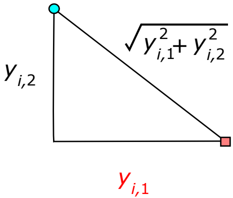
Figure: Three dimensional Gaussian density.

Mathematics

What is the density of probability mass?

$$y_{i,k} \sim \mathcal{N}\left(0, \sigma^2\right)$$

$$\implies y_{i,k}^2 \sim \sigma^2 \chi_1^2$$



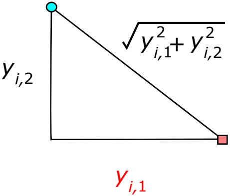
Square of sample from Gaussian is scaled chi-squared density

Mathematics

What is the density of probability mass?

$$y_{i,k} \sim \mathcal{N}\left(0, \sigma^2\right)$$

$$\implies y_{i,k}^2 \sim \mathcal{G}\left(\frac{1}{2}, \frac{1}{2\sigma^2}\right)$$



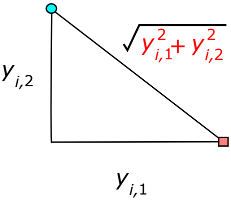
Chi squared density is a variant of the gamma density with shape parameter $a = \frac{1}{2}$, rate parameter $b = \frac{1}{2\sigma^2}$, $\mathcal{G}\left(x|a, b\right) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$.

Mathematics

What is the density of probability mass?

$$y_{i,k} \sim \mathcal{N}\left(0, \sigma^2\right)$$

$$\implies y_{i,1}^2 + y_{i,2}^2 \sim \mathcal{G}\left(\frac{2}{2}, \frac{1}{2\sigma^2}\right)$$



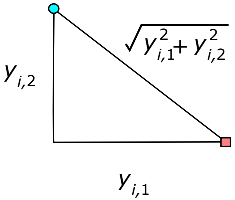
Addition of gamma random variables with the same rate is gamma with sum of shape parameters ($y_{i,k}$s are independent)

Mathematics

What is the density of probability mass?

$$\sum_{k=1}^{p} y_{i,k}^2 \sim \mathcal{G}\left(\frac{p}{2}, \frac{1}{2\sigma^2}\right)$$

$$\implies \left\langle \sum_{k=1}^{p} y_{i,k}^2 \right\rangle = p\sigma^2$$



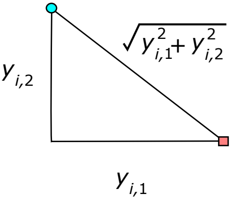
Addition of gamma random variables with the same rate is gamma with sum of shape parameters ($y_{i,k}$s are independent)

Mathematics

What is the density of probability mass?

$$\frac{1}{p}\sum_{k=1}^{p} y_{i,k}^2 \sim \mathcal{G}\left(\frac{p}{2}, \frac{p}{2\sigma^2}\right)$$

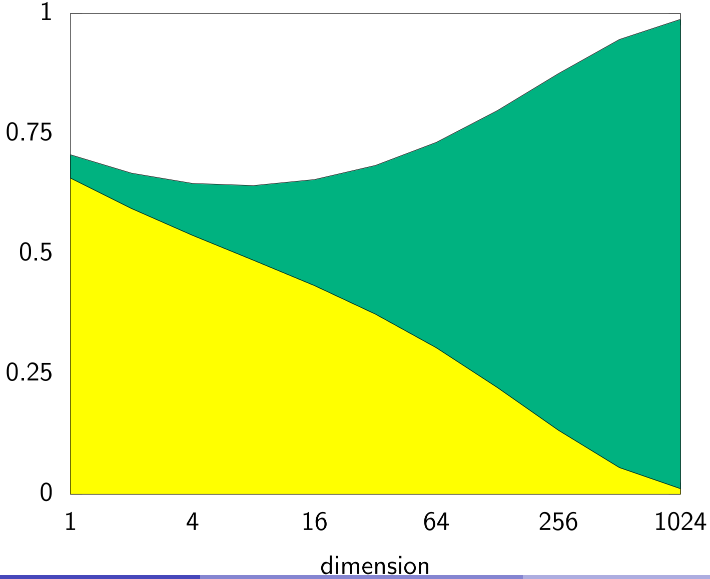
$$\implies \left\langle \frac{1}{p}\sum_{k=1}^{p} y_{i,k}^2 \right\rangle = \sigma^2$$



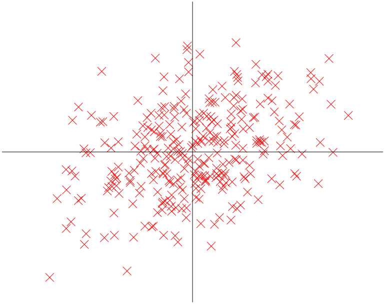
Scaling of gamma density scales the rate parameter

Where is the Mass?

- Squared distances are gamma distributed.



Looking at Gaussian Samples



Interpoint Distances

- The other effect in high dimensions is all points become equidistant.
- Can show this for Gaussians with a similar proof to the above,

$$y_{i,k} \sim \mathcal{N}\left(0, \sigma_k^2\right) \qquad y_{j,k} \sim \mathcal{N}\left(0, \sigma_k^2\right)$$

$$y_{i,k} - y_{j,k} \sim \mathcal{N}\left(0, 2\sigma_k^2\right)$$

$$\left(y_{i,k} - y_{j,k}\right)^2 \sim \mathcal{G}\left(\frac{1}{2}, \frac{1}{4\sigma_k^2}\right)$$

Interpoint Distances

- The other effect in high dimensions is all points become equidistant.
- Can show this for Gaussians with a similar proof to the above,

$$y_{i,k} \sim \mathcal{N}\left(0, \sigma_k^2\right) \qquad y_{j,k} \sim \mathcal{N}\left(0, \sigma_k^2\right)$$

$$y_{i,k} - y_{j,k} \sim \mathcal{N}\left(0, 2\sigma_k^2\right)$$

$$\left(y_{i,k} - y_{j,k}\right)^2 \sim \mathcal{G}\left(\frac{1}{2}, \frac{1}{4\sigma_k^2}\right)$$

For spherical Gaussian, $\sigma_k^2 = \sigma^2$

$$\sum_{k=1}^{p} \left(y_{i,k} - y_{j,k}\right)^2 \sim \mathcal{G}\left(\frac{p}{2}, \frac{1}{4\sigma^2}\right)$$

$$\frac{1}{p}\sum_{k=1}^{p} \left(y_{i,k} - y_{j,k}\right)^2 \sim \mathcal{G}\left(\frac{p}{2}, \frac{p}{4\sigma^2}\right)$$

Dimension normalized distance between points is drawn from a gamma. Mean is $2\sigma^2$. Variance is $\frac{8\sigma^2}{p}$.

Interpoint Distances

- The other effect in high dimensions is all points become equidistant.
- Can show this for Gaussians with a similar proof to the above,

$$y_{i,k} \sim \mathcal{N}\left(0, \sigma_k^2\right) \qquad y_{j,k} \sim \mathcal{N}\left(0, \sigma_k^2\right)$$

$$y_{i,k} - y_{j,k} \sim \mathcal{N}\left(0, 2\sigma_k^2\right)$$

$$\left(y_{i,k} - y_{j,k}\right)^2 \sim \mathcal{G}\left(\frac{1}{2}, \frac{1}{4\sigma_k^2}\right)$$

For spherical Gaussian, $\sigma_k^2 = \sigma^2$

$$\sum_{k=1}^{p} \left(y_{i,k} - y_{j,k}\right)^2 \sim \mathcal{G}\left(\frac{p}{2}, \frac{1}{4\sigma^2}\right)$$

$$\frac{1}{p}\sum_{k=1}^{p} \left(y_{i,k} - y_{j,k}\right)^2 \sim \mathcal{G}\left(\frac{p}{2}, \frac{p}{4\sigma^2}\right)$$

Dimension normalized distance between points is drawn from a gamma. Mean is $2\sigma^2$. Variance is $\frac{8\sigma^2}{p}$.

Central Limit Theorem and Non-Gaussian Case

- We can compute the density of squared distance *analytically* for spherical, independent Gaussian data.
- More generally, for *independent* data, the *central limit theorem* applies.

Central Limit Theorem and Non-Gaussian Case

- We can compute the density of squared distance *analytically* for spherical, independent Gaussian data.
- More generally, for *independent* data, the *central limit theorem* applies.
  - The mean squared distance in high dimensional space is the mean of the variances.

Central Limit Theorem and Non-Gaussian Case

- We can compute the density of squared distance *analytically* for spherical, independent Gaussian data.
- More generally, for *independent* data, the *central limit theorem* applies.
  - The mean squared distance in high dimensional space is the mean of the variances.
  - The variance about the mean scales as p^{-1}.

Central Limit Theorem and Non-Gaussian Case

- We can compute the density of squared distance *analytically* for spherical, independent Gaussian data.
- More generally, for *independent* data, the *central limit theorem* applies.
  - The mean squared distance in high dimensional space is the mean of the variances.
  - The variance about the mean scales as p^{-1}.

Summary until now

- In high dimensions if individual dimensions are *independent* the distributions behave counter intuitively.
- All data sits at one standard deviation from the mean.
- The densities of squared distances can be analytically calculated for the Gaussian case.
- For non-Gaussian *independent* systems we can invoke the central limit theorem.
- Next we will consider example data sets and see how their interpoint distances are distributed.

Sanity Check

Data sampled from independent Gaussian distribution

- If dimensions are independent, we expect low variance, Gaussian behavior for the distribution of squared distances.

Distance distribution for a Gaussian with $p = 1000$, $n = 1000$

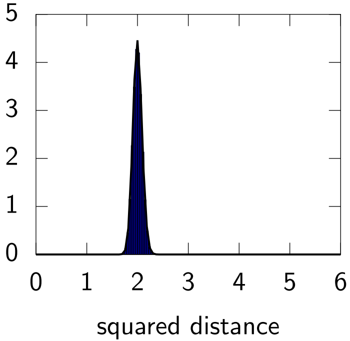


Figure: A good match betwen theory and the samples for a 1000 dimensional Gaussian distribution.

Sanity Check

Same data generation, but fewer data points.

- If dimensions are independent, we expect low variance, Gaussian behaviour for the distribution of squared distances.

Distance distribution for a Gaussian with $p = 1000$, $n = 100$

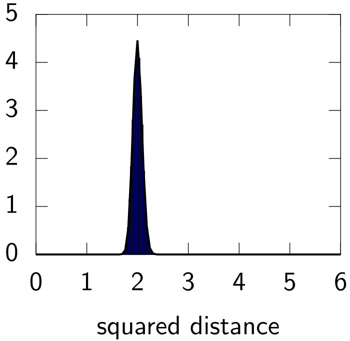
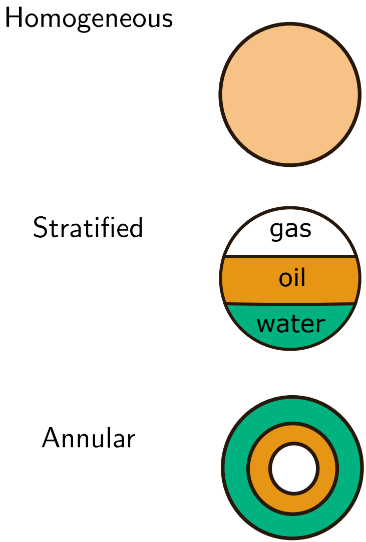


Figure: A good match betwen theory and the samples for a 1000 dimensional Gaussian distribution.

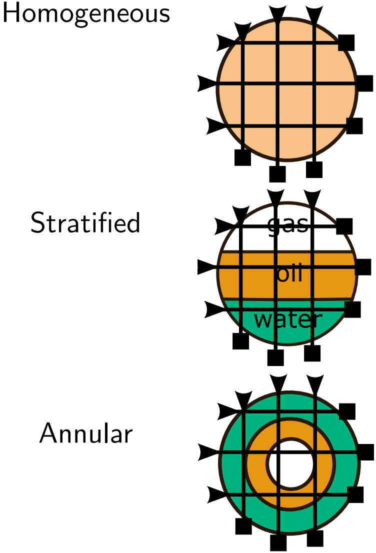
Oil Data

- Simulated measurements from an oil pipeline (Bishop 93)
- Pipeline contains oil, water and gas.
- Three phases of flow in pipeline—homogeneous, stratified and annular.
- Gamma densitometry sensors arranged in a configuration around pipeline.



Oil Data

- Simulated measurements from an oil pipeline (Bishop 93)
- Pipeline contains oil, water and gas.
- Three phases of flow in pipeline—homogeneous, stratified and annular.
- Gamma densitometry sensors arranged in a configuration around pipeline.



Oil Data

- 12 simulated measurements of oil flow in a pipe.
- Nature of flow is dependent on relative proportion of oil, water and gas.

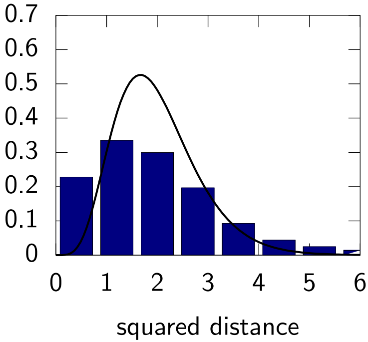
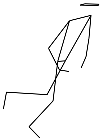


Figure: Interpoint squared distance distribution for oil data with $p = 12$ (variance of squared distances is 1.98 vs predicted 0.667).

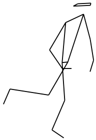
Stick Man Data

- $n = 55$ frames of motion capture.
- *xyz* locations of 34 points on the body.
- $p = 102$ dimensional data.
- "Run 1" available from `http://accad.osu.edu/research/mocap/mocap_data.htm`.

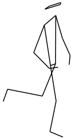
Changing



Angle



of Run



Stick Man

- Motion capture data inter point distance histogram.

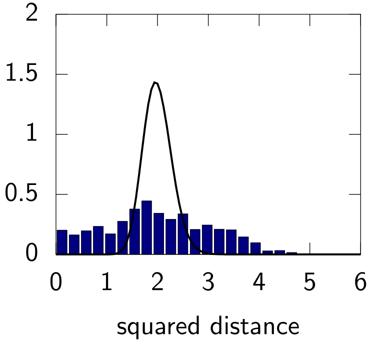


Figure: Interpoint squared distance distribution for stick man data with $p = 102$ (variance of squared distances is 1.09 vs predicted 0.0784).

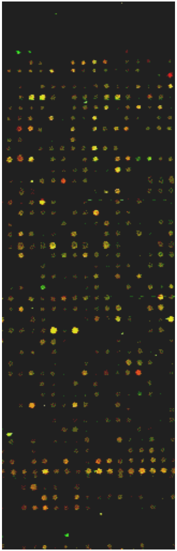
Microarray Data

- Gene expression measurements reflecting the cell cycle in yeast (Spellman 98)
- $p = 6,178$ Genes measured for $n = 77$ experiments
- Data available from `http://genome-www.stanford.edu/cellcycle/data/rawdata/individual.htm`.

Yeast

Cell

Cycle



Microarray Data

- Spellman yeast cell cycle.

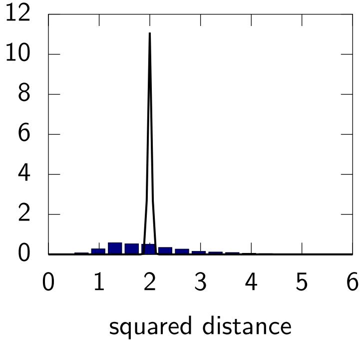


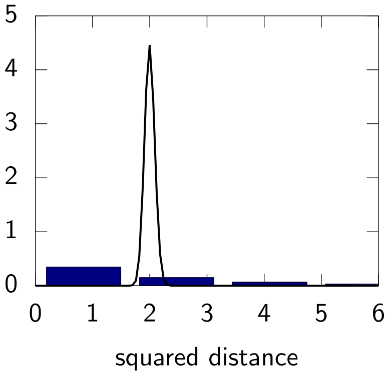
Figure: Interpoint squared distance distribution for Spellman microarray data with $p = 6178$ (variance of squared distances is 0.694 vs predicted 0.00129).

Where does practice depart from our theory?

- The situation for real data does not reflect what we expect.
- Real data exhibits greater variances on interpoint distances.
  - Somehow the real data seems to have a smaller effective dimension.
- Let's look at another $p = 1000$.

1000-D Gaussian

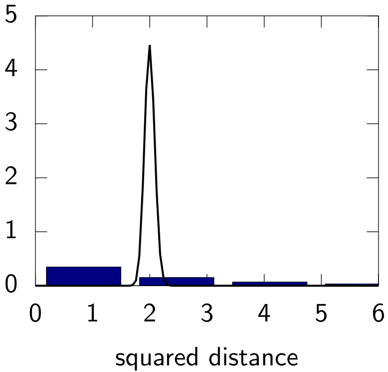
Distance distribution for a different Gaussian with $p = 1000$



1. Gaussian has a specific low rank covariance matrix $\mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}$.
2. Take $\sigma^2 = 1e-2$ and sample $\mathbf{W} \in \Re^{1000\times 2}$ from $\mathcal{N}(0,1)$.

1000-D Gaussian

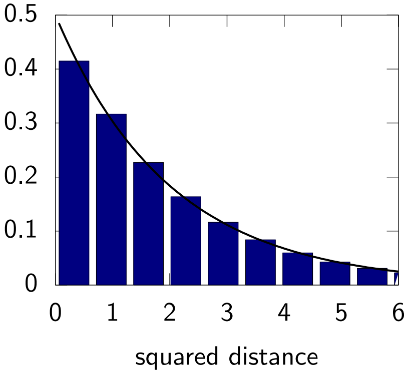
Distance distribution for a different Gaussian with $p = 1000$



1. Gaussian has a specific low rank covariance matrix $\mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}$.
2. Take $\sigma^2 = 1e-2$ and sample $\mathbf{W} \in \Re^{1000\times 2}$ from $\mathcal{N}(0, 1)$.

1000-D Gaussian

Distance distribution for a different Gaussian with $p = 1000$



1. Gaussian has a specific low rank covariance matrix $\mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}$.
2. Take $\sigma^2 = 1e-2$ and sample $\mathbf{W} \in \Re^{1000\times 2}$ from $\mathcal{N}(0, 1)$.
3. Theoretical curve taken assuming dimensionality of 2.

Linear Probabilistic Dimensionality Reduction

Where does this Low Rank Covariance Matrix Come From?

- It arises from a low dimensional approximation for the data set.
- Probabilistic PCA (Tipping 99, Roweis 97)

Linear Probabilistic Dimensionality Reduction

Where does this Low Rank Covariance Matrix Come From?

- It arises from a low dimensional approximation for the data set.
- Probabilistic PCA (Tipping 99, Roweis 97)
  - Linear Mapping from q-dimensional latent space to p-dimensional data space.

Linear Probabilistic Dimensionality Reduction

Where does this Low Rank Covariance Matrix Come From?

- It arises from a low dimensional approximation for the data set.
- Probabilistic PCA (Tipping 99, Roweis 97)
  - Linear Mapping from q-dimensional latent space to p-dimensional data space.
  - Corrupt the mapping by independent Gaussian noise.

Linear Probabilistic Dimensionality Reduction

Where does this Low Rank Covariance Matrix Come From?

- It arises from a low dimensional approximation for the data set.
- Probabilistic PCA (Tipping 99, Roweis 97)
  - Linear Mapping from q-dimensional latent space to p-dimensional data space.
  - Corrupt the mapping by independent Gaussian noise.
  - Marginalise latent variables using Gaussian prior.

Linear Probabilistic Dimensionality Reduction

Where does this Low Rank Covariance Matrix Come From?

- It arises from a low dimensional approximation for the data set.
- Probabilistic PCA (Tipping 99, Roweis 97)
  - Linear Mapping from q-dimensional latent space to p-dimensional data space.
  - Corrupt the mapping by independent Gaussian noise.
  - Marginalise latent variables using Gaussian prior.

A bit more Notation

q— dimension of latent/embedded space
p— dimension of data space
n— number of data points

$$\text{data, } \mathbf{Y} = \left[\mathbf{y}_{1,:}, \ldots, \mathbf{y}_{n,:}\right]^\top = \left[\mathbf{y}_{:,1}, \ldots, \mathbf{y}_{:,p}\right] \in \Re^{n \times p}$$
$$\textit{centred} \text{ data, } \hat{\mathbf{Y}} = \left[\hat{\mathbf{y}}_{1,:}, \ldots, \hat{\mathbf{y}}_{n,:}\right]^\top = \left[\hat{\mathbf{y}}_{:,1}, \ldots, \hat{\mathbf{y}}_{:,p}\right] \in \Re^{n \times p}, \ \hat{\mathbf{y}}_{i,:} = \mathbf{y}_{i,:} - \boldsymbol{\mu}$$
$$\text{latent variables, } \mathbf{X} = \left[\mathbf{x}_{1,:}, \ldots, \mathbf{x}_{n,:}\right]^\top = \left[\mathbf{x}_{:,1}, \ldots, \mathbf{x}_{:,q}\right] \in \Re^{n \times q}$$
$$\text{mapping matrix, } \mathbf{W} \in \Re^{p \times q}$$

$\mathbf{a}_{i,:}$ is a vector from the ith row of a given matrix $\mathbf{A}$
$\mathbf{a}_{:,j}$ is a vector from the jth row of a given matrix $\mathbf{A}$

Reading Notation

$\mathbf{X}$ and $\mathbf{Y}$ are *design matrices*

- **Data covariance** given by $\frac{1}{n}\hat{\mathbf{Y}}^\top\hat{\mathbf{Y}}$

$$\mathrm{cov}\left(\mathbf{Y}\right) = \frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{y}}_{i,:}\hat{\mathbf{y}}_{i,:}^\top = \frac{1}{n}\hat{\mathbf{Y}}^\top\hat{\mathbf{Y}} = \mathbf{S}.$$

- **Inner product matrix** given by $\mathbf{Y}\mathbf{Y}^\top$

$$\mathbf{K} = \left(k_{i,j}\right)_{i,j}, \qquad k_{i,j} = \mathbf{y}_{i,:}^\top\mathbf{y}_{j,:}$$

Linear Dimensionality Reduction

- Find a lower dimensional plane embedded in a higher dimensional space.
- The plane is described by the matrix $\mathbf{W} \in \Re^{p \times q}$.

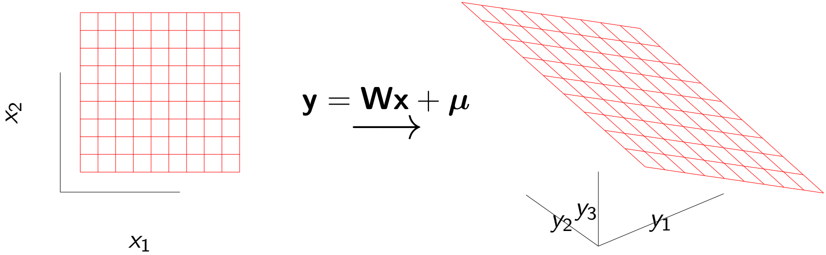
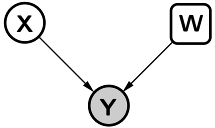


Figure: Mapping a two dimensional plane to a higher dimensional space in a linear way. Data are generated by corrupting points on the plane with noise.

Linear Latent Variable Model

Probabilistic PCA

- Linear-Gaussian relationship between latent variables and data, $\mathbf{y}_{i,:} = \mathbf{W}\mathbf{x}_{i,:} + \boldsymbol{\mu} + \boldsymbol{\eta}_{i,:}$.
- $\mathbf{X}$ are 'nuisance' variables.

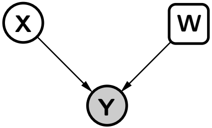


$$p\left(\mathbf{Y}|\mathbf{X},\mathbf{W}\right) = \prod_{i=1}^{n} \mathcal{N}\left(\mathbf{y}_{i,:}|\mathbf{W}\mathbf{x}_{i,:} + \boldsymbol{\mu}, \sigma^2\mathbf{I}\right)$$

Linear Latent Variable Model

Probabilistic PCA

- Linear-Gaussian relationship between latent variables and data, $\mathbf{y}_{i,:} = \mathbf{W}\mathbf{x}_{i,:} + \boldsymbol{\mu} + \boldsymbol{\eta}_{i,:}$.
- $\mathbf{X}$ are 'nuisance' variables.
- Latent variable model approach:

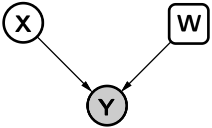


$$p(\mathbf{Y}|\mathbf{X},\mathbf{W}) = \prod_{i=1}^{n} \mathcal{N}\left(\mathbf{y}_{i,:}|\mathbf{W}\mathbf{x}_{i,:} + \boldsymbol{\mu}, \sigma^2\mathbf{I}\right)$$

Linear Latent Variable Model

Probabilistic PCA

- Linear-Gaussian relationship between latent variables and data, $\mathbf{y}_{i,:} = \mathbf{W}\mathbf{x}_{i,:} + \boldsymbol{\mu} + \boldsymbol{\eta}_{i,:}$.
- $\mathbf{X}$ are 'nuisance' variables.
- Latent variable model approach:
  1. Define Gaussian prior over *latent space*, $\mathbf{X}$.

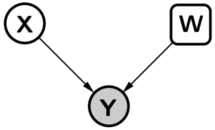


$$p(\mathbf{Y}|\mathbf{X},\mathbf{W}) = \prod_{i=1}^{n} \mathcal{N}\left(\mathbf{y}_{i,:}|\mathbf{W}\mathbf{x}_{i,:} + \boldsymbol{\mu}, \sigma^2\mathbf{I}\right)$$

Linear Latent Variable Model

Probabilistic PCA

- Linear-Gaussian relationship between latent variables and data, $\mathbf{y}_{i,:} = \mathbf{W}\mathbf{x}_{i,:} + \boldsymbol{\mu} + \boldsymbol{\eta}_{i,:}$.
- $\mathbf{X}$ are 'nuisance' variables.
- Latent variable model approach:
  1. Define Gaussian prior over *latent space*, $\mathbf{X}$.
  2. Integrate out nuisance *latent variables*.



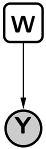
$$p(\mathbf{Y}|\mathbf{X},\mathbf{W}) = \prod_{i=1}^{n} \mathcal{N}\left(\mathbf{y}_{i,:}|\mathbf{W}\mathbf{x}_{i,:} + \boldsymbol{\mu}, \sigma^2\mathbf{I}\right)$$

$$p(\mathbf{X}) = \prod_{i=1}^{n} \mathcal{N}\left(\mathbf{x}_{i,:}|\mathbf{0},\mathbf{I}\right)$$

Linear Latent Variable Model

Probabilistic PCA

- Linear-Gaussian relationship between latent variables and data, $\mathbf{y}_{i,:} = \mathbf{W}\mathbf{x}_{i,:} + \boldsymbol{\mu} + \boldsymbol{\eta}_{i,:}$.
- $\mathbf{X}$ are 'nuisance' variables.
- Latent variable model approach:
  1. Define Gaussian prior over *latent space*, $\mathbf{X}$.
  2. Integrate out nuisance *latent variables*.
  3. Optimize likelihood wrt $\mathbf{W}$, $\boldsymbol{\mu}$.



$$p(\mathbf{Y}|\mathbf{X},\mathbf{W}) = \prod_{i=1}^{n} \mathcal{N}\left(\mathbf{y}_{i,:}|\mathbf{W}\mathbf{x}_{i,:} + \boldsymbol{\mu}, \sigma^2\mathbf{I}\right)$$

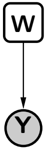
$$p(\mathbf{X}) = \prod_{i=1}^{n} \mathcal{N}\left(\mathbf{x}_{i,:}|\mathbf{0},\mathbf{I}\right)$$

$$p(\mathbf{Y}|\mathbf{W},\boldsymbol{\mu}) = \prod_{i=1}^{n} \mathcal{N}\left(\mathbf{y}_{i,:}|\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}\right)$$

Linear Latent Variable Model

Probabilistic PCA

- Linear-Gaussian relationship between latent variables and data, $\mathbf{y}_{i,:} = \mathbf{W}\mathbf{x}_{i,:} + \boldsymbol{\mu} + \boldsymbol{\eta}_{i,:}$.
- $\mathbf{X}$ are 'nuisance' variables.
- Latent variable model approach:
  1. Define Gaussian prior over *latent space*, $\mathbf{X}$.
  2. Integrate out nuisance *latent variables*.
  3. Optimize likelihood wrt $\mathbf{W}$, $\boldsymbol{\mu}$.

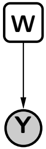


$$p\left(\mathbf{Y}|\mathbf{W},\boldsymbol{\mu}\right) = \prod_{i=1}^{n} \mathcal{N}\left(\mathbf{y}_{i,:}|\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^{\top} + \sigma^2\mathbf{I}\right)$$

Linear Latent Variable Model

Probabilistic PCA

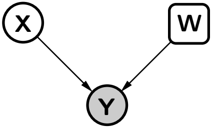
- Linear-Gaussian relationship between latent variables and data, $\mathbf{y}_{i,:} = \mathbf{W}\mathbf{x}_{i,:} + \boldsymbol{\mu} + \boldsymbol{\eta}_{i,:}$.
- $\mathbf{X}$ are 'nuisance' variables.
- Latent variable model approach:
  1. Define Gaussian prior over *latent space*, $\mathbf{X}$.
  2. Integrate out nuisance *latent variables*.
  3. Optimize likelihood wrt $\mathbf{W}$, $\boldsymbol{\mu}$.



$$p\left(\hat{\mathbf{Y}}|\mathbf{W}\right) = \prod_{i=1}^{n} \mathcal{N}\left(\hat{\mathbf{y}}_{i,:}|\mathbf{0}, \mathbf{W}\mathbf{W}^{\top} + \sigma^2\mathbf{I}\right)$$

Probabilistic PCA Solution

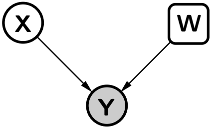
Probabilistic PCA Max. Likelihood Soln (Tipping 99)



$$p\left(\hat{\mathbf{Y}}|\mathbf{W}\right) = \prod_{i=1}^{n} \mathcal{N}\left(\hat{\mathbf{y}}_{i,:}|\mathbf{0}, \mathbf{W}\mathbf{W}^{\top} + \sigma^2\mathbf{I}\right)$$

Probabilistic PCA Solution

Probabilistic PCA Max. Likelihood Soln (Tipping 99)



$$p\left(\hat{\mathbf{Y}}|\mathbf{W}\right) = \prod_{j=1}^{p} \mathcal{N}\left(\hat{\mathbf{y}}_{i,:}|\mathbf{0},\mathbf{C}\right), \quad \mathbf{C} = \mathbf{W}\mathbf{W}^{\top} + \sigma^2\mathbf{I}$$

$$\log p\left(\hat{\mathbf{Y}}|\mathbf{W}\right) = -\frac{n}{2}\log|\mathbf{C}| - \frac{1}{2}\mathrm{tr}\left(\mathbf{C}^{-1}\hat{\mathbf{Y}}^{\top}\hat{\mathbf{Y}}\right) + \text{const.}$$

If $\mathbf{U}_q$ are first q principal eigenvectors of $n^{-1}\hat{\mathbf{Y}}^{\top}\hat{\mathbf{Y}}$ and the corresponding eigenvalues are $\mathbf{\Lambda}_q$,

$$\mathbf{W} = \mathbf{U}_q\mathbf{L}\mathbf{R}^{\top}, \qquad \mathbf{L} = \left(\mathbf{\Lambda}_q - \sigma^2\mathbf{I}\right)^{\frac{1}{2}}$$

where $\mathbf{R}$ is an arbitrary rotation matrix.

Details

PCA on Stick Man

- First two principal components of stick man data.

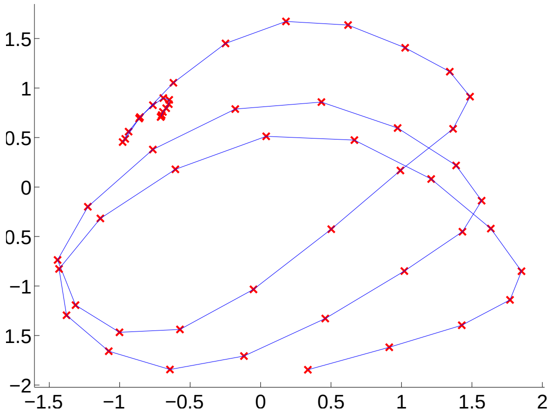


Figure: Stick man data projected onto their first two principal components. `demStickPpca1`.

PCA on Oil Data

- First two principal components of oil data.

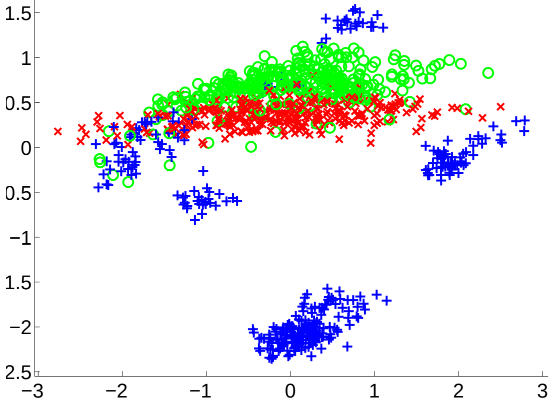


Figure: Oil data projected onto their first two principal components. `demOilPpca1`.

PCA on Microarray

- First two principal components of gene expression data.

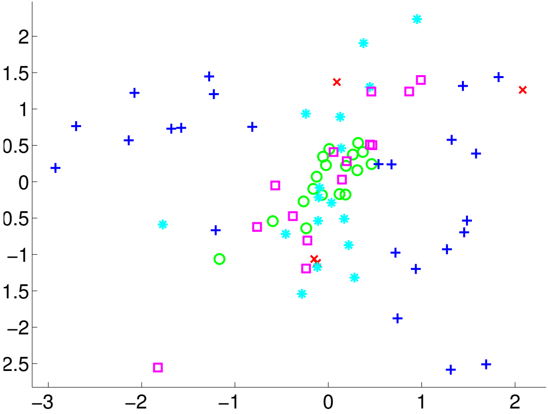


Figure: Microarray data projected onto their first two principal components. `demSpellmanPpca1`. Different symbols show different experiment groups (separate time series).

Why Probabilistic PCA?

- What is the point in probabilistic methods?
- Could we not just project with regular PCA?

Why Probabilistic PCA?

- What is the point in probabilistic methods?
- Could we not just project with regular PCA?
  - Integration within other models (*e.g.* mixtures of PCA (Tipping 97), temporal models).

Why Probabilistic PCA?

- What is the point in probabilistic methods?
- Could we not just project with regular PCA?
  - Integration within other models (*e.g.* mixtures of PCA (Tipping 97), temporal models).
  - Model selection through Bayesian treatment of parameters (Bishop 98)

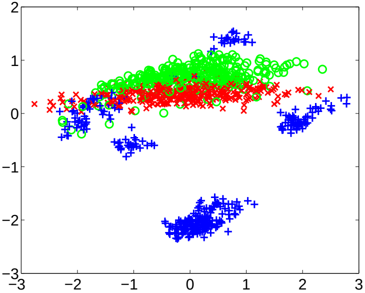
Why Probabilistic PCA?

- What is the point in probabilistic methods?
- Could we not just project with regular PCA?
  - Integration within other models (*e.g.* mixtures of PCA (Tipping 97), temporal models).
  - Model selection through Bayesian treatment of parameters (Bishop 98)
  - Marginalisation of missing data (Tipping 99)

Why Probabilistic PCA?

- What is the point in probabilistic methods?
- Could we not just project with regular PCA?
  - Integration within other models (*e.g.* mixtures of PCA (Tipping 97), temporal models).
  - Model selection through Bayesian treatment of parameters (Bishop 98)
  - Marginalisation of missing data (Tipping 99)

Oil and Missing Data



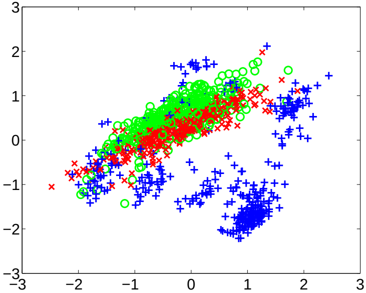
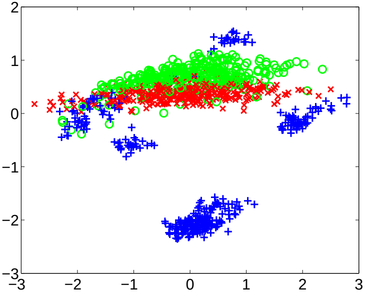


Figure: Projection of the oil data set on to $q = 2$ latent dimensions. *Left*: full data set with no missing data. *Right*: data set with 10% values missing at random.

Oil and Missing Data



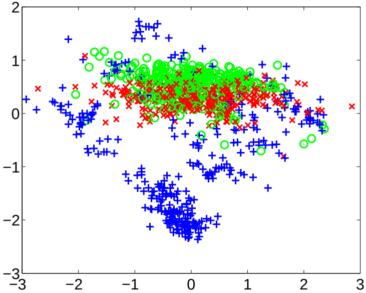


Figure: Projection of the oil data set on to $q = 2$ latent dimensions. *Left*: full data set with no missing data. *Right*: data set with 20% values missing at random.

Oil and Missing Data

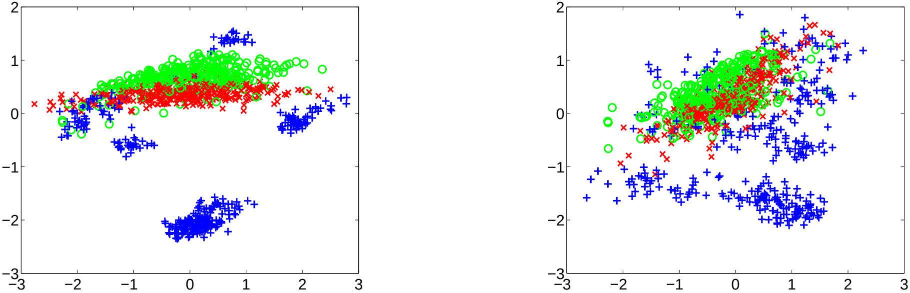


Figure: Projection of the oil data set on to $q = 2$ latent dimensions. *Left*: full data set with no missing data. *Right*: data set with 30% values missing at random.

Oil and Missing Data

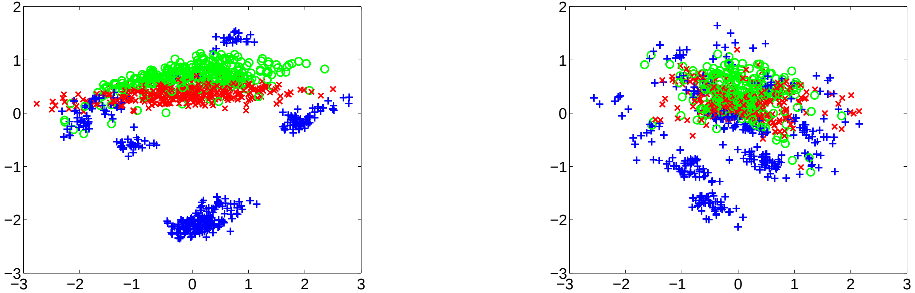


Figure: Projection of the oil data set on to $q = 2$ latent dimensions. *Left*: full data set with no missing data. *Right*: data set with 50% values missing at random.

Is (P)PCA Used in Computer Vision?

It's difficult not to find a paper that doesn't use it!

- EigenFaces: **y** is an image of a face (Sirovich & Kirby 87, Turk & Pentland 91)

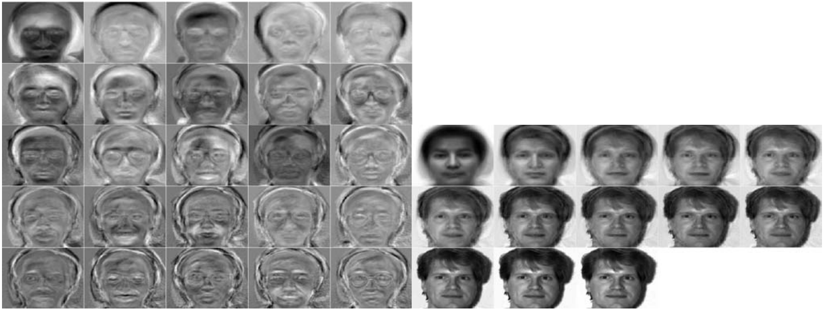
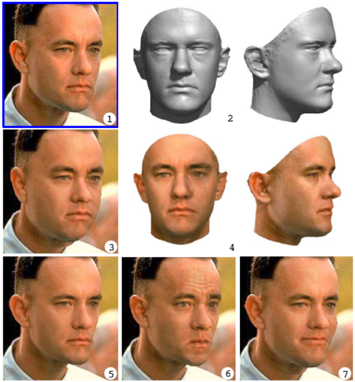


Figure: Yale faces: Image from C. de CORO

Is (P)PCA Used in Computer Vision?

It's difficult not to find a paper that doesn't use it!

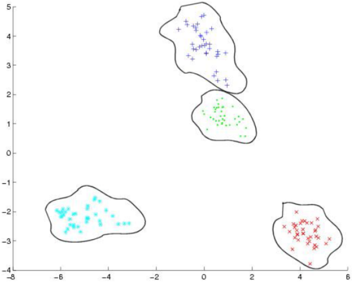
- EigenFaces: $\mathbf{y}$ is an image of a face (Sirovich & Kirby 87, Turk & Pentland 91)
- Morphable Model for the Synthesis of 3D Faces (Blanz & Vetter 99)



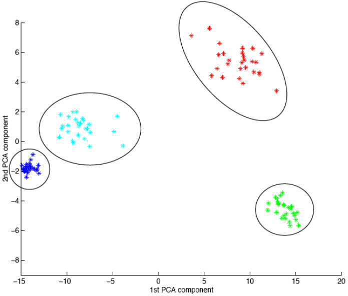
Is (P)PCA Used in Computer Vision?

It's difficult not to find a paper that doesn't use it!

- EigenFaces: **y** is an image of a face (Sirovich & Kirby 87, Turk & Pentland 91)
- Morphable Model for the Synthesis of 3D Faces (Blanz & Vetter 99)
- Tracking where **y** is the full motion (e.g., all poses for a full golf swing) (Siddenbladh et al. 02, Urtasun et al. 05)



(Walk)



(Run)

Is (P)PCA Used in Computer Vision?

It's difficult not to find a paper that doesn't use it!

- EigenFaces: $\mathbf{y}$ is an image of a face (Sirovich & Kirby 87, Turk & Pentland 91)
- Morphable Model for the Synthesis of 3D Faces (Blanz & Vetter 99)
- Tracking where $\mathbf{y}$ is the full motion (e.g., all poses for a full golf swing) (Siddenbladh et al. 02, Urtasun et al. 05)
- Object recognition: PCA-SIFT (Ke et al. 04)

Is (P)PCA Used in Computer Vision?

It's difficult not to find a paper that doesn't use it!

- EigenFaces: $\mathbf{y}$ is an image of a face (Sirovich & Kirby 87, Turk & Pentland 91)
- Morphable Model for the Synthesis of 3D Faces (Blanz & Vetter 99)
- Tracking where $\mathbf{y}$ is the full motion (e.g., all poses for a full golf swing) (Siddenbladh et al. 02, Urtasun et al. 05)
- Object recognition: PCA-SIFT (Ke et al. 04)
- Object detection: Deformable part-based models (Felzenbwald et al. 10)

Is (P)PCA Used in Computer Vision?

It's difficult not to find a paper that doesn't use it!

- EigenFaces: $\mathbf{y}$ is an image of a face (Sirovich & Kirby 87, Turk & Pentland 91)
- Morphable Model for the Synthesis of 3D Faces (Blanz & Vetter 99)
- Tracking where $\mathbf{y}$ is the full motion (e.g., all poses for a full golf swing) (Siddenbladh et al. 02, Urtasun et al. 05)
- Object recognition: PCA-SIFT (Ke et al. 04)
- Object detection: Deformable part-based models (Felzenbwald et al. 10)
- $\cdots$

Is (P)PCA Used in Computer Vision?

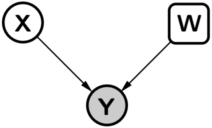
It's difficult not to find a paper that doesn't use it!

- EigenFaces: $\mathbf{y}$ is an image of a face (Sirovich & Kirby 87, Turk & Pentland 91)
- Morphable Model for the Synthesis of 3D Faces (Blanz & Vetter 99)
- Tracking where $\mathbf{y}$ is the full motion (e.g., all poses for a full golf swing) (Siddenbladh et al. 02, Urtasun et al. 05)
- Object recognition: PCA-SIFT (Ke et al. 04)
- Object detection: Deformable part-based models (Felzenbwald et al. 10)
- $\cdots$
- You probably have used it too! (Audience et al.)

Let's see what Neil has to say ...

Maximum Likelihood Solution

Probabilistic PCA Max. Likelihood Soln (Tipping 99)

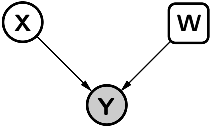


$$p(\mathbf{Y}|\mathbf{W}, \boldsymbol{\mu}) = \prod_{i=1}^{n} \mathcal{N}\left(\mathbf{y}_{i,:}|\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^{\top} + \sigma^2\mathbf{I}\right)$$

Gradient of log likelihood

Maximum Likelihood Solution

Probabilistic PCA Max. Likelihood Soln (Tipping 99)



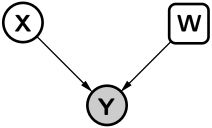
$$p\left(\hat{\mathbf{Y}}|\mathbf{W}\right) = \prod_{i=1}^{n} \mathcal{N}\left(\hat{\mathbf{y}}_{i,:}|\mathbf{0}, \mathbf{W}\mathbf{W}^{\top} + \sigma^2\mathbf{I}\right)$$

Gradient of log likelihood

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{W}} \log p\left(\hat{\mathbf{Y}}|\mathbf{W}\right) = -\frac{n}{2}\mathbf{C}^{-1}\mathbf{W} + \frac{1}{2}\mathbf{C}^{-1}\hat{\mathbf{Y}}^{\top}\hat{\mathbf{Y}}\mathbf{C}^{-1}\mathbf{W}$$

Maximum Likelihood Solution

Probabilistic PCA Max. Likelihood Soln (Tipping 99)

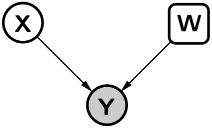


$$p\left(\hat{\mathbf{Y}}|\mathbf{W}\right) = \prod_{i=1}^{n} \mathcal{N}\left(\hat{\mathbf{y}}_{i,:}|\mathbf{0}, \mathbf{C}\right), \quad \mathbf{C} = \mathbf{W}\mathbf{W}^{\top} + \sigma^2\mathbf{I}$$

Gradient of log likelihood

Maximum Likelihood Solution

Probabilistic PCA Max. Likelihood Soln (Tipping 99)



$$p\left(\hat{\mathbf{Y}}|\mathbf{W}\right) = \prod_{i=1}^{n} \mathcal{N}\left(\hat{\mathbf{y}}_{i,:}|\mathbf{0}, \mathbf{C}\right), \quad \mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}$$

$$\log p\left(\hat{\mathbf{Y}}|\mathbf{W}\right) = -\frac{n}{2}\log|\mathbf{C}| - \frac{1}{2}\mathrm{tr}\left(\mathbf{C}^{-1}\hat{\mathbf{Y}}^\top\hat{\mathbf{Y}}\right) + \text{const.}$$

Gradient of log likelihood

Optimization

Seek fixed points

$$\mathbf{0} = -\frac{n}{2}\mathbf{C}^{-1}\mathbf{W} + \frac{1}{2}\mathbf{C}^{-1}\hat{\mathbf{Y}}^{\top}\hat{\mathbf{Y}}\mathbf{C}^{-1}\mathbf{W}$$

pre-multiply by $2\mathbf{C}$

$$\mathbf{0} = -n\mathbf{W} + \hat{\mathbf{Y}}^{\top}\hat{\mathbf{Y}}\mathbf{C}^{-1}\mathbf{W}$$

$$\frac{1}{n}\hat{\mathbf{Y}}^{\top}\hat{\mathbf{Y}}\mathbf{C}^{-1}\mathbf{W} = \mathbf{W}$$

Substitute W with singular value decomposition

$$\mathbf{W} = \mathbf{U}\mathbf{L}\mathbf{R}^{\top}$$

which implies

$$\begin{aligned}\mathbf{C} &= \mathbf{W}\mathbf{W}^{\top} + \sigma^2\mathbf{I}\\ &= \mathbf{U}\mathbf{L}^2\mathbf{U}^{\top} + \sigma^2\mathbf{I}\end{aligned}$$

Using matrix inversion lemma

$$\mathbf{C}^{-1}\mathbf{W} = \mathbf{U}\mathbf{L}\left(\sigma^2 + \mathbf{L}^2\right)^{-1}\mathbf{R}^{\top}$$

Solution given by

$$\frac{1}{n}\hat{\mathbf{Y}}^{\top}\hat{\mathbf{Y}}\mathbf{U} = \mathbf{U}\left(\sigma^2 + \mathbf{L}^2\right)$$

which is recognised as an eigenvalue problem.

- This implies that the columns of $\mathbf{U}$ are the eigenvectors of $\frac{1}{n}\hat{\mathbf{Y}}^{\top}\hat{\mathbf{Y}}$ and that $\sigma^2 + \mathbf{L}^2$ are the eigenvalues of $\frac{1}{n}\hat{\mathbf{Y}}^{\top}\hat{\mathbf{Y}}$.
- $l_i = \sqrt{\lambda_i - \sigma^2}$ where λ_i is the ith eigenvalue of $\frac{1}{n}\hat{\mathbf{Y}}^{\top}\hat{\mathbf{Y}}$.
- Further manipulation shows that if we constrain $\mathbf{W} \in \Re^{p\times q}$ then the solution is given by the largest q eigenvalues.

Probabilistic PCA Solution

- If $\mathbf{U}_q$ are first q principal eigenvectors of $n^{-1}\hat{\mathbf{Y}}^\top\hat{\mathbf{Y}}$ and the corresponding eigenvalues are $\mathbf{\Lambda}_q$,

$$\mathbf{W} = \mathbf{U}_q\mathbf{L}\mathbf{R}^\top, \quad \mathbf{L} = \left(\mathbf{\Lambda}_q - \sigma^2\mathbf{I}\right)^{\frac{1}{2}}$$

  where $\mathbf{R}$ is an arbitrary rotation matrix.
- Some further work shows that the *principal* eigenvectors need to be retained.
- The maximum likelihood value for σ^2 is given by the average of the discarded eigenvalues.

Return