3F3: Signal and Pattern Processing

Lecture 5: Dimensionality Reduction

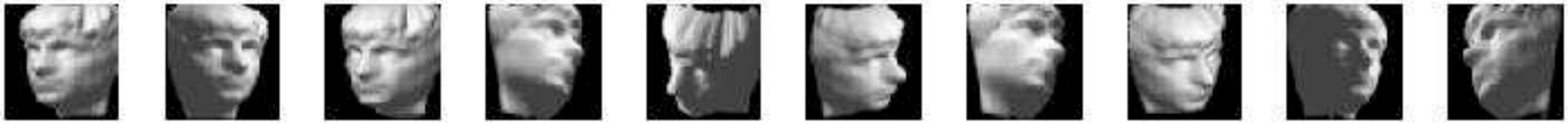
Zoubin Ghahramani
zoubin@eng.cam.ac.uk

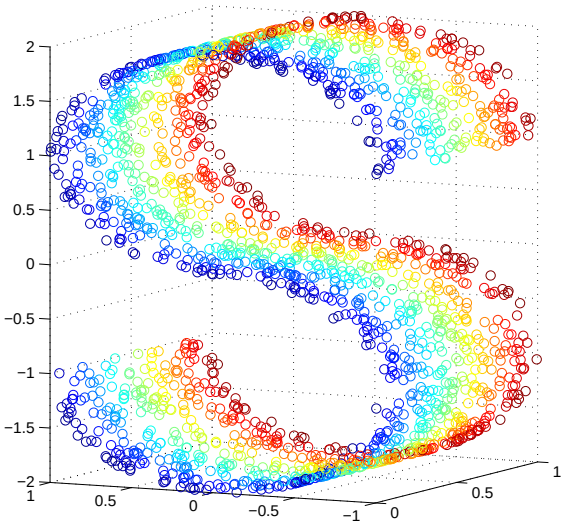
Department of Engineering
University of Cambridge

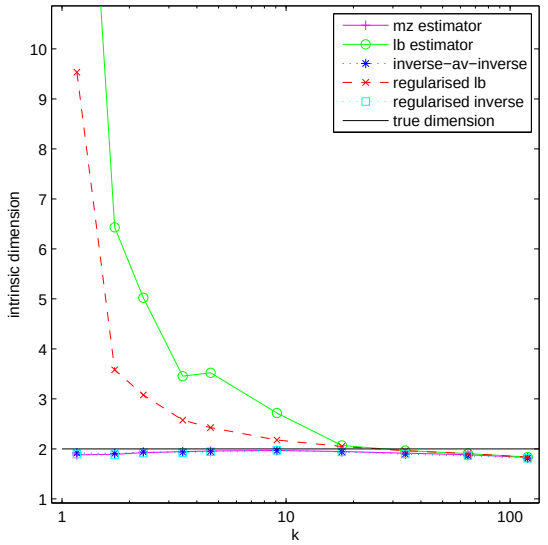
Lent Term

Dimensionality Reduction

Given some data, the goal is to discover and model the intrinsic dimensions of the data, and/or to project high dimensional data onto a lower number of dimensions that preserve the relevant information.







Principal Components Analysis (PCA)

Data Set $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ where $\mathbf{x}_n \in \Re^D$

Assume that the data is zero mean, $\frac{1}{N}\sum_n \mathbf{x}_n = 0$.

Principal Components Analysis (PCA) is a linear dimensionality reduction method which finds the linear projection(s) of the data which:

- maximise variance
- minimise squared reconstruction error
- have highest mutual information with the data under a Gaussian model
- are maximum likelihood parameters under a linear Gaussian factor model of the data

PCA: Direction of Maximum Variance

Let $y = \mathbf{w}^\top \mathbf{x}$. Find $\mathbf{w}$ such that $\mathrm{var}(y)$ is maximised for the data set $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. Since $\mathcal{D}$ is assumed zero mean, $\mathbb{E}_{\mathcal{D}}(y) = 0$. Using $y_n = \mathbf{w}^\top \mathbf{x}_n$ we optimise:

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \mathrm{var}(y) = \arg\max_{\mathbf{w}} \mathbb{E}_{\mathcal{D}}(y^2) = \arg\max_{\mathbf{w}} \frac{1}{N} \sum_n y_n^2$$

$$\begin{aligned} \frac{1}{N} \sum_n y_n^2 &= \frac{1}{N} \sum_n (\mathbf{w}^\top \mathbf{x}_n)^2 = \frac{1}{N} \sum_n \mathbf{w}^\top \mathbf{x}_n \mathbf{x}_n{}^\top \mathbf{w} \\ &= \mathbf{w}^\top \left(\frac{1}{N} \sum_n \mathbf{x}_n \mathbf{x}_n{}^\top \right) \mathbf{w} = \mathbf{w}^\top C \mathbf{w} \end{aligned}$$

where $C = \frac{1}{N} \sum_n \mathbf{x}_n \mathbf{x}_n{}^\top$ is the data covariance matrix. Clearly arbitrarily increasing the magnitude of $\mathbf{w}$ will increase $\mathrm{var}(y)$, so we will restrict outselves to *directions* $\mathbf{w}$ with unit norm, $\|\mathbf{w}\|^2 = \mathbf{w}^\top \mathbf{w} = 1$. Using a Lagrange multiplier λ to enforce this constraint:

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \ \mathbf{w}^\top C \mathbf{w} - \lambda(\mathbf{w}^\top \mathbf{w} - 1)$$

Solution $\mathbf{w}^*$ is the eigenvector with maximal eigenvalue of covariance matrix C.

Eigenvalues and Eigenvectors

λ is an eigenvalue and $\mathbf{z}$ is an eigenvector of A if:

$$A\mathbf{z} = \lambda\mathbf{z}$$

and $\mathbf{z}$ is a unit vector ($\mathbf{z}^\top\mathbf{z} = 1$).

Interpretation: the operation of A in direction $\mathbf{z}$ is a scaling by λ.

The K Principal Components are the K eigenvectors with the largest eigenvalues of the data covariance matrix (i.e. K directions with the largest variance).

Note: C can be decomposed:

$$C = USU^\top$$

where S is diag$(\sigma_1^2, \ldots, \sigma_D^2)$ and U is a an orthonormal matrix.

PCA: Minimising Squared Reconstruction Error

Solve the following **minimum reconstruction error** problem:

$$\min_{\{\alpha_n\},\mathbf{w}} \|\mathbf{x}_n - \alpha_n \mathbf{w}\|^2$$

Solving for α_n holding $\mathbf{w}$ fixed gives:

$$\alpha_n = \frac{\mathbf{w}^\top \mathbf{x}_n}{\mathbf{w}^\top \mathbf{w}}$$

Note if we rescale $\mathbf{w}$ to $\beta\mathbf{w}$ and α_n to α_n/β we get equivalent solutions, so there won't be a unique minimum. Let's constrain $\|\mathbf{w}\| = 1$ which implies $\mathbf{w}^\top \mathbf{w} = 1$. Plugging α_n into the original cost we get:

$$\min_{\mathbf{w}} \sum_n \|\mathbf{x}_n - (\mathbf{w}^\top \mathbf{x}_n)\mathbf{w}\|^2$$

Expanding the quadratic, and adding the Lagrange multiplier, the solution is again:

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \ \mathbf{w}^\top C \mathbf{w} - \lambda(\mathbf{w}^\top \mathbf{w} - 1)$$

PCA: Maximising Mutual Information

Problem: Given $\mathbf{x}$ and assuming that $P(\mathbf{x})$ is zero mean Gaussian, find $y = \mathbf{w}^\top \mathbf{x}$, with $\mathbf{w}$ a unit vector, such that the mutual information $I(\mathbf{x}; y)$ is maximised.

$$I(\mathbf{x}; y) = H(\mathbf{x}) + H(y) - H(\mathbf{x}, y) = H(y)$$

So we want to maximise the entropy of y. What is the entropy of a Gaussian?
Let $\mathbf{z} \sim \mathcal{N}(\mu, \Sigma)$, then:

$$H(\mathbf{z}) \quad = \quad -\int p(\mathbf{z}) \ln p(\mathbf{z}) \, d\mathbf{z} = \frac{1}{2} \ln |\Sigma| + \frac{D}{2}(1 + \ln 2\pi)$$

Therefore we want the distribution of y to have largest variance (in the multidimensional case, largest volume —i.e. det of covariance matrix).

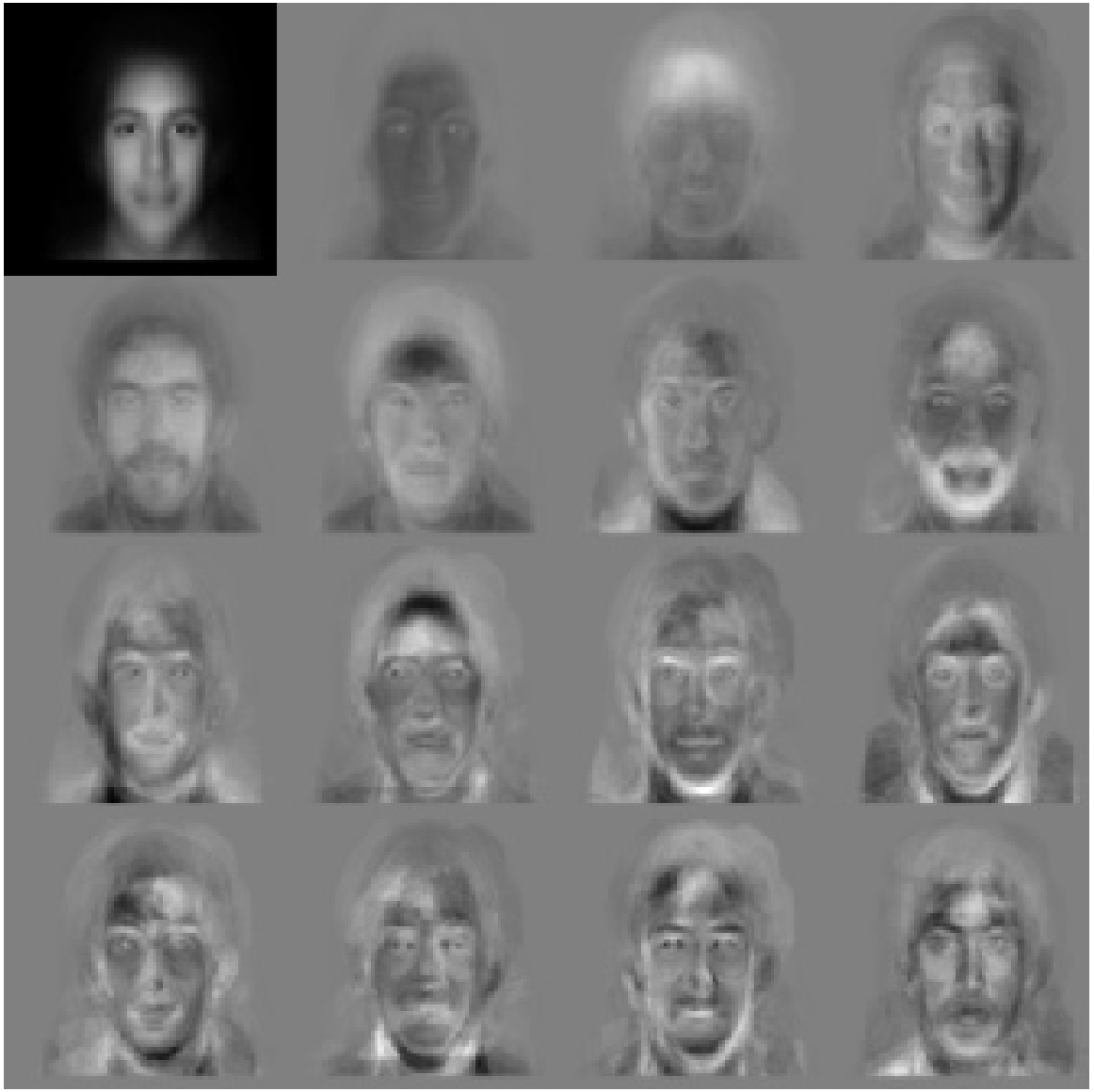
$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \ \mathrm{var}(y) \ \text{ subject to } \ \|\mathbf{w}\| = 1$$

Principal Components Analysis

The full multivariate case of PCA finds a sequence of K *orthogonal* directions $\mathbf{w}_1, \mathbf{w}_2, \ldots \mathbf{w}_K$.

Here $\mathbf{w}_1$ is the eigenvector with largest eigenvalue of C, $\mathbf{w}_2$ is the eigenvector with second largest eigenvalue and orthogonal to $\mathbf{w}_1$ (i.e. $\mathbf{w}_2^\top \mathbf{w}_1 = 0$), etc.

Example of PCA: Eigenfaces



from www-white.media.mit.edu/vismod/demos/facerec/basic.html

Nonlinear and Kernel PCA

There are many different ways of generalising PCA to find *nonlinear* directions of variation in the data.

A simple example (very similar to what we did with regression and classification!) is to map the data in some nonlinear way,

$$\mathbf{x} \rightarrow \boldsymbol{\phi}(\mathbf{x})$$

and then do PCA on the $\{\boldsymbol{\phi}(\mathbf{x}_1) \ldots \boldsymbol{\phi}(\mathbf{x}_N)\}$ vectors.

This is sometimes called "kernel PCA" since it can be completely defined in terms of the kernel functions $K(\mathbf{x}_n, \mathbf{x}_m) = \boldsymbol{\phi}(\mathbf{x}_n)^\top \boldsymbol{\phi}(\mathbf{x}_m)$, or alternatively in terms of a similarity metric on the inputs.

Summary

We have covered four key topics in machine learning and pattern recognition:

- Classification
- Regression
- Clustering
- Dimensionality Reduction

In each case, we see that these methods can be viewed as building probabilistic models of the data. We can start from simple linear models and build up to nonlinear models.

Appendix: Information, Probability and Entropy

Information is the reduction of uncertainty. How do we measure uncertainty?

Some axioms (informal):

- if something is certain its uncertainty $= 0$
- uncertainty should be maximum if all choices are equally probable
- uncertainty (information) should add for independent sources

This leads to a discrete random variable X having uncertainty equal to the entropy function:

$$H(X) = -\sum_{x \in \mathcal{X}} P(X = x) \log P(X = x)$$

measured in *bits* (**bi**nary digi**ts**) if the base 2 logarithm is used or *nats* (**na**tural digi**ts**) if the natural (base e) logarithm is used.

Appendix: Information, Probability and Entropy

- Surprise (for event $X = x$): $-\log P(X = x)$
- Entropy = average surprise: $H(X) = -\sum_{x \in \mathcal{X}} P(X = x) \log_2 P(X = x)$
- Conditional entropy

$$H(X|Y) = -\sum_x \sum_y P(x, y) \log_2 P(x|y)$$

- Mutual information

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X, Y)$$

- Independent random variables: $P(x, y) = P(x)P(y) \, \forall x \, \forall y$