02457 Non-linear signal processing

2017 - Lecture 2

Lars Kai Hansen
Technical University of Denmark
DTU Compute, Lyngby, Denmark

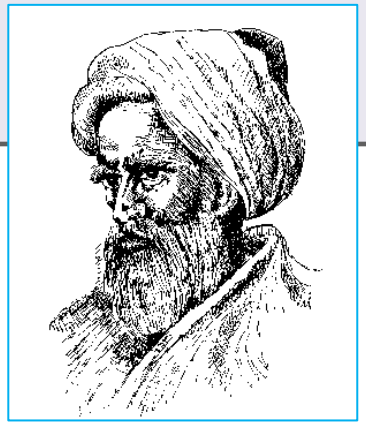
Outline lecture 2

- Densities and 1D Normal distribution
- Bayes' decision theory

- Multivariate data - podcast indexing example
- Multivariate normal distribution
- Matrices, eigenvalues, eigenvectors
- Correlations

- Features & the curse of dimensionality
- Principal components
- Principal components of functional brain images

The scientific approach



Alhazen - Hasan Ibn al-Haytham 965-1040

Alhazen year 1000:

"Read, be curious, experiment, communicate!" http://www-history.mcs.st-and.ac.uk/Biographies/Al-Haytham.html

Google 2006: *"Prior to launch, any innovation in AdSense goes through three phases: analysis, implementation, and experimentation. When evaluating a new idea that might improve ad targeting, we first analyze all of our existing data. AdSense generates an amazing amount of data every day -- a record of every ad impression and click occurring on every web publisher in our network –"*
Inside AdSense the experimental approach

Joaquin Quiñonero Candela - Director of Applied Machine Learning at Facebook

Facebook 2016: *"Randomized controlled trials are considered the "gold standard" of measurement because they ensure that control groups and test groups are comparable in makeup and allow analysts to isolate the causal impact of ad exposure."*

Demystifying Measurement: Why Methodology Matters

Joint probabilities

- Probabilities are limits of relative #counts
  (relative frequency)
- Event characterized by two "values" (x,y)
- We observe pairs (x,y)
- Examples
  - Height of a subject vs male, female
  - Image vs indoor, outdoor

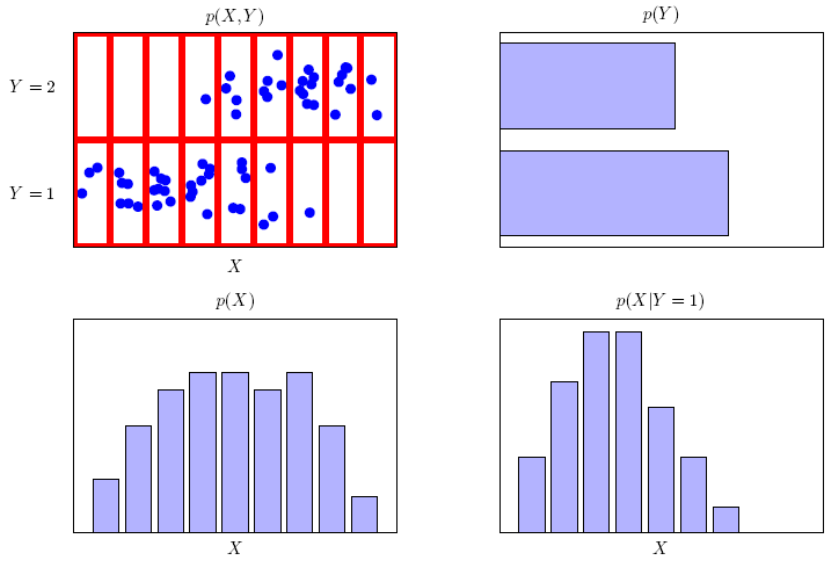


Figure 1.11 An illustration of a distribution over two variables, X, which takes 9 possible values, and Y, which takes two possible values. The top left figure shows a sample of 60 points drawn from a joint probability distribution over these variables. The remaining figures show histogram estimates of the marginal distributions $p(X)$ and $p(Y)$, as well as the conditional distribution $p(X|Y = 1)$ corresponding to the bottom row in the top left figure.

Conditional probability

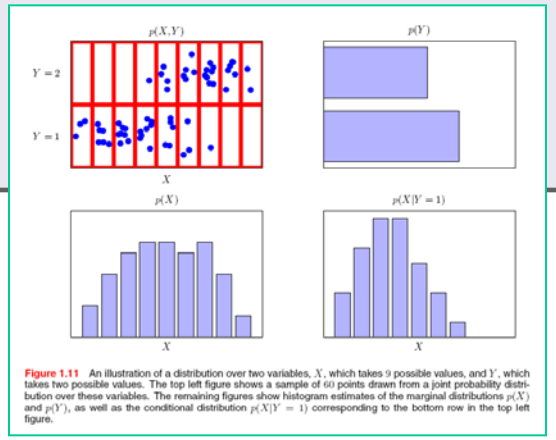
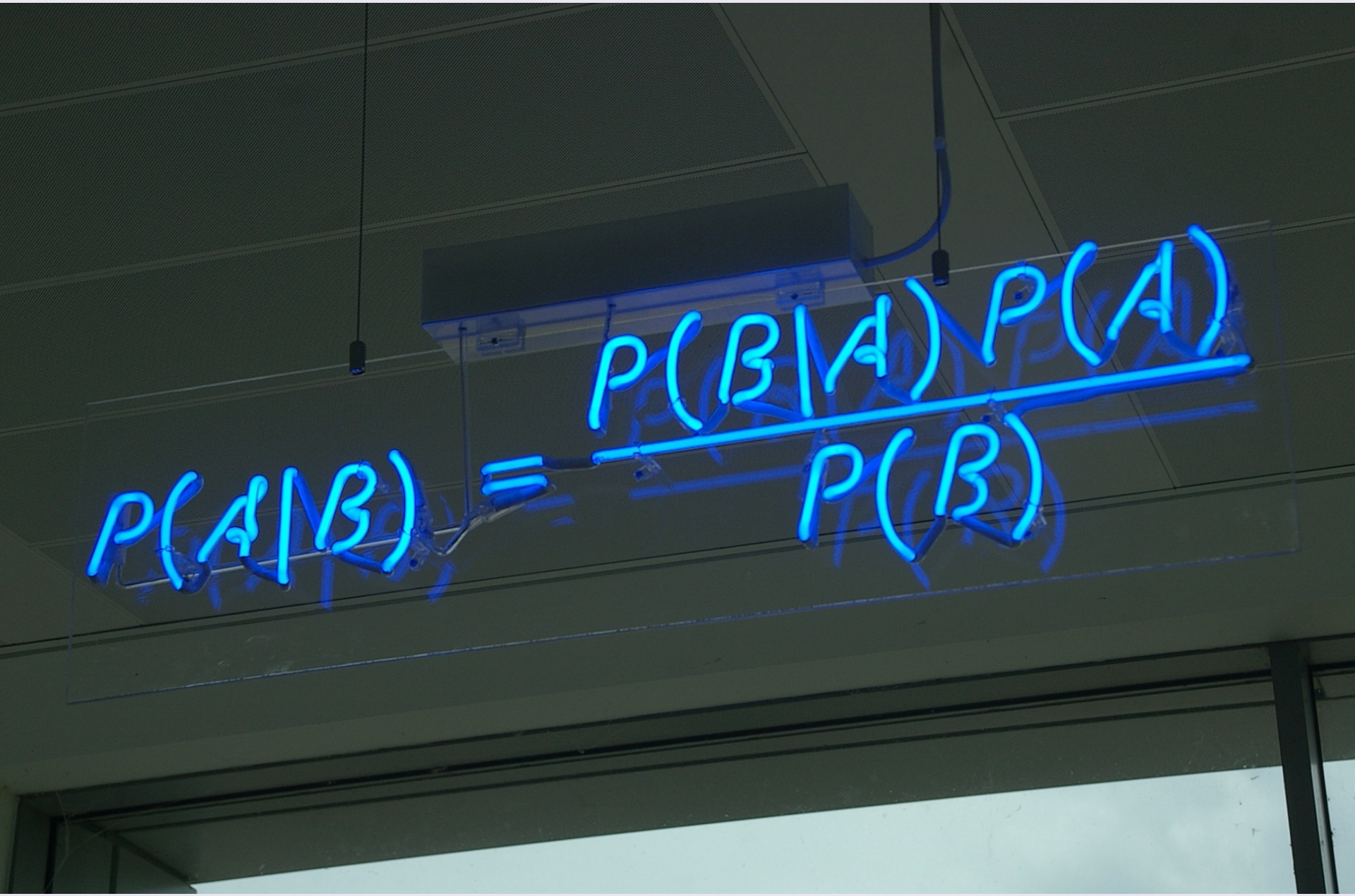


Figure 1.11 An illustration of a distribution over two variables, X, which takes 9 possible values, and Y, which takes two possible values. The top left figure shows a sample of 60 points drawn from a joint probability distribution over these variables. The remaining figures show histogram estimates of the marginal distributions $p(X)$ and $p(Y)$, as well as the conditional distribution $p(X|Y = 1)$ corresponding to the bottom row in the top left figure.

- Bayes' theorem

$$p(x,y) = \lim_{N\to\infty} \frac{N_{x,y}}{N} = \lim_{N\to\infty} \frac{N_{x,y}}{N_x} \frac{N_x}{N} = p(y \mid x)\, p(x)$$

$$p(x,y) = \lim_{N\to\infty} \frac{N_{x,y}}{N} = \lim_{N\to\infty} \frac{N_{x,y}}{N_y} \frac{N_y}{N} = p(x \mid y)\, p(y)$$


P(A|B) = P(B|A) P(A) / P(B)

Probability density functions

For smooth densities, events with finite probability are intervals

$$P(x \in [a,b]) = \int_a^b p(x)dx$$

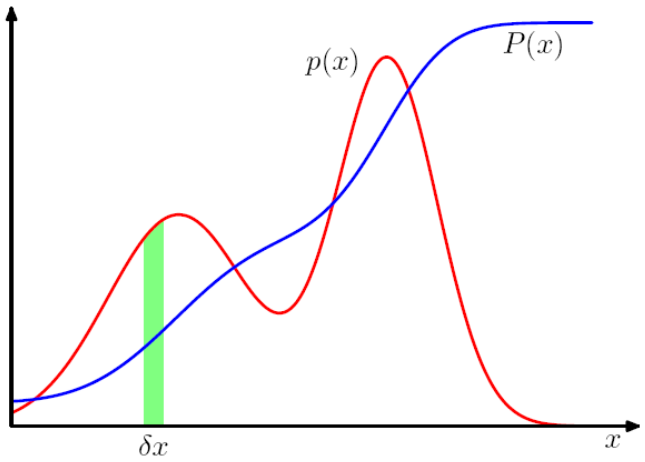


Figure 1.12 The concept of probability for discrete variables can be extended to that of a probability density $p(x)$ over a continuous variable x and is such that the probability of x lying in the interval $(x, x+\delta x)$ is given by $p(x)\delta x$ for $\delta x \to 0$. The probability density can be expressed as the derivative of a cumulative distribution function $P(x)$.

Expectations computed from the pdf

$$P(x \in \text{Domain of x}) = \int_{\text{Domain of x}} p(x)dx = 1$$

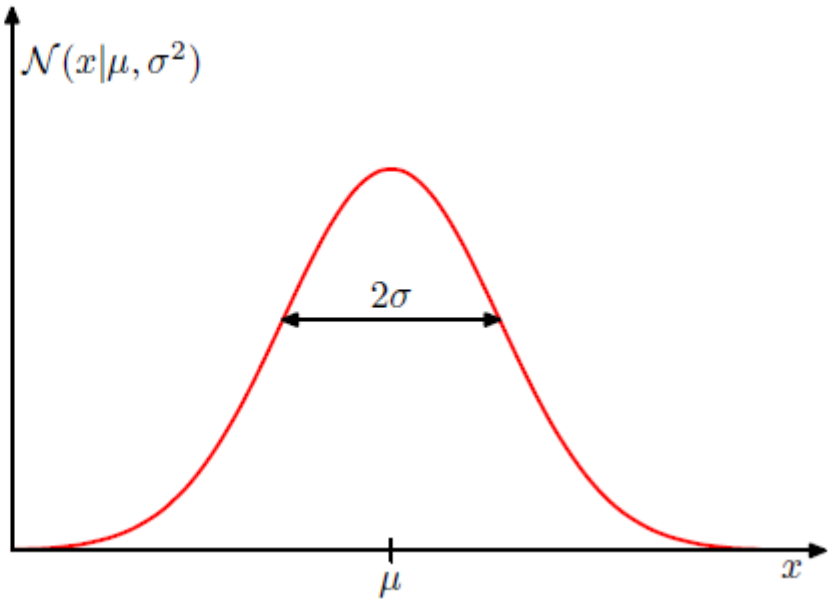
$$\mathcal{E}(f(x)) = \int_{\text{Domain of x}} f(x)p(x)dx$$

$$\mathcal{E}(x) \equiv \mu = \int_{\text{Domain of x}} xp(x)dx$$

The typical ”spread” of the data

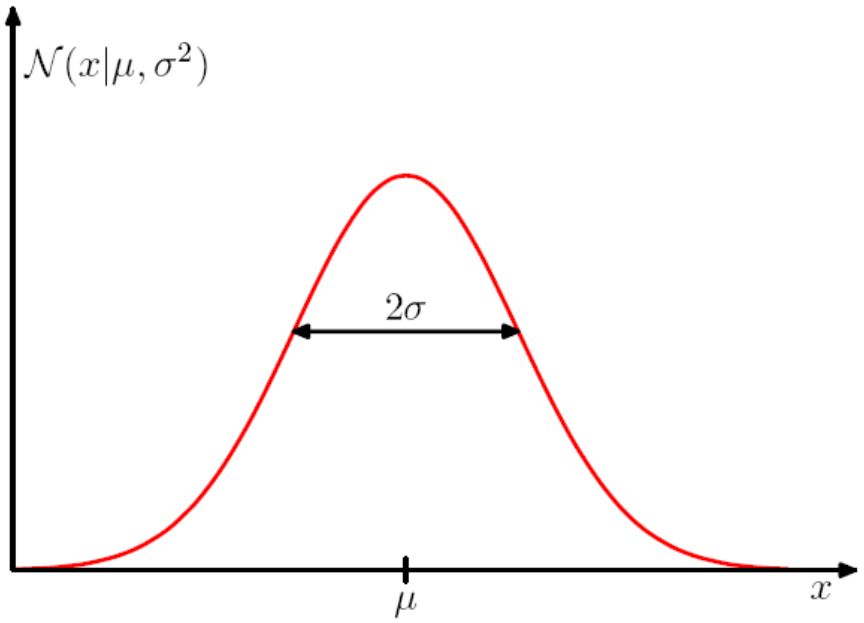
$$\sigma = \sqrt{\int_{\text{Domain of x}} (x-\mu)^2 p(x) dx}$$

Figure 1.13 Plot of the univariate Gaussian showing the mean μ and the standard deviation σ.



The normal distribution as a signal model

Figure 1.13 Plot of the univariate Gaussian showing the mean μ and the standard deviation σ.

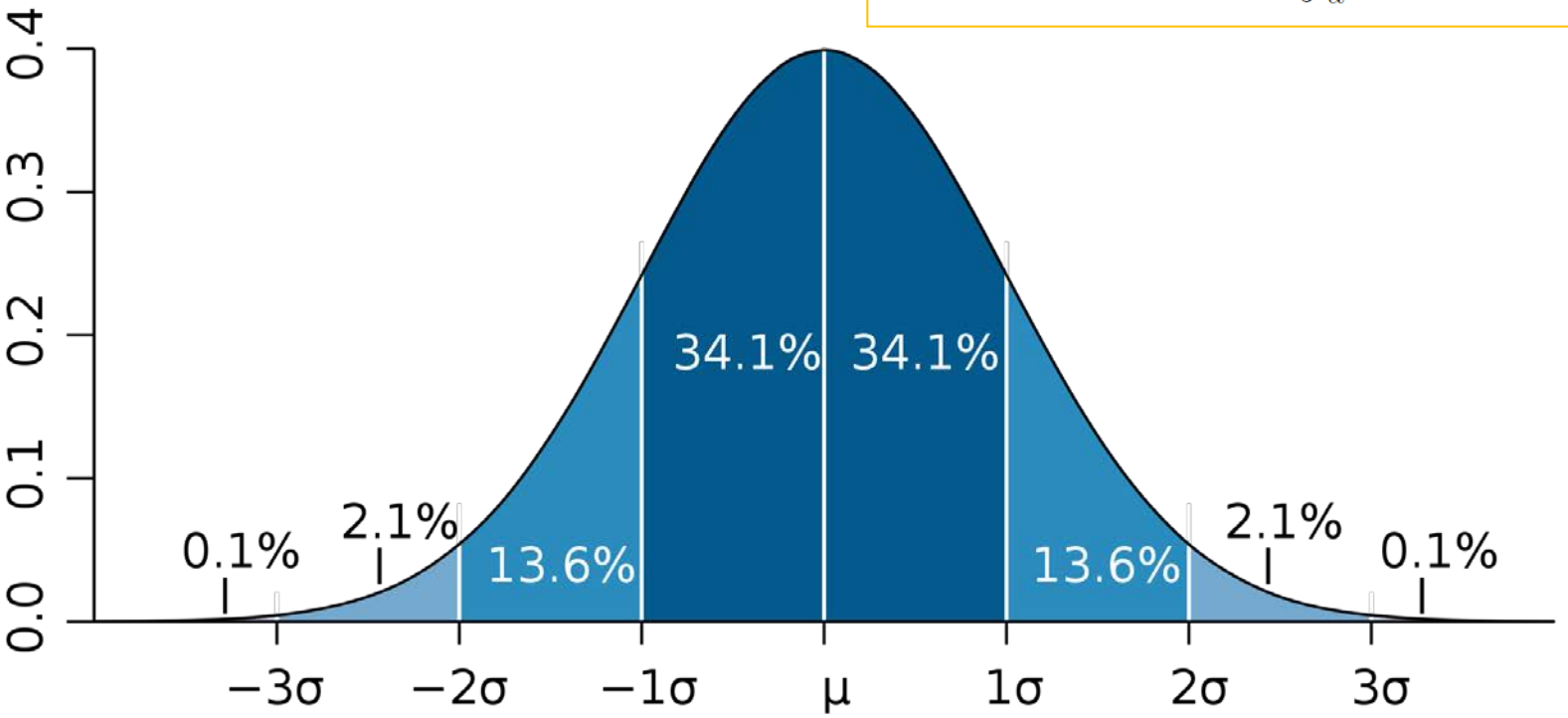


If events can be described as a "mean effect" with additive noise

$$p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

Probabilities under the normal pdf curve

$$P(x \in [a,b]) = \int_a^b p(x)dx$$



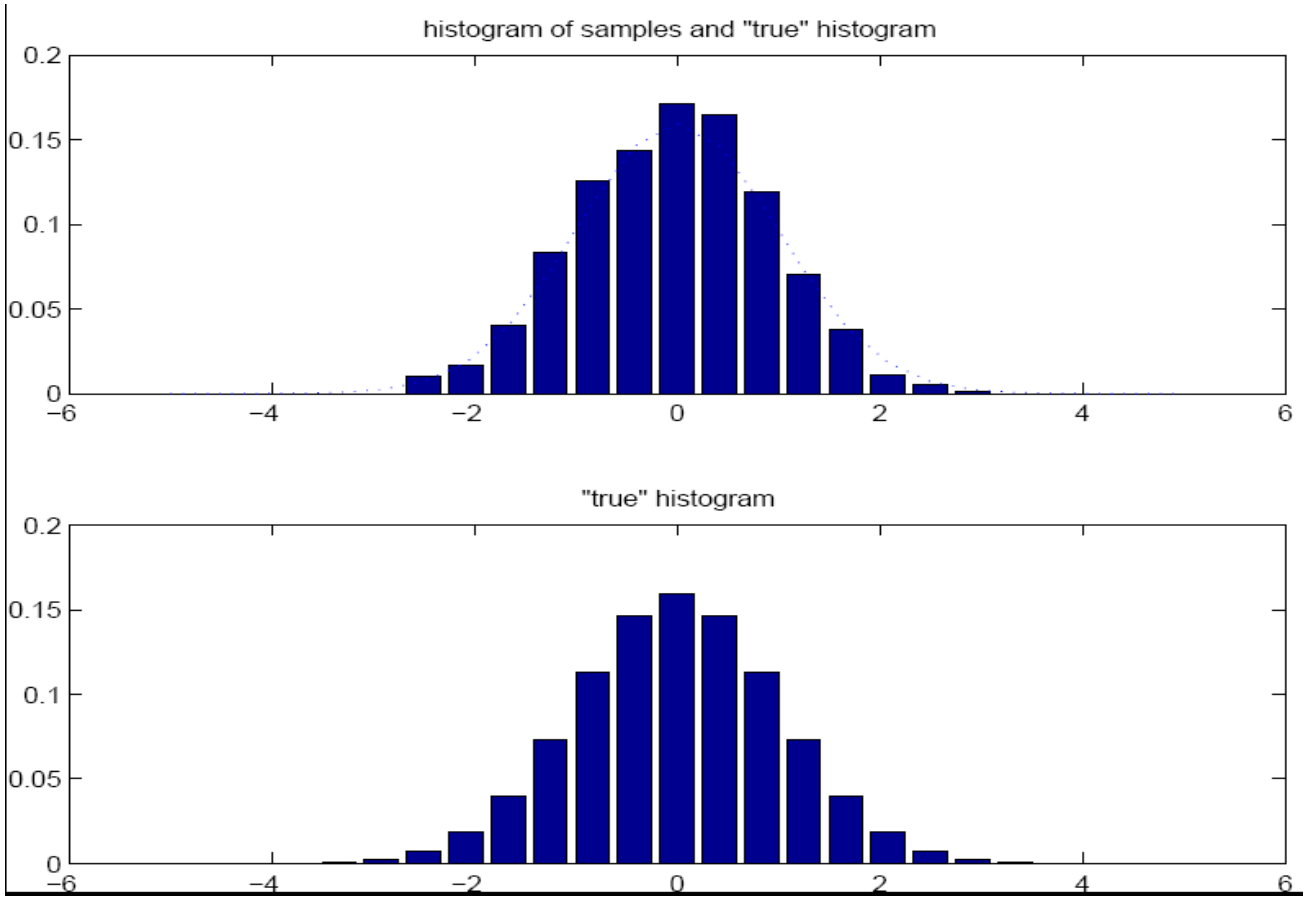
Parameters in the normal distribution

$$p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

$$\mu = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx$$

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx$$

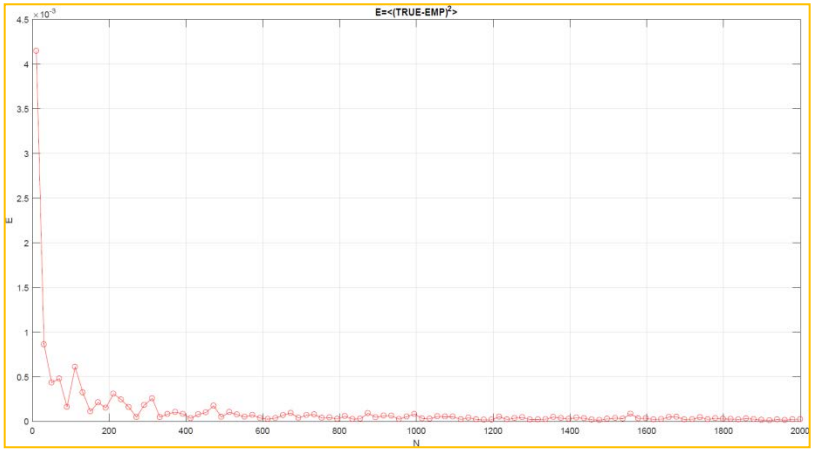
Models vs data

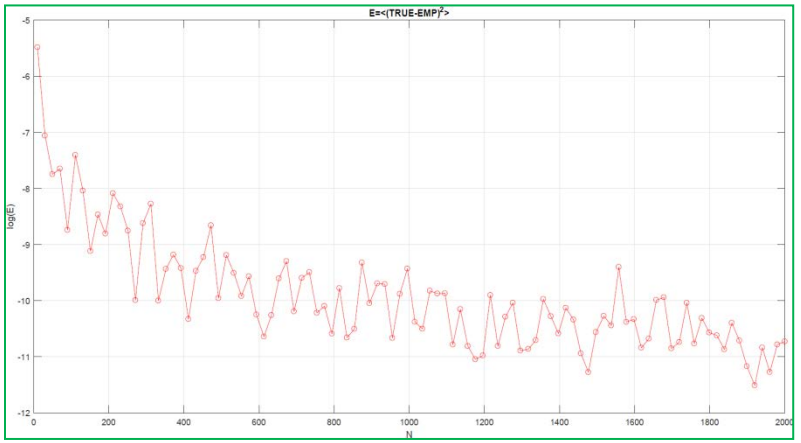


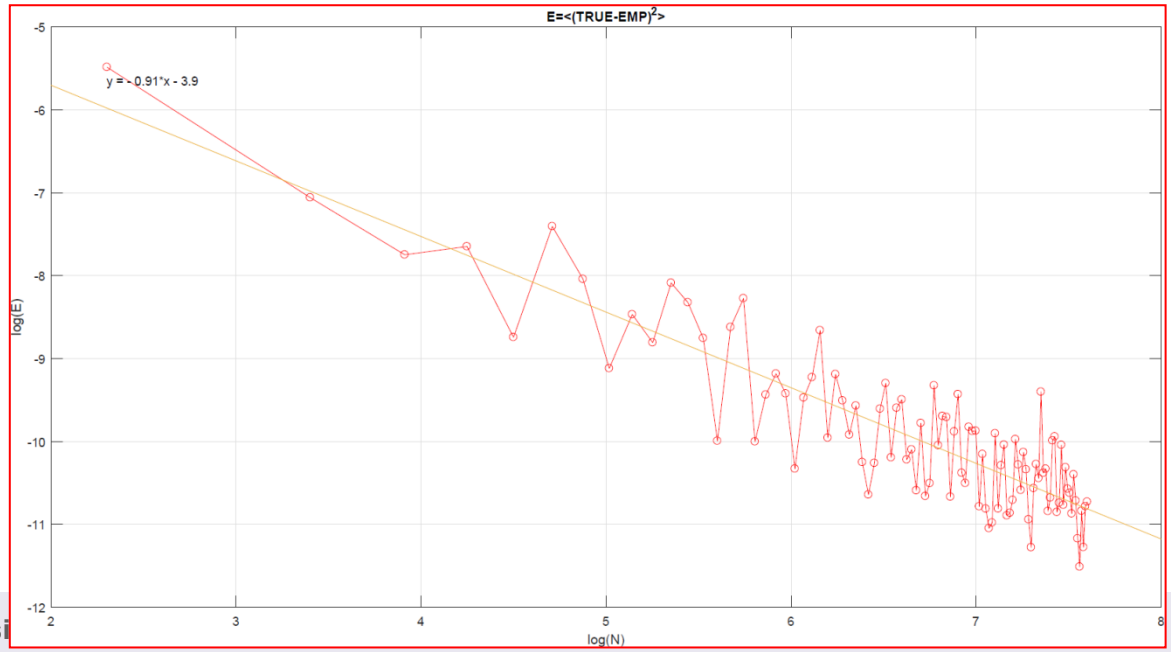
$$\mu = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx$$

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

Convergence of histograms







Signal detection problem

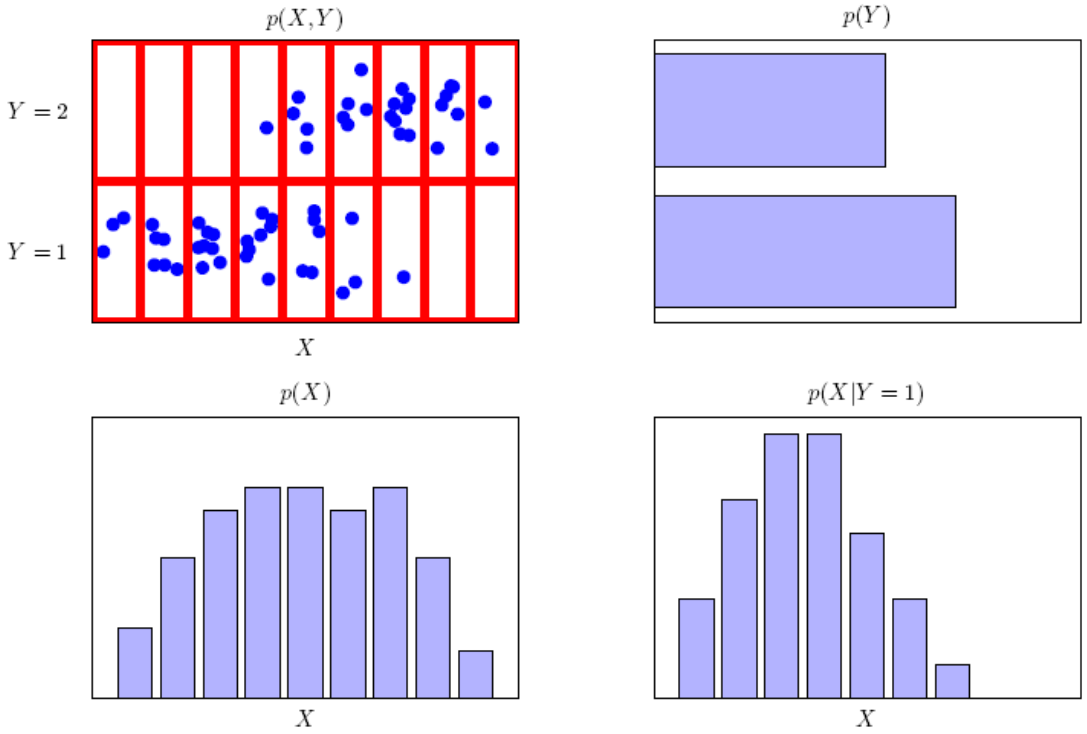


Figure 1.11 An illustration of a distribution over two variables, X, which takes 9 possible values, and Y, which takes two possible values. The top left figure shows a sample of 60 points drawn from a joint probability distribution over these variables. The remaining figures show histogram estimates of the marginal distributions $p(X)$ and $p(Y)$, as well as the conditional distribution $p(X|Y=1)$ corresponding to the bottom row in the top left figure.

$$P(\mathcal{C}_k, \mathbf{x}) = p(\mathbf{x}|\mathcal{C}_k)P(\mathcal{C}_k)$$
$$P(\mathcal{C}_k, \mathbf{x}) = P(\mathcal{C}_k|\mathbf{x})p(\mathbf{x})$$

$$P(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)P(\mathcal{C}_k)}{p(\mathbf{x})}$$

$$p(\mathbf{x}|\mathcal{C}_k) = \frac{P(\mathcal{C}_k|X^l)p(\mathbf{x})}{P(\mathcal{C}_k)}$$

$$\sum_{k=1}^{c} P(\mathcal{C}_k|\mathbf{x}) = 1$$

$$\sum_{k=1}^{c} p(\mathbf{x}|\mathcal{C}_k)P(\mathcal{C}_k) = p(\mathbf{x})$$

Signal detection: Bayes decision theory

- A signal detection system (or pattern classifier) provides a rule for assigning a measurement to a given signal category (class)
- Hence, a classifier divides measurement space (feature space) into disjoint regions $\mathcal{R}_1, \mathcal{R}_2, ..., \mathcal{R}_c$, such that measurements that fall into region $\mathcal{R}_k$ are assigned with class $\mathcal{C}_k$.
- Boundaries between regions are denoted decision surfaces or decision boundaries

Signal Detection: Bayes decision theory

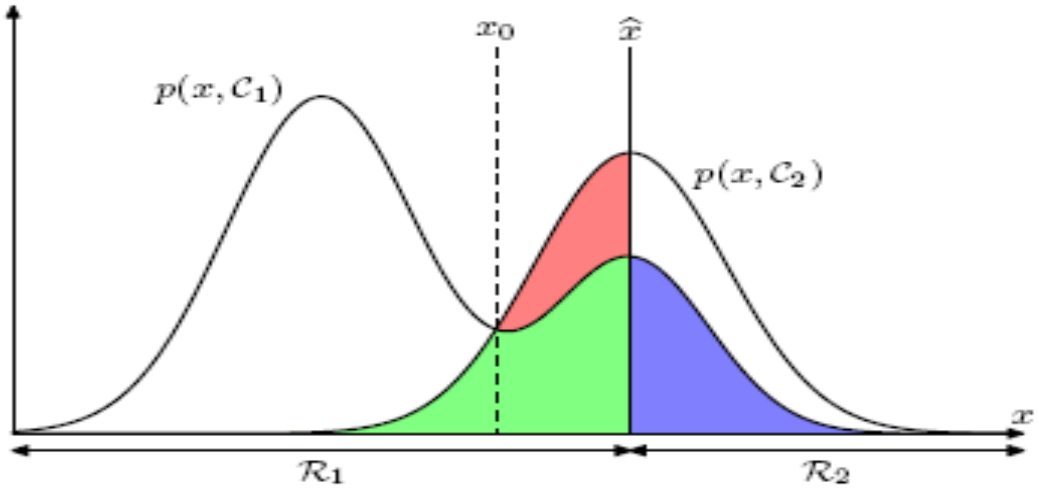


Figure 2: Schematic plot of the densities for a measured signal drawn from either of two populations $\mathcal{C}_1, \mathcal{C}_2$

$$
\begin{aligned}
P(\text{error}) &= P(x \in \mathcal{R}_2, \mathcal{C}_1) + P(x \in \mathcal{R}_1, \mathcal{C}_2) \\
&= P(x \in \mathcal{R}_2 | \mathcal{C}_1) P(\mathcal{C}_1) + P(x \in \mathcal{R}_1 | \mathcal{C}_2) P(\mathcal{C}_2) \\
&= \left(\int_{\mathcal{R}_2} p(x|\mathcal{C}_1) dx \right) P(\mathcal{C}_1) + \left(\int_{\mathcal{R}_1} p(x|\mathcal{C}_2) dx \right) P(\mathcal{C}_2)
\end{aligned}
$$

- The probability of error is minimized if we assign points to $\mathcal{R}_1$, whenever $p(x|\mathcal{C}_1)P(\mathcal{C}_1) > p(x|\mathcal{C}_2)P(\mathcal{C}_2)$

Signal detection: Use posterior for decision

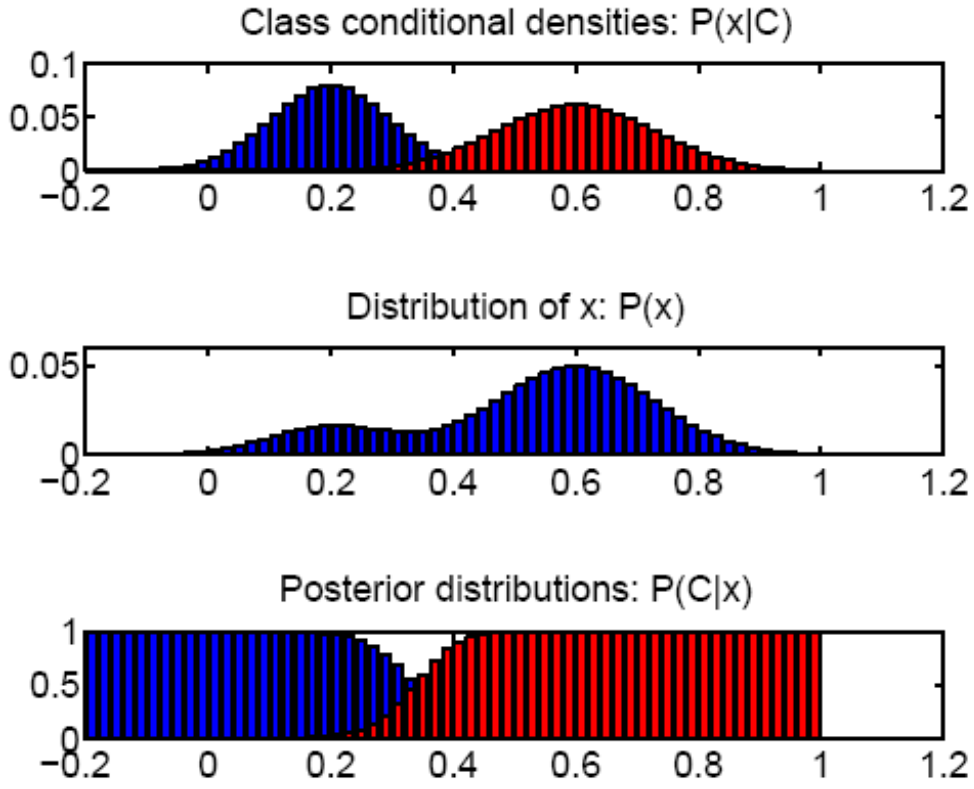


Figure 1: Schematic plot of the histograms for a measured signal drawn from either of t populations $\mathcal{C}_1, \mathcal{C}_2$, density of x, and the corresponding posteriors $P(\mathcal{C}|X)$'s

$$P(\mathcal{C}_k, \mathbf{x}) = p(\mathbf{x}|\mathcal{C}_k)P(\mathcal{C}_k)$$
$$P(\mathcal{C}_k, \mathbf{x}) = P(\mathcal{C}_k|\mathbf{x})p(\mathbf{x})$$

$$P(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)P(\mathcal{C}_k)}{p(\mathbf{x})}$$
$$p(\mathbf{x}|\mathcal{C}_k) = \frac{P(\mathcal{C}_k|X^l)p(\mathbf{x})}{P(\mathcal{C}_k)}$$

$$\sum_{k=1}^{c} P(\mathcal{C}_k|\mathbf{x}) = 1$$

$$\sum_{k=1}^{c} p(\mathbf{x}|\mathcal{C}_k)P(\mathcal{C}_k) = p(\mathbf{x})$$

Classification with asymmetric loss

$$
\begin{aligned}
R_k &= \sum_{j=1}^{c} L_{k,j} \int_{\mathcal{R}_j} p(\mathbf{x}|\mathcal{C}_k) d\mathbf{x} \\
R &= \sum_{k=1}^{c} R_k P(\mathcal{C}_k) \\
&= \sum_{j=1}^{c} \int_{\mathcal{R}_j} \sum_{k=1}^{c} L_{k,j} p(\mathbf{x}|\mathcal{C}_k) P(\mathcal{C}_k) d\mathbf{x}
\end{aligned}
$$

$$
\sum_{k=1}^{c} L_{k,j} p(\mathbf{x}|\mathcal{C}_k) P(\mathcal{C}_k) < \sum_{k=1}^{c} L_{k,i} p(\mathbf{x}|\mathcal{C}_k) P(\mathcal{C}_k)
$$

Machine learning: strategy for modeling objects in complex vector space representations (x)

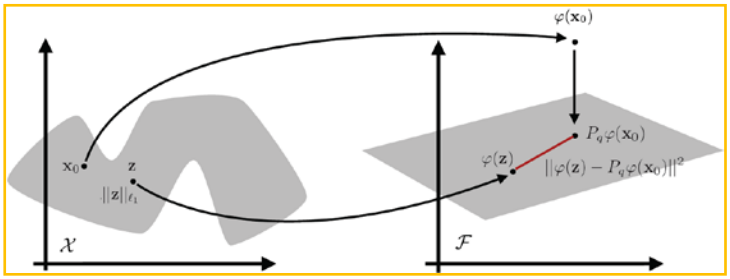
Measurements represented in vector spaces –
a point in a high-dimensional ”feature space”

Simplest learning idea: object similarity ~ spatial proximity
–similarity measures / distance metrics become important,

More realistic: proximity in complex manifolds representing the physical constraints
- ”long range” dependency

May need complex statistical models and lots of data (deep networks / big data)
to discover such representations from data

...e.g. ”objects” in image space occupy highly complex manifolds determined by
(approximate) invariances



Unsupervised learning in vector spaces

Identification of structure

- Preprocessing/ dim redux
- Missing data
- De-noising
- Novelty / outlier detection

Methods

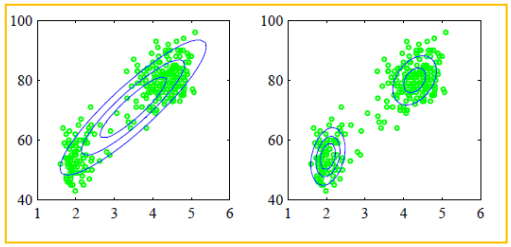
- Clustering
- Signal separation linear/non-linear

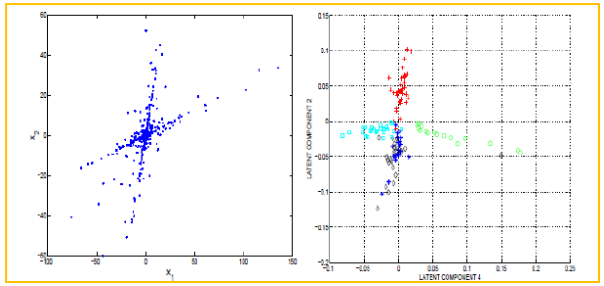
$$p(\mathbf{x}|\theta)$$

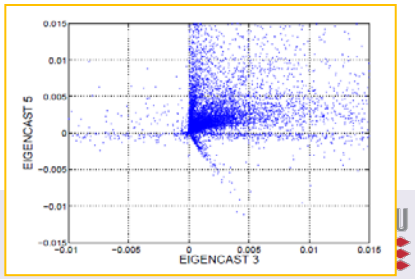
Measurement **Parameters**

Analysis by synthesis: generative models

Cautionary note:
Careful distinction between **prior** (say clustering assumption) and **posterior** (how well does the assumption hold...)
ICA: components are independent in prior but not necessarily in the posterior







Supervised Learning in vector spaces

Issues

Selection of model family

How to incorporate unlabeled samples?

Discriminative vs generative learning

A priori knowledge constraints/probabilistic Learning (ML, MAP, Bayes)

Outlier detection

- Erroneous label
- Unknown label

Model structure optimization

Performance issues

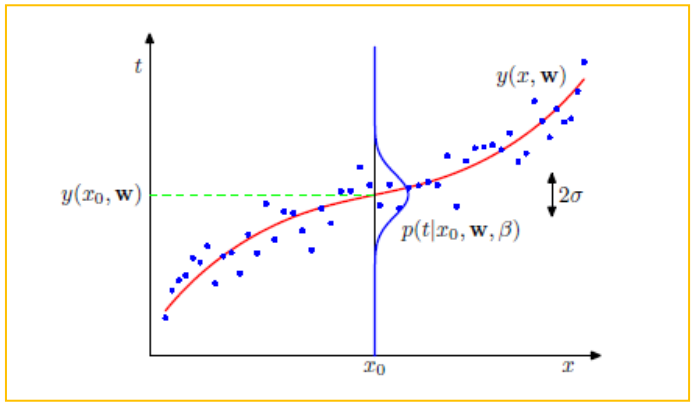
- Generalization
- Confidence
- Consensus methods
- Reproducibility

Visualization/interpretation

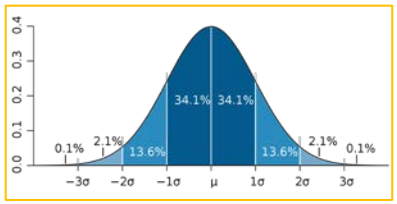
- Importance analysis
- Why should I trust you?

$$p(t|\mathbf{x},\theta)$$

Labels Measurement Parameters

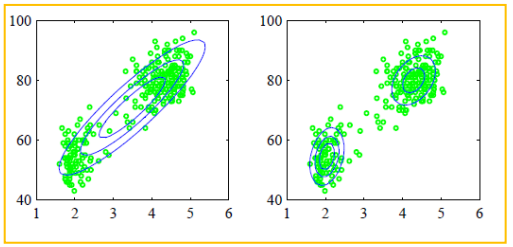


Basic terminology of model families



Parametric models

Models with fixed parametrization e.g. normal distribution

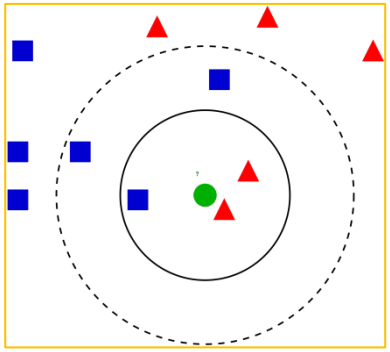


Semi-parametric modeling

Approach where we choose parametrization within a "family" of fixed parametrizations e.g. clustering

Non-parametric modeling

Models that grow with growing sample sizes e.g. nearest neighbor classification



Probability density functions in multivariate data: A text modeling example

Digital media trick: Vector space representation

Abstract representation - can be used for all digital media

Document is represented as a point in a high-dimensional ”feature space” – document similarity ~ spatial proximity

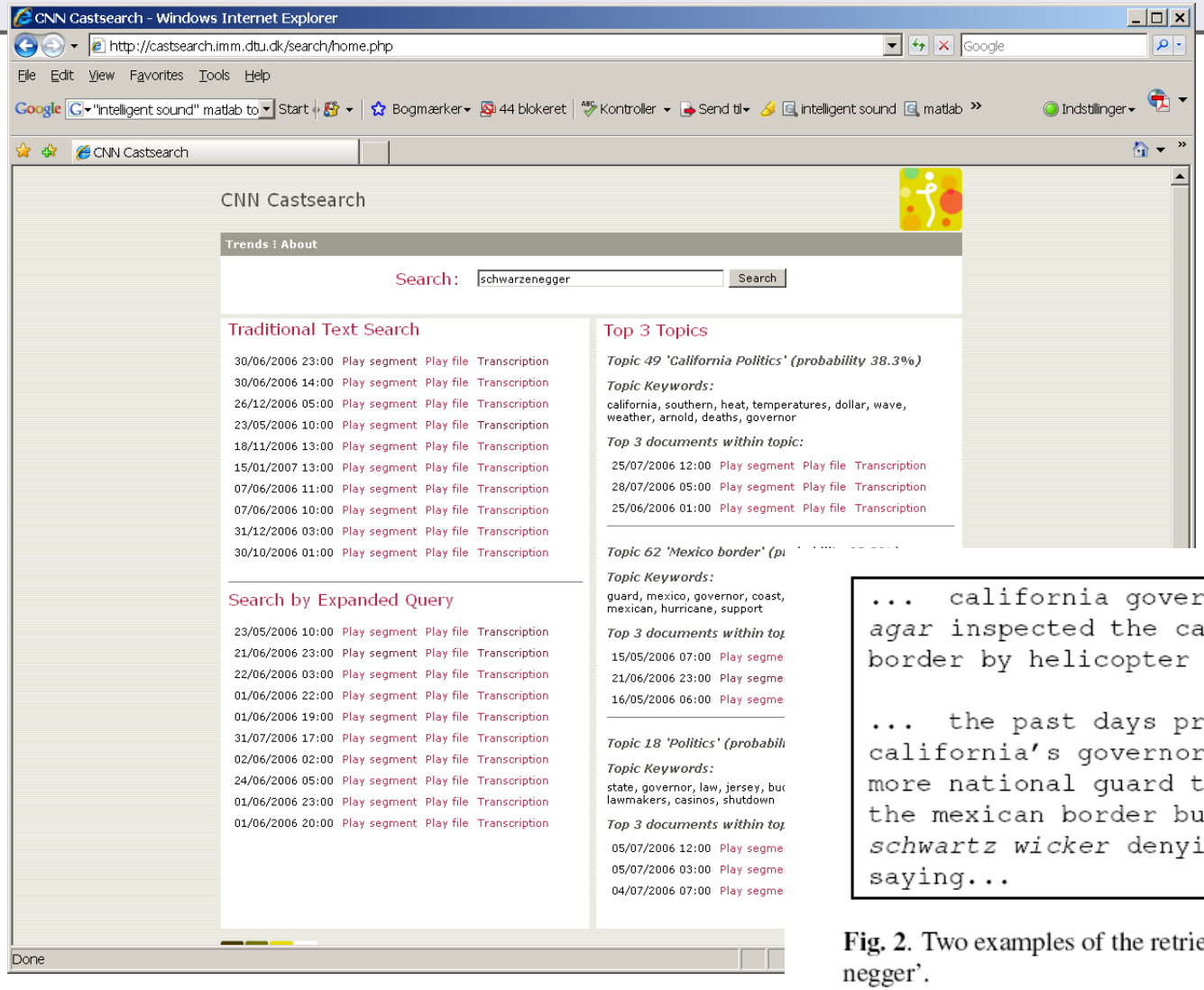
General features or “events”

- Social network: Social event involving a set of nodes
- Text: Bag of words (Term/keyword histograms),
- Image: Color histogram, texture measures, ”bag of features”
- Video: Object coordinates (tracking), active appearance models
- Sound: Spectrograms, cepstral coefficients, gamma tone filters

Document features are correlated, the pattern of correlation reflects ”associations”. Associations are context specific

Contexts can be identified unsupervised fashion by their feature associations = Latent semantics

castsearch.imm.dtu.dk (2007)



... california governor arnold's *fortson agar* inspected the california mexico border by helicopter wednesday to see ...

... the past days president bush asking california's governor for fifteen hundred more national guard troops to help patrol the mexican border but governor orville *schwartz wicker* denying the request saying...

Fig. 2. Two examples of the retrieved text for a query on 'schwarzenegger'.

CASTSEARCH - CONTEXT BASED SPEECH DOCUMENT RETRIEVAL

Lasse Lohilahti Mølgaard, Kasper Winther Jørgensen, and Lars Kai Hansen

Informatics and Mathematical Modelling
Technical University of Denmark Richard Petersens Plads
Building 321, DK-2800 Kongens Lyngby, Denmark

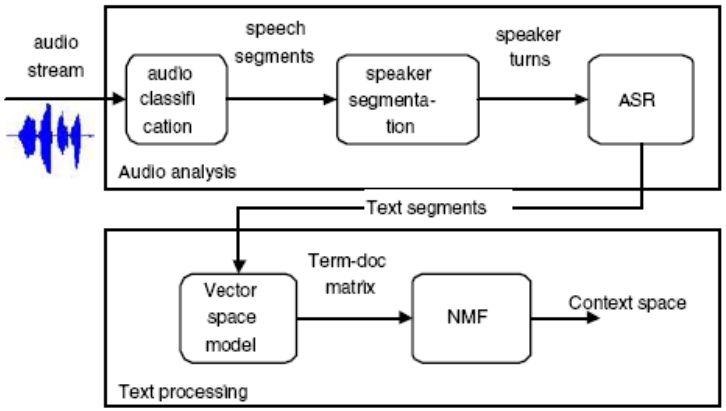
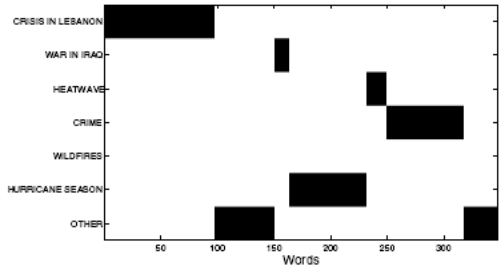
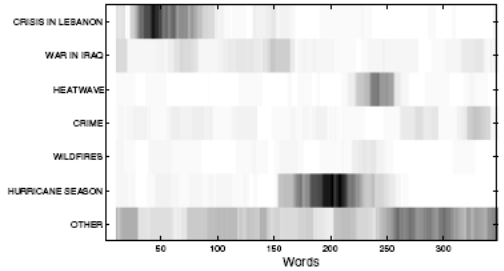


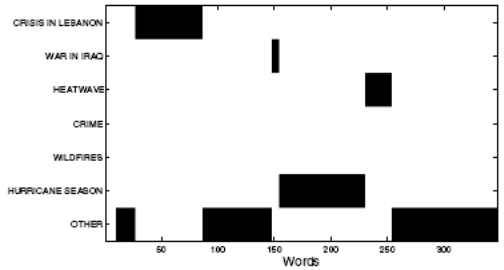
Fig. 1. The system setup. The audio stream is first processed using audio segmentation. Segments are then using an automatic speech recognition (ASR) system to produce text segments. The text is then processed using a vector representation of text and apply non-negative matrix factorization (NMF) to find a topic space.



(a) Manual segmentation.



(b) $p(k|d^*)$ for each context. Black means high probability.

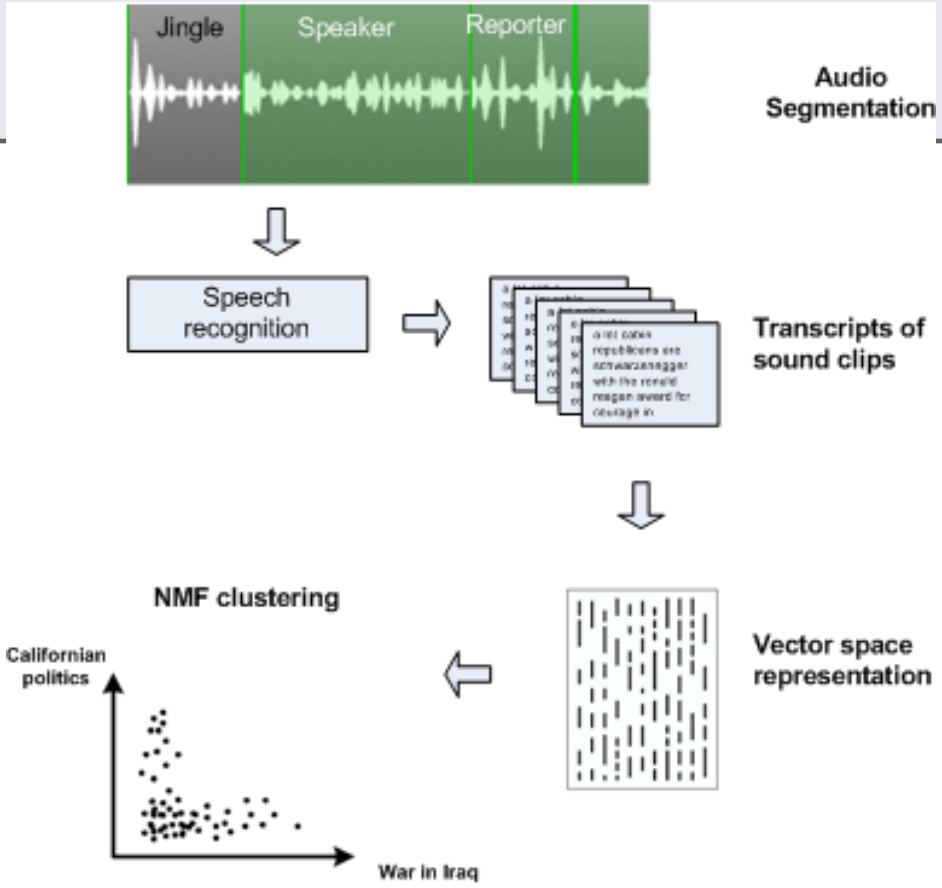


(c) The segmentation based on $p(k|d^*)$.

Fig. 3. Figure 3(a) shows the manual segmentation of the news show into 7 classes. Figure 3(b) shows the distribution $p(k|d^*)$ used to do the actual segmentation shown in figure 3(c). The NMF-segmentation is in general consistent with the manual segmentation. Though, the segment that is manually segmented as 'crime' is labeled 'other' by the NMF-segmentation

“Bag of Words”

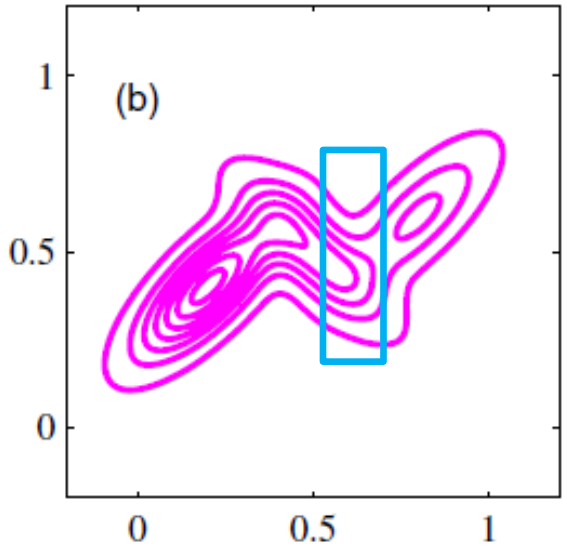
Terms	Documents								
	c1	c2	c3	c4	c5	m1	m2	m3	m4
computer	1	1	0	0	0	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
time	0	1	0	0	1	0	0	0	0
user	0	1	1	0	1	0	0	0	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0

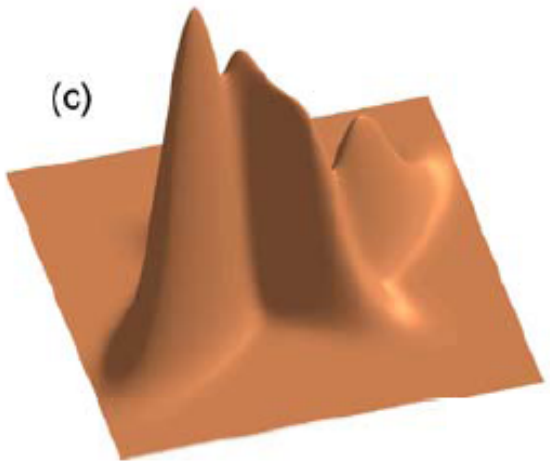


Very efficient for detection of context
Often leads to very high-dimensional learning problems
Correlations between words effectively reduces dimension

Multivariate processes

$$P(x_j \in [a_j, b_j] | j = 1, .., d) = \int_{a_1}^{b_1} ... \int_{a_d}^{b_d} p(\mathbf{x}) d\mathbf{x}$$





Multivariate normal distribution

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

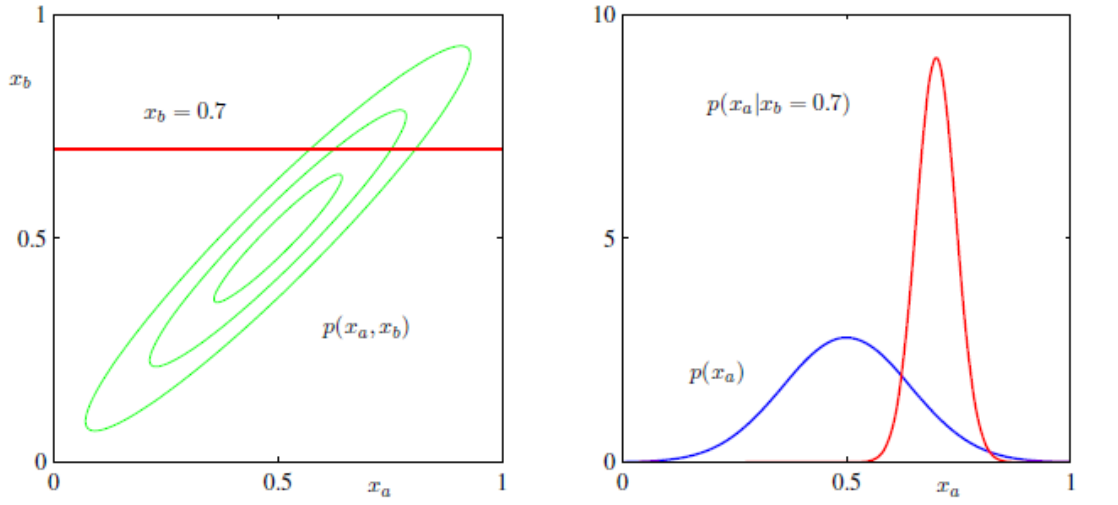


Figure 2.9 The plot on the left shows the contours of a Gaussian distribution $p(x_a, x_b)$ over two variables, and the plot on the right shows the marginal distribution $p(x_a)$ (blue curve) and the conditional distribution $p(x_a|x_b)$ for $x_b = 0.7$ (red curve).

Expectations in the normal distribution

$$\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \int \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \mathbf{x}\mathbf{x}^{\mathrm{T}}\,\mathrm{d}\mathbf{x}$$

$$= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \int \exp\left\{-\frac{1}{2}\mathbf{z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{z}\right\} (\mathbf{z}+\boldsymbol{\mu})(\mathbf{z}+\boldsymbol{\mu})^{\mathrm{T}}\,\mathrm{d}\mathbf{z}$$

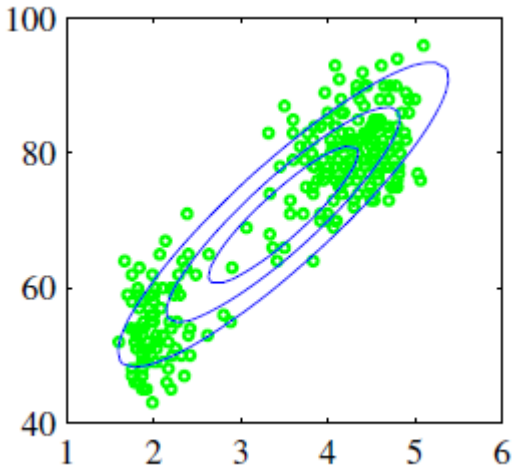
$$\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] = \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma}.$$

Dependency

$$\begin{aligned}\mathrm{cov}[x,y] &= \mathbb{E}_{x,y}\left[\{x-\mathbb{E}[x]\}\{y-\mathbb{E}[y]\}\right] \\ &= \mathbb{E}_{x,y}[xy]-\mathbb{E}[x]\mathbb{E}[y]\end{aligned}$$

$$\boldsymbol{\Sigma} = \int_{\text{Domain of } \mathbf{x}} (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\top} p(\mathbf{x}) d\mathbf{x}$$

$$\boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n-\boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}}$$



Real symmetric matrices recap

Appendix C. Properties of Matrices

$$\mathbf{A}\mathbf{u}_i = \lambda_i \mathbf{u}_i \qquad |\mathbf{A} - \lambda_i \mathbf{I}| = 0 \qquad \mathbf{A} = \sum_{i=1}^{M} \lambda_i \mathbf{u}_i \mathbf{u}_i^{\mathrm{T}}$$

$$\mathbf{u}_i^{\mathrm{T}} \mathbf{u}_j = I_{ij} \qquad \mathbf{A}^{-1} = \sum_{i=1}^{M} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^{\mathrm{T}}.$$

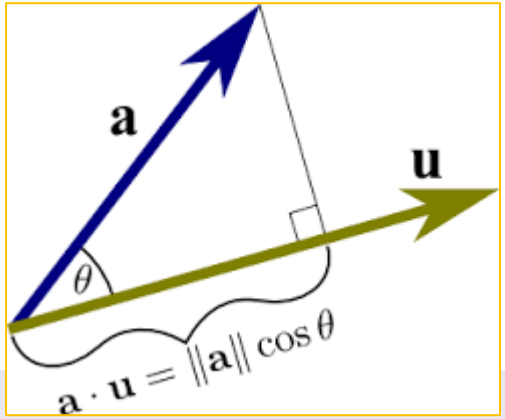
We can take the eigenvectors $\mathbf{u}_i$ to be the columns of an $M \times M$ matrix $\mathbf{U}$, which from orthonormality satisfies

$$\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}. \tag{C.37}$$

$$\mathbf{A}\mathbf{U} = \mathbf{U}\mathbf{\Lambda}$$

$$\mathbf{U}^{\mathrm{T}}\mathbf{A}\mathbf{U} = \mathbf{\Lambda}$$

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathrm{T}}$$



(Symmetric) matrices recap

Appendix C. Properties of Matrices

$$(\mathbf{AB})^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}} \qquad \mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}. \qquad (\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}.$$

$$\mathrm{Tr}(\mathbf{AB}) = \mathrm{Tr}(\mathbf{BA}).$$

$$|\mathbf{A}| = \sum (\pm 1) A_{1i_1} A_{2i_2} \cdots A_{Ni_N} \qquad |\mathbf{A}| = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}.$$

$$|\mathbf{AB}| = |\mathbf{A}||\mathbf{B}| \qquad \left|\mathbf{A}^{-1}\right| = \frac{1}{|\mathbf{A}|} \qquad |\mathbf{A}| = \prod_{i=1}^{M} \lambda_i.$$

$$\mathbf{A}\mathbf{u}_i = \lambda_i \mathbf{u}_i \qquad |\mathbf{A} - \lambda_i \mathbf{I}| = 0 \qquad \mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j = I_{ij} \qquad \mathbf{AU} = \mathbf{U}\mathbf{\Lambda}$$

$$\mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w} \geqslant 0$$ Positive semi-definite

Multivariate normal distribution

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

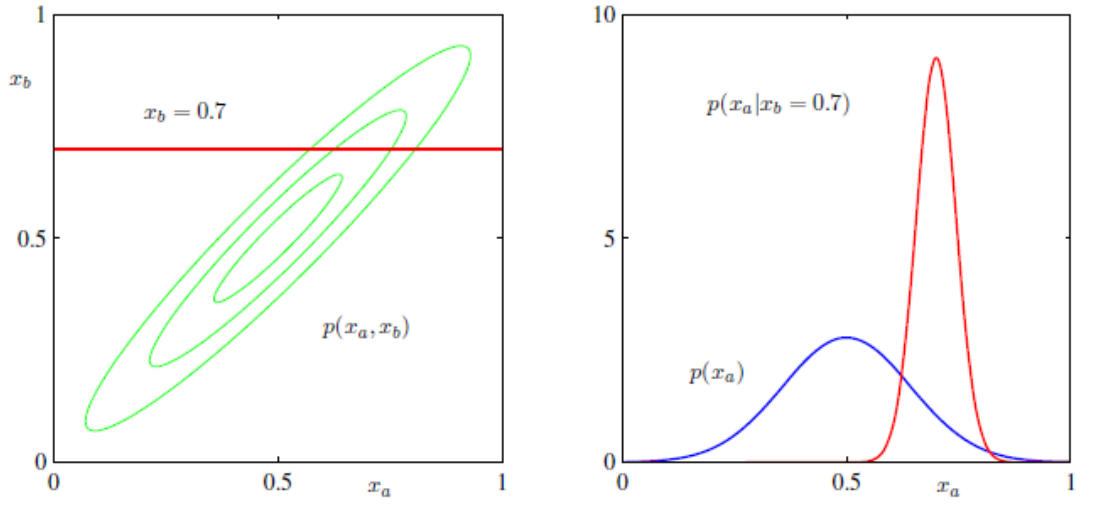


Figure 2.9 The plot on the left shows the contours of a Gaussian distribution $p(x_a, x_b)$ over two variables, and the plot on the right shows the marginal distribution $p(x_a)$ (blue curve) and the conditional distribution $p(x_a|x_b)$ for $x_b = 0.7$ (red curve).

Analysis of the covariance matrix

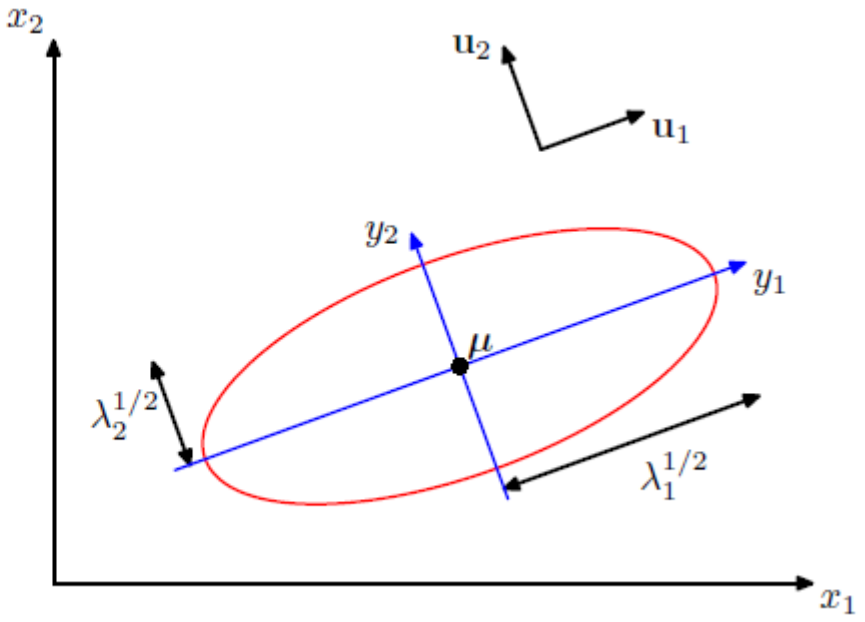


Figure 2.7 The red curve shows the elliptical surface of constant probability density for a Gaussian in a two-dimensional space $\mathbf{x} = (x_1, x_2)$ on which the density is $\exp(-1/2)$ of its value at $\mathbf{x} = \boldsymbol{\mu}$. The major axes of the ellipse are defined by the eigenvectors $\mathbf{u}_i$ of the covariance matrix, with corresponding eigenvalues λ_i.

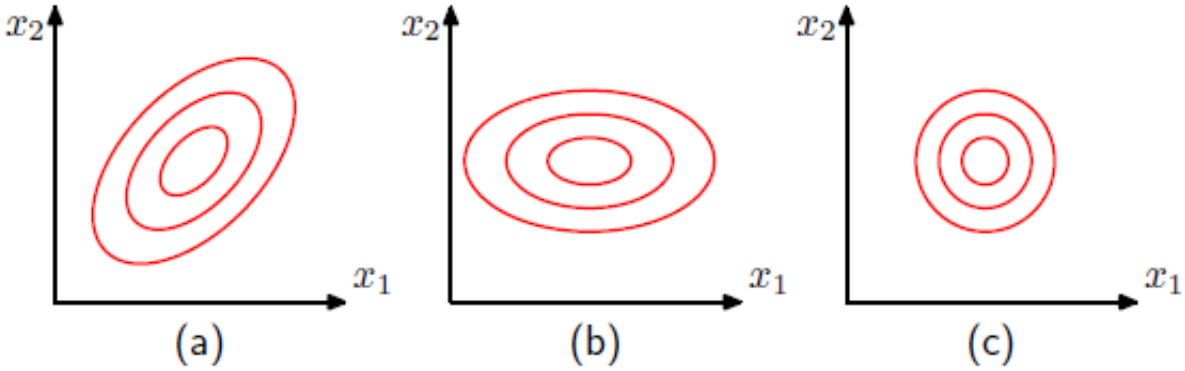
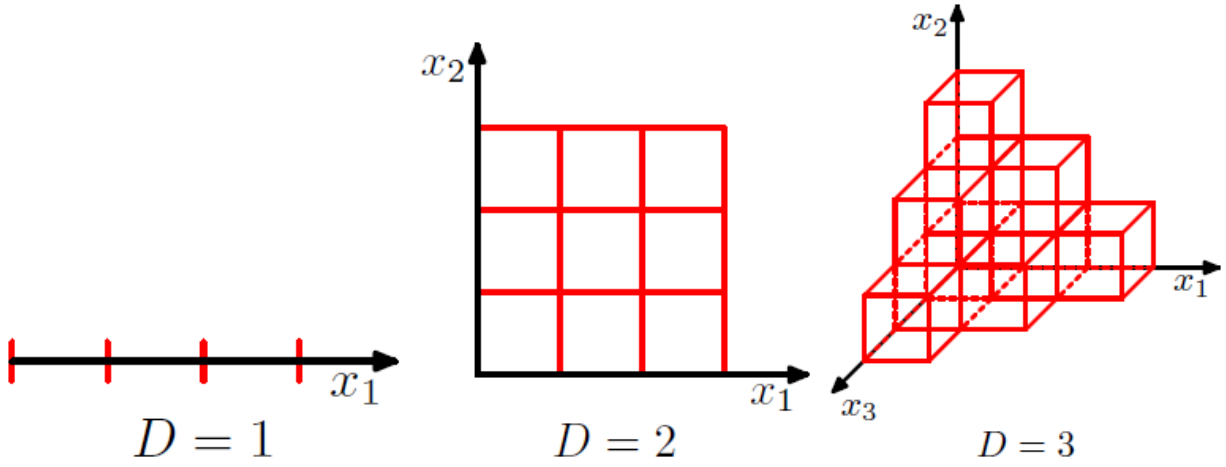


Figure 2.8 Contours of constant probability density for a Gaussian distribution in two dimensions in which the covariance matrix is (a) of general form, (b) diagonal, in which the elliptical contours are aligned with the coordinate axes, and (c) proportional to the identity matrix, in which the contours are concentric circles.

Features and the curse of dimensionality

- To specify a map (e.g. a discriminant function) on a d-dimensional space by dividing the relevant parts of the this space into L cells pr. dimension requires L^d cells.



Principal components (Bishop chapter 12)

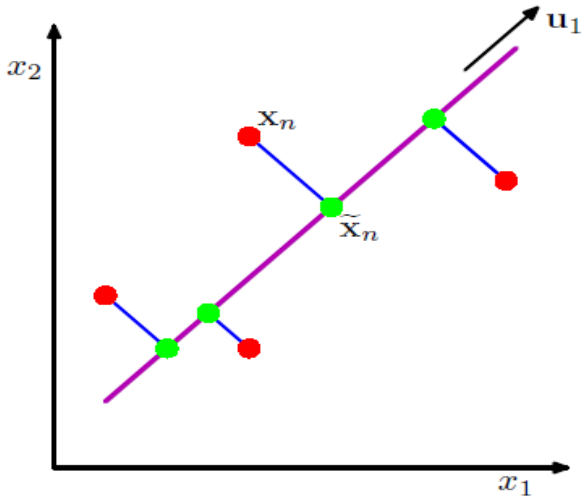
Consider a data set of observations $\{\mathbf{x}_n\}$ where $n = 1, \ldots, N$ $\quad \overline{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n$

Variance of projection on unit vector $\quad \frac{1}{N}\sum_{n=1}^{N}\{\mathbf{u}_1^{\mathrm{T}}\mathbf{x}_n - \mathbf{u}_1^{\mathrm{T}}\overline{\mathbf{x}}\}^2 = \mathbf{u}_1^{\mathrm{T}}\mathbf{S}\mathbf{u}_1$

where $\mathbf{S}$ is the data covariance matrix defined by

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \overline{\mathbf{x}})(\mathbf{x}_n - \overline{\mathbf{x}})^{\mathrm{T}}.$$

Figure 12.2 Principal component analysis seeks a space of lower dimensionality, known as the principal subspace and denoted by the magenta line, such that the orthogonal projection of the data points (red dots) onto this subspace maximizes the variance of the projected points (green dots). An alternative definition of PCA is based on minimizing the sum-of-squares of the projection errors, indicated by the blue lines.



Principal components

We now maximize the projected variance $\mathbf{u}_1^{\mathrm{T}}\mathbf{S}\mathbf{u}_1$ with respect to $\mathbf{u}_1$. Clearly, this has to be a constrained maximization to prevent $\|\mathbf{u}_1\| \to \infty$. The appropriate constraint comes from the normalization condition $\mathbf{u}_1^{\mathrm{T}}\mathbf{u}_1 = 1$. To enforce this constraint, we introduce a Lagrange multiplier that we shall denote by λ_1, and then make an unconstrained maximization of

$$\mathbf{u}_1^{\mathrm{T}}\mathbf{S}\mathbf{u}_1 + \lambda_1\left(1 - \mathbf{u}_1^{\mathrm{T}}\mathbf{u}_1\right). \tag{12.4}$$

By setting the derivative with respect to $\mathbf{u}_1$ equal to zero, we see that this quantity will have a stationary point when

$$\mathbf{S}\mathbf{u}_1 = \lambda_1\mathbf{u}_1 \tag{12.5}$$

which says that $\mathbf{u}_1$ must be an eigenvector of $\mathbf{S}$. If we left-multiply by $\mathbf{u}_1^{\mathrm{T}}$ and make use of $\mathbf{u}_1^{\mathrm{T}}\mathbf{u}_1 = 1$, we see that the variance is given by

$$\mathbf{u}_1^{\mathrm{T}}\mathbf{S}\mathbf{u}_1 = \lambda_1 \tag{12.6}$$

Principal components of handwritten digits

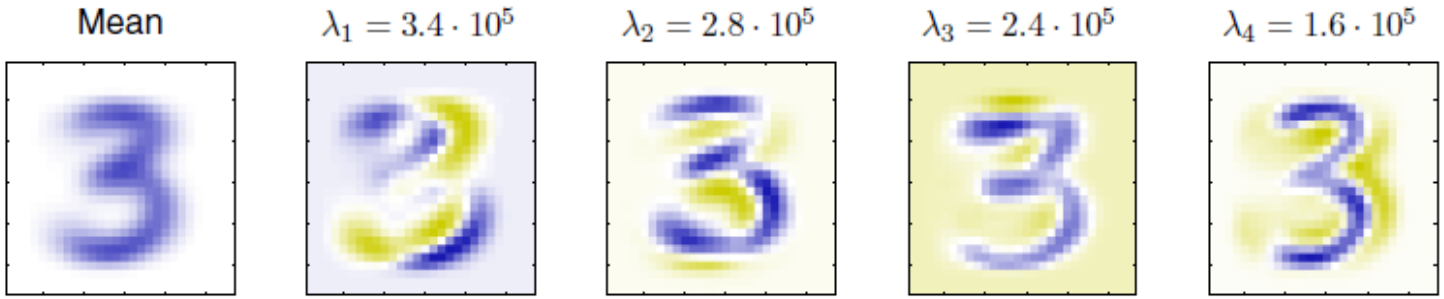


Figure 12.3 The mean vector $\overline{\mathbf{x}}$ along with the first four PCA eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_4$ for the off-line digits data set, together with the corresponding eigenvalues.

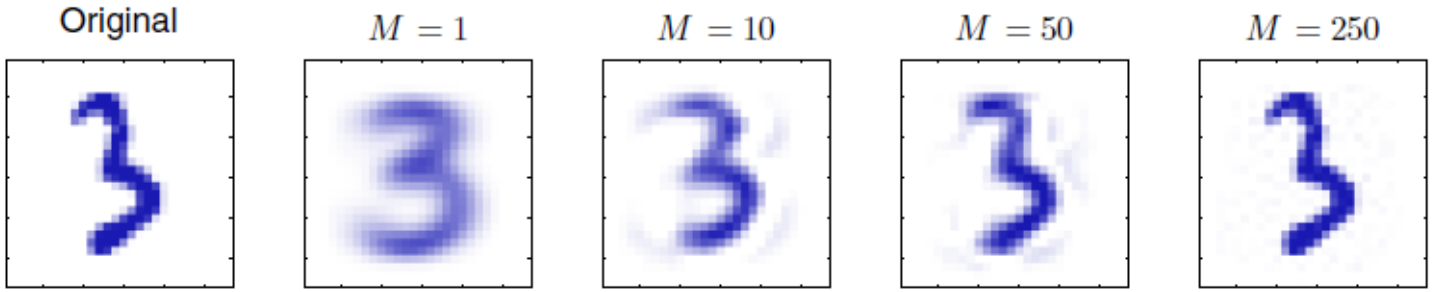
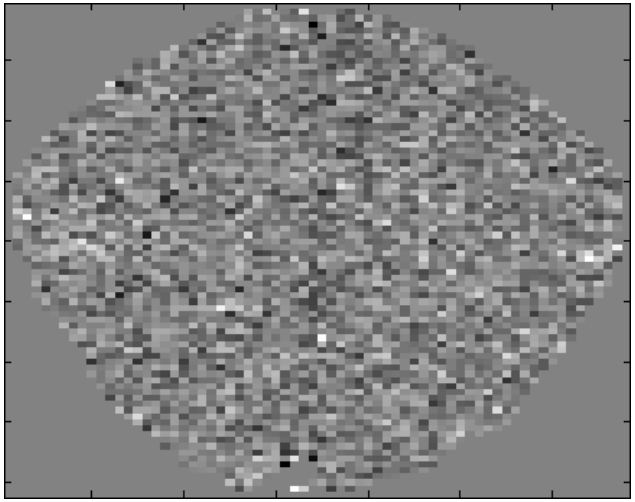


Figure 12.5 An original example from the off-line digits data set together with its PCA reconstructions obtained by retaining M principal components for various values of M. As M increases the reconstruction becomes more accurate and would become perfect when $M = D = 28 \times 28 = 784$.

Functional Magnetic Resonance Imaging

- Indirect measure of neural activity – hemodynamics
- Invented in 1992
- A cloudy window to the human brain
- Challenges:
  - Signals are multi-dimensional mixtures
  - No simple relation between measures and brain state - ”what is signal and what is noise”?



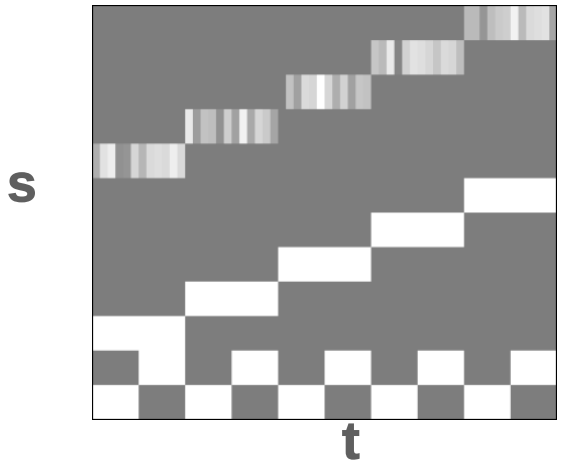
TR = 333 ms

Multivariate neuroimaging models

Neuroimaging aims at extracting the mutual information between stimulus and response.

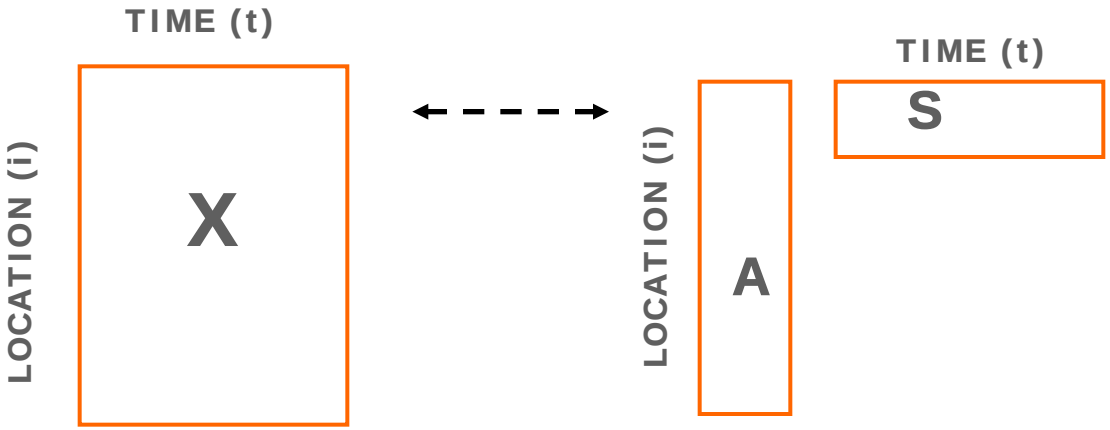
- Stimulus: Macroscopic variables, ”design matrix” … *s(t)*
- Response: Micro/meso-scopic variables, the neuroimage … *x(t)*
- Mutual information is stored in the joint distribution … *p(x,s).*

Often s(t) is assumed known….unsupervised methods consider s(t) or parts of s(t) ”hidden”…..



Factor models

- Represent a datamatrix by a low-dimensional approximation



$$X(i,t) \approx \sum_{k=1}^{K} A(i,k)S(k,t)$$

PCA of fMRI

NeuroImage 9, 534–544 (1999)
Article ID nimg.1998.0425, available online at http://www.idealibrary.com on IDEAL®

Generalizable Patterns in Neuroimaging: How Many Principal Components?

Lars Kai Hansen,*,1 Jan Larsen,* Finn Årup Nielsen,* Stephen C. Strother,†,‖ Egill Rostrup,‡ Robert Savoy,§ Nicholas Lange,¶ John Sidtis,‖ Claus Svarer,** and Olaf B. Paulson**

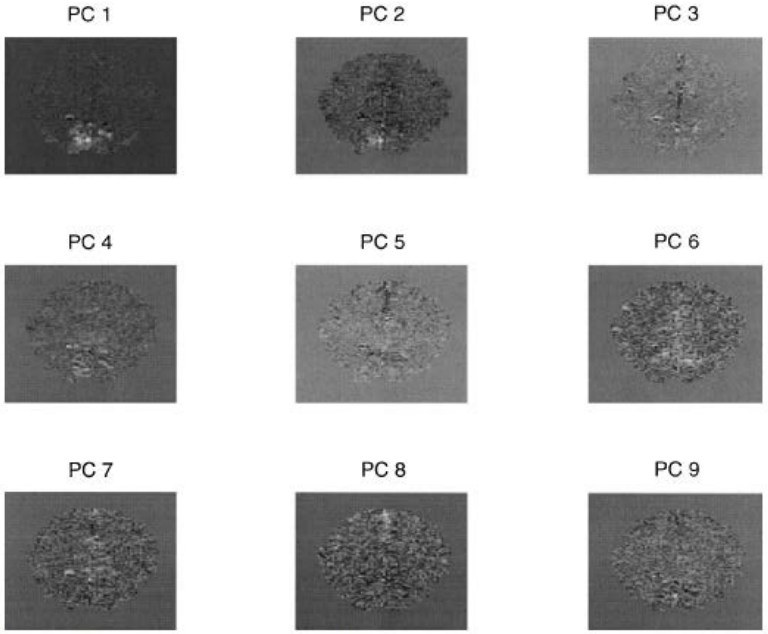


FIG. 6. Data set II. Covariance eigenimages corresponding to the nine most significant principal components. The eigenimage corresponding to the dominating first PC is focused in visual areas. Using the bias/variance trade-off curves in Fig. 4, we find that only the eigenimages corresponding to the first three components generalize.

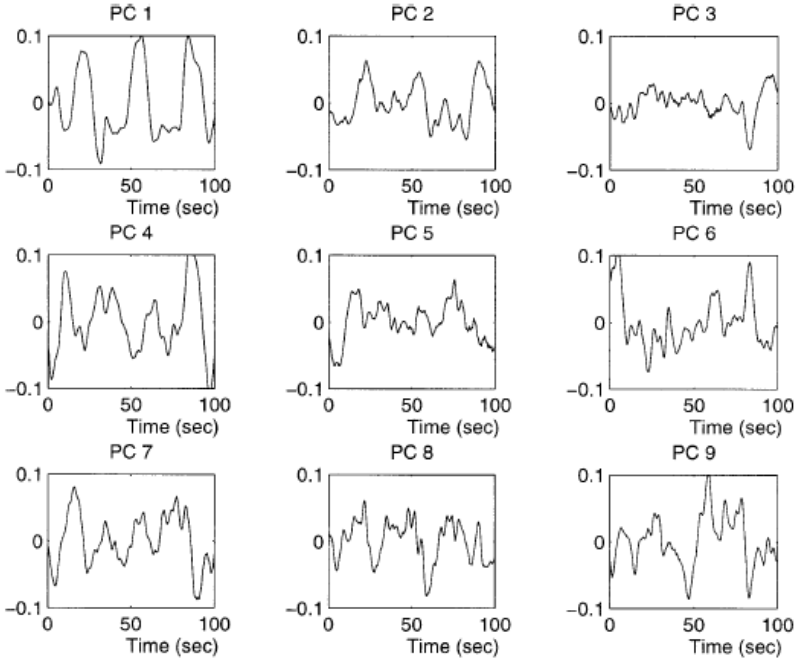
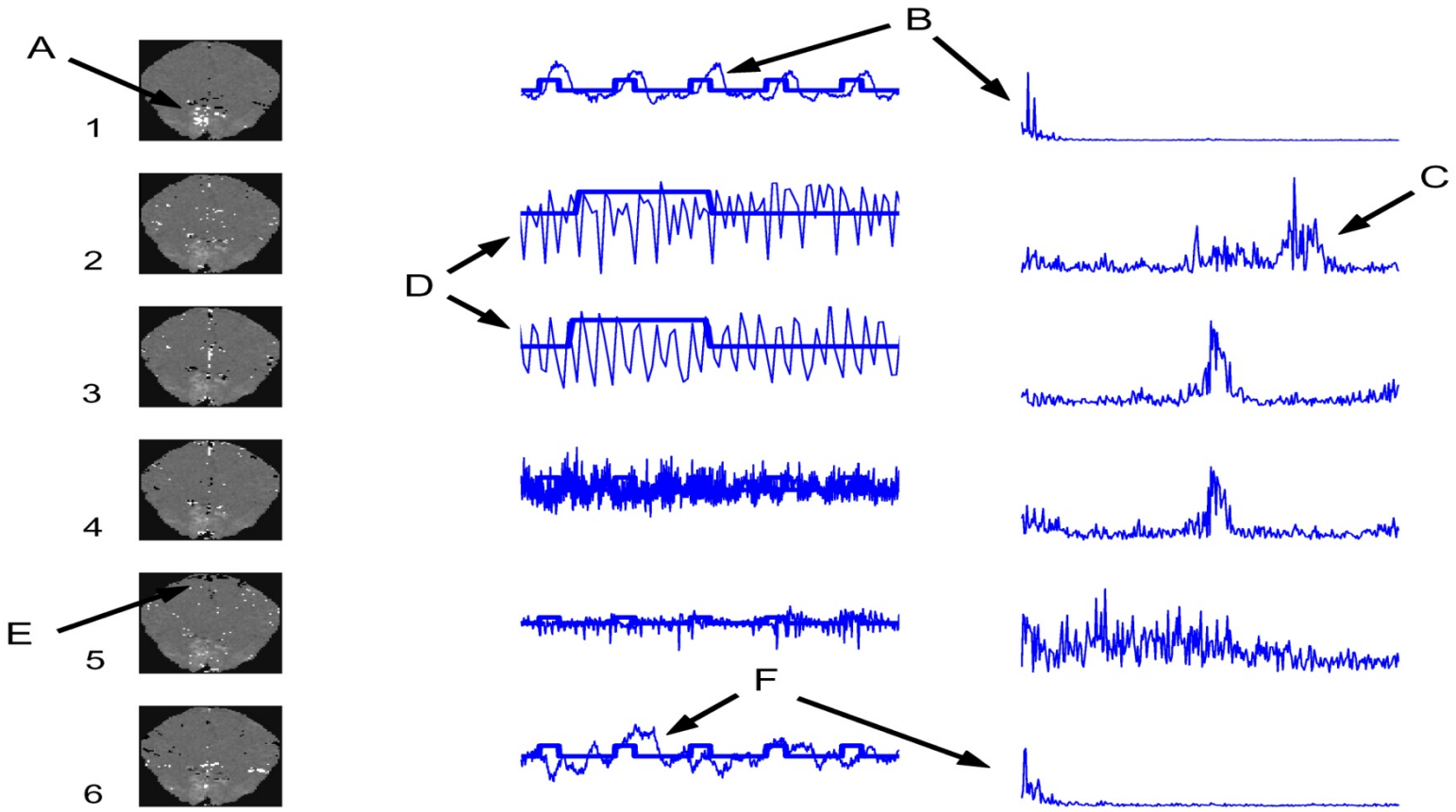


FIG. 5. Data set II. The principal components corresponding to the nine largest covariance eigenvalues for a sequence of 300 fMRI scans of a visually stimulated subject. Stimulation takes places at scan times $\tau = 8$–25 s, $\tau = 42$–59 s, and $\tau = 75$–92 s relative to the start of this three-run sequence. Scan time sampling interval was TR = 0.33 s. The sequences have been smoothed for presentation, reducing noise and high-frequency physiological components. Note that most of the response is captured by the first principal component, showing a strong response to all three periods of stimulation. Using the generalization error estimates in Fig. 4, we find that only the time sequences corresponding to the first three components generalize.

ICA: Assume S(k,t), S(k',t) statistically independent



(McKeown, Hansen, Sejnowski, Curr. Op. in Neurobiology (2003)