DD2434 Machine Learning, Advanced Course

Lecture 10: Non-Gaussian and Discrete Latent Variable Models

Hedvig Kjellström
hedvig@kth.se
https://www.kth.se/social/course/DD2434/

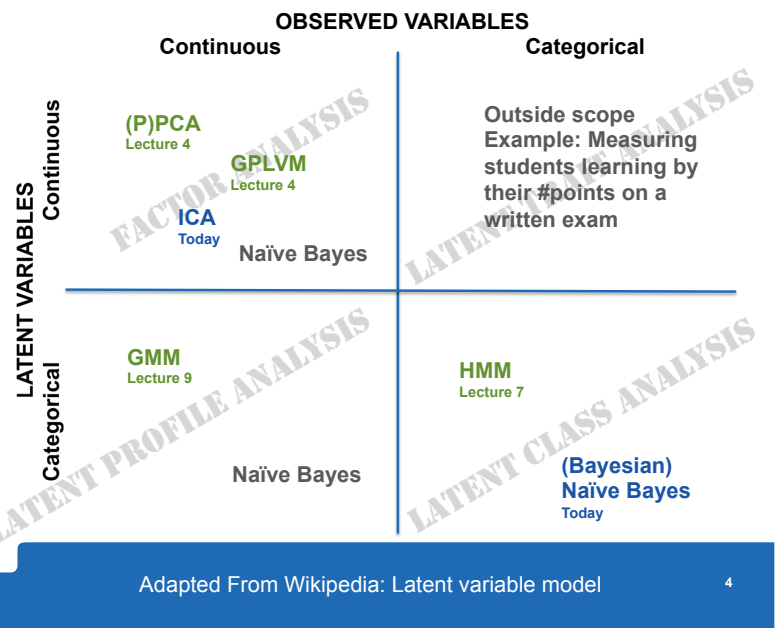
Today

Taxonomy of latent variable models

Independent Component Analysis (Bishop 12.4.1, Hyvärinen & Oja)

Bayesian Naïve Bayes (Bishop 8.2.2)

Taxonomy of Latent Variable Models

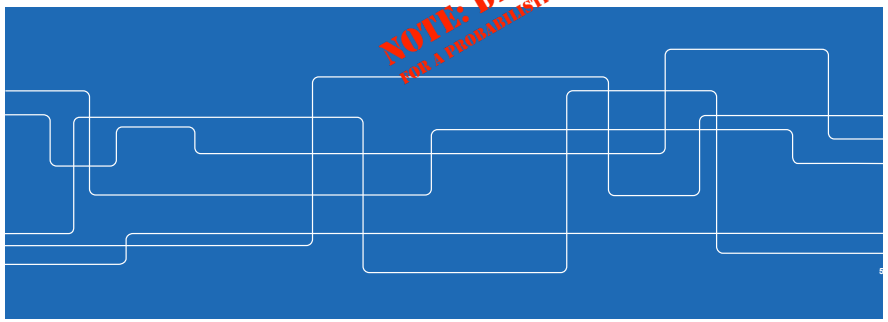


Adapted From Wikipedia: Latent variable model

Independent Component Analysis (ICA)

Bishop Section 12.4.1
Hyvärinen and Oja

NOTE: DETERMINISTIC METHOD
FOR A PROBABILISTIC VERSION, SEE (BECKMANN AND SMITH 2004)



From Lecture 4: Different prior assumptions

PPCA: Assume that latent space is Gaussian distributed

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{m}_z, \Sigma_z)$$

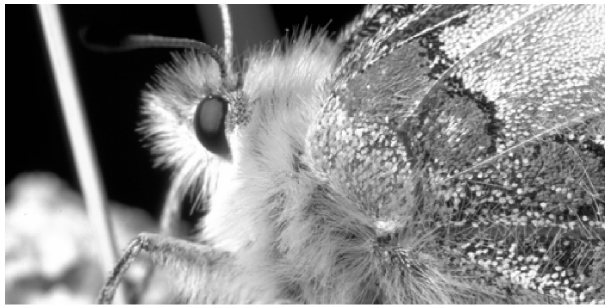
Good idea if we do not know anything about the data

ICA: Assume that all dimensions in the latent space independent of each other

$$p(\mathbf{s}) = \prod_j p(s_j)$$

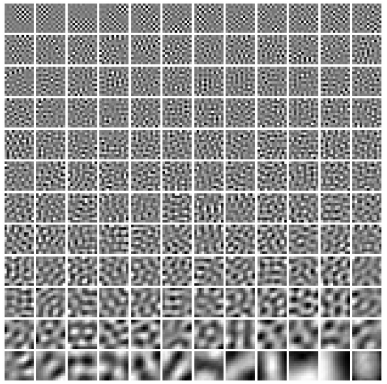
Discuss with your neighbor: What does this assumption mean? Can you come up with any examples of data where this is a good assumption?

Example: Image compression



Slide from Hyvärinen

Learn PCA basis for image patches (eigenvectors in 16x16D space)

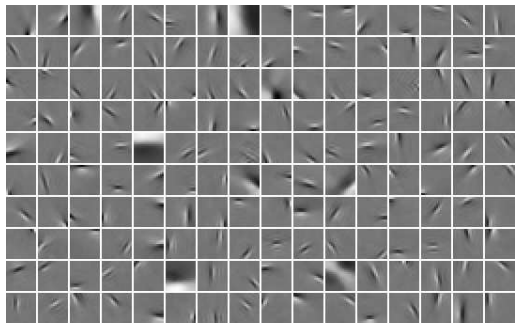


Slide from Hyvärinen

Throw away 90% of the (smallest) eigendimensions

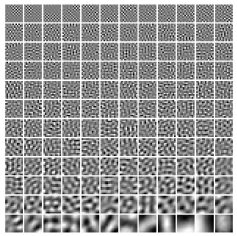
Result: More “fuzzy” image

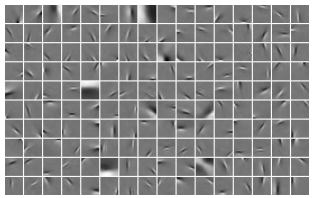
The corresponding ICA basis (independent components in 16x16D space)



Discuss with your neighbor for 5 mins

PCA vs ICA





Which requires the most bases to achieve the same image compression quality? Why?

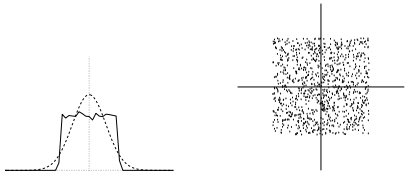
Measure of independence: non-Gaussianity

Basic intuitive principle of ICA estimation.

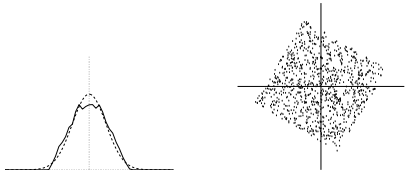
(Sloppy version of) the Central Limit Theorem (Donoho, 1982).

- Consider a linear combination $\mathbf{w}^T\mathbf{x} = \mathbf{q}^T\mathbf{s}$
- $q_i s_i + q_j s_j$ is more gaussian than s_i.
- *Maximizing the nongaussianity* of $\mathbf{q}^T\mathbf{s}$, we can find s_i.
- Also known as projection pursuit.

Illustration: sub-Gaussianity

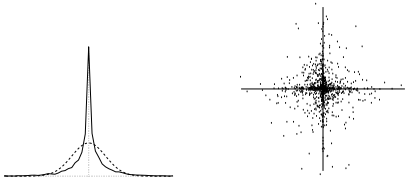


Marginal and joint densities, uniform distributions.

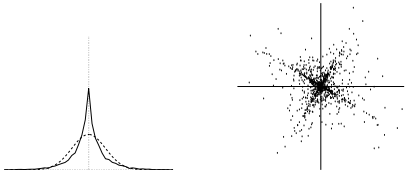


Marginal and joint densities, whitened mixtures of uniform ICs

Illustration: super-Gaussianity



Marginal and joint densities, supergaussian distributions.



Whitened mixtures of supergaussian ICs

How measure non-Gaussianity?

Kurtosis as nongaussianity measure.

- Problem: how to measure nongaussianity?
- Definition:

$$\text{kurt}(x) = E\{x^4\} - 3(E\{x^2\})^2$$

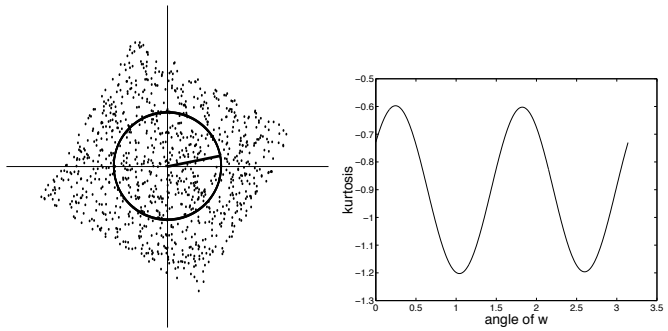
- if variance constrained to unity, essentially 4th moment.
- Simple algebraic properties because it's a cumulant:

$$\text{kurt}(s_1 + s_2) = \text{kurt}(s_1) + \text{kurt}(s_2)$$

$$\text{kurt}(\alpha s_1) = \alpha^4 \,\text{kurt}(s_1)$$

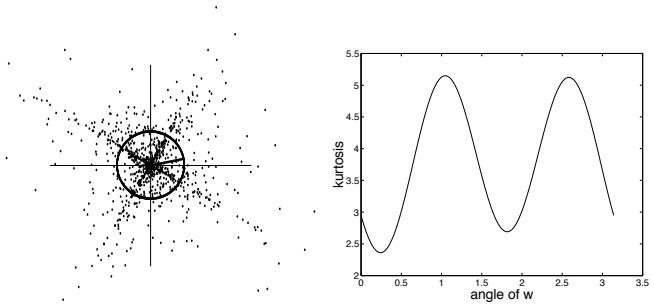
- zero for gaussian RV, non-zero for most nongaussian RV's.
- positive vs. negative kurtosis have typical forms of pdf.

Kurtosis: sub-Gaussianity



Case of negative kurtosis. Kurtosis is minimized, and its absolute value maximized, in the directions of the independent components.

Kurtosis: super-Gaussianity



Kurtosis as a function of the direction of projection. For positive kurtosis, kurtosis (and its absolute value) are maximized in the directions of the independent components.

Basic ICA estimation procedure

1. Whiten the data to give $\mathbf{z}$.
2. Set iteration count $i = 1$.
3. Take a random vector $\mathbf{w}_i$.
4. Maximize nongaussianity of $\mathbf{w}_i^T \mathbf{z}$,
   under constraints $\|\mathbf{w}_i\|^2 = 1$ and $\mathbf{w}_i^T \mathbf{w}_j = 0, j < i$
   (by a suitable algorithm, see later)
5. increment iteration count by 1, go back to 3

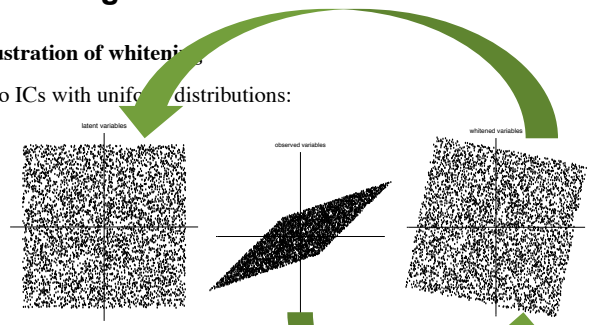
Alternatively: maximize all the $\mathbf{w}_i$ in parallel, keeping them orthogonal.

Whitening = PCA

ICA then finds the most independent basis

Illustration of whiteni

Two ICs with unif distributions:



Original variables, observed mixtures whitened mixtures.

Cf. gaussian density: symmetric in all directions.

PCA whitens the observed space (for a Gaussian distribution this is enough!)

Why kurtosis is not optimal

- Sensitive to outliers:
  Consider a sample of 1000 values with unit var, and one value equal to 10.
  Kurtosis equals at least $10^4/1000 - 3 = 7$.
- For supergaussian variables, statistical performance not optimal even without outliers.
- Other measures of nongaussianity should be considered.

Differential entropy as nongaussianity measure

- Generalization of ordinary discrete Shannon entropy:

$$H(x) = E\{-\log p(x)\}$$

- for fixed variance, maximized by gaussian distribution.
- often normalized to give negentropy

$$J(x) = H(x_{gauss}) - H(x)$$

- Good statistical properties, but computationally difficult.

Overview of ICA estimation principles.

- Most approaches can be interpreted as maximizing the nongaussianity of ICs.
- Basic choice: the nonquadratic function in the nongaussianity measure:
  - kurtosis: fourth power
  - entropy/likelihood: log of density
  - approx of entropy: $G(s) = \log\cosh s$ or others.
- One-by-one estimation vs. estimation of the whole model.
- Estimates constrained to be white vs. no constraint

Algorithms (1). Adaptive gradient methods

- Gradient methods for one-by-one estimation straightforward.
- Stochastic gradient ascent for likelihood (Bell-Sejnowski 1995)

$$\Delta\mathbf{W} \propto (\mathbf{W}^{-1})^T + g(\mathbf{W}\mathbf{x})\mathbf{x}^T$$

  with $g = (\log p_s)'$. Problem: needs matrix inversion!
- Better: natural/relative gradient ascent of likelihood (Amari et al, 1996, Cardoso and Laheld, 1994)

$$\Delta\mathbf{W} \propto [\mathbf{I} + g(\mathbf{y})\mathbf{y}^T]\mathbf{W}$$

  with $\mathbf{y} = \mathbf{W}\mathbf{x}$. Obtained by multiplying gradient by $\mathbf{W}^T\mathbf{W}$.

Algorithms (2). The FastICA fixed-point algorithm (Hyvärinen 1997,1999)

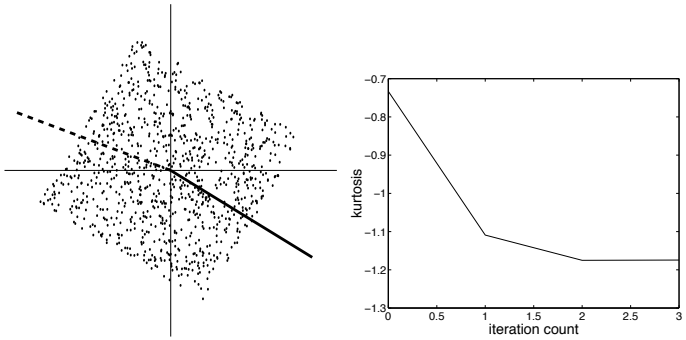
- An approximate Newton method in block (batch) mode.
- No matrix inversion, but still quadratic (or cubic) convergence.
- No parameters to be tuned.
- For a single IC (whitened data)

$$\mathbf{w} \leftarrow E\{\mathbf{x}g(\mathbf{w}^T\mathbf{x})\} - E\{g'(\mathbf{w}^T\mathbf{x})\}\mathbf{w}, \text{ normalize } \mathbf{w}$$

  where g is the derivative of G.
- For likelihood:

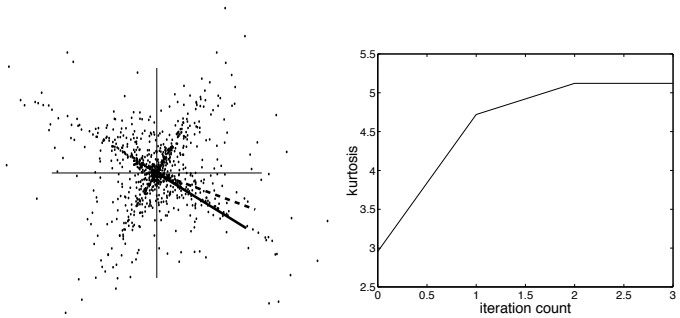
$$\mathbf{W} \leftarrow \mathbf{W} + \mathbf{D}_1[\mathbf{D}_2 + E\{g(\mathbf{y})\mathbf{y}^T\}]\mathbf{W}, \text{ orthonormalize } \mathbf{W}$$

Convergence: sub-Gaussianity



Convergence of FastICA. Vectors after 1 and 2 iterations, values of kurtosis.

Convergence: super-Gaussianity



Convergence of FastICA (2). Vectors after 1 and 2 iterations, values of kurtosis.

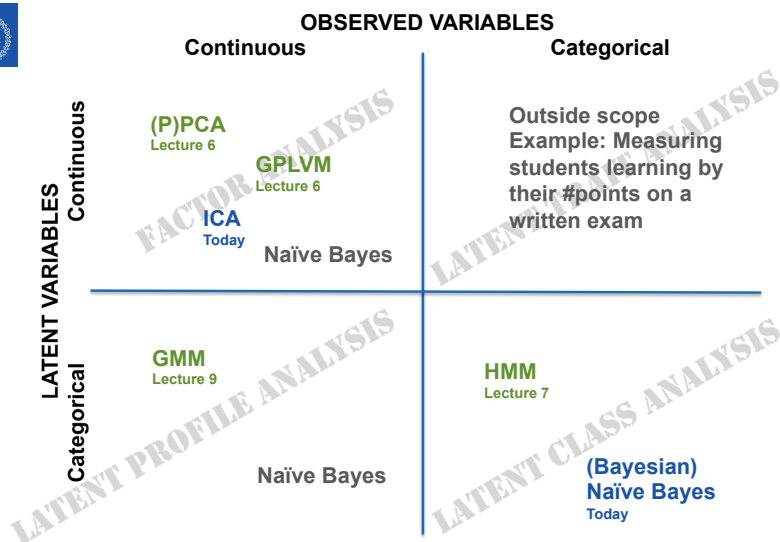
ICA summary

Very simple model: linear non-Gaussian latent variable model

- Gives sparse representations

Estimation not so simple due to non-Gaussianity: objective functions not quadratic like with Gaussians

- Estimation by maximizing non-Gaussianity of independent components.
- Equivalent to maximum likelihood or minimizing mutual information

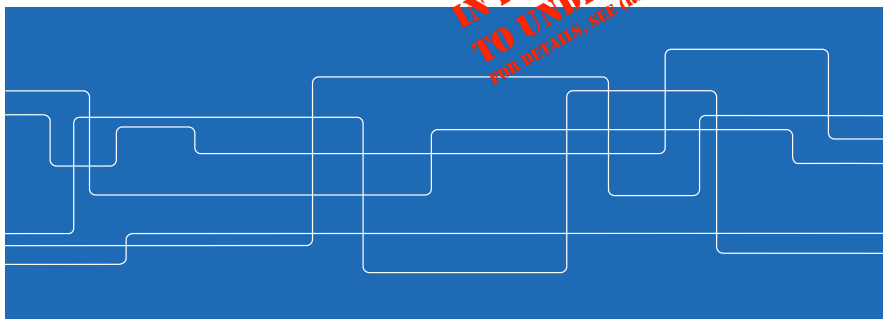


Bayesian Naïve Bayes (aka Dirichlet-Multinomial Classifier)

Bishop Section 8.2.2

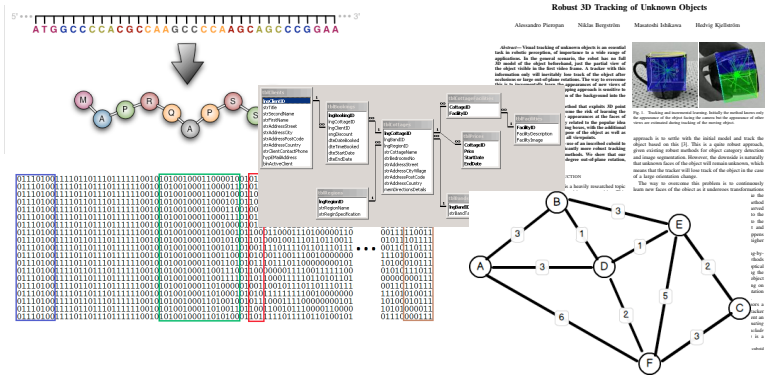
WARNING: NOT A GOOD MODEL IN ITSELF, ONLY AS A PRETEST TO UNDERSTANDING LDA

FOR DEFAULTS, SEE (RENNIE ET AL., ICML 2003)



Latent Variable Models for Discrete Data

Discrete data:

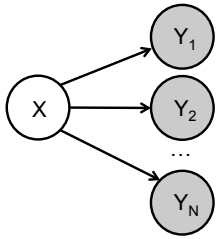


Naïve Bayes model

Recap from the Machine Learning, basic course

y_i conditionally independent given x:

$p(\mathbf{y}|x) = \Pi p(y_i|x)$



Assume y_i independent and identically distributed (i.i.d.)

Then this notation can be written more compactly in…

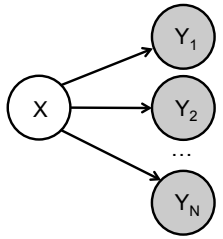
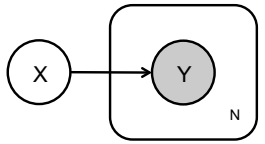


plate notation:



Dice Roll as an Example of Multinomial Distribution

Suppose that we observe $\mathcal{D} = \{x_1, ..., x_N\}$ where $x_i \in \{1, ..., K\}, K = 6$

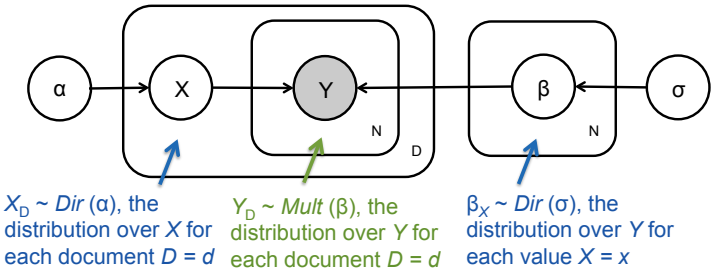
The rolls are independent so the likelihood is

$$p(\mathcal{D}|\theta) = \prod_{k=1}^{K} \theta_k^{N_k}$$

where N_k is the number of times the dice turned up *k*
This is a **Multinomial** distribution.

Making the Naïve Bayes model Bayesian: Adding priors to all variables

Dirichlet distribution is conjugate prior to Multinomial distribution



X_D ~ *Dir* (α), the distribution over *X* for each document *D = d*

Y_D ~ *Mult* (β), the distribution over *Y* for each document *D = d*

β_X ~ *Dir* (σ), the distribution over *Y* for each value *X = x*

Lecture 11: LDA, a more sophisticated version of this Bayesian Naïve Bayes model!

What is next?

Project groups are published on the home page. See if you are listed, otherwise email me before **November 30**! Talk to your project group and select papers before **December 3**.

Continue with Assignment 2, deadline **December 16**.

Next on the schedule
Tue 1 Dec 10:15-12:00 M2
Lecture 2: Bag of Words, Topic Models
Hedvig Kjellström
Readings: Blei and Lafferty

Have a nice weekend! ☺