Factor and Independent Component Analysis

Michael Gutmann

Probabilistic Modelling and Reasoning (INFR11134)
School of Informatics, University of Edinburgh

Spring semester 2018

Recap

- Model-based learning from data
- Observed data as a sample from an unknown data generating distribution
- Learning using parametric statistical models and Bayesian models,
- Their relation to probabilistic graphical models
- Likelihood function, maximum likelihood estimation, and the mechanics of Bayesian inference
- Classical examples to illustrate the concepts

Applications of factor and independent component analysis

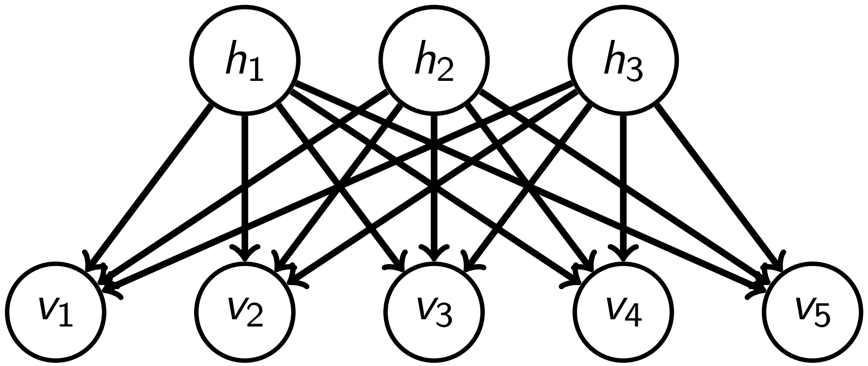
- Factor analysis and independent component analysis are two classical methods for data analysis.
- The origins of factor analysis (FA) are attributed to a 1904 paper by psychologist Charles Spearman. It is used in fields such as
  - Psychology, e.g intelligence research
  - Marketing
  - Wide range of physical and biological sciences

  ⋮
- Independent component analysis (ICA) has mainly been developed in the 90s. It can be used where FA can be used. Popular applications include
  - Neuroscience (brain imaging, spike sorting) and theoretical neuroscience
  - Telecommunications (de-convolution, blind source separation)
  - Finance (finding hidden factors)

  ⋮

Directed graphical model underlying FA and ICA

FA: factor analysis   ICA: independent component analysis



- The visibles $\mathbf{v} = (v_1, \ldots, v_D)$ are independent from each other given the latents $\mathbf{h} = (h_1, \ldots, h_H)$, but generally dependent under the marginal $p(\mathbf{v})$.
- Explains statistical dependencies between (observed) v_i through (unobserved) h_i.
- Different assumptions on $p(\mathbf{v}|\mathbf{h})$ and $p(\mathbf{h})$ lead to different statistical models, and data analysis methods with markedly different properties.

Program

1. Factor analysis

2. Independent component analysis

Program

1. Factor analysis
   - Parametric model
   - Ambiguities in the model (factor rotation problem)
   - Learning the parameters by maximum likelihood estimation
   - Probabilistic principal component analysis as special case
2. Independent component analysis

Parametric model for factor analysis

- In factor analysis (FA), all random variables are Gaussian.
- Importantly, the number of latents H is assumed smaller than the number of visibles D.
- Latents: $p(\mathbf{h}) = \mathcal{N}(\mathbf{h}; \mathbf{0}, \mathbf{I})$ (uncorrelated standard normal)
- Conditional $p(\mathbf{v}|\mathbf{h}; \boldsymbol{\theta})$ is Gaussian

$$p(\mathbf{v}|\mathbf{h}; \boldsymbol{\theta}) = \mathcal{N}(\mathbf{v}; \mathbf{F}\mathbf{h} + \mathbf{c}, \boldsymbol{\Psi})$$

Parameters $\boldsymbol{\theta}$ are

- Vector $\mathbf{c} \in \mathbb{R}^D$: sets the mean of $\mathbf{v}$
- $\mathbf{F} = (\mathbf{f}_1, \ldots \mathbf{f}_H)$: $D \times H$ matrix with $D > H$
  Columns $\mathbf{f}_i$ are called "factors", its elements the "factor loadings".
- $\boldsymbol{\Psi}$: diagonal matrix $\boldsymbol{\Psi} = \text{diag}(\Psi_1, \ldots, \Psi_D)$

Tuning parameter: the number of factors H

Parametric model for factor analysis

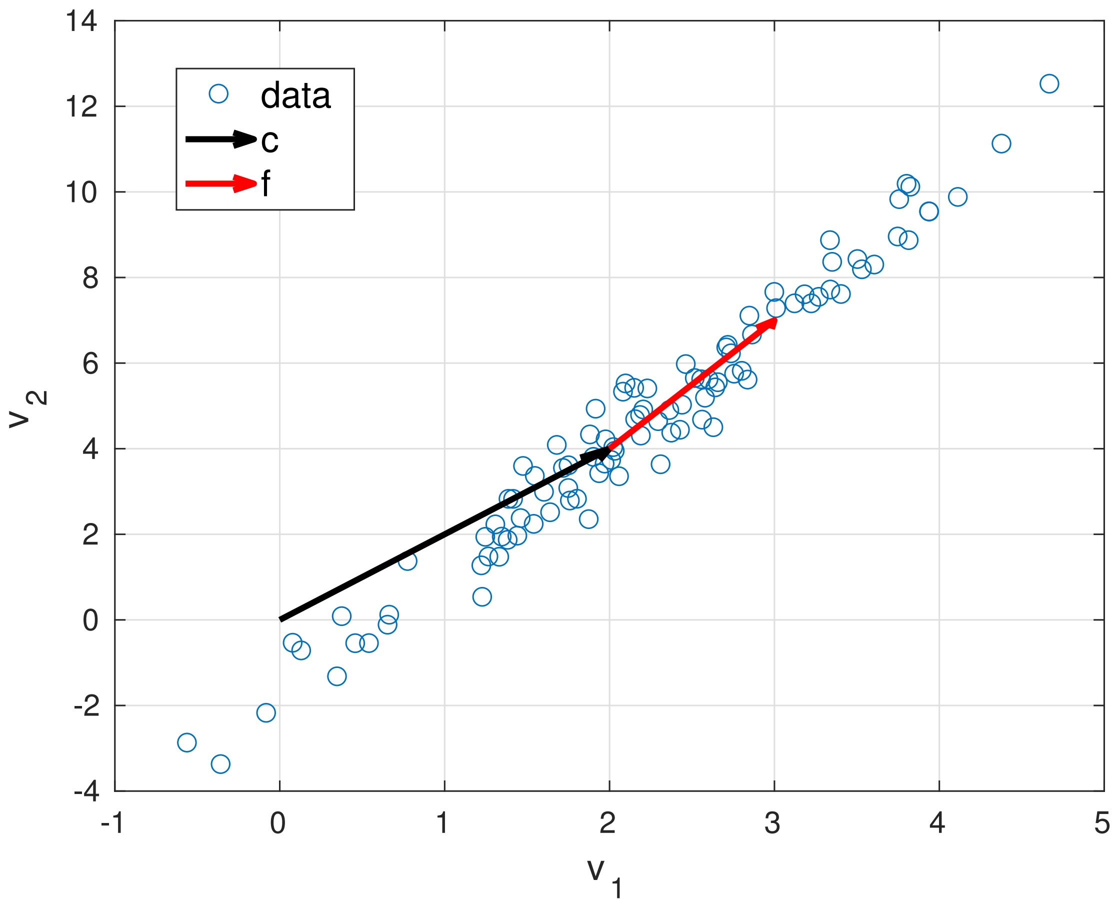
- $p(\mathbf{v}|\mathbf{h};\boldsymbol{\theta}) = \mathcal{N}(\mathbf{v};\mathbf{Fh}+\mathbf{c},\boldsymbol{\Psi})$ is equivalent to

$$\begin{aligned} \mathbf{v} &= \mathbf{Fh} + \mathbf{c} + \boldsymbol{\epsilon} \\ &= \sum_{i=1}^{H} \mathbf{f}_i h_i + \mathbf{c} + \boldsymbol{\epsilon} \qquad\qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}; 0, \boldsymbol{\Psi}) \end{aligned}$$

- Data generation: Add $H < D$ factors weighted by h_i to the constant vector $\mathbf{c}$, and corrupt the "signal" $\mathbf{Fh}+\mathbf{c}$ by additive Gaussian noise.
- $\mathbf{Fh}$ spans a H dimensional subspace of $\mathbb{R}^D$

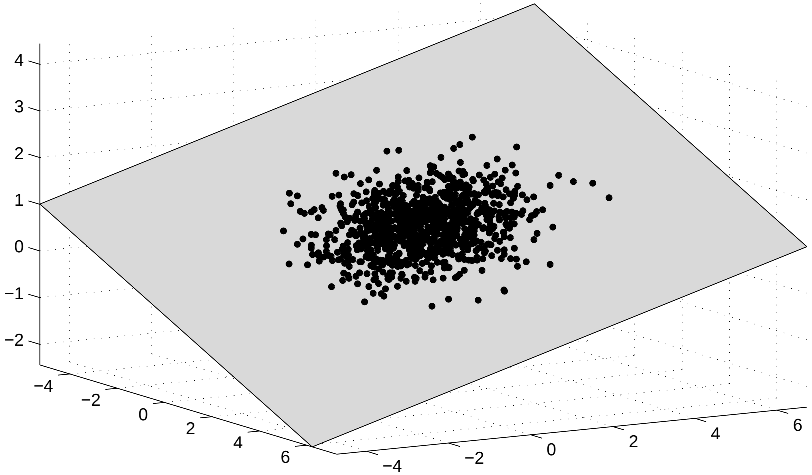
Interesting structure of the data is contained in a subspace

Example for $D = 2, H = 1$.

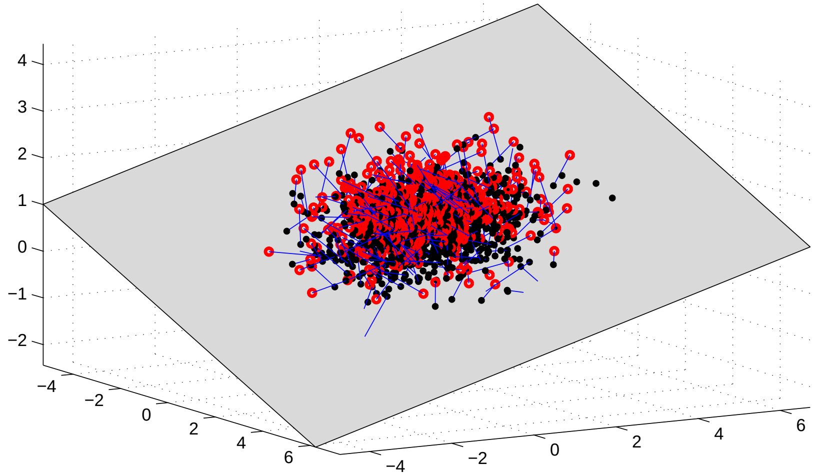


Interesting structure of the data is contained in a subspace

Example for $D = 3, H = 2$ ("pancake" in the 3D space)



Black points: $\mathbf{Fh} + \mathbf{c}$



Red points: $\mathbf{Fh} + \mathbf{c} + \boldsymbol{\epsilon}$

(points below the plane not shown)

(Figures courtesy of David Barber)

Basic results that we need

- If $\mathbf{x}$ has density $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_x, \mathbf{C}_x)$, $\mathbf{z}$ density $\mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_z, \mathbf{C}_z)$, and $\mathbf{x} \perp\!\!\!\perp \mathbf{z}$ then $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{z}$ has density

$$\mathcal{N}(\mathbf{y}; \mathbf{A}\boldsymbol{\mu}_x + \boldsymbol{\mu}_z, \mathbf{A}\mathbf{C}_x\mathbf{A}^\top + \mathbf{C}_z)$$

  (see e.g. Barber Result 8.3)

- An orthonormal (orthogonal) matrix $\mathbf{R}$ is a square matrix for which the transpose $\mathbf{R}^\top$ equals the inverse $\mathbf{R}^{-1}$, i.e.

$$\mathbf{R}^\top = \mathbf{R}^{-1} \qquad \text{or} \qquad \mathbf{R}^\top\mathbf{R} = \mathbf{R}\mathbf{R}^\top = \mathbf{I}$$

  (see e.g. Barber Appendix A.1)

- Orthonormal matrices rotate points.

Factor rotation problem

- Using the basic results, we obtain

$$\begin{aligned} \mathbf{v} &= \mathbf{F}\mathbf{h} + \mathbf{c} + \boldsymbol{\epsilon} \\ &= \mathbf{F}(\mathbf{R}\mathbf{R}^\top)\mathbf{h} + \mathbf{c} + \boldsymbol{\epsilon} \\ &= (\mathbf{F}\mathbf{R})(\mathbf{R}^\top\mathbf{h}) + \mathbf{c} + \boldsymbol{\epsilon} \\ &= (\mathbf{F}\mathbf{R})\tilde{\mathbf{h}} + \mathbf{c} + \boldsymbol{\epsilon} \end{aligned}$$

- Since $p(\mathbf{h}) = \mathcal{N}(\mathbf{h}; \mathbf{0}, \mathbf{I})$ and $\mathbf{R}$ is orthonormal, $p(\tilde{\mathbf{h}}) = \mathcal{N}(\tilde{\mathbf{h}}; \mathbf{0}, \mathbf{I})$, and the two models

$$\mathbf{v} = \mathbf{F}\mathbf{h} + \mathbf{c} + \boldsymbol{\epsilon} \qquad\qquad \mathbf{v} = (\mathbf{F}\mathbf{R})\tilde{\mathbf{h}} + \mathbf{c} + \boldsymbol{\epsilon}$$

produce data with exactly the same distribution.

Factor rotation problem

- Two estimates $\hat{\mathbf{F}}$ and $\hat{\mathbf{F}}\mathbf{R}$ explain the data equally well.
- Estimation of the factor matrix $\mathbf{F}$ is not unique.
- With the Gaussianity assumption on $\mathbf{h}$, there is a rotational ambiguity in the factor analysis model.
- The columns of $\mathbf{F}$ and $\mathbf{FR}$ span the same subspace, so that the FA model is best understood to define a subspace of the data space.
- The individual columns of $\mathbf{F}$ (factors) carry little meaning by themselves.
- There are post-processing methods that choose $\mathbf{R}$ *after* estimation of $\mathbf{F}$ so that the columns of $\mathbf{FR}$ have some desirable properties to aid interpretation, e.g. that they have as many zeros as possible (sparsity).

Likelihood function

- We have seen that the FA model can be written as

$$\mathbf{v} = \mathbf{F}\mathbf{h} + \mathbf{c} + \boldsymbol{\epsilon} \qquad \mathbf{h} \sim \mathcal{N}(\mathbf{h}; \mathbf{0}, \mathbf{I}) \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}; \mathbf{0}, \boldsymbol{\Psi})$$

  with $\boldsymbol{\epsilon} \perp\!\!\!\perp \mathbf{h}$

- From the basic results on multivariate Gaussians: $\mathbf{v}$ is Gaussian with mean and variance equal to

$$\mathbb{E}[\mathbf{v}] = \mathbf{c} \qquad \mathbb{V}[\mathbf{v}] = \mathbf{F}\mathbf{F}^\top + \boldsymbol{\Psi}$$

- Likelihood is given by likelihood for multivariate Gaussian (see Barber Section 21.1)
- But due to the form of the covariance matrix of $\mathbf{v}$, closed form solution is not possible and iterative methods are needed (see Barber Section 21.2, not examinable).

Probabilistic principal component analysis as special case

- In FA, the variances Ψ_i of the additive noise $\boldsymbol{\epsilon}$ can be different for each dimension.
- Probabilistic principal component analysis (PPCA) is obtained for

$$\Psi_i = \sigma^2 \qquad\qquad \mathbf{\Psi} = \sigma^2 \mathbf{I}$$

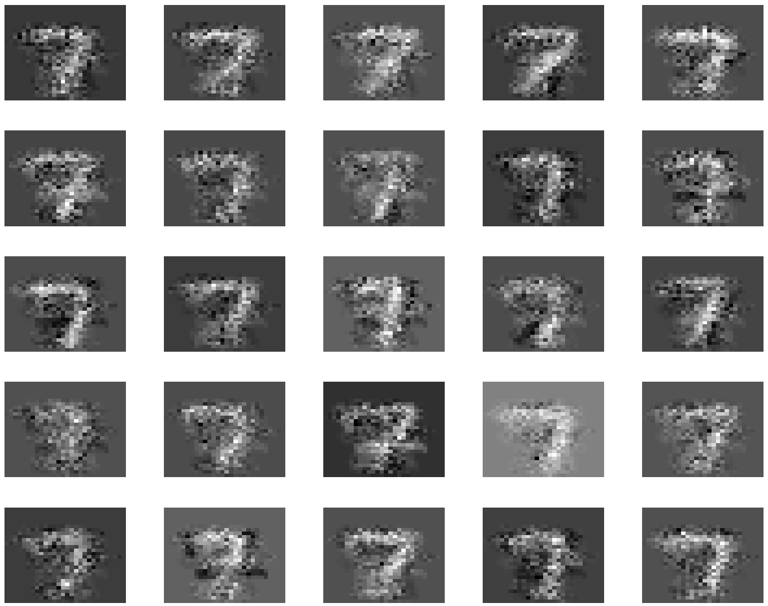
- FA has a richer description of the additive noise than PCA.

Comparison of FA and PPCA (Based on a slide from David Barber)

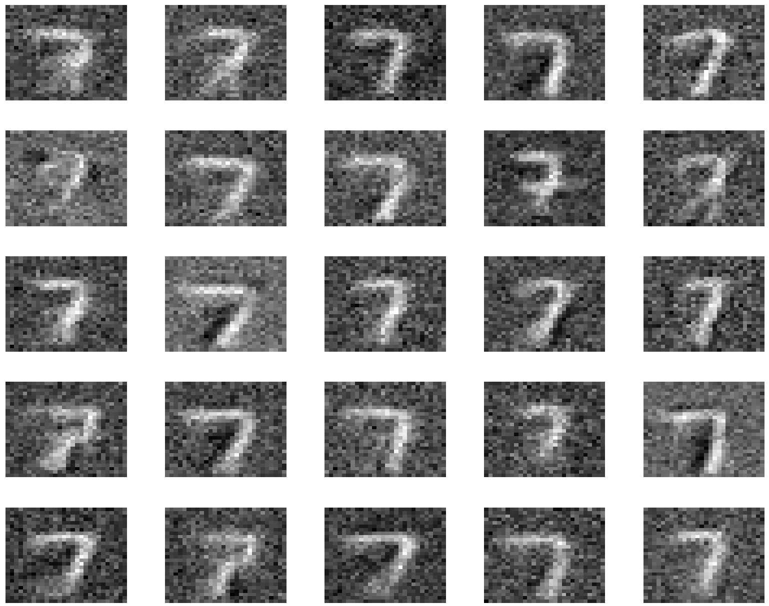
The parameters were estimated from handwritten "7s" for FA and PPCA. After learning, samples can be drawn from the model via

$$\mathbf{v} = \hat{\mathbf{F}}\mathbf{h} + \hat{\mathbf{c}} + \boldsymbol{\epsilon} \qquad \boldsymbol{\epsilon} \sim \begin{cases} \mathcal{N}(\boldsymbol{\epsilon}; \mathbf{0}; \hat{\boldsymbol{\Psi}}) & \text{for FA} \\ \mathcal{N}(\boldsymbol{\epsilon}; \mathbf{0}; \hat{\sigma}^2\mathbf{I}) & \text{for PPCA} \end{cases}$$

Figures below show samples. Note how the noise variance for FA depends on the pixel, being zero for pixels on the boundary of the image.



(a) Factor Analysis



(b) PPCA

Program

1. Factor analysis
   - Parametric model
   - Ambiguities in the model (factor rotation problem)
   - Learning the parameters by maximum likelihood estimation
   - Probabilistic principal component analysis as special case
2. Independent component analysis

Program

1. Factor analysis

2. Independent component analysis
    - Parametric model
    - Ambiguities in the model
    - sub-Gaussian and super-Gaussian pdfs
    - Learning the parameters by maximum likelihood estimation

Parametric model for independent component analysis

- In ICA, unlike in FA, the latents are assumed to be non-Gaussian. (one latent can be assumed to be Gaussian)
- The latents h_i are assumed to be statistically independent

$$p_{\mathbf{h}}(\mathbf{h}) = \prod_i p_h(h_i)$$

- Conditional $p(\mathbf{v}|\mathbf{h};\boldsymbol{\theta})$ is generally Gaussian

$$p(\mathbf{v}|\mathbf{h};\boldsymbol{\theta}) = \mathcal{N}(\mathbf{v};\mathbf{F}\mathbf{h}+\mathbf{c},\boldsymbol{\Psi}) \qquad \text{or} \qquad \mathbf{v} = \mathbf{F}\mathbf{h}+\mathbf{c}+\boldsymbol{\epsilon}$$

  Called "noisy" ICA
- The number of latents H can be larger than D ("overcomplete" case) or smaller ("undercomplete" case).
- We here consider the widely used special case where the noise is zero and $H = D$.

Parametric model for independent component analysis

In ICA, the matrix $\mathbf{F}$ is typically denoted by $\mathbf{A}$ and called the "mixing" matrix. The model is

$$\mathbf{v} = \mathbf{A}\mathbf{h} \qquad\qquad p_{\mathbf{h}}(\mathbf{h}) = \prod_{i=1}^{D} p_h(h_i)$$

where the h_i are typically assumed to have zero mean and unit variance.

Ambiguities

- Denote the columns of $\mathbf{A}$ by $\mathbf{a}_i$.
- From

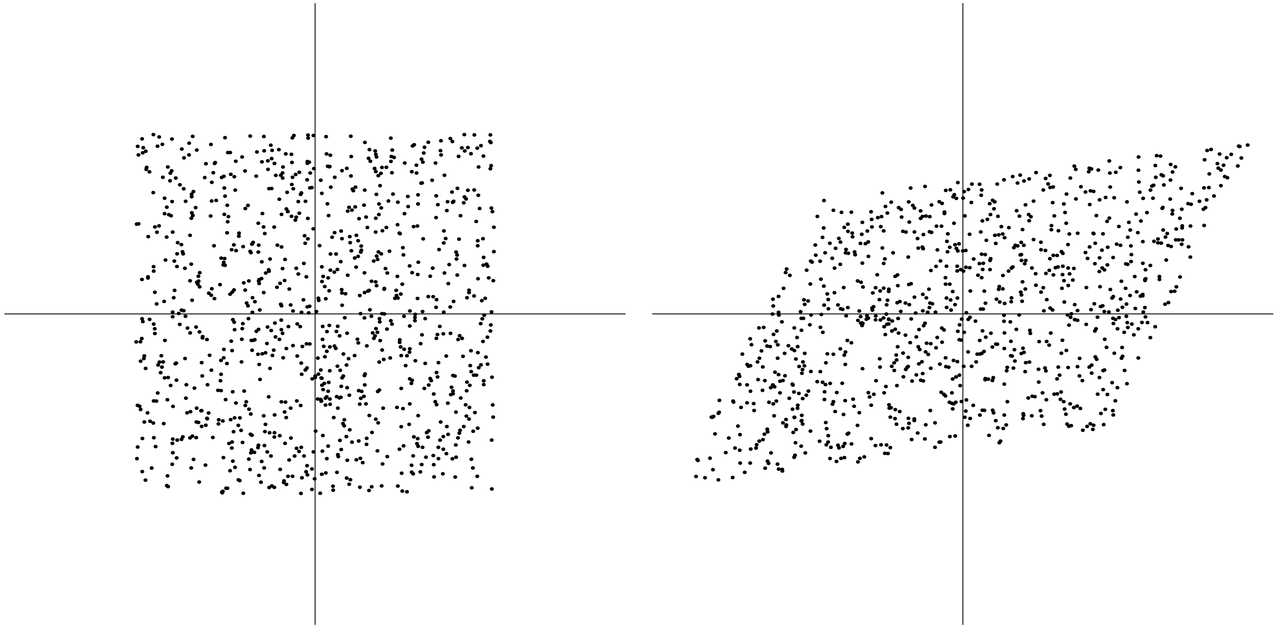
$$\mathbf{v} = \mathbf{A}\mathbf{h} = \sum_{i=1}^{D} \mathbf{a}_i h_i = \sum_{k=1}^{D} \mathbf{a}_{i_k} h_{i_k} = \sum_{i=1}^{D} (\mathbf{a}_i \alpha_i) \frac{1}{\alpha_i} h_i$$

  it follows that the ICA model has an ambiguity regarding the ordering of the columns of $\mathbf{A}$ and their scaling.
- The unit variance assumption on the latents fixes the scaling but not the ordering ambiguity.
- Note: for non-Gaussian latents, there is no rotational ambiguity.

Non-Gaussian latents: variables with sub-Gaussian pdfs

- Sub-Gaussian pdf: (assume variables have mean zero) pdf that is less peaked at zero than a Gaussian of the same variance.
- Example: pdf of a uniform distribution

Samples (h_1, h_2) Samples (v_1, v_2)



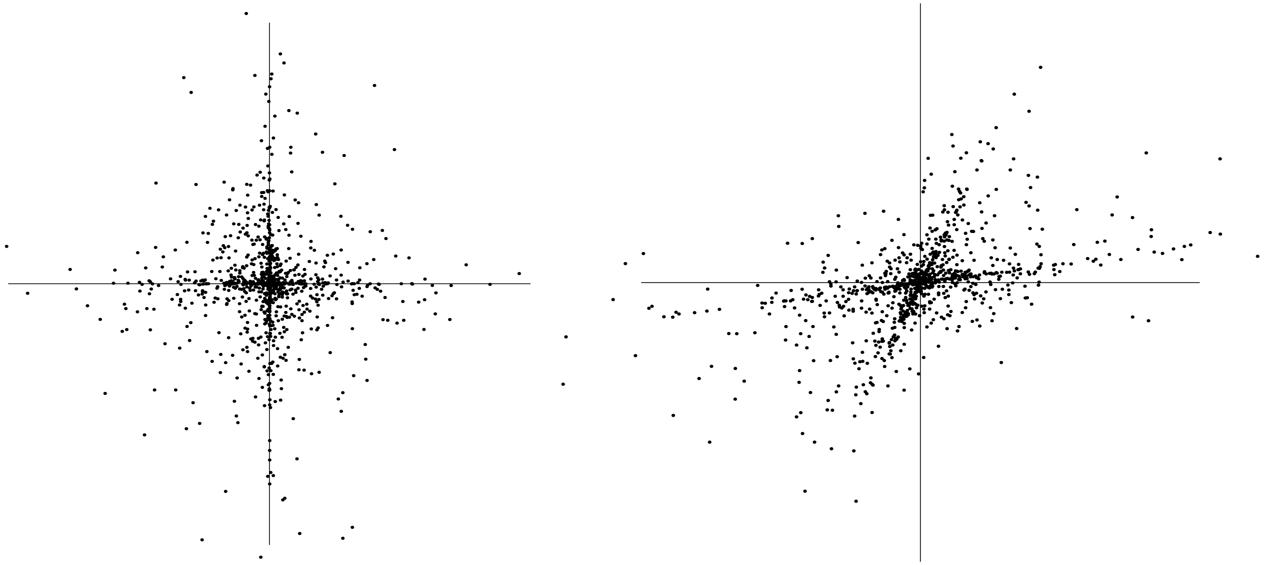
Horizontal axes: h_1 and v_1. Vertical axes h_2 and v_2. Not in the same scale

(Figures 7.5 and 7.6 from *Independent Component Analysis* by Hyvärinen, Karhunen, and Oja).

Non-Gaussian latents: variables with super-Gaussian pdfs

- Super-Gaussian pdf: (assume variables have mean zero) pdf that is more peaked at zero than a Gaussian of the same variance.
- Example: pdf of a Laplace distribution (see Def 8.24 in Barber)

Samples (h_1, h_2) Samples (v_1, v_2)



Horizontal axes: h_1 and v_1. Vertical axes h_2 and v_2. Not in the same scale

(Figures 7.8 and 7.9 from *Independent Component Analysis* by Hyvärinen, Karhunen, and Oja).

Distribution of the visibles

- The mapping $\mathbf{h} \mapsto \mathbf{v} = \mathbf{A}\mathbf{h}$ is deterministic and invertible. By the laws of transformation of random variables
$$p(\mathbf{v}; \mathbf{A}) = p_{\mathbf{h}}(\mathbf{A}^{-1}\mathbf{v})|\det \mathbf{A}^{-1}|$$
(see e.g. Barber Result 8.1)
- Denote the inverse of $\mathbf{A}$ by $\mathbf{B}$
$$\mathbf{A}^{-1}\mathbf{v} = \mathbf{B}\mathbf{v} = \begin{pmatrix} \mathbf{b}_1\mathbf{v} \\ \vdots \\ \mathbf{b}_D\mathbf{v} \end{pmatrix}$$
where the $\mathbf{b}_1, \ldots, \mathbf{b}_D$ are the *row* vectors of the matrix $\mathbf{B}$.
- Given the independence of the latents, we thus have
$$\begin{aligned} p(\mathbf{v}; \mathbf{A}) &= p_{\mathbf{h}}(\mathbf{A}^{-1}\mathbf{v})|\det \mathbf{A}^{-1}| = p_{\mathbf{h}}(\mathbf{B}\mathbf{v})|\det \mathbf{B}| \\ &= \left[\prod_{j=1}^{D} p_h(\mathbf{b}_j\mathbf{v})\right]|\det \mathbf{B}| \end{aligned}$$

Likelihood function

- Since the mapping from $\mathbf{A}$ to $\mathbf{B}$ is invertible. We can write the likelihood function in terms of the matrix $\mathbf{B}$,
- Given iid data $\mathcal{D} = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$, we obtain

$$L(\mathbf{B}) = \prod_{i=1}^{n} \left[\prod_{j=1}^{D} p_h(\mathbf{b}_j \mathbf{v}_i) \right] |\det \mathbf{B}|$$

- The log-likelihood is given by

$$\ell(\mathbf{B}) = \sum_{i=1}^{n} \sum_{j=1}^{D} \log p_h(\mathbf{b}_j \mathbf{v}_i) + n \log |\det \mathbf{B}|$$

- Can be optimised using gradient ascent (slow) or with more powerful methods (see Barber 21.6)

The likelihood and the distribution of the latents

$\ell(\mathbf{B}) = \sum_{i=1}^{n} \sum_{j=1}^{D} \log p_h(\mathbf{b}_j \mathbf{v}_i) + n \log |\det \mathbf{B}|$

- $\mathbf{B}$ and hence the mixing $\mathbf{A}$ can be uniquely estimated, up to the scaling and order ambiguity, as long as the p_h are non-Gaussian (see Barber 21.6) (one latent Gaussian is allowed).
- Non-Gaussianity assumption on the latents solves the "factor rotation" problem in FA.
- The pdf p_h of the latents enter the (log) likelihood.
- If not known, they have to be estimated, which is difficult.
- It turns out that learning whether p_h is super-Gaussian or sub-Gaussian is enough. (not examinable, Section 9.1.2 of *Independent Component Analysis* by Hyvärinen, Karhunen, and Oja)

Program recap

1. Factor analysis
   - Parametric model
   - Ambiguities in the model (factor rotation problem)
   - Learning the parameters by maximum likelihood estimation
   - Probabilistic principal component analysis as special case

2. Independent component analysis
   - Parametric model
   - Ambiguities in the model
   - sub-Gaussian and super-Gaussian pdfs
   - Learning the parameters by maximum likelihood estimation