Linear Factor Models

Sargur N. Srihari
srihari@cedar.buffalo.edu

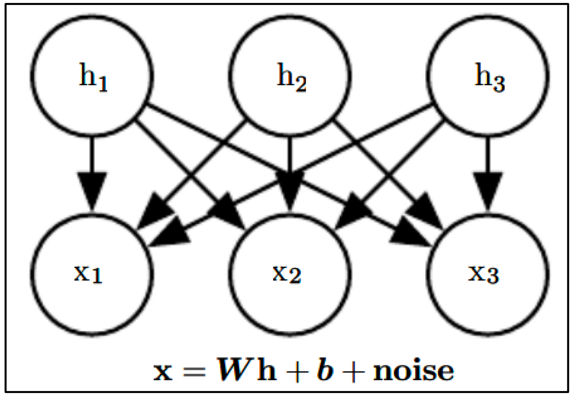
Topics in Linear Factor Models

- Linear factor model definition

1. Probabilistic PCA and Factor Analysis
2. Independent Component Analysis (ICA)
3. Slow Feature Analysis
4. Sparse Coding
5. Manifold Interpretation of PCA

Probabilistic Model with Latent Variables

- A linear factor model describes a data generating process for $\boldsymbol{x}$ that includes latent variables $\boldsymbol{h}$, where $\boldsymbol{x}$ is a linear function of $\boldsymbol{h}$
- A linear factor model:



$\boldsymbol{h} \sim p(\boldsymbol{h})$ with $\boxed{p(\boldsymbol{h}) = \prod_i p(h_i)}$

The noise is Gaussian and diagonal (independent across dimensions)

- Different models such as probabilistic PCA, factor analysis or ICA make different choices about the form of *noise* and *prior* $p(\boldsymbol{h})$

Factor Analysis

- The latent variable prior is a unit variance Gaussian $\boldsymbol{h} \sim N(\boldsymbol{h};0,\mathrm{I})$
  - Observed variables x_i are conditionally independent given $\boldsymbol{h}$
  - Noise is drawn from $\psi=\mathrm{diag}\ (\sigma^2)$
    - with $\sigma^2=[\sigma_1{}^2,.. \sigma_n{}^2]$
  - It can be shown that $\boldsymbol{x}$ is just a multivariate normal random variable $\boldsymbol{x} \sim N(\boldsymbol{x};\boldsymbol{b},\mathrm{WW}^{\mathrm{T}}+\psi)$

Probabilistic PCA

- A slightly modified factor analysis model
- Assume equal conditional variances: $\sigma^2 = \sigma_1{}^2 = .. = \sigma_n{}^2$
- Thus $\boldsymbol{x} \sim N(\boldsymbol{x};\ \boldsymbol{b}, \mathrm{WW}^{\mathrm{T}} + \sigma^2 \mathrm{I})$
  - Or equivalently $\boldsymbol{x} = \mathrm{W}\boldsymbol{h} + \boldsymbol{b} + \sigma \boldsymbol{z}$

    where $\boldsymbol{z} \sim \boldsymbol{N}(\boldsymbol{z};\ \mathbf{0}, \mathbf{I})$ is Gaussian noise
  - Iterative EM can be used to estimate W and σ^2
  - Most variations are captured by the latent variables $\boldsymbol{h}$, upto some small residual reconstruction error σ^2

Probabilistic PCA becomes PCA as $\sigma \rightarrow 0$

Independent Component Analysis

- Among oldest representation algorithms
- Approach seeks to separate an observed signal into many underlying signals that are scaled and added together to form the observed data
  - Signals are intended to be fully independent rather than merely decorrelated from each other
    - Independence is stronger than zero covariance
      - Ex: We sample $\boldsymbol{x}$ from $[-1,1]$. We choose s to be 1 with probability 0.5, otherwise $s=0$. We generate random variable y by assigning $y=sx$. Clearly x and y are not independent, since y is generated from x. But x and y have zero covariance.

An ICA model

- Prior over underlying factors $p(\boldsymbol{h})$ fixed ahead of time
- Model deterministically generates $\boldsymbol{x} = W\boldsymbol{h}$
  - Use nonlinear change of variables to determine $p(\boldsymbol{x})$

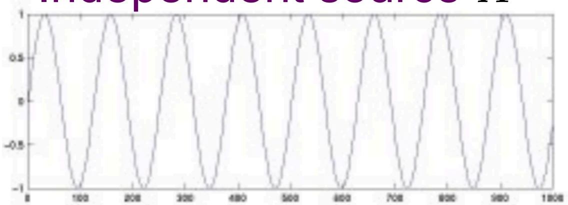
$$\boxed{p_x(x) = p_y(g(x)) \left| \frac{\partial g(x)}{\partial x} \right|}$$

- Learning proceeds using maximum likelihood
- By choosing independent $p(\boldsymbol{h})$ we can recover underlying factors that are close to independent
  - Used to recover low level signals that are mixed together

ICA signal separation

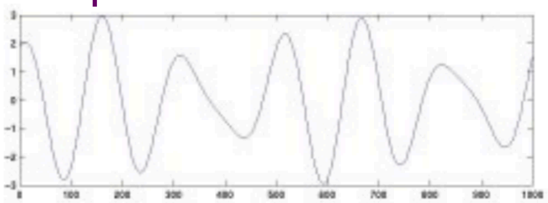
- Each example is one moment in time
- Each x_i is a sensor observation of mixed signals
- Each h_i is one estimate of the original signals

Independent source A

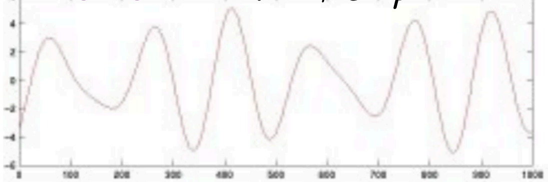


Independent source B

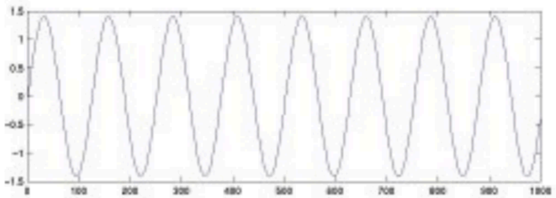
Top: A -$2B$

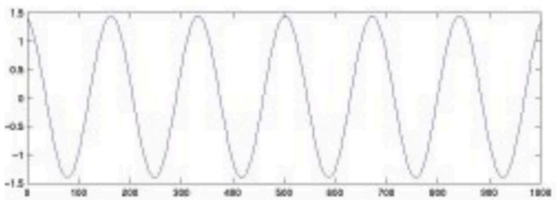


Bottom:$1.7A+3.4B$



Recovered signals





Choice of $p(\boldsymbol{h})$ in ICA

- All ICA variants require $p(\boldsymbol{h})$ be non-Gaussian
  - This is because if $p(\boldsymbol{h})$ is an independent prior with Gaussian components then W is not identifiable
- This is different from probabilistic PCA and factor analysis, where $p(\boldsymbol{h})$ is Gaussian
- Typical choice is $p(h_i)=[d/dh_i]\sigma(h_i)$
  - Have larger peaks near 0 than does Gaussian
    - So ICA is learning sparse features

Generalization of ICA

- Just as PCA can be generalized to nonlinear autoencoders
- ICA can be generalized to a nonlinear generative model
  - In which we use a nonlinear function f to generate observed data

Slow Feature Analysis

- Linear factor model that uses information from time signals to learn invariant features
- Motivation: important features change slowly
- Ex: running zebra:
  - overall location: doesn't change,
  - position changes slowly, stripes change quickly.
- Performed by adding a term to the loss function

$$\lambda \sum_t L(f(\boldsymbol{x}^{(t+1)}), f(\boldsymbol{x}^{(t)}))$$

  - where f is feature extractor to be regularized,
  - λ is the strength of the regularization term,
  - L is a loss function, e.g., mean squared difference