Investigations of Gaussian Products-of-Experts Models

Felix Agakov
MSc Artificial Intelligence

Division of Informatics
The University of Edinburgh

September 2000

The Presentation:

1. Introduction
   (a) Latent Variable Modeling
   (b) Mixture Models
   (c) Products of Experts (PoE)
2. Results
   (a) Constrained Gaussians (hyper-disks)
   (b) Probabilistic Principal Component Analyzers (hyper-ellipsoids)
   (c) Contrastive Divergence Learning (simple and efficient learning rule for Gaussian models)
3. Applications

1.1. Latent Variable Modeling: Intro

Data:

- observable (visible)
- unobservable, missing, unknown (**latent**)

Important question: how to explain [classify] the observations?

Examples:

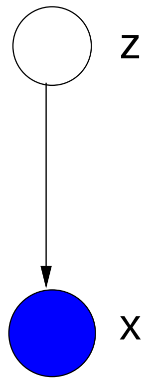
- given a visible symptom, how to explain the hidden disease?
- given a process, how to determine its state?

The answer is not always evident, because there are entities that cannot be observed

[but very often cause the observations].

Latent Variable Modeling: Tasks

To explain the data one can apply probabilistic inference



1. Direct approach:

   If $\mathbf{z}$ is hidden, $\mathbf{x}$ is visible, what are

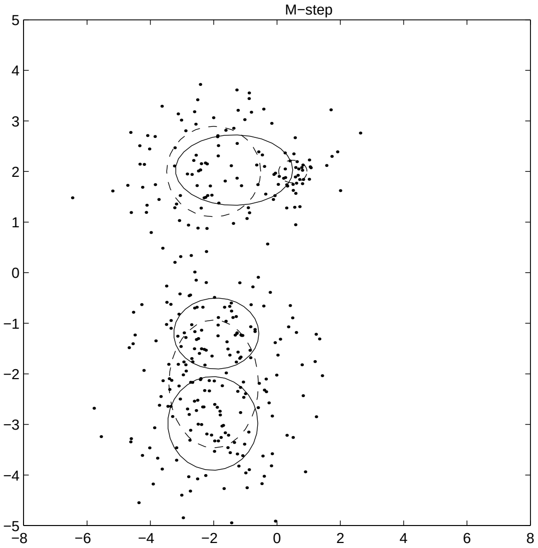
   - $P(\mathbf{z}|\mathbf{x})$: - classification
   - $P(\mathbf{z}_\tau|\mathbf{x}_t), \tau > t$: - prediction

2. Generative approach:

   We can hope to understand the underlying structure of the data better by **reconstructing** it (modeling the data-generating distribution).

   NB: here we model the data to explain it

1.2. Mixture models: Intro



How to estimate a non-trivial density in $\mathcal{R}^d$?

- assign high probabilities to the observed data
- assign low probabilities to the irrelevant points
- may want to represent a complicated probability density as a combination of simpler **expert** probabilities

Mixture models approximate a non-trivial probability density as a **weigted summation** of simpler probabilities:

$$P(\mathbf{x}) = \sum_{i=1}^{m} P(\mathbf{x}|z_i)P(z_i)$$

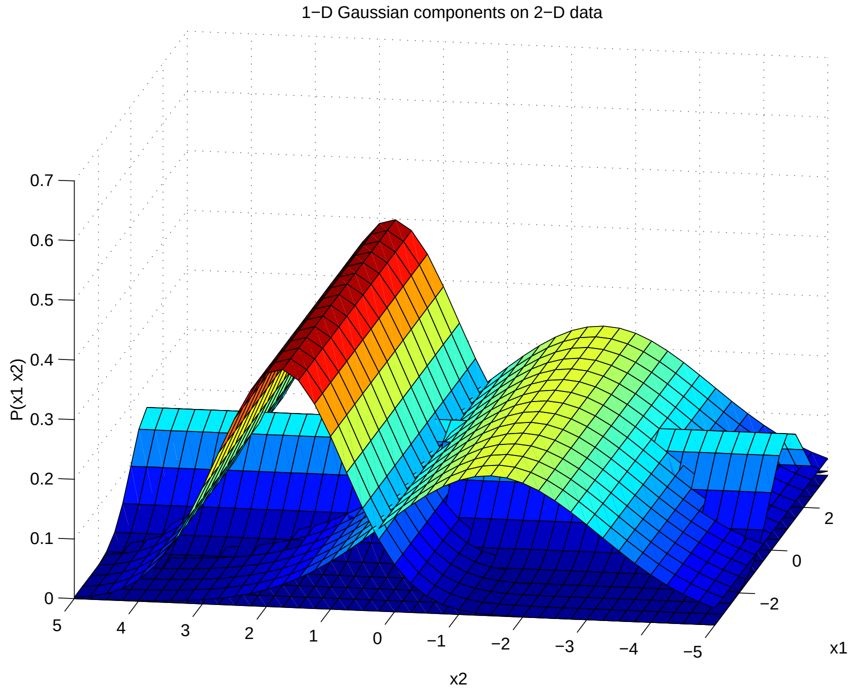
Mixture Models: Problems

- Parameters of mixture components can converge to local optima
- Experts assign high probabilities to irrelevant regions of the data space

Therefore,

- Mixture models are vague, especially in the higher dimensions

Not very good for fitting local patterns of the data



1.3. PoE: an alternative

Geoffrey Hinton, 1999

Products of Experts (PoE) models approximate a non-trivial probability density as a **normalized multiplication** of simpler probabilities:

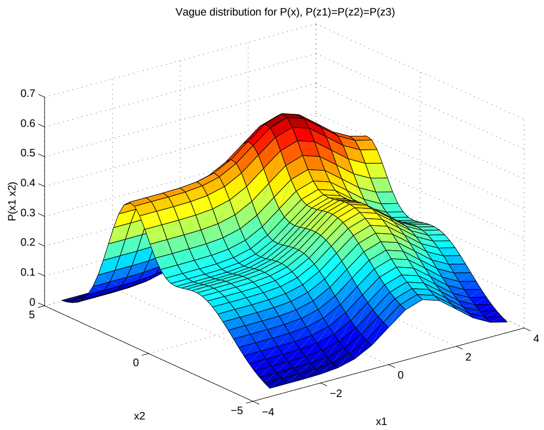
$$P(\mathbf{x}) = \frac{\prod_{i=1}^{m} P_i(\mathbf{x}|\mathbf{\Theta_i})}{\int_{\acute{\mathbf{x}}} \prod_{i=1}^{m} P_i(\acute{\mathbf{x}}|\mathbf{\Theta_i})\mathrm{d}\acute{\mathbf{x}}}$$

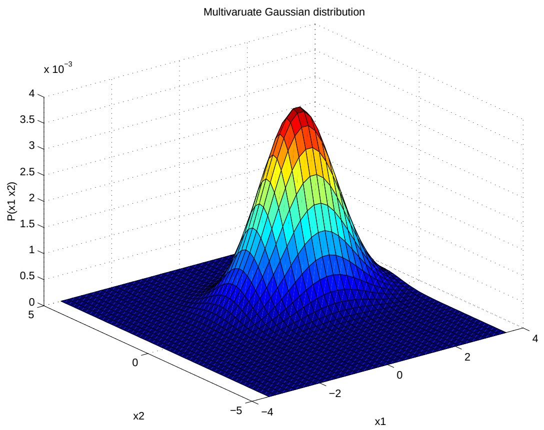
$P_i(\mathbf{x}|\mathbf{\Theta_i})$ - probability of $\mathbf{x}$ under the model with parameters $\mathbf{\Theta_i}$.

- Yields much sharper distribution

- Each component may assign high probability to irrelevant data, as long as this probability is small under other components

Product of Experts: Mixtures of PoEs





- Mixture models - weighted summation - OR

- PoE models - normalized multiplication - AND

- Mixtures of PoE models - DNF (disjunction of conjunctions)

Hope to get a better approximation of non-trivial densities

2. Results

Principal goal: Investigation of Products of Gaussian experts

1. Two new models:

   - Product of constrained Gaussians (multiplication of "hyper-disks");

   - Product of probabilistic principal component analyzers (multiplication of elongated hyper-ellipsoids)

     Challenges: Complicated structure of the covariance in the visible space, highly complicated analytic results.

2. Analytically confirmed correctness of the contrastive divergence learning rule (Hinton 2000) - simple, computationally efficient iteraive learning

Results: Principal Advantages

- Probabilistic analogues of non-probabilistic techniques (e.g. principal and minor component analysis);

- Definition of likelihood for model comparison;

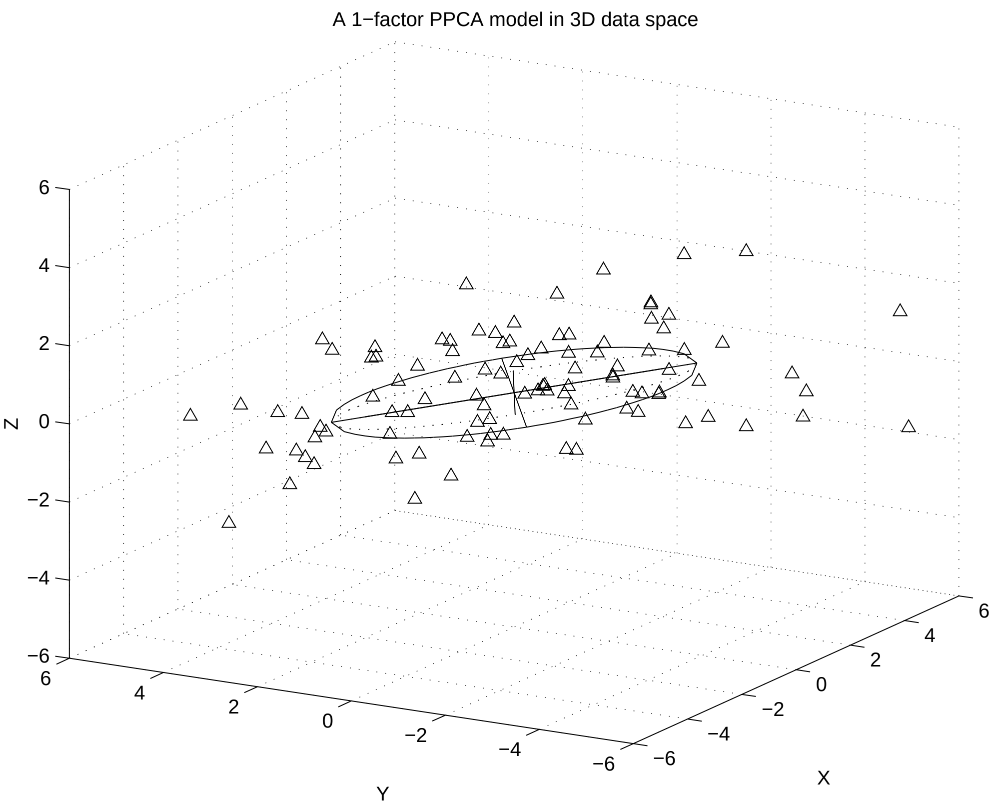
- Incorporation of the prior knowledge of the process parameters;

- Incorporation of models into mixtures for modeling more complicated data sets;

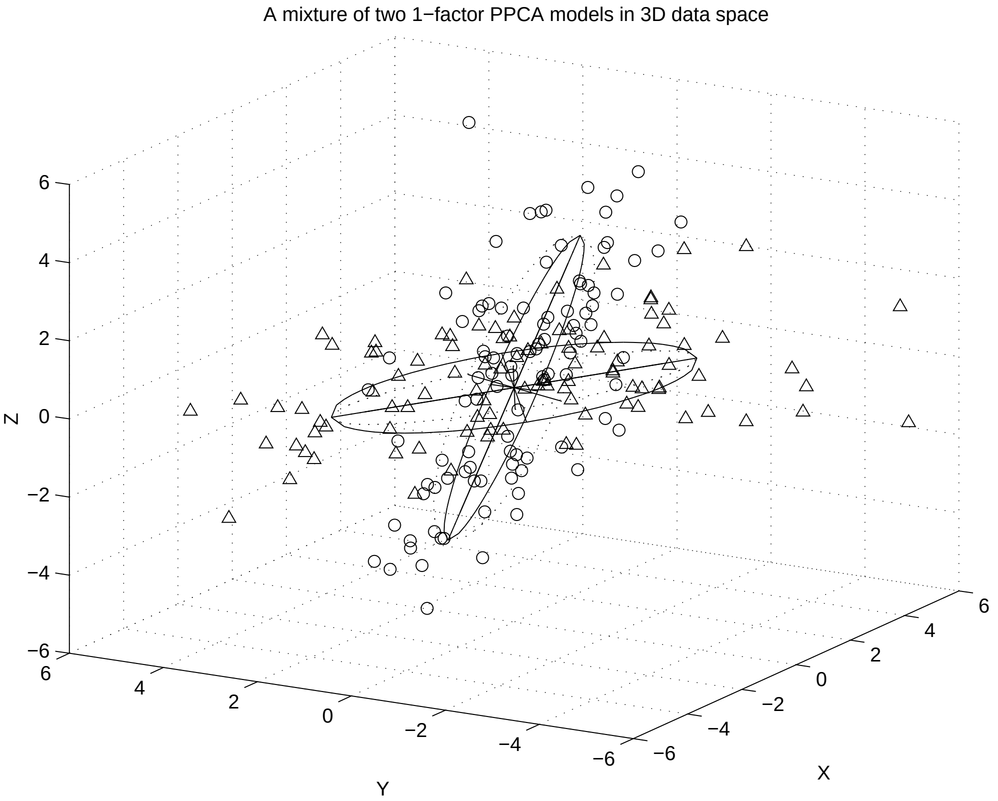
- Decrease in the number of parameters to estimate;

- Efficient and fast learning

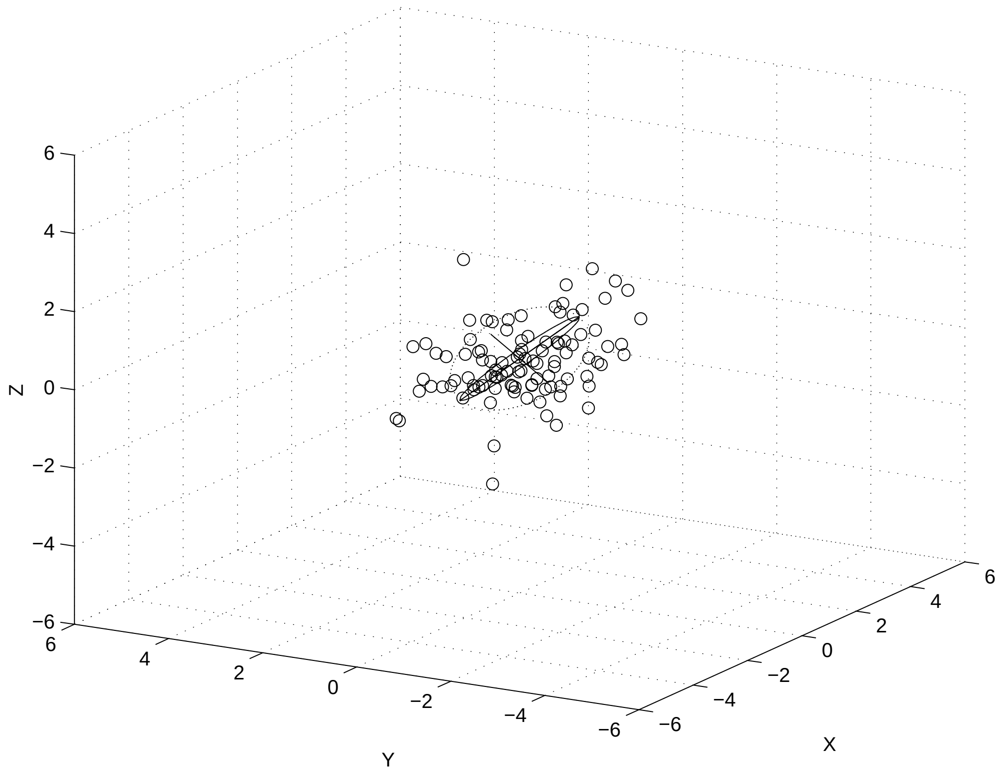
A 1-factor PPCA model in $\mathcal{R}^3$



A mixture of PPCAs in $\mathcal{R}^3$



A product of PPCAs in $\mathcal{R}^3$



3. Applications

As a **density estimation** technique, PoE and mixtures of PoE can be used for:

- **classification**
- **time series forecasting**
- dimensionality reduction
- data visualization

Some of the practical applications include:

- diagnostic tasks
- signal processing
- motion detection
- image recognition
- speech recognition

Hope: to create better models for solving **classification** and **prediction** tasks.