Investigations of Gaussian Products-of-Experts Models

MSc Presentation

Felix Agakov
Division of Informatics
The University of Edinburgh

October 2000

The Presentation:

1. Examples of latent variable modeling

2. Mixture models, PoE, mixtures of PoE

3. Results

   - Product of constrained Gaussians [*PoCG*]

   - Product of Probabilistic principal component analyzers [*PoPPCA*]

   - Contrastive divergence learning [*CDL*]

4. Further goals

5. Applications

Examples: latent variable modeling

Data:

- observable (visible)
- unobservable, missing, unknown (hidden)

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

Examples:

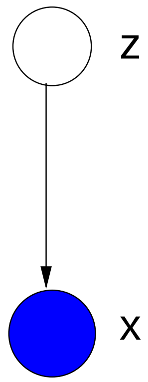
- given a picture, how to explain what class it belongs to?
- given a sentence, how to guess who could have generated it (a child, a foreigner, a computer program)?

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

[but very often cause the observations].

Examples: latent variable modeling - II

To explain the data, can apply probabilistic reasoning



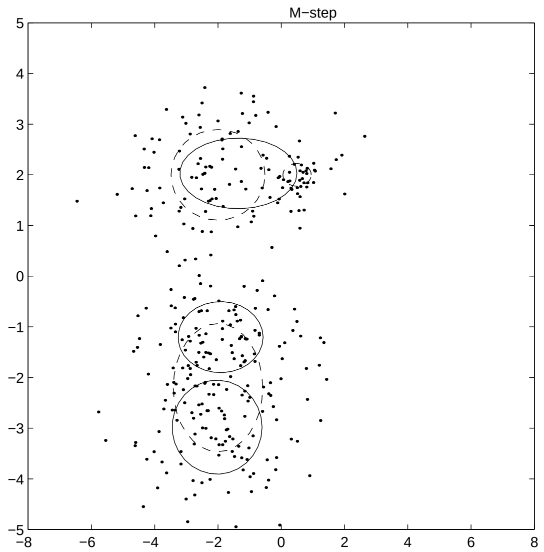
If $\mathbf{z}$ is hidden, $\mathbf{x}$ is visible, what are

- $P(\mathbf{z}|\mathbf{x})$? - classification

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

Alternatively, we can hope to understand the underlying structure of the data by modeling the data-generating distribution.

Mixture models



How to estimate a non-trivial density?

- want to assign high probabilities to the observed data
- do not want to assign high probabilities to the irrelevant points in $\mathcal{R}^d$

Mixture models - a way of estimating non-trivial probability densities.

Hidden variable ($\mathbf{z}$) represents a component (expert), generating a data point

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

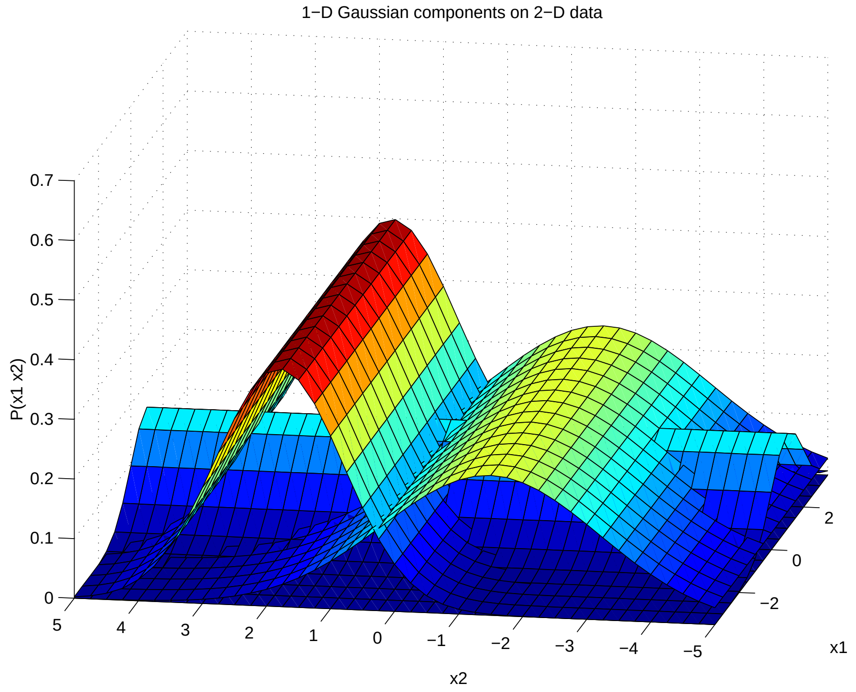
However:

- 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



PoE: an alternative

Geoffrey Hinton, 1999

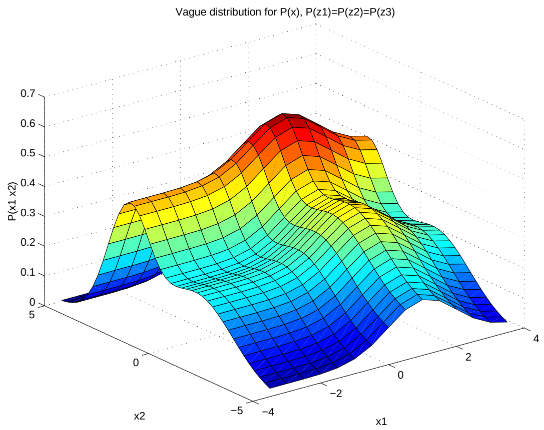
$$P(\mathbf{x}) = \frac{\prod_{i=1}^{m} P_i(\mathbf{x}|\Theta_\mathbf{i})}{\int_{\acute{\mathbf{x}}} \prod_{i=1}^{m} P_i(\acute{\mathbf{x}}|\Theta_\mathbf{i}) \mathrm{d}\acute{\mathbf{x}}}$$

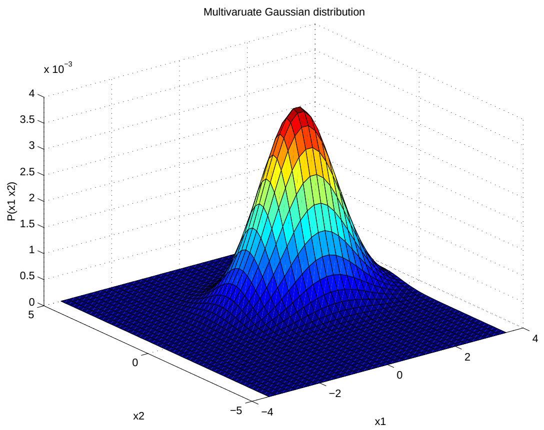
$P_i(\mathbf{x}|\Theta_\mathbf{i})$ - probability of $\mathbf{x}$ under the model with parameters $\Theta_\mathbf{i}$. $\acute{\mathbf{x}}$ - "fantasy" data, generated by the PoE model.

- Yields much sharper distribution

- Each component may assign a high probability to irrelevant data, as long as it 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

This Work

Principal goal: Investigation of the PoE model for Gaussian experts

- Analytic Part:
  - Analytic solution for ML parameters of *PoCG*
  - Analytic solution for ML parameters of *PoPPCA*
  - Investigation of the *CD learning rule* for Gaussians (comparison of maximizing the log-likelihood with maximizing its approximation)
- Experimental Part: Confirmation of the analytic results;
- Implementation: The ExLab library.

Product of Gaussians

For the i^{th} expert

$$p_i(\mathbf{x}|\Theta_\mathbf{i}) \propto \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_i)^T\mathbf{C}_i^{-1}(\mathbf{x}-\mu_i)\right\}.$$

Then

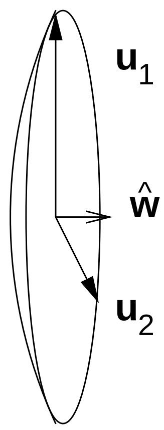
$$\begin{aligned} p(\mathbf{x}) &\propto \prod_{i=1}^{m}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_\mathbf{i})^\mathbf{T}\mathbf{C}_\mathbf{i}^{-1}(\mathbf{x}-\mu_\mathbf{i})\right\} \\ &\propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{m}(\mathbf{x}-\mu_\mathbf{i})^\mathbf{T}\mathbf{C}_\mathbf{i}^{-1}(\mathbf{x}-\mu_\mathbf{i})\right\} \end{aligned}$$

Completing the square,

$$\begin{aligned} p(\mathbf{x}) &\propto N(\mu_\Sigma, \mathbf{C}_\Sigma) \\ \mathbf{C}_\Sigma^{-1} &= \sum_{i=1}^{m}\mathbf{C}_i^{-1} \end{aligned}$$

PoCG-I

- Probability contours: "hyper-disks" *equally* elongated in $d-1$ directions and contracted in 1 direction (in $\mathcal{R}^d$).



$\hat{\mathbf{w}}$ is the direction of contraction;

σ_0^2 - the distortion variance ($\sigma_0^2 > \sigma_{\hat{w}}^2$).

- 1 expert: $\mathbf{C}^{-1} = \sigma_0^{-2}\mathbf{I_d} + \mathbf{w}\mathbf{w}^\mathbf{T}$,

  $\mathbf{w}$ is collinear with the direction of contraction.

- m experts: $\mathbf{C}_\Sigma^{-1} = m\sigma_0^{-2}\mathbf{I_d} + \mathbf{W}\mathbf{W}^\mathbf{T}$.

Corresponds to the inverse Factor Analysis.

PoCG-II

Results:

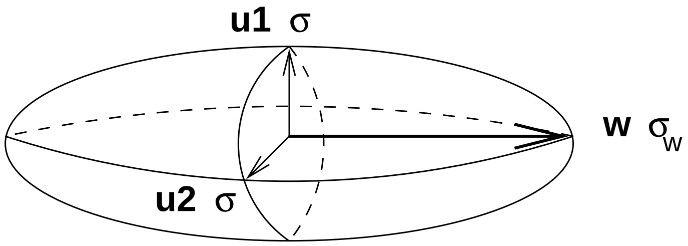
- Found analytically that the ML expert weights $\mathbf{W}_{ML}$ correspond to **minor** eigenvectors of the sample covariance.
- The distortion variance σ_0^2 is averaged over variances in the directions of elongation of the sample covariance.
- I.e., *PoCG* gives rise to **probabilistic minor component analysis**

Advantages of probabilistic models (T&B, 1997):

- the measure of likeliness of data under the model
- comparison with other dimensionality reduction techniques
- incorporating the priors
- considering mixtures

PoPPCA - I

- Probability contours: hyper-ellipsoids elongated in one direction and *equally* contracted in $d-1$ directions (in $\mathcal{R}^d$).



$\mathbf{w}$ - the direction of elongation;

σ^2 - the noise variance ($\sigma^2 < \sigma_w^2$).

- 1 expert: $\mathbf{C} = \sigma^2 \mathbf{I_d} + \mathbf{w}\mathbf{w}^{\mathbf{T}}$;

- m experts: $\mathbf{C}_{\boldsymbol{\Sigma}}^{-1} = m\sigma^{-2}\mathbf{I_d} - \sigma^{-\mathbf{2}}\mathbf{W}\Gamma\mathbf{W}^{\mathbf{T}}$.

Γ is the scaling factor ensuring that $\mathbf{C}_{\boldsymbol{\Sigma}}^{-1}$ is positive-definite (so that $\mathbf{C}_{\boldsymbol{\Sigma}}$ is a valid covariance of the Gaussian distribution).

PoPPCA - II

Results:

- The general solution for the ML expert weights is complicated due to the scaling factor Γ (unless $\Gamma = \alpha \mathbf{I}_m$).

- Ignoring the constraints on weight magnitudes, the optimal weights of the m-expert PoPPCA either:

  – correspond to principal eigenvectors of the sample covariance (m-PPCA)

    or

  – are imaginary (shows limitations of the model).

Analytic solutions may be difficult or impossible (so we may consider sampling techniques).

CDL in Gaussian models - I

- Gaussian: The joint density ($\mathbf{0}$-mean):

$$p(\mathbf{x}, \mathbf{z}) \propto \exp\left\{-\frac{\mathbf{1}}{\mathbf{2}}[\mathbf{x}^{\mathbf{T}}\mathbf{z}^{\mathbf{T}}]\mathbf{C}^{-\mathbf{1}}\begin{bmatrix}\mathbf{x}\\ \mathbf{z}\end{bmatrix}\right\},$$

- Boltzmann Machines (BM):

  The *energy* of the state $[\mathbf{x}, \mathbf{z}]$:

$$E(\mathbf{x}, \mathbf{z}) = \frac{\beta}{2}[\mathbf{x}^T\ \mathbf{z}^T]\mathbf{W}\begin{bmatrix}\mathbf{x}\\ \mathbf{z}\end{bmatrix};$$

  The joint density:

$$p(\mathbf{x}, \mathbf{z}) \propto \exp\{-\mathbf{E}(\mathbf{x}, \mathbf{z})\}$$

Notice the correspondence of **W** and $\mathbf{C}^{-1}$.

Can apply Boltzmann learning (optimization of the log-likelihood via sampling) to determine the inverse covariance $\mathbf{C}^{-1}$.

CDL in Gaussian models - II

So, we can apply *BL* to learn $\mathbf{C}^{-1}$. But:

- Estimating the true log-likelihood is difficult due to the re-normalizing factor.

- Need to sample form the unknown PoE distribution $p(\mathbf{x}|\Theta_\mathbf{1}, \ldots, \Theta_\mathbf{m})$ (approximated by the equilibrium distribution of the Markov chain).

- Slow, results in high variance, may lead to incorrect approximation of $\partial\mathcal{L}/\partial\Theta_\mathbf{i}$.

Hinton (2000) suggests **contrastive divergence learning** (fast, showing low variance in experiments).

- *BL:* maximizing the log-likelihood $\mathcal{L}^\infty$;

- *CDL:* maximizing approximation $\mathcal{L}^{(i)}$ of the log-likelihood $\mathcal{L}^\infty$ after i complete iterations of Gibbs sampling.

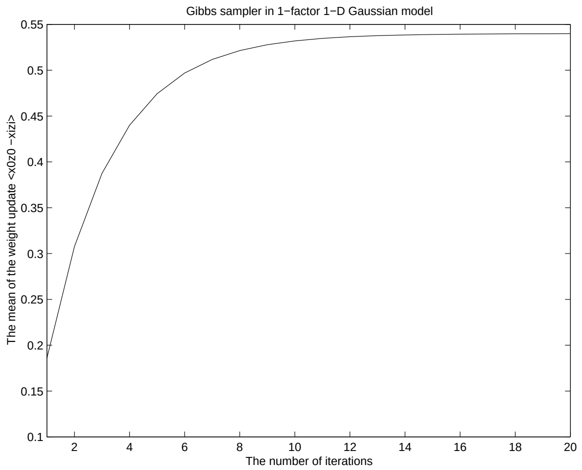
CDL in Gaussian models - III

Investigated Hinton's claim for a 1-D 1-factor linear Gaussian latent variable model.

Results:

- Both *CDL* and *BL* have the same optimal solutions;
- *CDL* introduces a systematic bias by underestimating the absolute value of the gradient

$$\left|\frac{\partial \mathcal{L}^{(i)}}{\partial w}\right| < \left|\frac{\partial \mathcal{L}^{\infty}}{\partial w}\right|.$$



- Found the moments of the distribution of the parameter update.

CDL and *BL* have similar solutions for the considered Gaussian model.

Applications

PoE, mixtures of PoE:

- clustering, classification
- dimensionality reduction
- DSP [e.g., signal extraction]
- data visualization
- image recognition
- language processing

Further goals

- Considering extensions of *PoCG* and *PoPPCA*;

- Further investigation of the CD learning rule and variance of the parameter update;

- Mixtures of PoE models;

- Other PoE architectures.

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