Latent Variable Models

Pradeep Ravikumar

Co-instructor: Manuela Veloso

Machine Learning 10-701

Unobserved (also called Latent) Variables

- In a Gaussian mixture model, the cluster label **z** for any point **x** was never observed

- Some other such models with systematically unobserved (also called latent) variables:
  - Mixture Models
  - Factor Analysis
  - Hidden Markov Models (HMMs)
  - State Space Models (Kalman Filters)

Unobserved Variables

Unobserved Variables

- Should we have always-unobserved variables in our models?

Unobserved Variables

- Should we have always-unobserved variables in our models?

- **Logical Positivism:** Denies the meaningfulness of what you cannot observe, in scientific reasoning

Unobserved Variables

- Should we have always-unobserved variables in our models?

- **Logical Positivism:** Denies the meaningfulness of what you cannot observe, in scientific reasoning

- On the other hand, such unobserved variables are widely used in scientific settings:

Unobserved Variables

- Should we have always-unobserved variables in our models?

- **Logical Positivism:** Denies the meaningfulness of what you cannot observe, in scientific reasoning

- On the other hand, such unobserved variables are widely used in scientific settings:

  - Group of patients with a certain syndrome

Unobserved Variables

- Should we have always-unobserved variables in our models?

- **Logical Positivism:** Denies the meaningfulness of what you cannot observe, in scientific reasoning

- On the other hand, such unobserved variables are widely used in scientific settings:

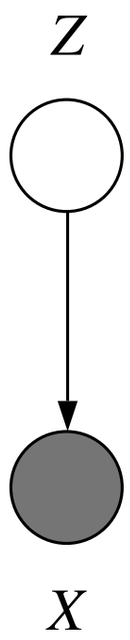
  - Group of patients with a certain syndrome

  - Group of animals as a distinct species

Mixture Models

- Simplest class of models with unobserved r.v.s are **mixture models**
  - Latent variables are discrete
- We will look at two subclasses:
  - Unconditional Mixture Models: Clustering (density estimation)
  - Conditional Mixture Models: Prediction (regression, classification)

Unconditional Mixture Models



- Mixture Model represented as a graphical model

- Latent Variable Z is a multinomial node taking one of K values

Unconditional Mixture Models

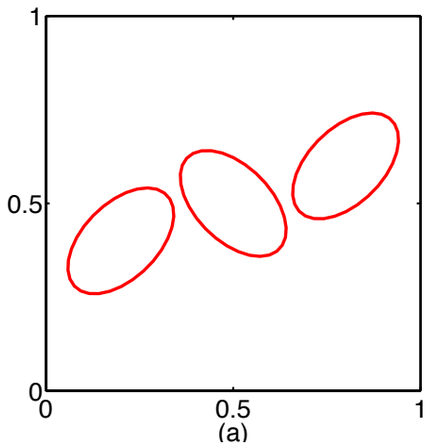
$$
\begin{aligned}
p(x \mid \theta) &= \sum_i p(Z^i = 1 \mid \pi_i) p(x \mid Z^i = 1, \theta_i) \\
&= \sum_i \pi_i p(x \mid Z^i = 1, \theta_i).
\end{aligned}
$$

where $\theta = (\pi_1, \ldots, \pi_K, \theta_1, \ldots, \theta_K)$ and where the π_i are constrained to sum to one.

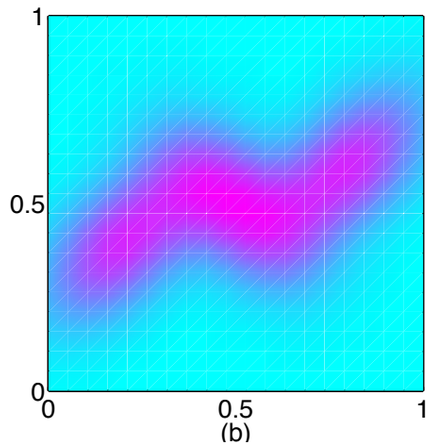
- Distribution of Z: Multinomial

- Different distributions of P(x|z) yield different mixture models

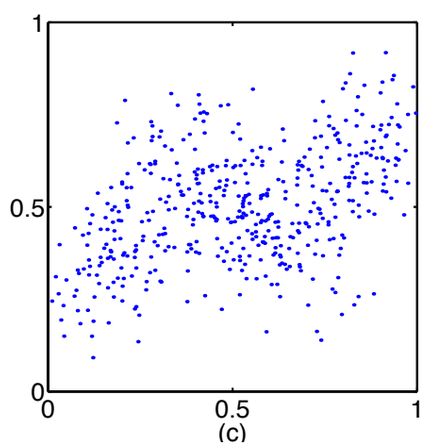
Gaussian Mixture Models



Contours of density

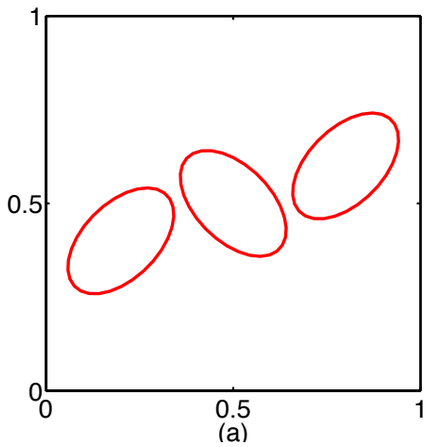


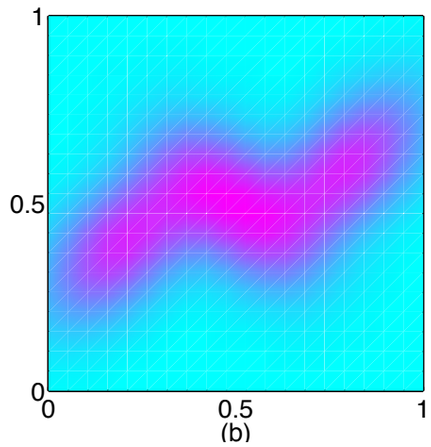
Density

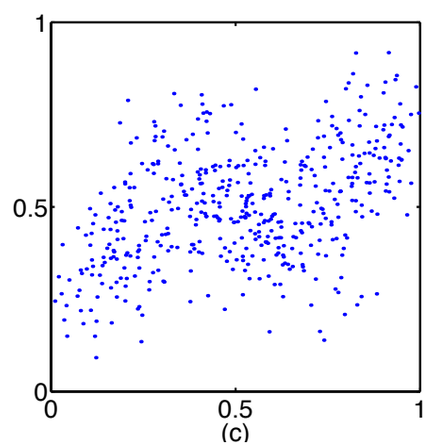


500 samples from Density

Gaussian Mixture Models





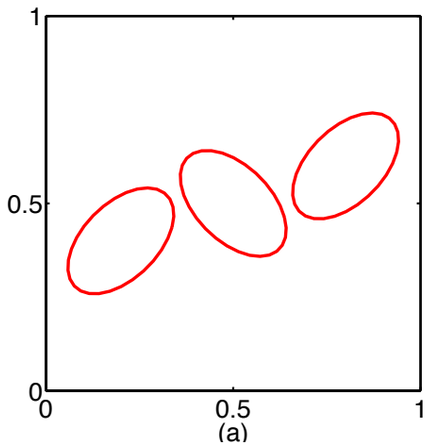


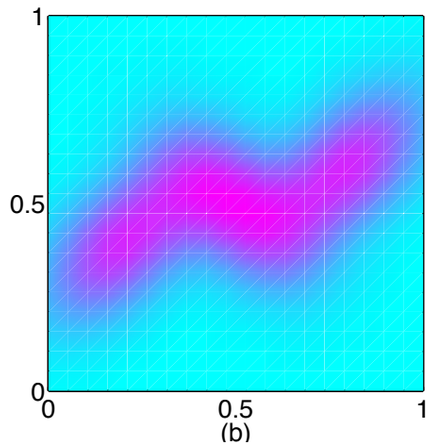
Contours of density Density 500 samples from Density

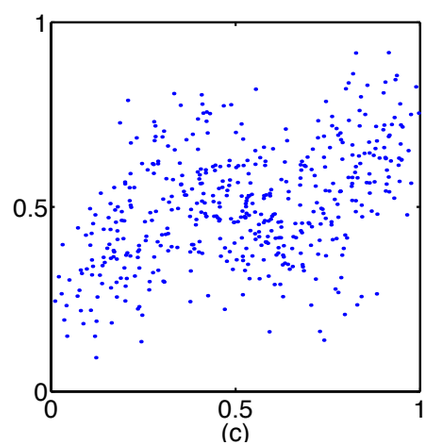
$$
\begin{aligned}
p(x \mid \theta) &= \sum_i p(Z^i = 1 \mid \pi_i) p(x \mid Z^i = 1, \theta_i) \\
&= \sum_i \pi_i p(x \mid Z^i = 1, \theta_i).
\end{aligned}
$$

$$
p(x \mid \theta) = \sum_i \pi_i \frac{1}{(2\pi)^{m/2} |\Sigma_i|^{1/2}} \exp\left\{ -\frac{1}{2} (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) \right\}.
$$

Gaussian Mixture Models







Contours of density Density 500 samples from Density

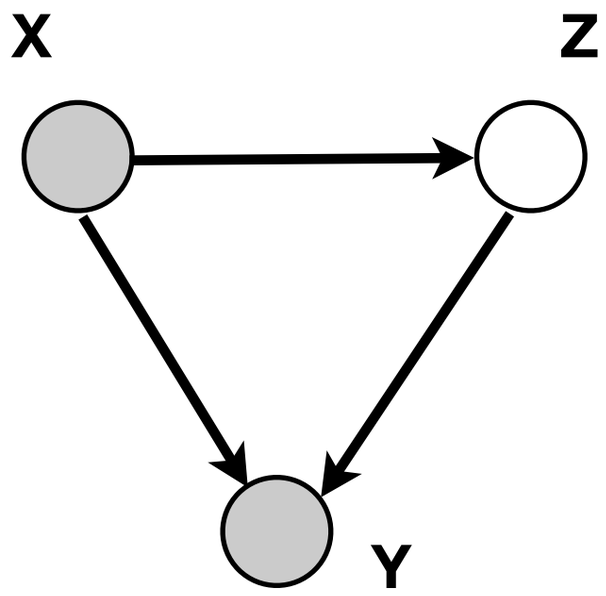
$$p(x \mid \theta) = \sum_i p(Z^i = 1 \mid \pi_i) p(x \mid Z^i = 1, \theta_i)$$

$$= \sum_i \pi_i p(x \mid Z^i = 1, \theta_i).$$

$$p(x \mid \theta) = \sum_i \pi_i \frac{1}{(2\pi)^{m/2} |\Sigma_i|^{1/2}} \exp\left\{ -\frac{1}{2} (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) \right\}.$$

$$p(x \mid \theta) = \sum_i \pi_i \mathcal{N}(x \mid \mu_i, \Sigma_i)$$

Conditional Mixture Models



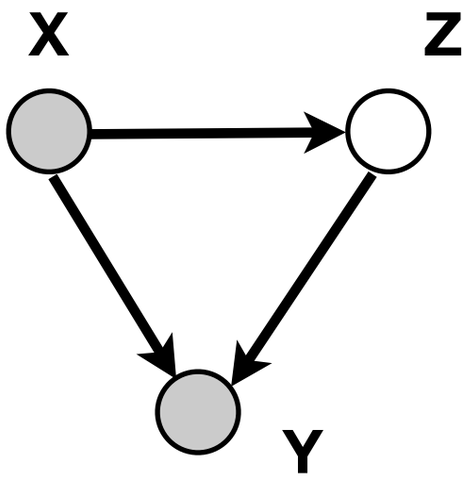
$$p(Z^i = 1 \,|\, x, \xi) = \frac{e^{\xi_i^T x}}{\sum_j e^{\xi_j^T x}},$$

$$p(y \,|\, x, \theta) = \sum_i p(Z^i = 1 \,|\, x, \xi) p(y \,|\, Z^i = 1, x, \theta_i),$$

Gaussian Conditional Mixture Model

- Just a mixture of linear regressions:

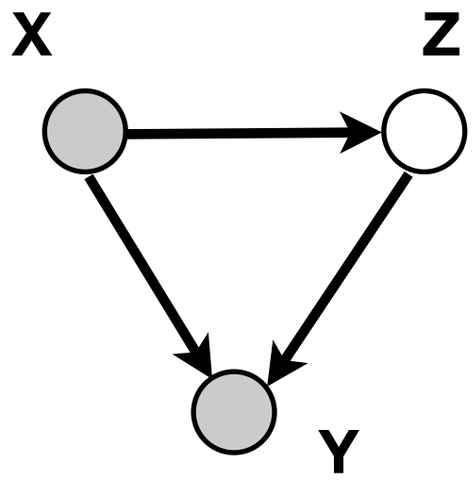
$$p(y \mid x, \theta) = \sum_i \pi_i(x, \xi) \mathcal{N}(y \mid \beta_i^T x, \sigma_i^2),$$



Logistic Conditional Mixture Model

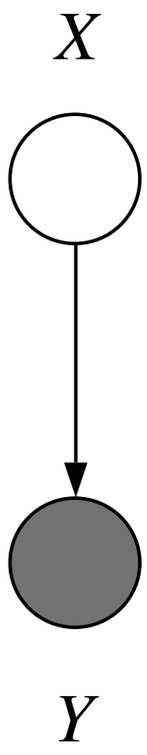
- Just a mixture of logistic regressions:

$$p(y \mid x, \theta) = \sum_i \pi_i(x, \xi) \mu(\theta_i^T x)^y (1 - \mu(\theta_i^T x))^{1-y},$$



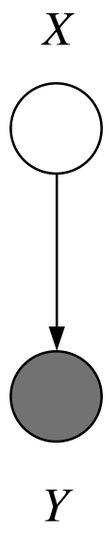
Latent Variables: The continuous case

The continuous case



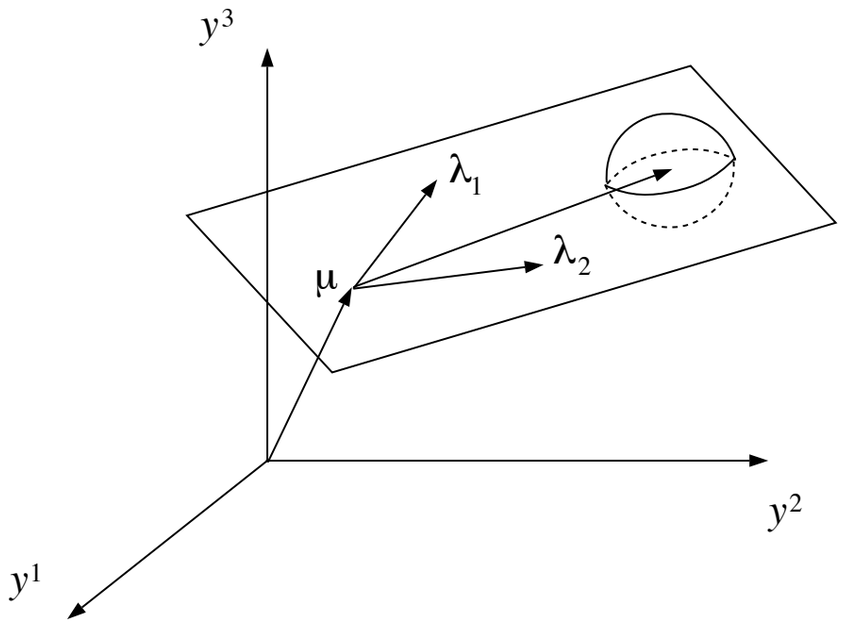
- The unobserved variable X is continuous
- Y could be {continuous, discrete}
- Not clustering anymore: we can’t think of distribution over Y as a convex combination of finitely many (cluster) components

Continuous Latent Variables



- Suppose Y lies in $\mathbb{R}^p$, where p is large
- But suppose Y has a “simpler” story to tell:
  - A point X in a lower-dimensional subspace of $\mathbb{R}^p$
  - Y is Gaussian distributed around X

Continuous Latent Variables



Suppose the lower-dimensional subspace of $\mathbb{R}^p$ has basis functions $\{\Lambda_j\}$, so that a point in this subspace can be expressed as $\sum_j \Lambda_j x_j = \Lambda x$, where Λ has columns as the basis vectors $\{\Lambda_j\}$.

Suppose the subspace is translated away from the origin by μ, so that we have a hyperplane, with points $\mu + \Lambda x$.

We then observe y which is distributed as a Gaussian, with mean as a point $\mu + \Lambda x$.

Continuous Latent Variables

Suppose the lower-dimensional subspace of $\mathbb{R}^p$ has basis functions $\{\Lambda_j\}$, so that a point in this subspace can be expressed as $\sum_j \Lambda_j x_j = \Lambda x$, where Λ has columns as the basis vectors $\{\Lambda_j\}$.

Suppose the subspace is translated away from the origin by μ, so that we have a hyperplane, with points $\mu + \Lambda x$.

We then observe y which is distributed as a Gaussian, with mean as a point $\mu + \Lambda x$.

The coefficients x are unobserved!

Suppose that we endow the coefficients X with a Gaussian density: the resulting model yields "Factor Analysis".

Factor Analysis uses the basis vectors $\{\Lambda_j\}$ as latent factors explaining the data (e.g. IQ)

Factor Analysis

$$
\begin{aligned}
X &\sim \mathcal{N}(0, I) \\
Y|X = x &\sim \mathcal{N}(\mu + \Lambda x, \Psi)
\end{aligned}
$$

Factor Analysis

$$
\begin{aligned}
X &\sim \mathcal{N}(0, I) \\
Y|X = x &\sim \mathcal{N}(\mu + \Lambda x, \Psi) \\
Y &= \mu + \Lambda \mathbf{X} + W,\ W \sim \mathcal{N}(0, \Psi)
\end{aligned}
$$

Factor Analysis

$$
\begin{aligned}
X &\sim \mathcal{N}(0, I) \\
Y|X = x &\sim \mathcal{N}(\mu + \Lambda x, \Psi) \\
Y &= \mu + \Lambda \mathbf{X} + W,\ W \sim \mathcal{N}(0, \Psi)
\end{aligned}
$$

$$
\begin{aligned}
E(Y) &= E(\mu + \Lambda X + W) \\
&= \mu + \Lambda EX + EW \\
&= \mu,
\end{aligned}
$$

Factor Analysis

$$
\begin{aligned}
X &\sim \mathcal{N}(0, I) \\
Y|X = x &\sim \mathcal{N}(\mu + \Lambda x, \Psi) \\
Y &= \mu + \Lambda \mathbf{X} + W,\ W \sim \mathcal{N}(0, \Psi)
\end{aligned}
$$

$$
\begin{aligned}
E(Y) &= E(\mu + \Lambda X + W) \\
&= \mu + \Lambda EX + EW \\
&= \mu,
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Var}(Y) &= E[(\mu + \Lambda X + W - \mu)(\mu + \Lambda X + W - \mu)^T] \\
&= E[(\Lambda X + W)(\Lambda X + W)^T] \\
&= \Lambda E(XX^T)\Lambda^T + E(WW^T) \\
&= \Lambda\Lambda^T + \Psi.
\end{aligned}
$$

Factor Analysis

$$
\begin{aligned}
X &\sim \mathcal{N}(0, I) \\
Y|X = x &\sim \mathcal{N}(\mu + \Lambda x, \Psi) \\
Y &= \mu + \Lambda \mathbf{X} + W,\ W \sim \mathcal{N}(0, \Psi)
\end{aligned}
$$

$$
\begin{aligned}
E(Y) &= E(\mu + \Lambda X + W) \\
&= \mu + \Lambda EX + EW \\
&= \mu,
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Var}(Y) &= E[(\mu + \Lambda X + W - \mu)(\mu + \Lambda X + W - \mu)^T] \\
&= E[(\Lambda X + W)(\Lambda X + W)^T] \\
&= \Lambda E(XX^T)\Lambda^T + E(WW^T) \\
&= \Lambda\Lambda^T + \Psi.
\end{aligned}
$$

Marginal distribution of observed variables Y is Gaussian!

$$Y \sim \mathcal{N}(\mu, \Lambda\Lambda^T + \Psi)$$

Factor Analysis

- Given data D = {y_1, ... y_N}, estimate the parameters!

Factor Analysis

- Given data D = {y_1, ... y_N}, estimate the parameters!

- Maximize the log-likelihood:

$$l(\theta \mid D) = -\frac{N}{2} \log |\Lambda\Lambda^T + \Psi| - \frac{1}{2} \left\{ \sum_n (y_n - \mu)^T (\Lambda\Lambda^T + \Psi)^{-1} (y_n - \mu) \right\},$$

Using the marginal distribution of Y

Factor Analysis

- Given data D = {y_1, ... y_N}, estimate the parameters!

- Maximize the log-likelihood:

$$l(\theta \mid D) = -\frac{N}{2} \log |\Lambda\Lambda^T + \Psi| - \frac{1}{2} \left\{ \sum_n (y_n - \mu)^T (\Lambda\Lambda^T + \Psi)^{-1} (y_n - \mu) \right\},$$

- Easy to optimize for \mu : $\hat{\mu}_{ML} = \frac{1}{N} \sum_n y_n,$

Factor Analysis

- Given data D = {y_1, ... y_N}, estimate the parameters!

- Maximize the log-likelihood:

$$l(\theta \mid D) = -\frac{N}{2} \log |\Lambda\Lambda^T + \Psi| - \frac{1}{2} \left\{ \sum_n (y_n - \mu)^T (\Lambda\Lambda^T + \Psi)^{-1} (y_n - \mu) \right\},$$

- Easy to optimize for \mu : $\hat{\mu}_{ML} = \frac{1}{N} \sum_n y_n,$

- Solution for \Gamma cannot be unique under orthogonal transformations:

$$\Lambda R (\Lambda R)^T = \Lambda R R^T \Lambda^T = \Lambda\Lambda^T.$$

Factor Analysis

- Given data D = {y_1, ... y_N}, estimate the parameters!

- Maximize the log-likelihood:

$$l(\theta \mid D) = -\frac{N}{2} \log |\Lambda\Lambda^T + \Psi| - \frac{1}{2} \left\{ \sum_n (y_n - \mu)^T (\Lambda\Lambda^T + \Psi)^{-1} (y_n - \mu) \right\},$$

- Easy to optimize for \mu : $\hat{\mu}_{ML} = \frac{1}{N} \sum_n y_n,$

- The objective as a function of the rest of the parameters is non-convex, and difficult to optimize: reminiscent of the discrete mixture model case!

  - No closed form expressions in particular

Factor Analysis

- Given data D = {y_1, ... y_N}, estimate the parameters!

- Maximize the log-likelihood:

$$l(\theta \mid D) = -\frac{N}{2} \log |\Lambda\Lambda^T + \Psi| - \frac{1}{2} \left\{ \sum_n (y_n - \mu)^T (\Lambda\Lambda^T + \Psi)^{-1} (y_n - \mu) \right\},$$

- Easy to optimize for \mu : $\hat{\mu}_{ML} = \frac{1}{N} \sum_n y_n,$

- The objective as a function of the rest of the parameters is non-convex, and difficult to optimize: reminiscent of the discrete mixture model case!

  - No closed form expressions in particular

  - Expectation Maximization (EM) !

Dynamic Latent Models

- Consider dynamic generalization of the two-node case, with the two node pair copied in a spatial array, connecting successive state nodes in the array

- Dynamic generalization of discrete mixture models is **Hidden Markov Models**

- Dynamic generalization of factor analysis: **State Space Models**

  - Machinery of Kalman Filters

- HMMs and Kalman Filters developed in separate communities, commonalities not observed much later