Probabilistic PCA and its extensions

Kai Wang

UBC ·10/19/2017

Motivations for probabilistic methods

- Modeling for uncertainty

$$y = f\left(u, \theta\right) + e$$

deterministic model stochastic model

- State/parameters estimations

interval estimation/confidence interval

$$p\left(\theta \mid y\right) \propto p\left(y \mid \theta\right) p\left(\theta\right)$$

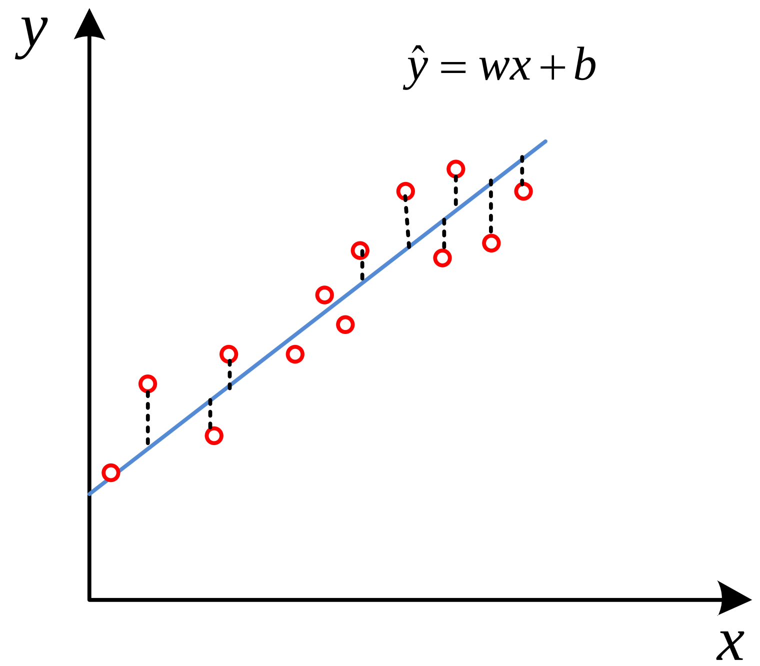
- Extensions to complex models

Example

Least squares

Linear regression with a dataset $\left\{x_n, y_n\right\}$

$$\min \frac{1}{N}\sum_n \left\| y_n - wx_n - b \right\|_2$$



Maximum likelihood

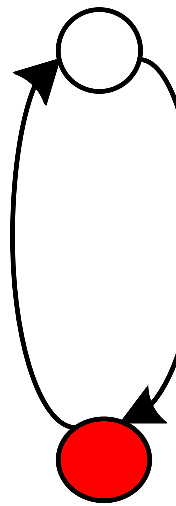
$$y = \underline{wx + b} + e, \quad e \sim N\left(0, \sigma^2\right)$$

deterministic model stochastic model

$$\ln p\left(e_1, \ldots, e_n, \ldots\right) = \sum_n \ln p\left(e_n\right) = \sum_n \ln \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{\left(y_n - wx_n - b\right)^2}{2\sigma^2}}$$

Latent variable models

- **Latent variables** are those that are not directly observed but are rather inferred from other variables that are observed (directly measured).



Latent variable models

Examples of latent variable models

- Principal component analysis
- Partial least squares
- Canonical variate analysis
- Factor analysis
- Independent component analysis
- Slow feature analysis

Probabilistic latent variables models(PLVM)

Basic PLVM

	i.i.d	dynamic
continuous	Probabilistic PCA	Linear dynamic systems
discrete	Gaussian mixtures	Hidden Markov models

Complex PLVM

Factor analysis; Particle filters; Locally weighted PLVM
Mixtures of probabilistic PCA; Switched linear dynamic systems

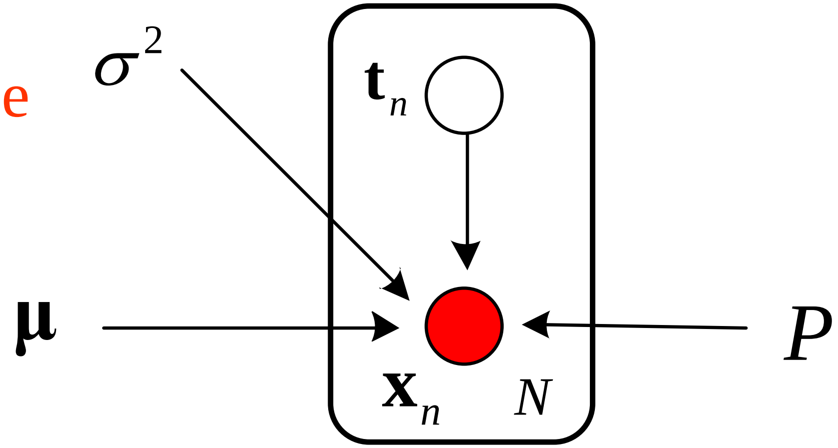
Probabilistic PCA(PPCA)

For each sample $\mathbf{x}$ with *m* dimensions, there is a *k*-dimentional latent variables such that

$$\mathbf{x} = P_{m\times k}\mathbf{t} + \boldsymbol{\mu} + \mathbf{e}$$

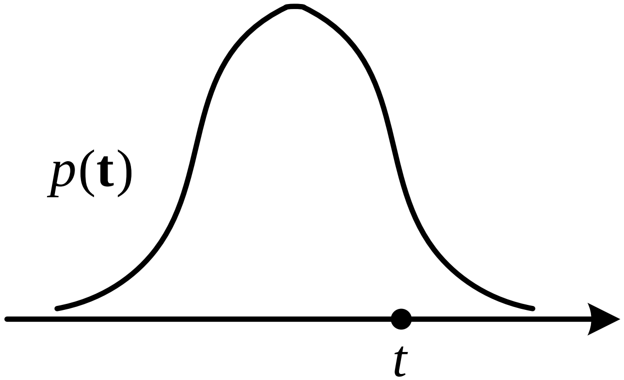
latent variables bias residuals

$$\mathbf{t} \sim N\left(\mathbf{0}, I\right) \qquad \mathbf{e} \sim N\left(\mathbf{0}, \sigma^2 I\right)$$



PPCA

The probability models:

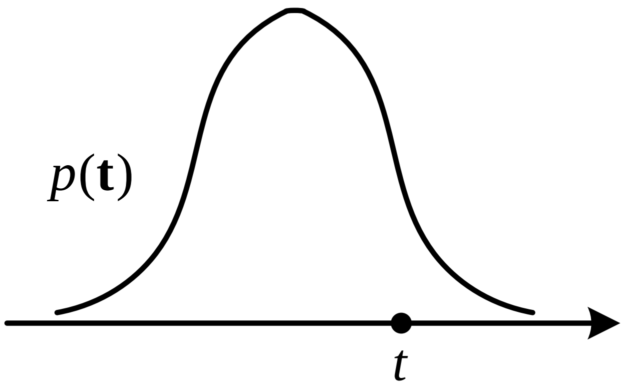


$$p(\mathbf{t}) = \mathcal{N}(\mathbf{t} \mid 0, \mathrm{I})$$

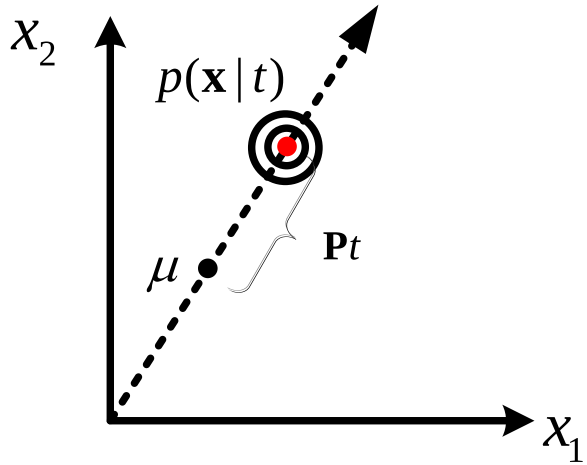
First we define a Gaussian prior distribution $p(\mathbf{t})$ over the latent variable $\mathbf{t}$

Probabilistic PCA

The probability models:



$$p(\mathbf{t}) = \mathcal{N}(\mathbf{t} \mid 0, \mathrm{I})$$



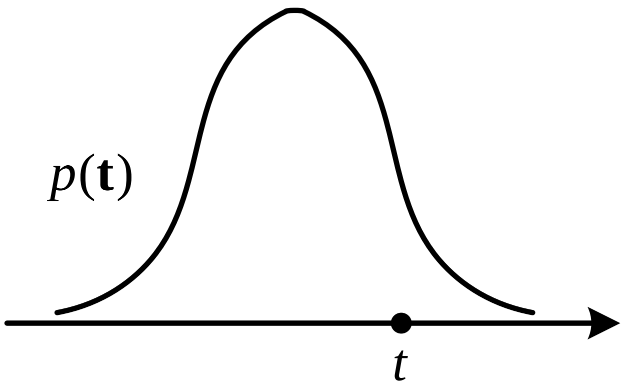
$$p\left(\mathbf{x} \mid \mathbf{t}\right) = N\left(P\mathbf{t} + \boldsymbol{\mu}, \sigma^2 I\right)$$

First we define a Ga
prior distribution $p($
the latent variable $\mathbf{t}$

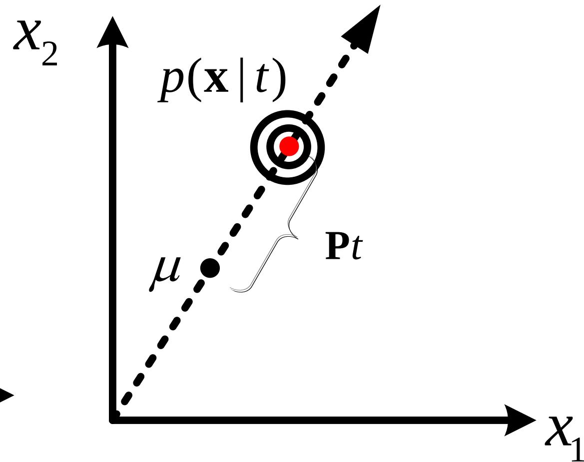
Then the conditional distribution of the observed variable can be formed as $p(\mathbf{x} \mid \mathbf{t})$

Probabilistic PCA

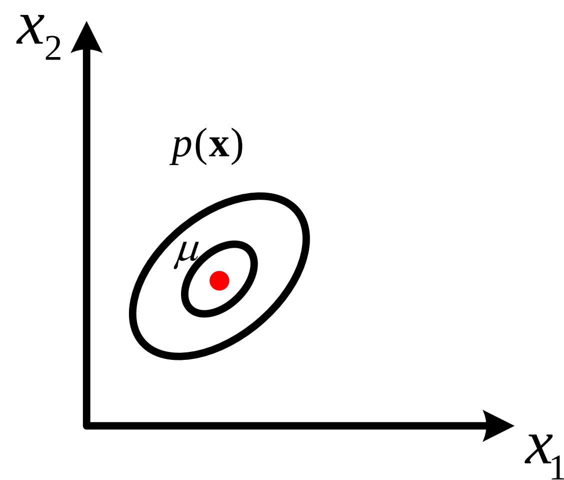
The probability models:



$$p(\mathbf{t}) = \mathcal{N}(\mathbf{t} \mid 0, \mathrm{I})$$



$$p(\mathbf{x} \mid \mathbf{t}) = N\left(P\mathbf{t} + \boldsymbol{\mu}, \sigma^2 I\right)$$



$$p(\mathbf{x}) = N(\boldsymbol{\mu}, M)$$

$$M = PP^T + \sigma^2 I$$

First we define a Ga
prior distribution $p($
the latent variable $\mathbf{t}$

Then the
of the obs
formed as

Finally, the marginal distribution $p(\mathbf{x})$ of the observed variable $\mathbf{x}$ can be expressed by $p(\mathbf{x}) = \int p(\mathbf{x} \mid \mathbf{t}) p(\mathbf{t}) d\mathbf{t}$

Relation with least squares

$$\min \frac{1}{N} \sum_{n} \left\| \mathbf{x}_n - P\mathbf{t}_n \right\|_2$$

$$s.t.\ P^T P = I$$

Relation with least squares

$$\min \sum_n \left\| \mathbf{x}_n - P\mathbf{t}_n \right\|_2$$

$$s.t.\ P^T P = I$$

orthonormal $\qquad p(\mathbf{t}) = N(0, I)$

Relation with least squares

Isotropic covariance assigns the same weight for each variable

$$p(\mathbf{x} \mid \mathbf{t}) = N\left(P\mathbf{t} + \boldsymbol{\mu}, \sigma^2 I\right)$$

$$\propto \exp\left(-\frac{1}{\sigma^2} (\mathbf{x} - P\mathbf{t} - \boldsymbol{\mu})^T (\mathbf{x} - P\mathbf{t} - \boldsymbol{\mu}) \right)$$

$$\min \sum_n \left\| \mathbf{x}_n - P\mathbf{t}_n \right\|_2$$

$$s.t.\ P^T P = I$$

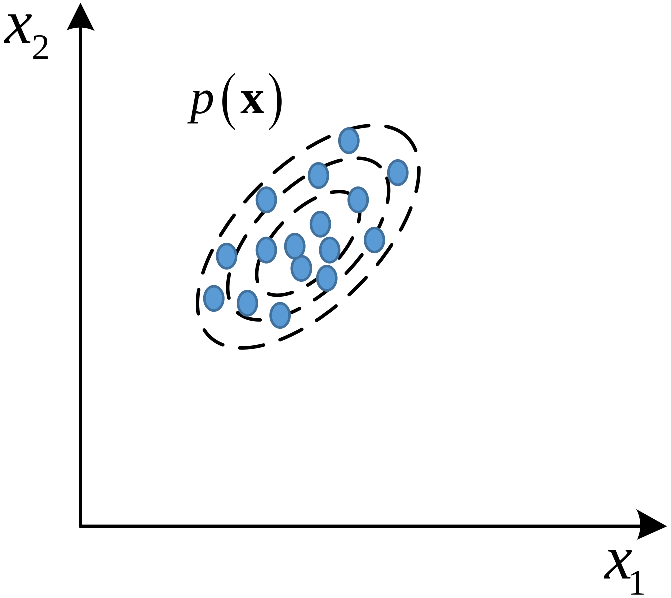
orthonormal

$$p(\mathbf{t}) = N(0, I)$$

Parameters estimation

Maximum likelihood:

Parameters: $\boldsymbol{\theta} = \left(P, \boldsymbol{\mu}, \sigma^2\right)$



$$p(\mathbf{x}) = N(\boldsymbol{\mu}, M)$$

$$M = PP^T + \sigma^2 I$$

Given a data set $X = \{\mathbf{x}_n\}, n = 1, \ldots, N$

the corresponding log likelihood function is

$$\ln p(X \mid \boldsymbol{\theta}) = \sum_{n=1}^{N} \ln p(\mathbf{x}_n \mid \boldsymbol{\theta})$$

$$= -\frac{Nm}{2} \ln(2\pi) - \frac{N}{2} \ln|M| - \frac{1}{2} \sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu})^T M^{-1} (\mathbf{x}_n - \boldsymbol{\mu})$$

Parameters estimation

- There is indeed a closed-form solution, even though the objective function is very complex.

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \qquad \sigma^2 = \frac{1}{m-k}\sum_{i=k+1}^{m}\lambda_i \qquad P = U_{1:\mathrm{k}}(\Lambda_{1:k} - \sigma^2\,\mathrm{I})^{\frac{1}{2}}$$

$$S = \frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{x}_n - \overline{\mathbf{x}}\right)\left(\mathbf{x}_n - \overline{\mathbf{x}}\right)^T = U\Lambda U^T, \Lambda = diag\left(\lambda_1, \ldots, \lambda_m\right)$$

- Regular PCA is a limiting case of PPCA, taken as the limit as the covariance noise becomes infinitely small $\sigma^2 \rightarrow 0$

Latent variables inference

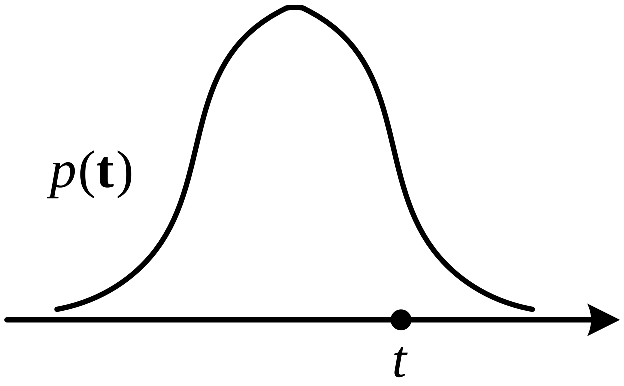
Given a PPCA model, the latent variables $\mathbf{t}$ can be inferred from their corresponding observed variables $\mathbf{x}$.

Prior distribution $\quad p(\mathbf{t}) = N(0, I)$

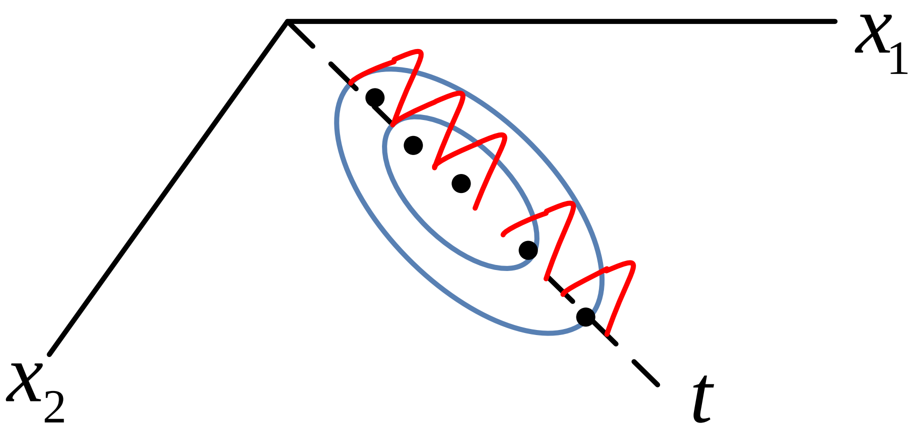
Likelihood $\quad p(\mathbf{x} \mid \mathbf{t}) = N\left(P\mathbf{t} + \boldsymbol{\mu}, \sigma^2 I\right)$

Posterior distribution $\quad p(\mathbf{t} \mid \mathbf{x}) = N\left(W^{-1}P^T(\mathbf{x} - \boldsymbol{\mu}), \sigma^{-2} W\right)$

$$W = P^T P + \sigma^2 I$$



$p(\mathbf{t}) = \mathcal{N}(\mathbf{t} \mid 0, \mathrm{I})$



PPCA and EM algorithm

$$p(\mathbf{x}) = N(\boldsymbol{\mu}, M)$$

$$M = PP^T + \sigma^2 I$$

$$\ln p(X \mid \boldsymbol{\theta}) = \sum_{n=1}^{N} \ln p(\mathbf{x}_n \mid \boldsymbol{\theta})$$

$$= -\frac{Nm}{2}\ln(2\pi) - \frac{N}{2}\ln|M| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^T M^{-1}(\mathbf{x}_n - \boldsymbol{\mu})$$

- The solution with respect to the maximum likelihood of the marginal distribution $p(X)$ is difficult to obtain.
- SVD on high-dimentional matrix is costly.
- All PLVM do not have closed-form solutions except PPCA.

Expectation maximization (EM) algorithm is recommended to find the maximum likelihood solution of latent variables models in an iterative manner.

PPCA and EM algorithm

Maximize the lower bound, referring to the expectation of $\ln p(\mathbf{x}, \mathbf{t} \mid \boldsymbol{\theta})$ with respect to the posterior distribution over the latent variables

$$\mathrm{E}\left(\ln p\left(X, T \mid \boldsymbol{\theta} \right)_{< p(T|X,\boldsymbol{\theta})>} \right) \le \ln p\left(X \mid \boldsymbol{\theta} \right)$$

$$\ln p\left(X, T \mid \boldsymbol{\theta} \right) = \sum_n \left\{ \ln p\left(\mathbf{x}_n \mid \mathbf{t}_n, \boldsymbol{\theta} \right) + \ln p\left(\mathbf{t}_n \right) \right\}$$

E-step

$$\mathrm{E}\left(\ln p\left(X, T \mid \boldsymbol{\theta}^{new} \right)_{< p\left(T|X,\boldsymbol{\theta}^{old}\right)>} \right)$$

M-step

$$\arg\max_{\boldsymbol{\theta}^{new}} \mathrm{E}\left(\ln p\left(X, T \mid \boldsymbol{\theta}^{new} \right)_{< p\left(T|X,\boldsymbol{\theta}^{old}\right)>} \right)$$

Extensions #1

PPCA and missing data

PPCA can deal with randomly missing values with EM algorithm

$$\mathbf{x} = \left(\mathbf{x}_o, \mathbf{x}_m\right)$$

observed variables missing variables

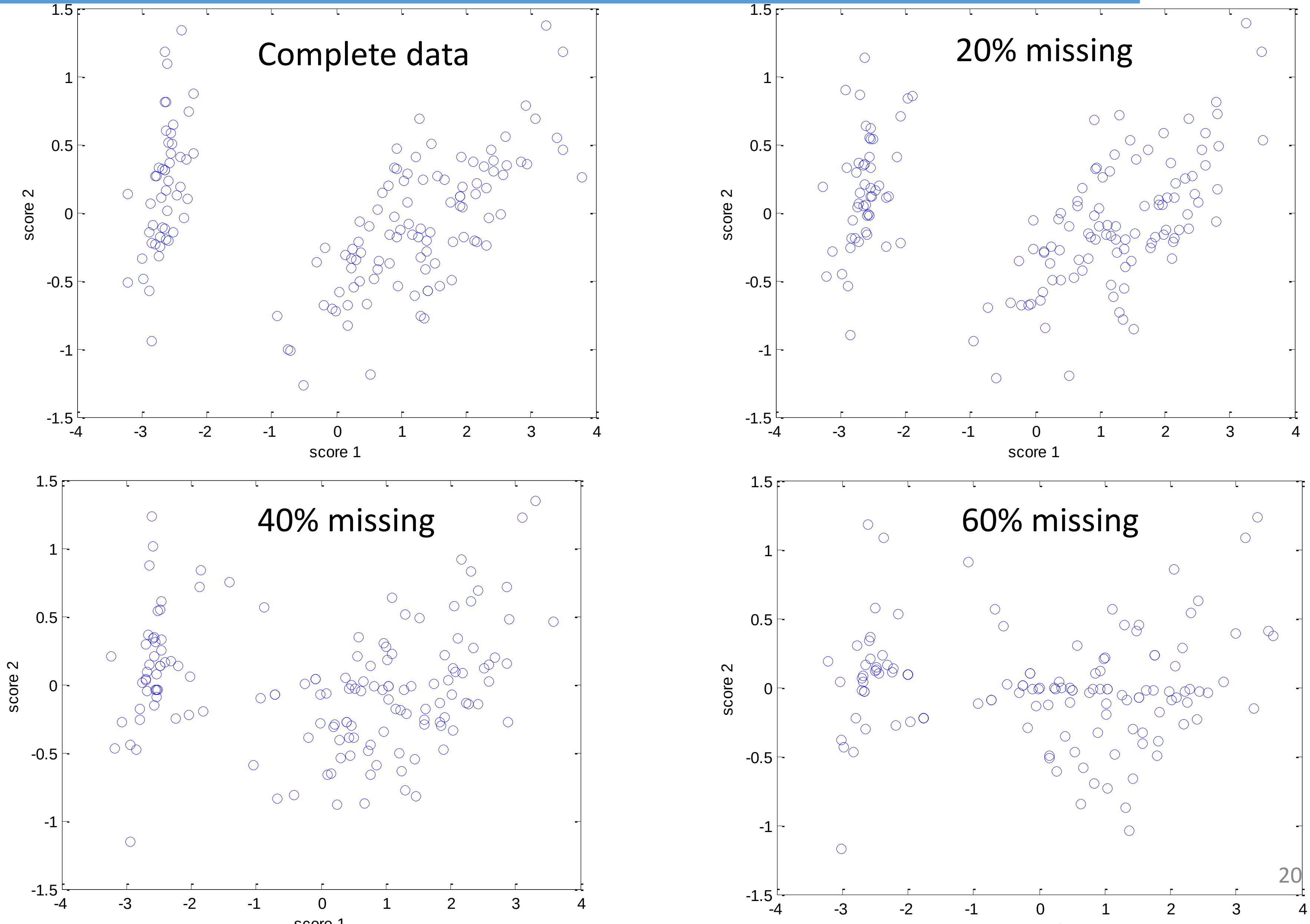
$$\begin{bmatrix} 1 & 3 & - \\ 1 & - & 2 \\ 3 & 2 & 1 \end{bmatrix}$$

Take the missing variables as the role of "latent variables", then a new extended latent variables are $\left(\mathbf{x}_m, \mathbf{t}\right)$

Log likelihood function of each sample

$$\ln p\left(\mathbf{x}_o, \mathbf{x}_m, \mathbf{t} \mid \boldsymbol{\theta}\right) = \ln\left\{ p\left(\mathbf{x}_o \mid \cancel{\mathbf{x}_m}, \mathbf{t}, \boldsymbol{\theta}\right) + p\left(\mathbf{x}_m \mid \mathbf{t}, \boldsymbol{\theta}\right) + p\left(\mathbf{t}\right) \right\}$$

Extensions #1



Extensions #2

Factor analysis

- The isotropic constraint in noise variance is strong in some cases.
- Factor analysis takes the noise distribution as a diagonal covariance such that

$$\mathbf{x} = P_{m\times k}\mathbf{t} + \boldsymbol{\mu} + \mathbf{e}$$

$$\mathbf{e} \sim N\left(\mathbf{0}, \mathrm{diag}\left(\sigma_1^2, \ldots, \sigma_m^2\right)\right)$$

$$p\left(\mathbf{x} \mid \mathbf{t}\right) = N\left(P\mathbf{t} + \boldsymbol{\mu}, \mathrm{diag}\left(\sigma_1^2, \ldots, \sigma_m^2\right)\right)$$

- There is no closed-form solution. EM algorithm should be used.

Extensions #3

Mixtures of PPCA

- A total of Q sub-PPCA are incorporated.
- Discrete latent variable z represents the Q possible states each sample may have

Prior distribution

$$p(z=q)=\pi_q, \sum_{q=1}^{Q}\pi_q=1$$

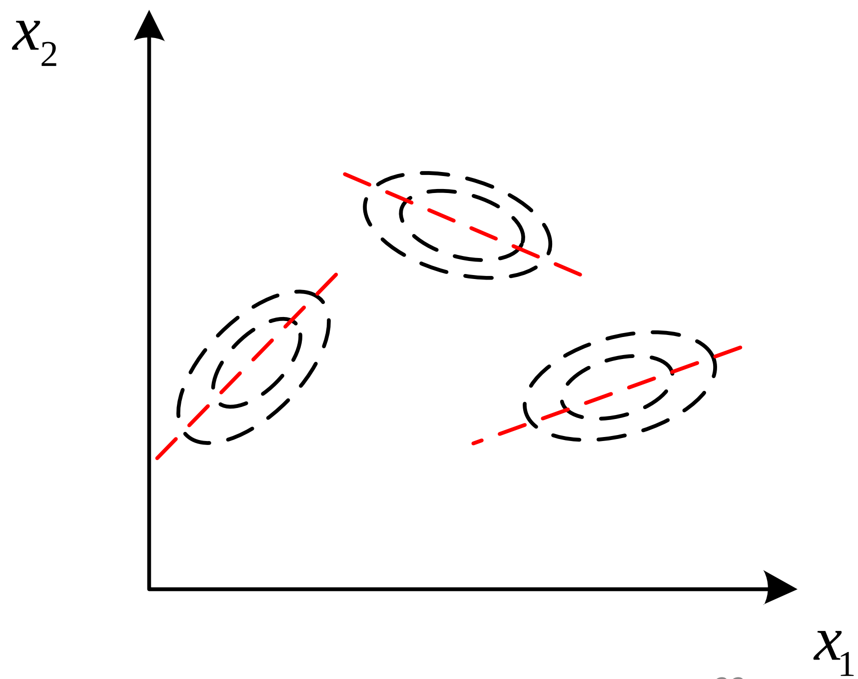
$$p(\mathbf{t})=N(0,I)$$

Conditional distribution

$$p(\mathbf{x}\,|\,\mathbf{t},\mathrm{z},\boldsymbol{\theta}_q)=N\left(P_q\mathbf{t}+\boldsymbol{\mu}_q,\sigma_q^2 I\right)$$

EM algorithm

$$\mathrm{E}\left\{\ln p\left(\mathbf{x},\mathbf{t},z\,|\,\boldsymbol{\theta}^{new}\right)_{\left\langle p\left(z,\mathbf{t}|\mathbf{x},\boldsymbol{\theta}^{old}\right)\right\rangle}\right\}$$



Extensions #4

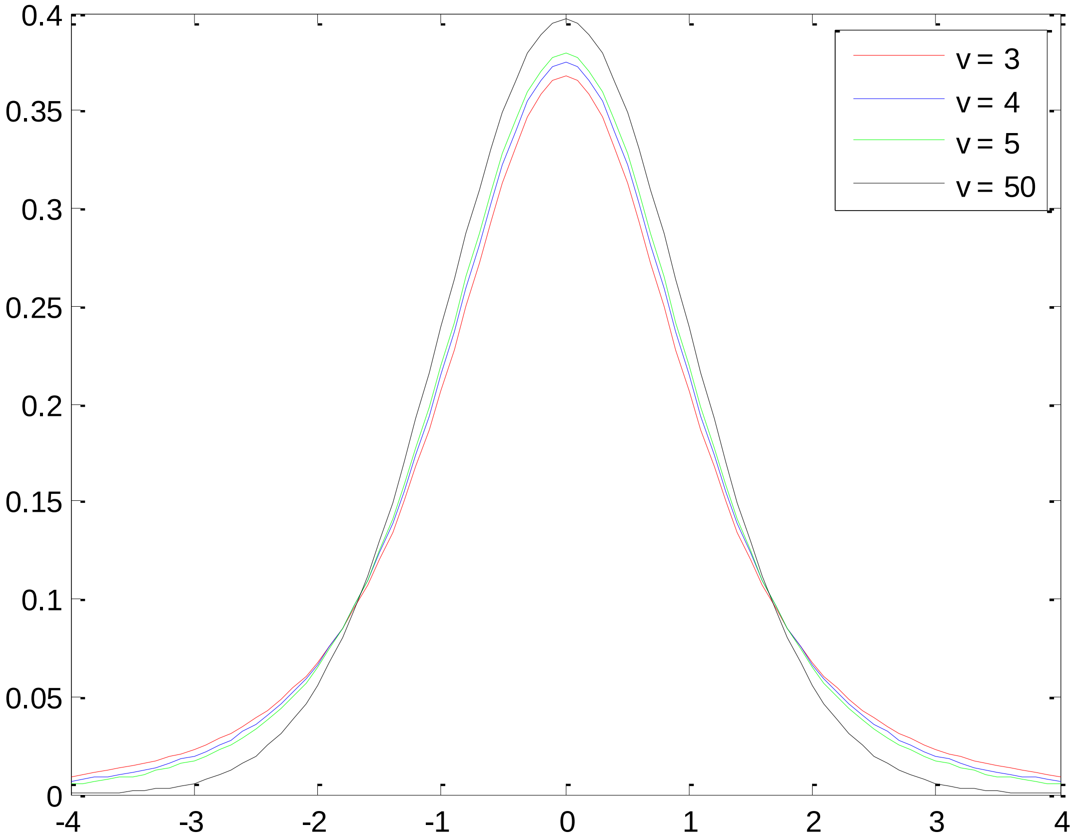
Robust PPCA）(Student t's distribution)

$$S\left(x \mid \mu, \sigma^2, v\right) \propto \left[1 + \frac{1}{v}\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-\frac{v+1}{2}}$$

μ mean

σ^2 scale factor

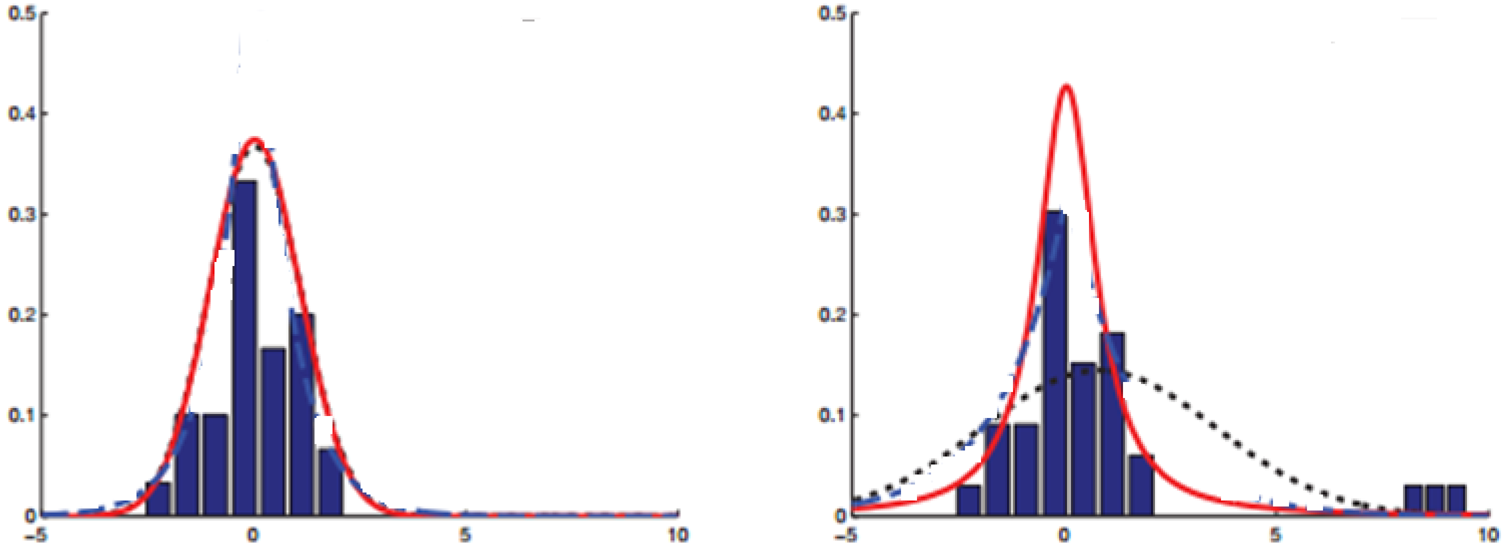
v degree of freedom



- Student t distribution will reduce to Gaussian if $v \to \infty$
- Student t distribution has a heavy tail when v is small to tolerate more outliers.

Extensions #4

Robust PPCA



$$\mathbf{x} = P_{m\times k}\mathbf{t} + \boldsymbol{\mu} + \mathbf{e}$$

$$p(\mathbf{t}) = S(\mathbf{t} \mid \mathbf{0}, I, v)$$

$$p(\mathbf{x} \mid \mathbf{t}) = S(\mathbf{x} \mid P\mathbf{t} + \boldsymbol{\mu}, \sigma^2 I, v)$$

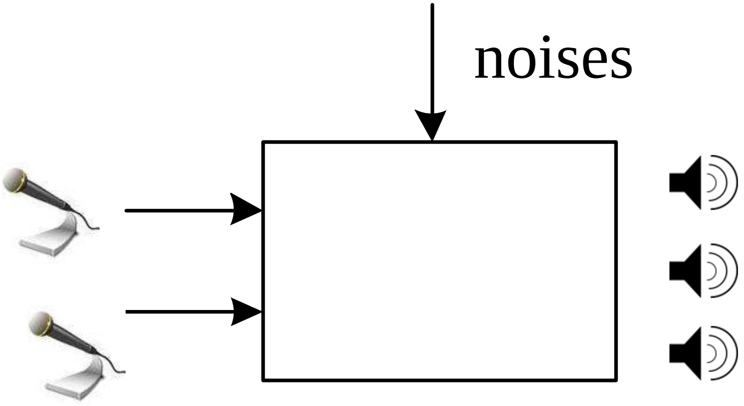
Summaries

- PPCA synthesizes the merits from both PCA and probabilistic theory and is a generalization of regular PCA.
- EM algorithm takes responsibility for almost all probabilistic latent variable models and missing data.
- Extensions are generally implemented by adjusting the distribution of latent variables or noises.
- More extensions about PPCA like locally weighted PPCA and variational Bayesian PCA.

Thanks for your attention

Latent variable models

- Latent variables sometimes have a physical meaning but cannot be measured for some reasons.



- In most cases, latent variables are abstract concepts presenting the inner states.

$$x(k+1) = Ax(k) + w(k)$$
$$y(k) = Cx(k) + v(k)$$

Parameters estimation

- There is indeed a closed-form solution, even though the objective function is very complex.

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \qquad \sigma^2 = \frac{1}{m-k}\sum_{i=k+1}^{m}\lambda_i \qquad P = U_{1:\mathrm{k}}(\Lambda_{1:k} - \sigma^2\,\mathrm{I})^{\frac{1}{2}}$$

$$S = \frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{x}_n - \overline{\mathbf{x}}\right)\left(\mathbf{x}_n - \overline{\mathbf{x}}\right)^T = U\Lambda U^T, \Lambda = diag\left(\lambda_1,\ldots,\lambda_m\right)$$

Can the PPCA capture the variance along the principal axes?

$$\mathbf{u}_i^T\left(PP^T + \sigma^2 I\right)\mathbf{u}_i = \left(\lambda_i - \sigma^2\right) + \sigma^2 = \lambda_i$$

Parameters estimation

- There is indeed a closed-form solution, even though the objective function is very complex.

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \qquad \sigma^2 = \frac{1}{m-k}\sum_{i=k+1}^{m}\lambda_i \qquad P = U_{1:\mathrm{k}}(\Lambda_{1:k} - \sigma^2 \mathrm{I})^{\frac{1}{2}}$$

$$S = \frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{x}_n - \overline{\mathbf{x}}\right)\left(\mathbf{x}_n - \overline{\mathbf{x}}\right)^T = U\Lambda U^T, \Lambda = diag\left(\lambda_1, \ldots, \lambda_m\right)$$

Can the PPCA capture the variance along the principal axes?

$$\mathbf{u}_i^T\left(PP^T + \sigma^2 I\right)\mathbf{u}_i = \left(\lambda_i - \sigma^2\right) + \sigma^2 = \lambda_i$$

What's the variance of PPCA along the residual axes?

$$\mathbf{u}_i^T\left(PP^T + \sigma^2 I\right)\mathbf{u}_i = \sigma^2$$