Machine learning

MSc Social, Cognitive, and Affective Neuroscience SoSe 2020

Prof. Dr. Dirk Ostwald

(4) Gaussian models

Overview

A generative perspective on factor analysis, PCA, and ICA.

- "Generative" here means probabilistic and model-based.

An introduction to the expectation-maximization (EM) algorithm.

- A general approach for parameter estimation in latent variable models.
- A modern take on EM from the perspective of ELBO maximization.
- A first step towards understanding variational inference.

A unifying perspective on inference and learning in probabilistic models.

- Natural generalization to HMMs, Kalman filters, Bayesian filters.
- A first steps towards understanding contemporary brain theories.

  $\rightarrow$ Free energy principle, active inference, agent-based behavioral models.

Bibliographic remarks

We discuss Gaussian models following Roweis and Ghahramani (1999). Dempster et al. (1977) provide a comprehensive introduction to the EM algorithm. The application of the EM algorithm in the context of factor analysis models is discussed in Rubin and Thayer (1982). Probabilistic PCA is reviewed in Tipping and Bishop (1999) and Roweis (1998) and dates back to work by Lawley (1953). PCA was originally proposed by Pearson (1901) and further refined by Hotelling (1933). A "model-free" introduction to independent component analysis is provided by Hyvärinen and Oja (2000) and Hyvärinen et al. (2001). Fundamental theorems for Gaussian distributions are included in an Appendix.

Foundations

Inference and learning

Factor analysis

Probabilistic principal component analysis

Principal component analysis

Independent component analysis

Foundations

Inference and learning

Factor analysis

Probabilistic principal component analysis

Principal component analysis

Independent component analysis

Definition (Multivariate Gaussian distribution)

Let X be an n-dimensional random vector with outcome set $\mathbb{R}^n$ and PDF

$$p : \mathbb{R}^n \to \mathbb{R},\ x \mapsto p(x) := (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right). \tag{1}$$

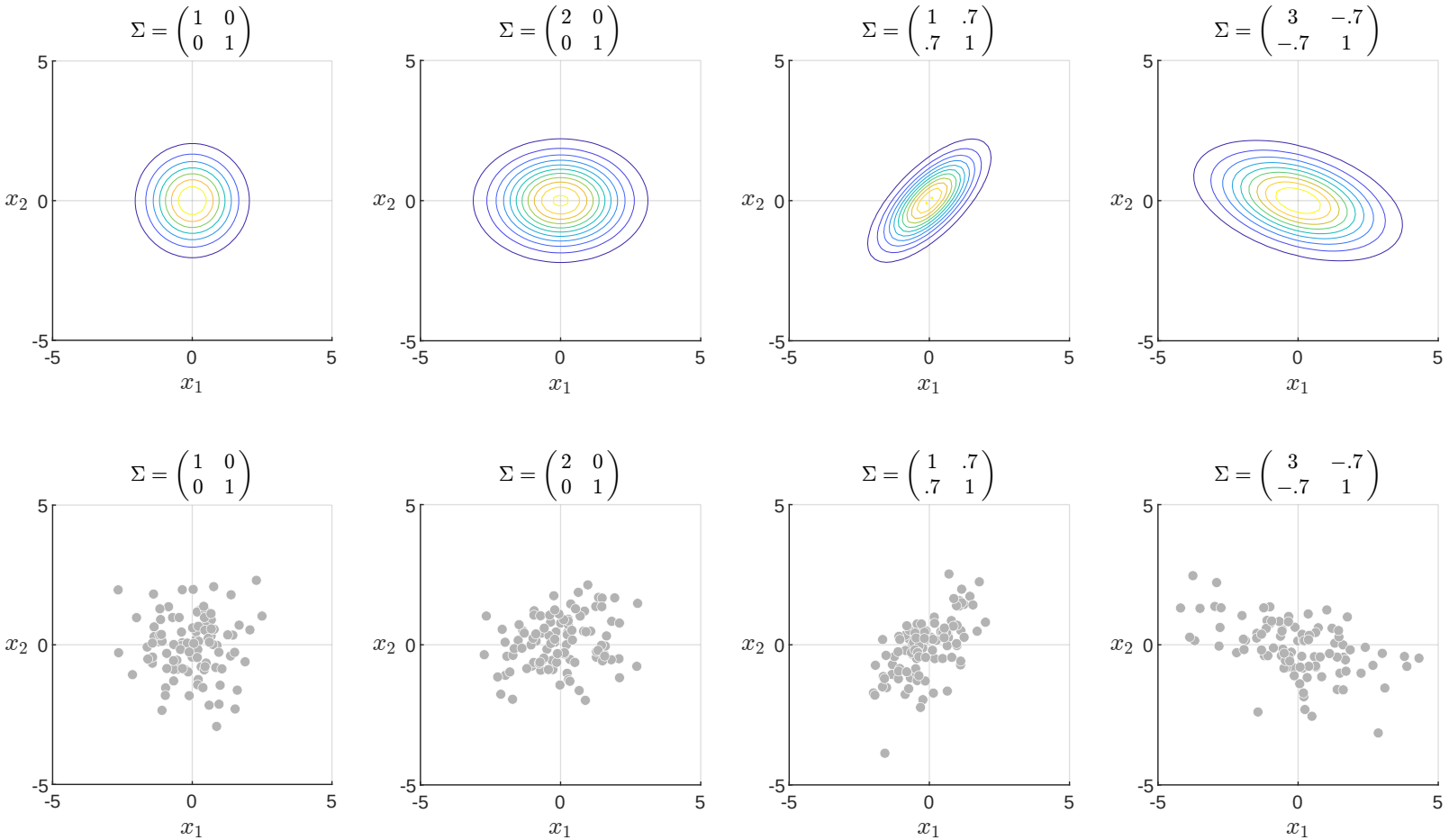
Then X is said to be distributed according to a *multivariate (or n-dimensional) Gaussian distribution* with *expectation parameter* $\mu \in \mathbb{R}^n$ and positive-definite *covariance matrix parameter* $\Sigma \in \mathbb{R}^{n \times n}$, for which we write $X \sim N(\mu, \Sigma)$. We abbreviate the PDF of a multivariate Gaussian distribution by

$$N(x; \mu, \Sigma) := (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right). \tag{2}$$

Remarks

- The parameter $\mu \in \mathbb{R}^n$ specifies the location of highest probability density in $\mathbb{R}^n$.
- The diagonal elements of Σ specify the width of the distribution w.r.t. $X_1, ..., X_n$.
- The i, jth off-diagonal element of Σ specifies the covariation of X_i and X_j.
- The term $(2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}}$ is the normalization constant for the exponential term.

Visual intuition



Definition (Linear Gaussian model)

Let x denote an unobservable k-dimensional random vector and let y denote an observable m-dimensional random vector. Then a probabilistic model of the form

$$p(x, y) = p(y|x)p(x), \tag{3}$$

where

$$p(y|x) := N(y; Bx, R) \text{ and } p(x) := N(x; 0_k, I_k) \tag{4}$$

with $B \in \mathbb{R}^{m \times k}$ and $R \in \mathbb{R}^{m \times m}$ p.d. is called a *linear Gaussian model* (LGM). The parameter set of an LGM is $\theta := \{B, R\}$ and we will write LGMs as

$$p_\theta(x, y) = N(y; Bx, R)N(x; 0_k, I_k). \tag{5}$$

Remarks

- An LGM is a special linear Gaussian state space model (LGSSM).
- x is also called *state variable* or *latent variable*, y is also called *data*.
- In hierarchical form, an LGM can be written as

$$\begin{aligned} x &= \xi & \xi &\sim N(0_k, I_k) \\ y &= Bx + \eta & \eta &\sim N(0_m, R). \end{aligned} \tag{6}$$

- ξ is called *state noise*, η is called *observation noise*.
- Sampling an LGM yields realizations $\left(x^{(i)}, y^{(i)}\right)$ for $i = 1, ..., n$.
- The $y^{(i)}$ model observed data, the $x^{(i)}$ model unobserved "virtual" data.

Theorem (LGM marginal observed data distribution)

An LGM

$$p_\theta(x, y) = N(y; Bx, R)N(x; 0_k, I_k) \tag{7}$$

generates the marginal observed data distribution (marginal likelihood)

$$p_\theta(y) = N(y; 0_m, BB^T + R). \tag{8}$$

Proof

We first note that with the LGM joint distribution theorem (Appendix), it follows directly that the joint distribution of x and y is

$$p_\theta(x, y) = N\left(\begin{pmatrix} x \\ y \end{pmatrix}; \begin{pmatrix} 0_k \\ 0_m \end{pmatrix}, \begin{pmatrix} I_k & B^T \\ B & BB^T + R \end{pmatrix}\right). \tag{9}$$

But then it follows immediately with the the same theorem, that the marginal distribution of y is

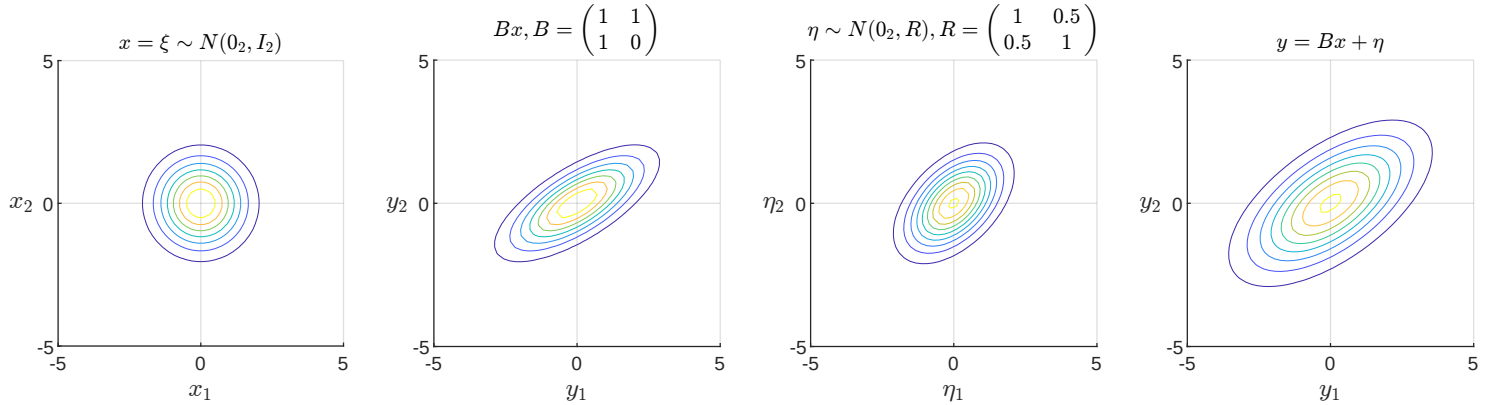
$$p_\theta(y) = N\left(y; 0_m, BB^T + R\right). \tag{10}$$

□

Remarks

- LGMs model mean-centered multivariate Gaussian distributed data sets.
- "Modelling" here means to describe the data covariance $\mathbb{C}(y)$.
- The description comes in the form of transformations of a spherical Gaussian.
- B and R mechanistically generate/explain the data covariance structure.
- B and R may offer a more parsimonious explanation than $\mathbb{C}(y)$ *per se*.
- Different LGMs have different data covariance explanation potentials.

Visual intuition



(1) The latent x is distributed according to a spherical ball in k-dimensional space.

(2) The ball is stretched and rotated into p-dimensional space by the matrix B.

(3) In p-dimensional space the ball looks like a "pancake".

(4) The pancake is convolved with the covariance of the observation noise η.

Special LGM cases

- The typical aim of LGM modelling is to capture the data covariance structure.
- The data covariance structure can be captured by adjusting B and R.
- Special LGM cases then correspond to additional constraints on R.

  $\Rightarrow$ Factor analysis constrains R to be diagonal.

  $\Rightarrow$ Probabilistic principal component analysis constrains R to be spherical.

  $\Rightarrow$ Principal component analysis constrains R to be zero.
- Independent component analysis introduces additional nonlinearities.

Definition (LGM data set joint and marginal distributions)

Let

$$Y := \begin{pmatrix} y^{(1)} & \cdots & y^{(n)} \end{pmatrix} \in \mathbb{R}^{m \times n} \text{ and } X := \begin{pmatrix} x^{(1)} & \cdots & x^{(n)} \end{pmatrix} \in \mathbb{R}^{k \times n} \tag{11}$$

denote an observed data matrix and the corresponding unobserved latent variable realization matrix, respectively. Under the assumption of the identical distribution and independence of the joint realizations $\left(x^{(i)}, y^{(i)}\right)$ for $i = 1, ..., n$, the *LGM data set joint distribution* is given by

$$p_\theta(X, Y) = \prod_{i=1}^{n} p_\theta\left(x^{(i)}, y^{(i)}\right) = \prod_{i=1}^{n} N\left(y^{(i)}; Bx^{(i)}, R\right) N\left(x^{(i)}; 0_k, I_k\right) \tag{12}$$

and the *LGM observed data set marginal distribution* is given by

$$p_\theta(Y) = \prod_{i=1}^{n} N\left(y^{(i)}; 0_m, BB^T + R\right). \tag{13}$$

Foundations

Inference and learning

Factor analysis

Probabilistic principal component analysis

Principal component analysis

Independent component analysis

Inference

Given a fixed value of the LGM parameter set θ and an observed data set Y, what is the distribution of the latent random vectors and their most likely values?

$\Rightarrow$ Conditional Gaussian distributions theorem.

Learning

Given an observed data set Y, what is the parameter value θ that maximizes the marginal likelihood function

$$L : \Theta \to \mathbb{R}_{\geq 0}, \theta \mapsto L(\theta) := p_\theta(Y) = \int p_\theta(X, Y)\, dX, \tag{14}$$

or, equivalently, the log marginal likelihood function

$$\ell : \Theta \to \mathbb{R}, \theta \mapsto \ell(\theta) := \ln p_\theta(Y) = \ln \int p_\theta(X, Y)\, dX \ ? \tag{15}$$

$\Rightarrow$ Expectation-maximization algorithm.

Theorem (LGM inference)

Let

$$p_\theta(x, y) = N(y; Bx, R)N(x; 0, I_k),$$

denote an LGM. Then the conditional distribution of x given y is

$$p_\theta(x|y) = N\left(x; B^T(BB^T + R)^{-1}y, I_k - B^T(BB^T + R)^{-1}B\right). \tag{16}$$

Remarks

- The conditional distribution of x is a Gaussian distribution.
- The most likely value of x given y is $\hat{x} := \mu_{x|y} = B^T(BB^T + R)^{-1}y$.
- The uncertainty associated with this value is $\Sigma_{x|y} := I_k - B^T(BB^T + R)^{-1}B$.

Proof

We first recall that with the LGM joint distribution theorem (Appendix), it follows directly that the joint distribution of x and y is

$$p_\theta(x,y) = N\left(\begin{pmatrix} x \\ y \end{pmatrix}; \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} I_k & B^T \\ B & BB^T + R \end{pmatrix}\right) \tag{17}$$

and that the marginal distribution of y is

$$p_\theta(y) = N\left(y; 0, BB^T + R\right). \tag{18}$$

With the conditional Gaussian distributions theorem, we then have by defining

$$\mu_x := 0, \mu_y := 0, \Sigma_{xx} := I_k, \Sigma_{xy} := B^T, \Sigma_{yx} := B, \text{ and } \Sigma_{yy} := BB^T + R \tag{19}$$

that

$$p_\theta(x|y) = N\left(x; \mu_{x|y}, \Sigma_{x|y}\right), \tag{20}$$

where

$$\mu_{x|y} = \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y) = B^T(BB^T + R)^{-1}y, \tag{21}$$

and where

$$\Sigma_{x|y} = \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx} = I_k - B^T(BB^T + R)^{-1}B. \tag{22}$$

□

Theorem (LGM data set inference)

Let $p_\theta(X, Y)$ denote an LGM data set joint distribution. Then the conditional distribution of X given Y is given by

$$p_\theta(X|Y) = \prod_{i=1}^{n} N\left(x^{(i)}; B^T(BB^T + R)^{-1}y^{(i)}, I_k - B^T(BB^T + R)^{-1}B\right). \tag{23}$$

Proof

With the LGM data set joint and marginal distributions, we have

$$p_\theta(X|Y) = \frac{p_\theta(X, Y)}{p_\theta(Y)} = \frac{\prod_{i=1}^{n} p_\theta(x^{(i)}, y^{(i)})}{\prod_{i=1}^{n} p_\theta\left(y^{(i)}\right)} = \prod_{i=1}^{n} \frac{p_\theta(x^{(i)}, y^{(i)})}{p_\theta(y^{(i)})} = \prod_{i=1}^{n} p_\theta\left(x^{(i)}|y^{(i)}\right).$$

The theorem then follows immediately with the LGM inference theorem.

□

Theorem (Evidence lower bound)

For a data set Y, let $\ln p_\theta(Y)$ denote the log data set marginal distribution of an LGM. Then for any distribution $q(X)$, it holds that

$$\ln p_\theta(Y) \geq \int q(X) \ln p_\theta(X, Y)\, dX - \int q(X) \ln q(X)\, dX =: \text{ELBO}(q(X), \theta).$$

$\text{ELBO}(q(X), \theta)$ is called the *evidence lower bound*.

Proof

With Jensen's inequality (Appendix), we have

$$\ln p_\theta(Y) := \ln \int p_\theta(X, Y)\, dX = \ln \int q(X) \left(\frac{p_\theta(X, Y)}{q(X)} \right) dX \geq \int q(X) \ln \left(\frac{p_\theta(X, Y)}{q(X)} \right) dX.$$

Hence, we also have

$$\begin{aligned} \ln p_\theta(Y) &\geq \int q(X) \ln \left(\frac{p_\theta(X, Y)}{q(X)} \right) dX \\ &= \int q(X) \left(\ln p_\theta(X, Y) - \ln q(X) \right) dX \\ &= \int q(X) \ln p_\theta(X, Y)\, dX - \int q(X) \ln q(X)\, dX. \end{aligned}$$

□

Remarks

- For a fixed data set Y, ELBO$(q(X), \theta)$ is a function of $q(X)$ and θ.
- The name evidence lower bound stems from the term "evidence" for $p_\theta(Y)$.
- In cognitive neuroimaging, the ELBO is referred to as "variational free energy".
- The importance of the ELBO extends far beyond LGMs:
  - The ELBO is at the heart of variational inference.
  - Variational inference is at the heart of contemporary brain theories.
- For introductions to variational inference, see Ostwald et al. (2014), Starke and Ostwald (2017), Blei et al. (2017), and Statistics for Data Science.

Definition (Expectation-maximization algorithm)

The iterative coordinate-wise maximization of the ELBO with respect to the distribution $q(X)$ and the parameter θ is called *expectation-maximization algorithm*. It takes the following general form:

EM algorithm

0. Initialization of $q^{(0)}(X)$ and $\theta^{(0)}$

For $k = 1, 2, ...$ until convergence

1. E Step $q^{(k)}(X) := \arg\max_{q(X)} \text{ELBO}\left(q(X), \theta^{(k-1)}\right)$
2. M Step $\theta^{(k)} := \arg\max_{\theta} \text{ELBO}\left(q^{(k)}(X), \theta\right)$

Remarks

- "E Step" is a misnomer, it is clearly also a maximization step.
- ... but the term "E Step" makes sense under exact expectation-maximization.

Theorem (Exact expectation-maximization algorithm)

Setting

$$q^{(k)}(X) := p_{\theta^{(k-1)}}(X|Y) \text{ for all } k = 1, 2, ... \tag{24}$$

in the E Step of the EM algorithm maximizes the ELBO with respect to $q(X)$ and is referred to as *exact E Step*. The ensuing algorithm takes the form

Exact EM algorithm

0. Initialization of $\theta^{(0)}$

For $k = 1, 2, ...$ until convergence

1. E Step $q^{(k)}(X) := p_{\theta^{(k-1)}}(X|Y)$
2. M Step $\theta^{(k)} := \arg\max_\theta \int p_{\theta^{(k-1)}}(X|Y) \ln p_\theta(X, Y)\, dX$

Remarks

- For LGMs, $p_{\theta^{(k-1)}}(X|Y)$ can be analytically evaluated $\Rightarrow$ Inference.
- The exact EM algorithm M Step for LGM parameter estimation $\Rightarrow$ Learning.

Proof

We first show that for $q^{(k)}(X) := p_{\theta^{(k-1)}}(X|Y)$ the ELBO assumes its maximal value $\ln p_{\theta^{(k-1)}}(Y)$:

$$\begin{aligned}
\text{ELBO}(p_{\theta^{(k-1)}}(X|Y), \theta) &= \int p_{\theta^{(k-1)}}(X|Y) \ln \left(\frac{p_{\theta^{(k-1)}}(X, Y)}{p_{\theta^{(k-1)}}(X|Y)} \right) dX \\
&= \int p_{\theta^{(k-1)}}(X|Y) \ln \left(\frac{p_{\theta^{(k-1)}}(Y) p_{\theta^{(k-1)}}(X|Y)}{p_{\theta^{(k-1)}}(X|Y)} \right) dX \\
&= \int p_{\theta^{(k-1)}}(X|Y) \ln p_{\theta^{(k-1)}}(Y) \, dX \\
&= \ln p_{\theta^{(k-1)}}(Y) \int p_{\theta^{(k-1)}}(X|Y) \, dX \\
&= \ln p_{\theta^{(k-1)}}(Y).
\end{aligned}$$

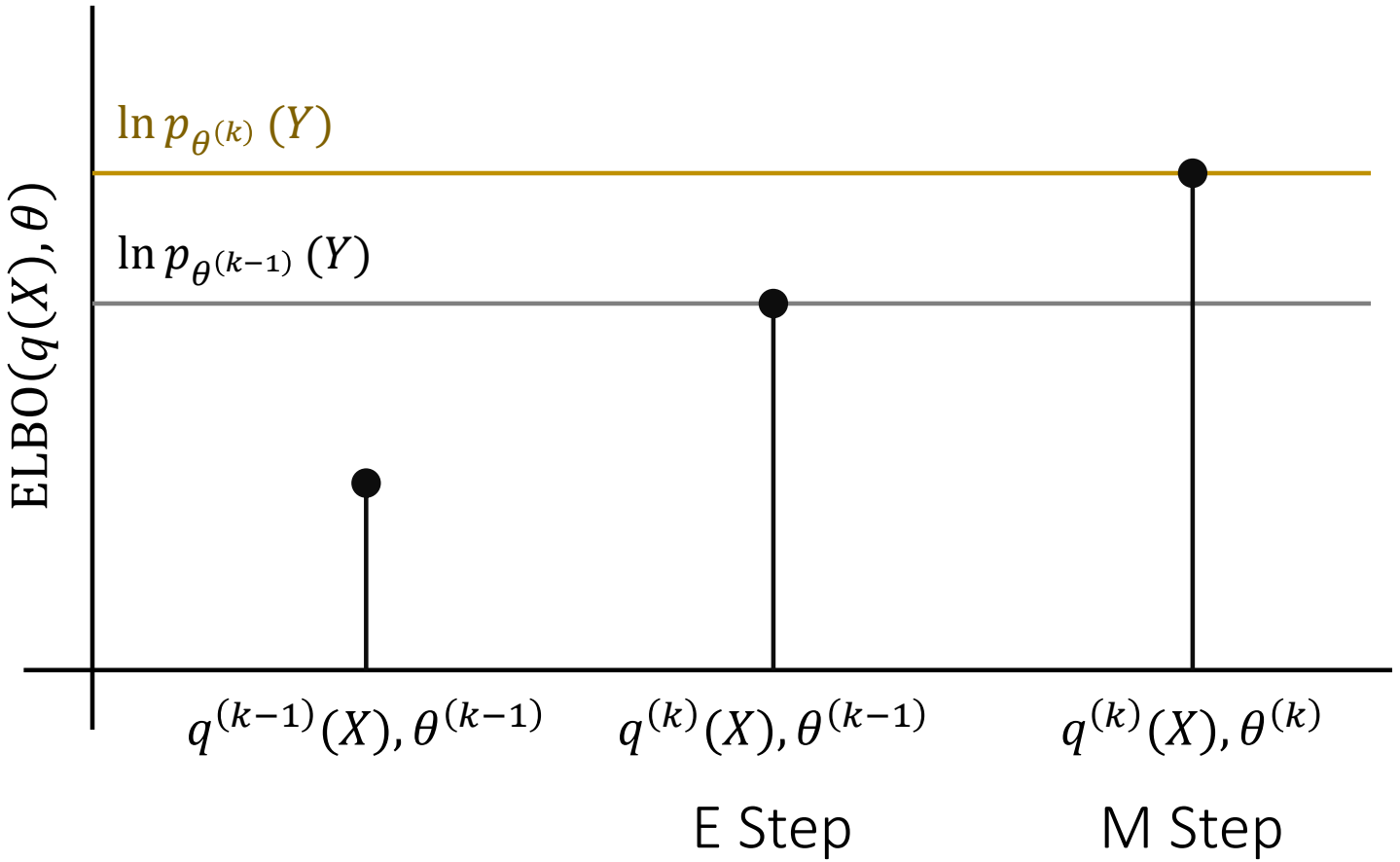
We next note that the ensuing maximization step has the form

$$\begin{aligned}
\theta^{(k)} &= \arg\max_{\theta} \text{ELBO} \left(p_{\theta^{(k-1)}}(X|Y), \theta \right) \\
&= \arg\max_{\theta} \int p_{\theta^{(k-1)}}(X|Y) \ln p_{\theta}(X, Y) \, dX - \int p_{\theta^{(k-1)}}(X|Y) \ln p_{\theta^{(k-1)}}(X|Y) \, dX.
\end{aligned}$$

Because the latter integral term on the right-hand side does not dependent on θ, but is fixed based on $\theta^{(k-1)}$, the form of the M Step of the exact EM algorithm follows immediately.

□

Visual intuition



Further remarks

- In words, the M Step on the kth iteration of the exact EM algorithm's iteration corresponds to the maximization of the expected log joint probability $p_\theta(X, Y)$ of X and Y with respect to θ, where the expectation is formed with respect to the conditional distribution of X given Y based on the parameter estimates $\theta^{(k-1)}$ obtained on the $(k-1)$th iteration of the exact EM algorithm.

- Somewhat surprisingly, via the inherent logic of the EM algorithm, maximization of the expected value

$$\mathbb{E}_{p_{\theta^{(k-1)}}(X|Y)} \left(\ln p_\theta(X, Y) \right) = \int p_{\theta^{(k-1)}}(X|Y) \ln p_\theta(X, Y) \, dX \tag{25}$$

  is thus guaranteed to also maximize (or at least keep constant) the actual quantity of interest, i.e., the marginal likelihood $\ln p_\theta(Y)$.

- For concrete algorithms and for specific LGMs, the expected value of eq. (25) has to be evaluated analytically as a function of $\theta^{(k-1)}$ and then be maximized with respect to θ either analytically or numerically to obtain the parameter estimate $\theta^{(k)}$.

Foundations

Inference and learning

Factor analysis

Probabilistic principal component analysis

Principal component analysis

Independent component analysis

Definition (Factor analysis model)

Let

$$p_\theta(x, y) = N(y; Bx, R)N(x; 0, I_k) \tag{26}$$

be an LGM with diagonal observation noise covariance matrix

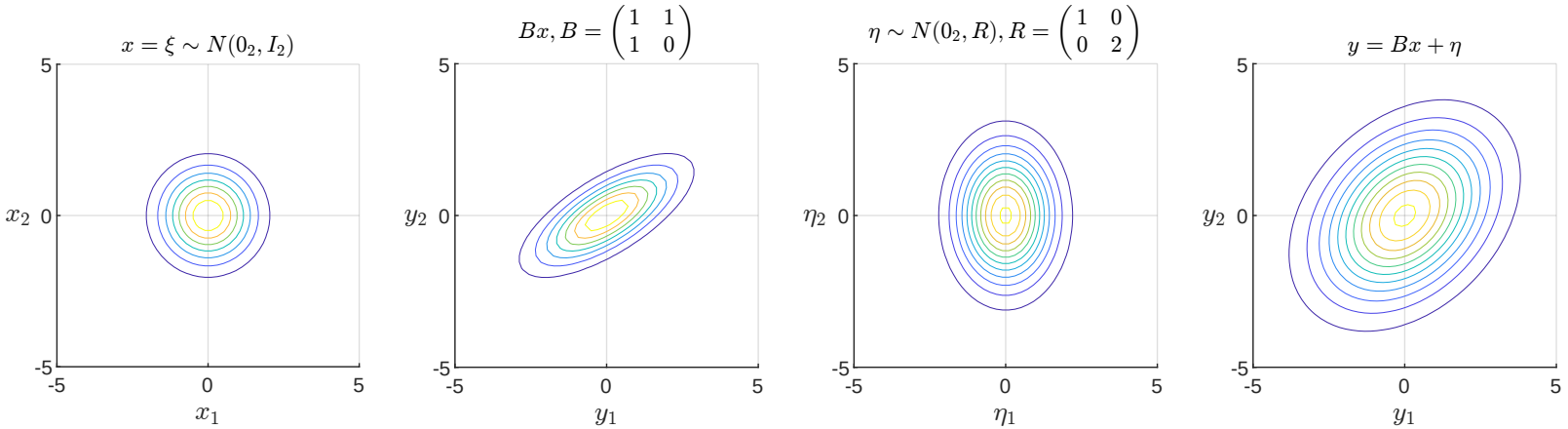
$$R = \mathsf{diag}(\sigma_1^2, \sigma_2^2, ..., \sigma_m^2) \in \mathbb{R}^{m \times m}, \sigma_i^2 > 0,\ i = 1, ..., m. \tag{27}$$

Then $p_\theta(x, y)$ is called a *factor analysis model*.

Remarks

- The latent variables $x_1, ..., x_k$ of a factor analysis model are called *factors*.
- The matrix B of a factor analysis model is called *factor loading matrix*.
- The diagonal elements of R are called *uniquenesses*.

Visual intuition



Factor analysis models attempt to explain data covariance structures by

(1) assigning all correlation structure between factors to B, and

(2) assigning all variance unique to each factor to R.

Further remarks

- Factor analysis does not treat covariance and variance identically.
- The data components $y_1, ..., y_m$ are conditionally independent given x.
- As the observation noise is assumed to be independent, the data correlational structure is assumed to be "special" or "meaningful".
- *Exploratory* and *confirmatory* factor analysis correspond to constraints on B.
- Spearman's *g factor* general intelligence theory is based on factor analysis.
- The factor analysis parameters B and R can be estimated using exact EM.

Factor analysis exact EM algorithm

0. Initialization of $B^{(0)}$ and $R^{(0)}$

For $k = 1, 2, ...$ until convergence

1. E Step

With $\tilde{B} := B^{(k-1)}$ and $\tilde{R} := R^{(k-1)}$ set

$$q^{(k)}(X) := \prod_{i=1}^{n} N\left(x^{(i)}; \hat{x}^{(i)}, \hat{\Sigma}^{(i)}\right), \tag{28}$$

where

$$\hat{x}^{(i)} := \tilde{B}^T(\tilde{B}\tilde{B}^T + \tilde{R})^{-1}y^{(i)} \text{ and } \hat{\Sigma}^{(i)} := I_k - \tilde{B}^T(\tilde{B}\tilde{B}^T + \tilde{R})^{-1}\tilde{B}. \tag{29}$$

2. M Step

Set

$$B^{(k)} := \sum_{i=1}^{n} y^{(i)}\hat{x}^{(i)^T}\left(\sum_{i=1}^{n}\hat{x}^{(i)}\hat{x}^{(i)^T} + \hat{\Sigma}^{(i)}\right)^{-1} \tag{30}$$

and

$$R^{(k)} := \frac{1}{n}\text{diag}\left(\sum_{i=1}^{n} y^{(i)}y^{(i)^T} - \sum_{i=1}^{n} y^{(i)}\hat{x}^{(i)^T}B^{(k)^T}\right). \tag{31}$$

Proof

The E Step of the algorithm follows directly with the LGM data set inference theorem. We thus focus on the derivation of the M Step. To this end, recall that the M Step of the exact EM algorithm takes the form

$$\theta^{(k)} := \arg\max_{\theta} \int p_{\theta^{(k-1)}}(X|Y) \ln p_{\theta}(X,Y)\, dX \tag{32}$$

For LGMs, the maximization of the expected joint likelihood with respect to θ can be carried out analytically and in the sense of the necessary condition for a maximum. To ease notation, we will write $\tilde{\theta} := \theta^{(k-1)}$ in the following and denote the expectation of a function f of the unobserved data X under the conditional distribution $p_{\tilde{\theta}}(X|Y)$ as a conditional expectation:

$$\mathbb{E}_{\tilde{\theta}}(f(X)|Y) := \mathbb{E}_{p_{\tilde{\theta}}(X|Y)}(f(X)) = \int f(X)\, p_{\tilde{\theta}}(X|Y)\, dX. \tag{33}$$

With these simplifications, we thus aim to evaluate

$$\frac{\partial}{\partial B}\mathbb{E}_{p_{\tilde{\theta}}(X|Y)}\left(\ln p_{\theta}(X,Y)\right) \text{ and } \frac{\partial}{\partial R}\mathbb{E}_{p_{\tilde{\theta}}(X|Y)}\left(\ln p_{\theta}(X,Y)\right), \tag{34}$$

set the results to zero, and solve for update equations for $B^{(k)}$ and $R^{(k)}$. We proceed in four steps: (1) We first use the IID data assumption and the linearity of conditional expectations and derivatives to simplify the problem of evaluating the expected data set joint likelihood and its partial derivatives for a single data point $(x^{(i)}, y^{(i)})$. We then (2) evaluate the conditional expectation and (3) evaluate the respective partial derivatives. By capitalizing on the results form the first step, we then (4) evaluate and simplify the ensuing parameter update equations.

Proof (cont.)

(1) Expected joint likelihood partial derivatives under IID data assumptions

We have

$$\begin{aligned}
\frac{\partial}{\partial B}\mathbb{E}_{p_{\tilde{\theta}}(X|Y)}\left(\ln p_\theta(X,Y)\right) &= \frac{\partial}{\partial B}\left(\mathbb{E}_{p_{\tilde{\theta}}(X|Y)}\left(\ln\prod_{i=1}^{n} p_\theta\left(x^{(i)},y^{(i)}\right)\right)\right)\\
&= \frac{\partial}{\partial B}\left(\mathbb{E}_{p_{\tilde{\theta}}(X|Y)}\left(\sum_{i=1}^{n}\ln p_\theta\left(x^{(i)},y^{(i)}\right)\right)\right)\\
&= \frac{\partial}{\partial B}\sum_{i=1}^{n}\mathbb{E}_{\prod_{i=1}^{n} p_{\tilde{\theta}}\left(x^{(i)},y^{(i)}\right)}\left(\ln p_\theta\left(x^{(i)},y^{(i)}\right)\right) \qquad (35)\\
&= \frac{\partial}{\partial B}\sum_{i=1}^{n}\mathbb{E}_{p_{\tilde{\theta}}\left(x^{(i)},y^{(i)}\right)}\left(\ln p_\theta\left(x^{(i)},y^{(i)}\right)\right)\\
&= \sum_{i=1}^{n}\frac{\partial}{\partial B}\mathbb{E}_{p_{\tilde{\theta}}\left(x^{(i)},y^{(i)}\right)}\left(\ln p_\theta\left(x^{(i)},y^{(i)}\right)\right),
\end{aligned}$$

and, following the same reasoning,

$$\frac{\partial}{\partial R}\mathbb{E}_{p_{\tilde{\theta}}(X|Y)}\left(\ln p_\theta(X,Y)\right) = \sum_{i=1}^{n}\frac{\partial}{\partial R}\mathbb{E}_{p_{\tilde{\theta}}\left(x^{(i)},y^{(i)}\right)}\left(\ln p_\theta\left(x^{(i)},y^{(i)}\right)\right). \qquad (36)$$

Proof (cont.)

(2) Expected joint likelihood for a single data point

For ease of notation, we omit the (i) superscript indexing the data realizations in this step. The expected log joint probability of x and y with respect to $p_{\tilde{\theta}}(x|y)$ then evaluates as follows:

$$
\begin{aligned}
&\mathbb{E}_{p_{\tilde{\theta}}(x|y)}\left(\ln p_\theta(x,y)\right)\\
&=\mathbb{E}_{\tilde{\theta}}\left(\ln p_\theta(x,y)|y\right)\\
&=\mathbb{E}_{\tilde{\theta}}\left(\ln\left(N(y;Bx,R)N(x;0,I_k)\right)|y\right)\\
&=\mathbb{E}_{\tilde{\theta}}\left(\ln N(y;Bx,R)+\ln N(x;0,I_k)|y\right)\\
&=\mathbb{E}\left(\ln\left((2\pi)^{-\frac{m}{2}}|R|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(y-Bx)^TR^{-1}(y-Bx)\right)\right)+\ln\left((2\pi)^{-\frac{k}{2}}|I_k|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}x^Tx\right)\right)|y\right)\\
&=\mathbb{E}_{\tilde{\theta}}\left(-\frac{m+k}{2}\ln 2\pi-\frac{1}{2}\ln|R|-\frac{1}{2}\left(y^TR^{-1}y-2y^TR^{-1}Bx+x^TB^TR^{-1}Bx\right)-\frac{1}{2}\ln 1-\frac{1}{2}x^Tx|y\right)\\
&=\mathbb{E}_{\tilde{\theta}}\left(-\frac{m+k}{2}\ln 2\pi-\frac{1}{2}\ln|R|-\frac{1}{2}y^TR^{-1}y+y^TR^{-1}Bx-\frac{1}{2}x^TB^TR^{-1}Bx-\frac{1}{2}x^Tx|y\right)\\
&=\mathbb{E}_{\tilde{\theta}}\left(-\frac{m+k}{2}\ln 2\pi-\frac{1}{2}\ln|R|-\frac{1}{2}y^TR^{-1}y+y^TR^{-1}Bx-\frac{1}{2}\mathsf{tr}\left(B^TR^{-1}Bxx^T\right)-\frac{1}{2}x^Tx|y\right)\\
&=-\frac{m+k}{2}\ln 2\pi-\frac{1}{2}\ln|R|-\frac{1}{2}y^TR^{-1}y+y^TR^{-1}B\mathbb{E}_{\tilde{\theta}}(x|y)-\frac{1}{2}\mathsf{tr}\left(B^TR^{-1}B\mathbb{E}_{\tilde{\theta}}\left(xx^T|y\right)\right)-\frac{1}{2}\mathbb{E}_{\tilde{\theta}}\left(x^Tx|y\right)
\end{aligned}
$$

where in the 7th equality, we made use of the fact that $x^TAx=\mathsf{tr}(Axx^T)$ for $x\in\mathbb{R}^n, A\in\mathbb{R}^{n\times n}$.

Proof (cont.)

(3) Partial derivatives

To evaluate the partial derivatives of the conditional expected joint likelihood with respect to the matrices B and R, we require the following identities from matrix calculus (e.g. Petersen and Pedersen, 2012):

$$\frac{\partial}{\partial X} A^T X B = A B^T, \frac{\partial}{\partial X} \mathrm{tr}(XA) = A^T, \frac{\partial}{\partial X} \mathrm{tr}(X^T A X B) = A X B + A^T X B^T, \frac{\partial}{\partial X} \ln |X| = \left(X^{-1}\right)^T . \tag{37}$$

We then have

$$\begin{aligned}
\frac{\partial}{\partial B} \mathbb{E}_{p_{\tilde{\theta}}(x|y)} \left(\ln p_\theta(x, y)\right) &= \frac{\partial}{\partial B} \left(y^T R^{-1} B \mathbb{E}_{\tilde{\theta}}(x|y) - \frac{1}{2} \mathrm{tr}\left(B^T R^{-1} B \mathbb{E}_{\tilde{\theta}}\left(x x^T | y\right)\right)\right) \\
&= \frac{\partial}{\partial B} \left(y^T R^{-1} B \mathbb{E}_{\tilde{\theta}}(x|y)\right) - \frac{1}{2} \frac{\partial}{\partial B} \mathrm{tr}\left(B^T R^{-1} B \mathbb{E}_{\tilde{\theta}}\left(x x^T | y\right)\right) \\
&= \left(y^T R^{-1}\right)^T \mathbb{E}_{\tilde{\theta}}(x|y)^T - \frac{1}{2} R^{-1} B \mathbb{E}_{\tilde{\theta}}\left(x x^T | y\right) - \frac{1}{2} \left(R^{-1}\right)^T B \mathbb{E}_{\tilde{\theta}}\left(x x^T | y\right)^T \\
&= R^{-1} y \mathbb{E}_{\tilde{\theta}}(x|y)^T - \frac{1}{2} R^{-1} B \mathbb{E}_{\tilde{\theta}}\left(x x^T | y\right) - \frac{1}{2} R^{-1} B \mathbb{E}_{\tilde{\theta}}\left(x x^T | y\right) \\
&= R^{-1} y \mathbb{E}_{\tilde{\theta}}(x|y)^T - R^{-1} B \mathbb{E}_{\tilde{\theta}}\left(x x^T | y\right) .
\end{aligned} \tag{38}$$

Proof (cont.)

Similarly, we have

$$
\begin{aligned}
&\frac{\partial}{\partial R^{-1}} \mathbb{E}_{p_{\tilde{\theta}}(x|y)} \left(\ln p_\theta(x, y)\right) \\
&= \frac{\partial}{\partial R^{-1}} \left(-\frac{1}{2} \ln |R| - \frac{1}{2} y^T R^{-1} y + y^T R^{-1} B \mathbb{E}_{\tilde{\theta}}(x|y) - \frac{1}{2} \mathsf{tr}\left(B^T R^{-1} B \mathbb{E}_{\tilde{\theta}}\left(x x^T | y\right)\right)\right) \\
&= -\frac{1}{2} \frac{\partial}{\partial R^{-1}} \ln |R| - \frac{1}{2} \frac{\partial}{\partial R^{-1}} y^T R^{-1} y + \frac{\partial}{\partial R^{-1}} y^T R^{-1} C \mathbb{E}_{\tilde{\theta}}(x|y) - \frac{1}{2} \frac{\partial}{\partial R^{-1}} \mathsf{tr}\left(R^{-1} C \mathbb{E}_{\tilde{\theta}}\left(x x^T | y\right) B^T\right) \\
&= \frac{1}{2} R - \frac{1}{2} y y^T + y \left(B \mathbb{E}_{\tilde{\theta}}(x|y)\right)^T - \frac{1}{2} \left(B \mathbb{E}_{\tilde{\theta}}\left(x x^T | y\right) B^T\right)^T \\
&= \frac{1}{2} R - \frac{1}{2} y y^T + y \mathbb{E}_{\tilde{\theta}}(x|y)^T B^T - \frac{1}{2} B \mathbb{E}_{\tilde{\theta}}\left(x x^T | y\right) B^T .
\end{aligned}
$$

(4) Parameter update equations

Re-substitution then yields for the partial derivative with respect to B

$$
\begin{aligned}
\frac{\partial}{\partial B} \mathbb{E}_{p_{\tilde{\theta}}(X|Y)} \left(\ln p_\theta(X, Y)\right) &= \sum_{i=1}^{n} \frac{\partial}{\partial B} \mathbb{E}_{p_{\tilde{\theta}}\left(x^{(i)}, y^{(i)}\right)} \left(\ln p_\theta\left(x^{(i)}, y^{(i)}\right)\right) \\
&= \sum_{i=1}^{n} R^{-1} y^{(i)} \mathbb{E}_{\tilde{\theta}}\left(x^{(i)} | y^{(i)}\right)^T - R^{-1} B \mathbb{E}_{\tilde{\theta}}\left(x^{(i)} x^{(i)^T} | y^{(i)}\right) \qquad (39) \\
&= R^{-1} \sum_{i=1}^{n} y^{(i)} \mathbb{E}_{\tilde{\theta}}\left(x^{(i)} | y^{(i)}\right)^T - R^{-1} B \sum_{i=1}^{n} \mathbb{E}_{\tilde{\theta}}\left(x^{(i)} x^{(i)^T} | y^{(i)}\right) .
\end{aligned}
$$

Proof (cont.)

Setting to zero and solving for $B^{(k)}$ then yields

$$\begin{aligned} R^{-1} B^{(k)} \sum_{i=1}^{n} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} x^{(i)^T} | y^{(i)} \right) &= R^{-1} \sum_{i=1}^{n} y^{(i)} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} | y^{(i)} \right)^T . \\ \Leftrightarrow B^{(k)} &= \sum_{i=1}^{n} y^{(i)} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} | y^{(i)} \right)^T \left(\sum_{i=1}^{n} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} x^{(i)^T} | y^{(i)} \right) \right)^{-1} . \end{aligned} \tag{40}$$

Similarly, re-substitution yields for the partial derivative with respect to R

$$\begin{aligned} & \frac{\partial}{\partial R^{-1}} \mathbb{E}_{p_{\tilde{\theta}}(X|Y)} \left(\ln p_\theta(X, Y) \right) \\ &= \sum_{i=1}^{n} \frac{\partial}{\partial R^{-1}} \mathbb{E}_{p_{\tilde{\theta}}\left(x^{(i)}, y^{(i)}\right)} \left(\ln p_\theta \left(x^{(i)}, y^{(i)} \right) \right) \\ &= \sum_{i=1}^{n} \frac{1}{2} R - \frac{1}{2} y^{(i)} y^{(i)^T} + y^{(i)} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} | y^{(i)} \right)^T B^T - \frac{1}{2} B \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} x^{(i)^T} | y^{(i)} \right) B^T \\ &= \frac{n}{2} R - \frac{1}{2} \sum_{i=1}^{n} y^{(i)} y^{(i)^T} + \sum_{i=1}^{n} y^{(i)} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} | y^{(i)} \right)^T B^T - \frac{1}{2} B \sum_{i=1}^{n} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} x^{(i)^T} | y^{(i)} \right) B^T . \end{aligned} \tag{41}$$

Proof (cont.)

Setting to zero and solving for $R^{(k)}$ then yields

$$\frac{n}{2} R^{(k)} = \frac{1}{2} \sum_{i=1}^{n} y^{(i)} y^{(i)T} - \sum_{i=1}^{n} y^{(i)} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} | y^{(i)} \right)^T B^T + \frac{1}{2} B \sum_{i=1}^{n} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} x^{(i)T} | y^{(i)} \right) B^T$$

$$R^{(k)} = \frac{1}{n} \sum_{i=1}^{n} y^{(i)} y^{(i)T} - \frac{2}{n} \sum_{i=1}^{n} y^{(i)} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} | y^{(i)} \right)^T B^T + \frac{1}{n} B \sum_{i=1}^{n} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} x^{(i)T} | y^{(i)} \right) B^T$$

Substitution of the update equation for B further yields

$$\begin{aligned} R^{(k)} =& \frac{1}{n} \sum_{i=1}^{n} y^{(i)} y^{(i)T} - \frac{2}{n} \sum_{i=1}^{n} y^{(i)} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} | y^{(i)} \right)^T B^{(k)T} \\ & + \frac{1}{n} \sum_{i=1}^{n} y^{(i)} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} | y^{(i)} \right)^T \left(\sum_{i=1}^{n} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} x^{(i)T} | y^{(i)} \right) \right)^{-1} \sum_{i=1}^{n} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} x^{(i)T} | y^{(i)} \right) B^{(k)T} \\ =& \frac{1}{n} \left(\sum_{i=1}^{n} y^{(i)} y^{(i)T} - 2 \sum_{i=1}^{n} y^{(i)} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} | y^{(i)} \right)^T B^{(k)T} + \sum_{i=1}^{n} y^{(i)} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} | y^{(i)} \right)^T B^{(k)T} \right) \\ =& \frac{1}{n} \left(\sum_{i=1}^{n} y^{(i)} y^{(i)T} - \sum_{i=1}^{n} y^{(i)} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} | y^{(i)} \right)^T B^{(k)T} \right) \end{aligned}$$

Factor analysis

Proof (cont.)

We thus obtained the parameter update equations

$$B^{(k)} = \sum_{i=1}^{n} y^{(i)} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} | y^{(i)} \right)^T \left(\sum_{i=1}^{n} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} x^{(i)^T} | y^{(i)} \right) \right)^{-1}$$

$$R^{(k)} = \frac{1}{n} \left(\sum_{i=1}^{n} y^{(i)} y^{(i)^T} - \sum_{i=1}^{n} y^{(i)} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} | y^{(i)} \right)^T B^{(k)^T} \right)$$

The computational forms of these update equations can be further simplified by noting that with

$$\mathbb{E}\left(x x^T | y \right) = \mu_{x|y} \mu_{x|y}^T + \Sigma_{x|y} \tag{42}$$

and the LGM inference theorem it holds that

$$\mathbb{E}_{\tilde{\theta}} \left(x^{(i)} | y^{(i)} \right) = \hat{x}^{(i)} \text{ and } \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} x^{(i)^T} | y^{(i)} \right) = \hat{x}^{(i)} \hat{x}^{(i)^T} + \hat{\Sigma}^{(i)} . \tag{43}$$

We thus obtain

$$B^{(k)} = \sum_{i=1}^{n} y^{(i)} \hat{x}^{(i)^T} \left(\sum_{i=1}^{n} \hat{x}^{(i)} \hat{x}^{(i)^T} + \hat{\Sigma}^{(i)} \right)^{-1}$$

$$R^{(k)} = \frac{1}{n} \left(\sum_{i=1}^{n} y^{(i)} y^{(i)^T} - \sum_{i=1}^{n} y^{(i)} \hat{x}^{(i)^T} B^{(k)^T} \right) .$$

Finally, enforcing the diagonality constraint on R can be achieved by setting

$$R^{(k)} := -\frac{1}{n} \text{diag} \left(\sum_{i=1}^{n} y^{(i)} y^{(i)^T} + \sum_{i=1}^{n} y^{(i)} \hat{x}^{(i)^T} B^{(k)^T} \right) \tag{44}$$

Foundations

Inference and learning

Factor analysis

Probabilistic principal component analysis

Principal component analysis

Independent component analysis

Definition (Probabilistic principal component analysis model)

Let

$$p_\theta(x, y) = N(y; Bx, R)N(x; 0_k, I_k) \tag{45}$$

be an LGM with spherical observation noise covariance matrix

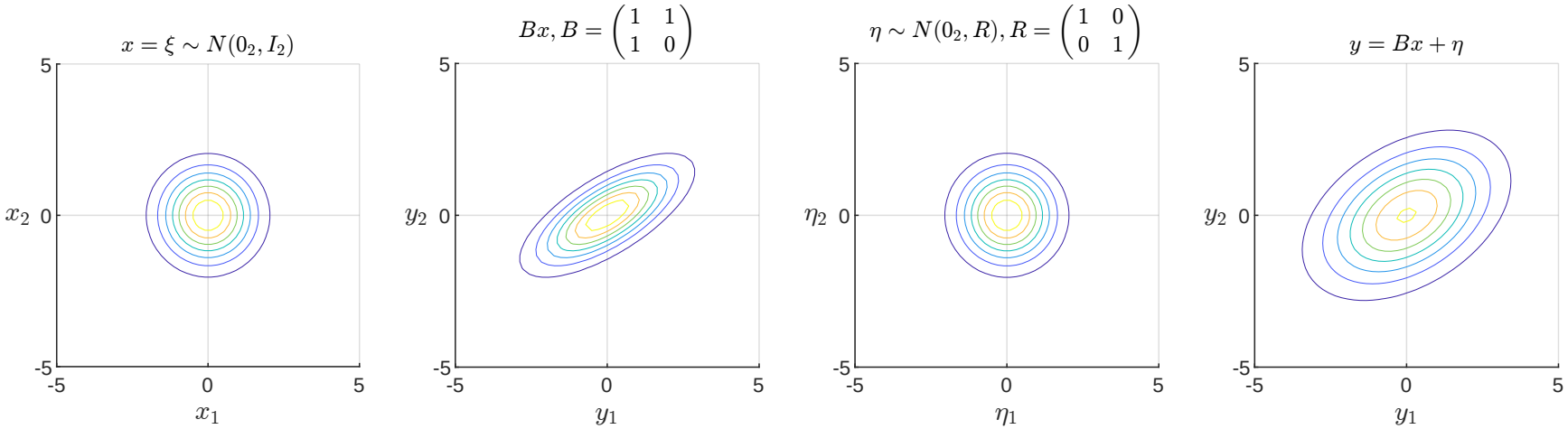
$$R := \sigma^2 I_m. \tag{46}$$

Then $p_\theta(x, y)$ is called a *probabilistic principal component analysis model (PPCA)*.

Remarks

- B establishes a relationship with classical PCA.
- σ^2 is referred to as the *global noise level*.
- B and σ^2 can be estimated by EM or direct marginal likelihood maximization.

PPCA model visual intuition



Theorem (PPCA model parameter)

Let

$$p_\theta(x, y) = N(y; Bx, \sigma^2 I_m) N(x; 0, I_k) \tag{47}$$

be a PPCA model and let

$$\mathbb{C}(y) = Q \Lambda Q^T \tag{48}$$

be the orthonormal decomposition of the covariance matrix of its associated marginal data distribution. Then the PPCA model parameter B can be written as

$$B = Q \left(\Lambda - \sigma^2 I_m\right)^{1/2}. \tag{49}$$

Remarks

- We called $\mathbb{C}(y) = Q \Lambda Q^T$ the principal component analysis of $\mathbb{C}(y)$.
- Q comprises the eigenvectors of $\mathbb{C}(y)$, Λ comprises the associated eigenvalues.
- We called the eigenvectors of $\mathbb{C}(y)$ the principal components of $\mathbb{C}(y)$.
- The columns of B are the principal components weighted by $\left(\Lambda - \sigma^2 I_m\right)^{1/2}$.

Proof

We first note that the marginal data distribution of the PPCA model is given by

$$p_\theta(y) = N(y; 0_m, BB^T + \sigma^2 I_m) \text{ and thus } \mathbb{C}(y) = BB^T + \sigma^2 I_m. \tag{50}$$

Substitution of $B = Q(\Lambda - \sigma^2 I_m)^{1/2}$ then yields

$$\begin{aligned}
\mathbb{C}(y) &= BB^T + \sigma^2 I_m \\
&= Q(\Lambda - \sigma^2 I_m)^{1/2}(Q(\Lambda - \sigma^2 I_m)^{1/2})^T + \sigma^2 I_m \\
&= Q(\Lambda - \sigma^2 I_m)^{1/2}((\Lambda - \sigma^2 I_m)^{1/2})^T Q^T + \sigma^2 I_m \\
&= Q(\Lambda - \sigma^2 I_m)^{1/2}(\Lambda - \sigma^2 I_m)^{1/2} Q^T + \sigma^2 I_m \\
&= Q(\Lambda - \sigma^2 I_m) Q^T + \sigma^2 I_m \\
&= (Q\Lambda - \sigma^2 Q I_m) Q^T + \sigma^2 I_m \\
&= Q\Lambda Q^T - \sigma^2 Q Q^T + \sigma^2 I_m \\
&= Q\Lambda Q^T - \sigma^2 I_m + \sigma^2 I_m \\
&= Q\Lambda Q^T.
\end{aligned} \tag{51}$$

Hence, we have the equivalency

$$\mathbb{C}(y) = Q\Lambda Q^T \Leftrightarrow B = Q(\Lambda - \sigma^2 I_m)^{\frac{1}{2}}. \tag{52}$$

□

PPCA exact EM algorithm

0. Initialization of $B^{(0)}$ and $R^{(0)}$

For $k = 1, 2, ...$ until convergence

1. E Step

With $\tilde{B} := B^{(k-1)}$ and $\tilde{R} := R^{(k-1)}$ set

$$q^{(k)}(X) := \prod_{i=1}^{n} N\left(x^{(i)}; \hat{x}^{(i)}, \hat{\Sigma}^{(i)}\right), \tag{53}$$

where

$$\hat{x}^{(i)} := \tilde{B}^T(\tilde{B}\tilde{B}^T + \tilde{R})^{-1}y^{(i)} \text{ and } \hat{\Sigma}^{(i)} := I_k - \tilde{B}^T(\tilde{B}\tilde{B}^T + \tilde{R})^{-1}\tilde{B}. \tag{54}$$

2. M Step

Set

$$B^{(k)} := \sum_{i=1}^{n} y^{(i)}\hat{x}^{(i)^T}\left(\sum_{i=1}^{n}\hat{x}^{(i)}\hat{x}^{(i)^T} + \hat{\Sigma}^{(i)}\right)^{-1} \tag{55}$$

and

$$R^{(k)} := \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{i=1}^{n} y^{(i)}y^{(i)^T} - \sum_{i=1}^{n} y^{(i)}\hat{x}^{(i)^T}B^{(k)^T}\right)_{ii} I_m. \tag{56}$$

Theorem (Direct marginal maximum likelihood estimation)

Let

$$p_\theta(x, y) = N(y; Bx, \sigma^2 I_m) N(x; 0, I_k) \tag{57}$$

denote a PPCA model and let $Y \in \mathbb{R}^{m \times n}$ denote a data set matrix obtained under IID sampling from the PPCA model. Let further

$$C = \frac{1}{n} Y Y^T \text{ and } C = Q \Lambda Q^T \tag{58}$$

denote the biased empirical data covariance matrix and its orthonormal decomposition, respectively. Finally, let $Q_q \in \mathbb{R}^{m \times q}$ and $\Lambda_q \in \mathbb{R}^{q \times q}$ denote the matrices created by columnwise concatenation of the leading eigenvectors of C (i.e., the eigenvectors with the largest associated eigenvalues) and their associated eigenvalues, respectively. Then maximum marginal likelihood estimators of B and σ^2 are given by

$$\hat{B} = Q_q (\Lambda_q - \sigma^2 I_m)^{1/2} \text{ and } \hat{\sigma}^2 = \frac{1}{m - l} \sum_{j=l+1}^{m} \lambda_j, \tag{59}$$

respectively.

Proof

We only show that $\hat{B}$ as defined in the theorem corresponds to maximum of the log marginal likelihood function. To this end, we closely follow the respective proof in **?**, and proceed in three steps: (1) We first rewrite the marginal data set log likelihood function of the PPCA model in a suitable manner. (2) We then evaluate its gradient with respect to B and (3) finally evaluate the resulting maximum marginal likelihood estimator.

(1) Log likelihood function

We first rewrite the marginal data set log likelihood function of the PPCA model

$$\ell : \Theta \to \mathbb{R}, \theta \mapsto \ell(\theta) := \ln p_\theta(Y) = \sum_{i=1}^{n} \ln N\left(y^{(i)}; 0_m, BB^T + \sigma^2 I_m\right) \tag{60}$$

with $\theta := \{B, \sigma^2\}$ in a way more amenable to direct maximization.
With the definitions of

$$\Sigma := BB^T + \sigma^2 I_m \text{ and } C := \frac{1}{n} \sum_{i=1}^{n} y^{(i)} y^{(i)^T} \tag{61}$$

and the trace operator properties

$$x^T A x = \mathrm{tr}(A x x^T) \text{ and } \mathrm{tr}(A) + \mathrm{tr}(B) = \mathrm{tr}(A + B), \tag{62}$$

(...)

Proof

(...) we have

$$
\begin{aligned}
\ln p_\theta(Y) &= \sum_{i=1}^{n} \ln N\left(y^{(i)}; 0_m, \Sigma\right) \\
&= \sum_{i=1}^{n} \ln\left((2\pi)^{-\frac{m}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} {y^{(i)}}^T \Sigma^{-1} y^{(i)}\right)\right) \\
&= -\frac{n}{2}\left(m \ln 2\pi + \ln|\Sigma| + \sum_{i=1}^{n} \frac{1}{n} {y^{(i)}}^T \Sigma^{-1} y^{(i)}\right) \\
&= -\frac{n}{2}\left(m \ln 2\pi + \ln|\Sigma| + \sum_{i=1}^{n} \mathsf{tr}\left(\Sigma^{-1} \frac{1}{n} y^{(i)} {y^{(i)}}^T\right)\right) \\
&= -\frac{n}{2}\left(m \ln 2\pi + \ln|\Sigma| + \mathsf{tr}\left(\sum_{i=1}^{n} \Sigma^{-1} \frac{1}{n} y^{(i)} {y^{(i)}}^T\right)\right) \\
&= -\frac{n}{2}\left(m \ln 2\pi + \ln|\Sigma| + \mathsf{tr}\left(\Sigma^{-1} \frac{1}{n} \sum_{i=1}^{n} y^{(i)} {y^{(i)}}^T\right)\right) \\
&= -\frac{n}{2}\left(m \ln 2\pi + \ln|\Sigma| + \mathsf{tr}\left(\Sigma^{-1} C\right)\right)
\end{aligned}
\tag{63}
$$

Proof

(2) Gradient of the log marginal likelihood function

With

$$\frac{\partial}{\partial X} \ln |X| = \left(X^{-1}\right)^T, \; \frac{\partial}{\partial X} X X^T = 2X, \text{ and } \frac{\partial}{\partial X} \mathrm{tr}(XA) = A^T, \tag{64}$$

the gradient of the log marginal likelihood function with respect to B evaluates to

$$\begin{aligned}
\frac{\partial}{\partial B} \ell(\theta) &= -\frac{n}{2} \frac{\partial}{\partial B} \left(m \ln 2\pi + \ln |BB^T + \sigma^2 I_m| + \mathrm{tr}\left(\left(BB^T + \sigma^2 I_m\right)^{-1} C \right) \right) \\
&= -\frac{n}{2} \frac{\partial}{\partial B} \ln |BB^T + \sigma^2 I_m| - \frac{n}{2} \frac{\partial}{\partial B} \mathrm{tr}\left(\left(BB^T + \sigma^2 I_m\right)^{-1} C \right) \\
&= -\frac{n}{2} 2 \left(\left(BB^T + \sigma^2 I_m\right)^{-1} B \right)^T + \frac{n}{2} 2 \left(BB^T + \sigma^2 I_m\right)^{-1} C \left(BB^T + \sigma^2 I_m\right)^{-1} B \\
&= n \left(-\Sigma^{-1} B + \Sigma^{-1} C \Sigma^{-1} B \right) \\
&= n \left(\Sigma^{-1} C \Sigma^{-1} B - \Sigma^{-1} B \right).
\end{aligned} \tag{65}$$

Proof

(3) Maximum marginal likelihood estimator evaluation

Setting the gradient of ℓ with respect to B to zero then yields

$$\Sigma^{-1}C\Sigma^{-1}\hat{B} - \Sigma^{-1}\hat{B} = 0 \Leftrightarrow \Sigma^{-1}\hat{B} = \Sigma^{-1}C\Sigma^{-1}\hat{B} \Leftrightarrow \hat{B} = C\Sigma^{-1}\hat{B}. \tag{66}$$

We consider solutions of this necessary condition for a stationary point of the log marginal likelihood function with $B \neq 0$ and $\Sigma \neq C$. To find these, we first express $\hat{B}$ in terms of its singular value decomposition

$$\hat{B} = ULV^T \tag{67}$$

where $U = (u_1, u_2, ..., u_l)$ is an $m \times q$ matrix of orthonormal column vectors, $L = \text{diag}(l_1, l_2, ..., l_q)$ is a $q \times q$ diagonal matrix of singular values, and V is a $q \times q$ orthogonal matrix. Substitution in the necessary condition for a stationary point then yields

$$CUL = U(\sigma^2 I_m + L^2)L. \tag{68}$$

For $l_j \neq 0$, eq. (68) implies that

$$Cu_j = (\sigma^2 + l_j^2)u_j \tag{69}$$

Hence, each column of U must be an eigenvector of C with corresponding eigenvalue $\lambda_j = \sigma^2 + l_j$, and thus

$$\lambda_j = \sigma^2 + l_j^2 \Leftrightarrow l_j^2 = \lambda_j - \sigma^2 \Leftrightarrow l_j = (\lambda_j - \sigma^2)^{\frac{1}{2}}. \tag{70}$$

Proof

(3) Maximum marginal likelihood estimator evaluation

For $l_j = 0$, u_j is arbitrary. Under the assumption that $l_j \neq 0$ for $j = 1, ..., m$, all potential solutions for $\hat{B}$ can thus be written in the form

$$\hat{B} = U_q \left(\Lambda_q - \sigma^2 I_m \right)^{\frac{1}{2}} R, \tag{71}$$

where U_q is a $m \times q$ matrix whose q columns are the eigenvectors of C, Λ_q is the diagonal matrix of the corresponding eigenvalues, and R is an arbitrary $q \times q$ orthogonal matrix, for example, $R = I_q$.

□

Foundations

Inference and learning

Factor analysis

Probabilistic principal component analysis

Principal component analysis

Independent component analysis

Definition (Principal component analysis model)

Let

$$p_\theta(x, y) = N(y; Bx, R)N(x; 0, I_k) \tag{72}$$

be an LGM with a limiting zero covariance matrix

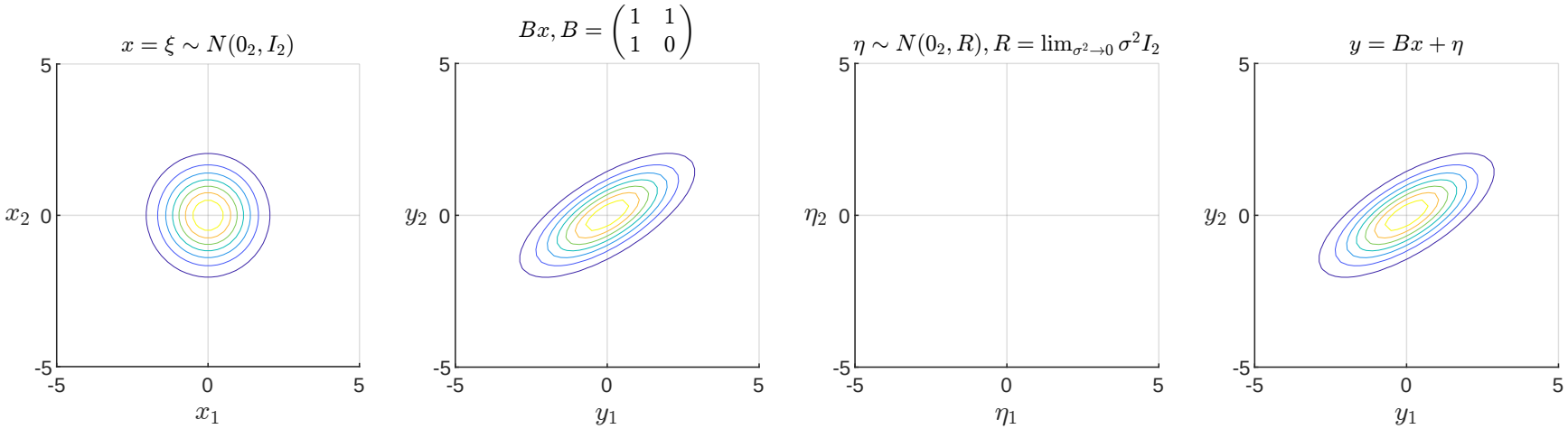
$$R := \lim_{\sigma^2 \to 0} \sigma^2 I_m \in \mathbb{R}^{m \times m}. \tag{73}$$

Then $p_\theta(x, y)$ is called a *principal component analysis model*.

Remarks

- B establishes a relationship with classical PCA.
- The observation noise is assumed to be zero.

PCA model visual intuition



Theorem (PCA model parameter)

Let

$$p_\theta(x, y) = N\left(y; Bx, \lim_{\sigma^2 \to 0} \sigma^2 I_m\right) N(x; 0_k, I_k) \tag{74}$$

be a PCA model and let

$$\mathbb{C}(y) = Q^T \Lambda Q \tag{75}$$

be the principal component analysis of its associated marginal data distribution. Then

$$B = Q\Lambda^{\frac{1}{2}}. \tag{76}$$

Proof

We first note that the marginal data distribution of the PCA model in the limit of $\sigma^2 \to 0$ given by

$$p_\theta(y) = N(y; 0_m, BB^T) \text{ and thus } \mathbb{C}(y) = BB^T. \tag{77}$$

We thus have

$$\mathbb{C}(y) = BB^T \Leftrightarrow Q\Lambda Q^T = BB^T \Leftrightarrow B = Q\Lambda^{\frac{1}{2}}. \tag{78}$$

□

Foundations

Inference and learning

Factor analysis

Probabilistic principal component analysis

Principal component analysis

Independent component analysis

Definition (Independent component analysis model)

Let x denote unobservable m-dimensional random vectors, let y denote an observable m-dimensional random vector, and let

$$g : \mathbb{R}^m \to \mathbb{R}^m, \xi \mapsto g(\xi) := \begin{pmatrix} g_1(\xi_1) \\ \vdots \\ g_m(\xi_m) \end{pmatrix} \text{ with } g_i := \gamma : \mathbb{R} \to \mathbb{R} \text{ for } i = 1, ..., m \tag{79}$$

denote a differentiable and bijective multivariate vector-valued function that operates componentwise on its argument. Then a probabilistic model of the form

$$p_\theta(x, y) = p(y|x)p(x), \tag{80}$$

where

$$p(x) = \frac{1}{|\det J^g(g^{-1}(x))|} N\left(g^{-1}(x); 0, I_m\right) \tag{81}$$

and

$$p(y|x) = N(y; Bx, R),\ B \in \mathbb{R}^{m \times m}, R = \lim_{\sigma^2 \to 0} \sigma^2 I_m \tag{82}$$

is called an *independent component analysis (ICA) model*.

Remarks

- In hierarchical form, the ICA model can be written as

$$\begin{aligned} x &= g(\xi), \qquad \xi \sim N(0, I_m),\ g : \mathbb{R}^m \to \mathbb{R}^m \\ y &= Bx + \eta, \quad \eta \sim N(0, R),\ R := \lim_{\sigma^2 \to 0} \sigma^2 I_m \end{aligned} \tag{83}$$

- The components of x are referred to as *independent components* (ICs).
- With $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^m$, there are as many ICs as data vector components.
- g is called *generative nonlinearity*.
- B is called *mixing matrix* and $W := B^{-1}$ is called *unmixing* matrix.
- The parameter of an ICA model is the unmixing matrix $\theta = \{W\}$.
- The form of $p(x)$ in eq. (81) results from the multivariate PDF transform theorem.
- ICA can be interpreted as linear generative model with nongaussian latent variable.
- ICA can be interpreted as nonlinear generative model with Gaussian latent variable.

Theorem (The multivariate probability density function transform)

Let ξ be an m-dimensional random vector with PDF $p_\xi(\xi)$, let $x = g(\xi)$ be an m-dimensional random vector for a differentiable and bijective multivariate vector-valued function $g : \mathbb{R}^m \to \mathbb{R}^m$, and let $g^{-1} : \mathbb{R}^m \to \mathbb{R}^m$ with $g^{-1}(x) = \xi$ denote the inverse of g. Let further

$$J^g(\xi) = \left(\frac{\partial}{\partial \xi_j} g_i(\xi) \right)_{i,j=1,\dots,m} \in \mathbb{R}^{m \times m} \tag{84}$$

denote the Jacobian of g at $\xi \in \mathbb{R}^m$, let $\det J^g(\xi)$ denote its determinant, and assume that $\det J^g(\xi) \neq 0$ for all $\xi \in \mathbb{R}^m$. Then the PDF of x is given by

$$p_x(x) := \begin{cases} \frac{1}{|\det J^g\left(g^{-1}(x)\right)|} p_\xi\left(g^{-1}(x)\right) & \text{for } x \in g(\mathbb{R}^m) \\ 0 & \text{for } x \in \mathbb{R}^m \setminus g(\mathbb{R}^k) \end{cases}. \tag{85}$$

Remark

- A formula for computing the PDF p_x of x if $x := g(\xi)$ with $\xi \sim p_\xi$.

Theorem (Independent components are independent)

Let

$$p_\theta(x, y) = p(y|x)p(x) \tag{86}$$

denote an ICA model for two m-dimensional random vectors x and y. Then the PDF

$$p(x) = \frac{1}{|\det J^g(g^{-1}(x))|} N\left(g^{-1}(x); 0, I_m\right) \tag{87}$$

factorizes according to

$$p(x) = \prod_{j=1}^{m} p_{x_j}(x_j) \text{ with } p_{x_j}(x_j) := \frac{1}{|\frac{\partial}{\partial \xi_j}\gamma(\gamma^{-1}(x_j))|} N\left(\gamma^{-1}(x_j); 0, 1\right). \tag{88}$$

The components of the random vector x are thus independent.

Remarks

- $p_{x_j} : \mathbb{R} \to \mathbb{R}_{\geq 0}$ denotes the PDF of $x_j, j = 1, ..., , m$.
- The independence of the $x_j, j = 1, ..., m$ results from
  (1) the assumed independence of the $\xi_j, j = 1, ..., m$.
  (2) the assumed component-wise operation of $g : \mathbb{R}^m \to \mathbb{R}^m$.

Proof

We first note that with the component-wise operation of g, its inverse also operates component-wise, i.e.

$$g(w) = (\gamma(w_1), ..., \gamma(w_m))^T \Rightarrow g^{-1}(x) = (\gamma^{-1}(x_1), ..., \gamma^{-1}(x_m)). \tag{89}$$

We next note that thus

$$N\left(g^{-1}(x); 0, I_m\right) = N\left((\gamma^{-1}(x_1), ..., \gamma^{-1}(x_m))^T; 0, I_m\right) = \prod_{j=1}^{m} N\left(\gamma^{-1}(x_j); 0, 1\right). \tag{90}$$

Finally, we note that the Jacobian matrix of g is a diagonal matrix, because

$$J^g(\xi) = \left(\frac{\partial}{\partial \xi_j} g_i(\xi)\right)_{1 \leq i,j \leq m} = \left(\frac{\partial}{\partial \xi_j} \gamma(\xi_i)\right)_{1 \leq i,j \leq m} \tag{91}$$

and thus for all $1 \leq i, j \leq m$

$$(J^g(\xi))_{ij} = \frac{\partial}{\partial \xi_j} \gamma(\xi_j) \text{ if } i = j \text{ and } (J^g(\xi))_{ij} = 0 \text{ if } i \neq j. \tag{92}$$

It thus follows that

$$\begin{aligned} p(x) &= \frac{1}{|\prod_{j=1}^{m} \frac{\partial}{\partial \xi_j} \gamma(\gamma^{-1}(x_j))|} \prod_{j=1}^{m} N\left(\gamma^{-1}(x_j); 0, 1\right) \\ &= \prod_{j=1}^{m} \frac{1}{|\frac{\partial}{\partial \xi_j} \gamma(\gamma^{-1}(x_j))|} N\left(\gamma^{-1}(x_j); 0, 1\right) =: \prod_{j=1}^{m} p_{x_j}(x_j). \end{aligned} \tag{93}$$

□

Inference and learning

- There exist a wide variety of closely related ICA estimation methods.
- We here consider the *infomax algorithm* (Bell and Sejnowski, 1995).
- A popular modification of infomax is *FastICA* (Hyvärinen, 1999).
- The infomax algorithm is a log likelihood gradient ascent algorithm.
- EM algorithms for ICA are subject of current research (e.g. Ablin et al., 2019).

  $\rightarrow$ Inference is complicated by the inherent ICA nonlinearities.

Theorem (Infomax algorithm for ICA parameter estimation)

Let $p_\theta(x, y)$ denote an ICA model. Then the log likelihood function for a data set $Y \in \mathbb{R}^{m \times n}$ of independently and identically distributed $y^{(i)} \in \mathbb{R}^m$ for $i = 1, \ldots, n$ is given by

$$\ell : \mathbb{R}^{m \times m} \to \mathbb{R}, W \mapsto \ell(W) := n \ln |\det W| + \sum_{i=1}^{n} \sum_{j=1}^{m} \ln p_{x_j} \left(w_j y^{(i)} \right), \tag{94}$$

where p_{x_j} denotes the PDF of the jth component of x and $w_j \in \mathbb{R}^{1 \times m}$ is the jth row of the unmixing matrix $W = B^{-1}$. The log likelihood function can be maximized using the

ICA infomax algorithm

0. Initialization of $W^{(0)}$, selection of a learning rate $\alpha > 0$

For $k = 0, 1, \ldots$ until convergence

1. Set

$$W^{(k+1)} := W^{(k)} + \alpha \nabla \ell \left(W^{(k)} \right), \text{ where } \nabla \ell \left(W^{(k)} \right) = n \left({W^{(k)}}^{-1} \right)^T + \sum_{i=1}^{n} f \left(W^{(k)} y^{(i)} \right) {y^{(i)}}^T$$

and f denotes the so-called *ICA learning nonlinearity*

$$f : \mathbb{R}^m \to \mathbb{R}^m, x \mapsto f(x) := \nabla \ln p_x(x). \tag{95}$$

Proof

To derive the infomax algorithm for ICA parameter estimation, we first evaluate the log likelihood function and then compute its gradient with respect to the unmixing matrix.

(1) Evaluation of the log likelihood function

By definition of the ICA model, we have

$$y = Bx \text{ with } x \sim p_x(x) = \prod_{j=1}^{m} p_{x_j}(x_j). \tag{96}$$

With the multivariate PDF transform theorem for linear functions (Appendix), it thus follows immediately that

$$p_y(y) = \frac{1}{|\det B|} p_x(B^{-1}y). \tag{97}$$

With $|\det B|^{-1} = |\det B^{-1}|$, $B^{-1} = W$, and the factorization property of p_x, it then follows that

$$p_y(y) = |\det W| p_x(Wy) = |\det W| \prod_{j=1}^{m} p_{x_j}(w_j y), \tag{98}$$

where w_j denotes the jth row of W.

Proof (cont.)

The log joint probability of n independent and identically distributed samples $y^{(i)}, i = 1, \ldots, n$ distributed according to the ICA model is thus given by

$$\begin{aligned}
\ell(W) &= \ln p_W\,(Y) \\
&= \ln \prod_{i=1}^{n} p_W\left(y^{(i)}\right) \\
&= \ln \left(\prod_{i=1}^{n} |\det W| \prod_{j=1}^{m} p_{x_j}\left(w_j y^{(i)}\right) \right) \\
&= n \ln |\det W| + \sum_{i=1}^{n} \sum_{j=1}^{m} \ln p_{x_j}\left(w_j y^{(i)}\right).
\end{aligned}$$

Proof (cont.)

(2) Evaluation of the log likelihood function gradient

With

$$\frac{\partial}{\partial X} \ln |\det X| = \left(X^{-1}\right)^T \text{ and } \frac{\partial}{\partial X} XA = A^T \tag{99}$$

we have

$$\begin{aligned}
\frac{\partial}{\partial W} \ell(W) &= \frac{\partial}{\partial W} \left(n \ln |\det W| + \sum_{i=1}^{n} \sum_{j=1}^{m} \ln p_{x_j} \left(w_j y^{(i)} \right) \right) \\
&= n \frac{\partial}{\partial W} \ln |\det W| + \sum_{i=1}^{n} \frac{\partial}{\partial W} \left(\ln \prod_{j=1}^{m} p_{x_j} \left(w_j y^{(i)} \right) \right) \\
&= n \left(W^{-1}\right)^T + \sum_{i=1}^{n} \frac{\partial}{\partial W} \ln p_x \left(W y^{(i)} \right) \\
&= n \left(W^{-1}\right)^T + \sum_{i=1}^{n} \nabla \ln p_x(x)|_{x=W y^{(i)}} \frac{\partial}{\partial W} \left(W y^{(i)} \right) \\
&= n \left(W^{-1}\right)^T + \sum_{i=1}^{n} \nabla \ln p_x(x)|_{x=W y^{(i)}} \, y^{(i)^T} \\
&= n \left(W^{-1}\right)^T + \sum_{i=1}^{n} f \left(W y^{(i)} \right) y^{(i)^T}
\end{aligned}$$

□

Example (ICA with hyperbolic tangent learning rule)

An often used "model-free" ICA learning non-linearity is multivariate vector-valued the hyperbolic tangent function (Hyvärinen and Oja, 2000)

$$f : \mathbb{R}^m \to \mathbb{R}^m, x \mapsto f(x) = \left(\phi_j(x_j)\right)_{j=1,\ldots,m} \text{ with } \phi_j(x_j) = -\tanh(x_j). \tag{100}$$

From a generative perspective, this choice of learning non-linearity implies independent component PDFs of the form

$$p_{x_j} : \mathbb{R} \to \mathbb{R}_{>0}, x_j \mapsto p_{x_j}(x_j) := \frac{1}{\pi \cosh(x_j)}, \tag{101}$$

as well as the component-specific generative nonlinearity

$$\gamma : \mathbb{R} \to \mathbb{R}, \xi \mapsto \gamma(\xi) := \ln\left(\tan\left(\frac{\pi}{4}\left(1 + \mathrm{erf}(\xi/\sqrt{2})\right)\right)\right). \tag{102}$$

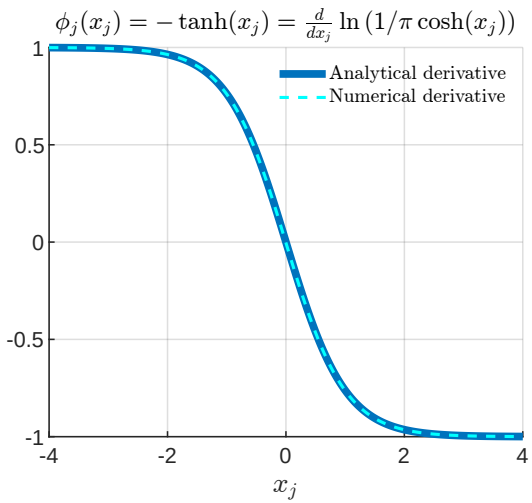
Proof

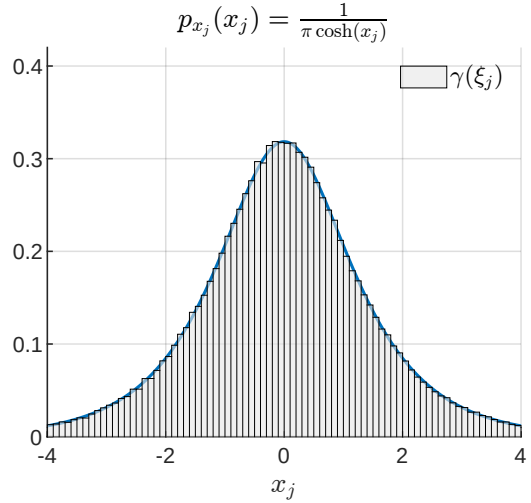
The result follows from the facts that

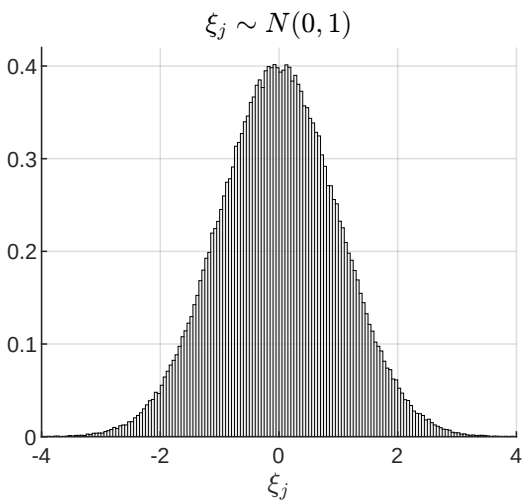
$$\frac{d}{dx}\left(\ln\left(\frac{1}{\pi\cosh(x)}\right)\right) = -\tanh(x) \text{ and } \frac{1}{|\frac{\partial}{\partial \xi_j}\gamma(\gamma^{-1}(x_j))|} N\left(\gamma^{-1}(x_j); 0, 1\right) = \frac{1}{\pi\cosh(x_j)}.$$

□

Hyperbolic tangent learning rule visual intuition







Foundations

Inference and learning

Factor analysis

Probabilistic principal component analysis

Principal component analysis

Independent component analysis

Summary

A generative perspective on factor analysis, PCA, and ICA.

- "Generative" here means probabilistic and model-based.

An introduction to the expectation-maximization (EM) algorithm.

- A general approach for parameter estimation in latent variable models.
- A modern take on EM from the perspective of ELBO maximization.
- A first step towards understanding variational inference.

A unifying perspective on inference and learning in probabilistic models.

- Natural generalization to HMMs, Kalman filters, Bayesian filters.
- A first steps towards understanding contemporary brain theories.

  $\rightarrow$ Free energy principle, active inference, agent-based behavioral models.

Appendix

Theorem (Joint Gaussian distributions)

Given an m-dimensional random vector X distributed according to a Gaussian distribution with PDF

$$p_X : \mathbb{R}^m \to \mathbb{R}_{>0},\ x \mapsto p_X(x) := N(x; \mu_x, \Sigma_{xx}) \text{ for } \mu_x \in \mathbb{R}^m, \Sigma_{xx} \in \mathbb{R}^{m \times m}, \tag{103}$$

a matrix $A \in \mathbb{R}^{n \times m}$, a vector $b \in \mathbb{R}^n$, and a n-dimensional random vector Y conditionally distributed according to a Gaussian distribution with conditional PDF

$$p_{y|X}(\cdot|x) : \mathbb{R}^n \to \mathbb{R}_{>0},\ y \mapsto p_{Y|X}(y|x) := N(y; AX + b, \Sigma_{yy}) \text{ for } \Sigma_{yy} \in \mathbb{R}^{n \times n} \tag{104}$$

the $m + n$-dimensional random vector (X, Y) is distributed according to a Gaussian distribution with joint PDF

$$p_{X,Y} : \mathbb{R}^{m+n} \to \mathbb{R}_{>0},\ (x, y) \mapsto p_{X,Y}(x, y) = N\left((x, y); \mu_{x,y}, \Sigma_{x,y}\right), \tag{105}$$

where $\mu_{x,y} \in \mathbb{R}^{n+m}$ and $\Sigma_{x,y} \in \mathbb{R}^{m+n \times m+n}$, and in particular

$$\mu_{x,y} = \begin{pmatrix} \mu_x \\ A\mu_x + b \end{pmatrix} \text{ and } \Sigma_{x,y} = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xx}A^T \\ A\Sigma_{xx} & \Sigma_{yy} + A\Sigma_{xx}A^T \end{pmatrix}. \tag{106}$$

Remark

- The parameters of $p(x, y)$ can be computed from the parameters of $p(x)$ and $p(y|x)$.

Theorem (Conditional Gaussian distributions)

Given an $m+n$-dimensional random vector (X, Y) distributed according to a Gaussian distribution with PDF

$$p_{X,Y} : \mathbb{R}^{m+n} \to \mathbb{R}_{>0}, (x, y) \mapsto p_{X,Y}(x, y) := N\left((x, y); \mu_{x,y}, \Sigma_{x,y}\right), \tag{107}$$

where

$$\mu_{x,y} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \Sigma_{x,y} = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}, \tag{108}$$

for $x, \mu_x \in \mathbb{R}^m$, $y, \mu_y \in \mathbb{R}^n$ and $\Sigma_{xx} \in \mathbb{R}^{m\times m}$, $\Sigma_{xy} \in \mathbb{R}^{m\times n}$, $\Sigma_{yy} \in \mathbb{R}^{n\times n}$, the distribution of X given Y has an m-dimensional conditional PDF

$$p_{X|Y}(\cdot|y) : \mathbb{R}^m \to \mathbb{R}_{>0}, x \mapsto p_{X|Y}(x|y) := N(x; \mu_{x|y}, \Sigma_{x|y}), \tag{109}$$

where

$$\mu_{x|y} = \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(Y - \mu_y) \in \mathbb{R}^m \tag{110}$$

and

$$\Sigma_{x|y} = \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx} \in \mathbb{R}^{m\times m}. \tag{111}$$

Remarks

- The parameters of $p(x|y)$ can be computed from the parameters of $p(x, y) = p(x)p(y|x)$.
- The parameters of $p(x|y)$ can be computed from the parameters of $p(x)$ and $p(y|x)$.

Theorem (Jensen's inequality)

Let x be a random variable and g be a convex function, i.e.

$$g(\lambda x_1 + (1-\lambda)x_2) \leq \lambda g(x_1) + (1-\lambda)g(x_2). \tag{112}$$

Then

$$\mathbb{E}(g(x)) \geq g(\mathbb{E}(x)). \tag{113}$$

Conversely, let g be a concave function, i.e.

$$g(\lambda x_1 + (1-\lambda)x_2) \geq \lambda g(x_1) + (1-\lambda)g(x_2). \tag{114}$$

Then

$$\mathbb{E}(g(x)) \leq g(\mathbb{E}(x)). \tag{115}$$

Remarks

- For convex g the function's graph lies below the straight line $g(x_1)$ to $g(x_2)$.
- For concave g the function's graph lies above the straight line $g(x_1)$ to $g(x_2)$.
- The logarithm is a concave function, hence $\mathbb{E}(\ln x) \leq \ln \mathbb{E}(x)$.

Proof

By adapting the proof of Casella and Berger (2012, Theorem 4.7.8), we show the inequality for the concave case. Let f be a tangent line at the point $g(\mathbb{E}(x))$, i.e. is a linear-affine function of the form $f(x) := ax + b$ for some $a, b \in \mathbb{R}$ with $f(\mathbb{E}(x)) = g(\mathbb{E}(x))$. Because g is concave, we have $g(x) \leq ax + b$ for all $x \in \mathbb{R}$ and thus als $g(x) \leq ax + b$. Hence,

$$\mathbb{E}(g(x)) \leq \mathbb{E}(ax + b) = a\mathbb{E}(x) + b = f(\mathbb{E}(x)) = g(\mathbb{E}(x)). \tag{116}$$

□

Theorem (The multivariate PDF transform for linear functions)

Let x be an m-dimensional random vector with PDF $p_x(x)$. Let $y = f(x)$ with

$$f(x) = Ax, A \in \mathbb{R}^{m \times m} \text{ and nonsingular.} \tag{117}$$

Then the PDF of y is given by

$$p_y : \mathbb{R}^n \to \mathbb{R}_{\geq 0}, y \mapsto p_y(y) = \frac{1}{|\det A|} p_x(A^{-1}y) \tag{118}$$

where $\det A$ and A^{-1} denote the determinant and the inverse of A, respectively.

Remark

- A formula for computing p_y if $y := Ax$ with $x \sim p_x$ and $A \in \mathbb{R}^{m \times m}$.

Proof

We first show that

$$f^{-1} : \mathbb{R}^m \to \mathbb{R}^m, y \mapsto f^{-1}(y) := A^{-1}y. \tag{119}$$

To this end, we note that

$$f^{-1}\left(f(x)\right) = A^{-1}Ax = x = \mathsf{id}_{\mathbb{R}^n}(x). \tag{120}$$

We next show that

$$J^f\left(f^{-1}(y)\right) = A. \tag{121}$$

To this end, we first note that

$$f_i(x) = \sum_{j=1}^{m} A_{ij}x_j. \tag{122}$$

Thus

$$J^f(x) = \left(\frac{\partial}{\partial x_j} f_i(x)\right)_{1\leq i,j\leq m} = \left(\sum_{j=1}^{m} \frac{\partial}{\partial x_j} A_{ij}x_j\right)_{1\leq i,j\leq m} = (A_{ij})_{1\leq i,j\leq m} = A. \tag{123}$$

□

References

Ablin, P., Gramfort, A., Cardoso, J.-F., and Bach, F. (2019). Stochastic algorithms with descent guarantees for ICA. *arXiv:1805.10054 [cs, stat]*.

Bell, A. J. and Sejnowski, T. (1995). An Information-Maximization Approach to Blind Separation and Blind Deconvolution. *Neural Computation*, 7:1129–1159.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Casella, G. and Berger, R. (2012). *Statistical Inference*. Duxbury.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Hotelling, H. (1933). Analysis of complext variables into principal components. *Journal of Educational Psychology*, 24:417–441 and 498–520.

Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis*. Wiley-Interscience, S.l.

Hyvärinen, A. and Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4-5):411–430.

Lawley, N. (1953). A modified method of estimation in factor analysis and some large sample results. *Nord. Psyko. Monogr. Ser*, 3:35–42.

Ostwald, D., Kirilina, E., Starke, L., and Blankenburg, F. (2014). A tutorial on variational Bayes for latent linear stochastic time-series models. *Journal of Mathematical Psychology*, 60:1–19.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Petersen, K. B. and Pedersen, M. S. (2012). The Matrixcookbook. page 72.

Roweis, S. (1998). EM Algorithms for PCA and SPCA. page 7.

Roweis, S. and Ghahramani, Z. (1999). A Unifying Review of Linear Gaussian Models. *Neural Computation*, 11(2):305–345.

Rubin, D. B. and Thayer, D. T. (1982). EM algorithms for ML factor analysis. *Psychometrika*, 47(1):69–76.

Starke, L. and Ostwald, D. (2017). Variational Bayesian Parameter Estimation Techniques for the General Linear Model. *Frontiers in Neuroscience*, 11.

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. page 12.