Probabilistic PCA using expectation maximization

Abhijit Chinchani

December 12, 2020

1 Introduction

Dimensionality reduction is the transformation of the observed high-dimensional data to low-dimensional latent space, which presumably correspond to the intrinsic hidden dimensions of the data and help represent meaningful properties of the data. Usually dimensionality reduction problems involve data observed using a large number of sensors. The activity of a number of sensors is highly correlated and the raw data is often sparse. In many experiments, the source or latent signal of interest is often low-dimensional compared to the number of sensors. Thus, it is desirable to reduce the dimensions and obtain only uncorrelated of independent latent dimensions. Another reason for working with low-dimensional data is the reduced computational complexity of the analysis. Dimensionality reduction techniques have applications in various fields like - uncovering underlying brain activity from observed cortical sensors in neuroscience, source localization in MIMO (Multi-input multi-output) systems in telecommunication, speech processing, bioinformatics, and many other applications involving signal processing and machine learning.

One such commonly used dimensionality technique is the Principal Component Analysis (PCA). PCA takes in a number of high-dimensional data vectors and finds direction vectors that maximize the variance of the data in that dimension. These direction vectors are known as the principal axes. The principal axes are orthogonal to each other. The data is then projected onto the principal axes to generate principal components which are uncorrelated. Let $X \in \mathbb{R}^{\gtrdot\curvearrowleft\ltimes}$ be the observed data, where m corresponds to the dimensions of X and n corresponds the number of samples. PCA involves performing eigen decomposition on the data matrix (X).

$$X = UDU^T$$
$$X = WZ$$

where, U is the unit orthogonal matrix containing the Eigen vectors or principal direction vectors, and D is a diagonal matrix containing the Eigen values. This can be rewritten in-terms of a principal direction matrix (W) and principal components (Z). The components are arranged based on their variance and usually components with larger variance are chosen for further analysis.

Although PCA is a very simple and powerful technique, it has certain limitations - it does not have noise model that can denoise uncorrelated noise, the model is rigid and cannot be changed to model the data better. In order to solve this issue, a probabilistic version of PCA (PPCA) was developed [TB99; Row97]. PPCA consists of a generative linear model that projects the data into low-dimensional latent space. Since, PPCA assigns probabilities to the data, performance of various models can be compared. If the dimensionality of data is large then PCA can be computationally intensive as it estimates all the dimensions; on the other hand PPCA can estimate the required number of latent dimensions more efficiently. PPCA can also deal with missing data. Since, PPCA is a probabilistic model, further extensions of the model are possible. It also has explicit noise model that can model uncorrelated observation noise and can also generate simulated data from the generative model. Expectation-maximization algorithm is widely used method to estimate PPCA parameters [TB99; Row97; Moo96; Wu83].

In Section 2 of this paper, we formulate the widely used PPCA model [TB99; Row97] and derive the equations for the EM algorithm. EM algorithm is a heuristic method that optimizes the maximum likelihood estimator. It estimates the parameters iteratively from a given random initial point. In Section 3, we prove that under certain restrictions, the EM algorithm converges to a local optima; and EM for PPCA converges to a global optima. These classical results [TB99; Row97; DLR77; Zan67; Rao+73; Wu83] are the foundation for various advanced extensions of PPCA. In Section 4, we briefly discuss few extensions of PPCA and research directions.

2 PPCA formulation

In this section, we mathematically formulate PPCA. Let $X \in \mathbb{R}^m$ be the random variable corresponding to the observable data and $[\bar{x}_1, \bar{x}_2, ..., \bar{x}_N]$ be a set of N observed data vectors. Let $Z \in \mathbb{R}^k$, $k \leq m$ be the random variable corresponding to the low dimensional latent variable. As described above, we are interested in orthogonal latent dimensions and hence model Z as Gaussian random variable with mean $\bar{0}$ and covariance as I.

$$f(Z) = \mathcal{N}(\bar{0}, I) \tag{1}$$

Where, f corresponds to the probability density function, $\mathcal{N}(\bar{u}, V)$ corresponds to multivariate Gaussian random variable with mean vector $\bar{u}$ and covariance matrix V. We then formulate the PPCA as linear projection of observed data onto a low-dimensional space as follows -

$$X \; WZ + \bar{\mu} + \epsilon \tag{2}$$

$$f(X|Z) = \mathcal{N}(WZ + \bar{\mu}, \sigma^2 I) \tag{3}$$

Where, W is the mixing matrix that transforms the data from latent space to the observed space. $\bar{\mu}$ is the mean of X and ϵ is the uncorrelated observation noise with zero mean and covariance, $\mathbb{E}[\epsilon\epsilon^T] = \sigma^2 I$. The density function of X is also Gaussian with mean and variance as follows -

$$\begin{aligned}
\mathbb{E}[X] &= \mathbb{E}[WZ] + \bar{\mu} + \mathbb{E}[\epsilon] = \bar{\mu} \\
var(X) &= \mathbb{E}[XX^T] - \mathbb{E}[X]\mathbb{E}[X]^T \\
&= \mathbb{E}[(WZ + \bar{\mu} + \epsilon)(WZ + \bar{\mu} + \epsilon)^T] - \bar{\mu}\bar{\mu}^T \\
&= W\mathbb{E}[ZZ^T]W^T + \bar{\mu}\bar{\mu}^T + \mathbb{E}[\epsilon\epsilon^T] - \bar{\mu}\bar{\mu}^T \\
&= WW^T + \sigma^2 I
\end{aligned}$$

Therefore,

$$f(X) = \mathcal{N}(\bar{\mu}, WW^T + \sigma^2 I) \tag{4}$$

2.1 Optimization using EM algorithm

In this section, we describe the optimization problem for PPCA. We aim to fit this model (3) to the observed data $[\bar{x}_1, \bar{x}_2, ..., \bar{x}_N]$, i.e., we fit the model parameters $\theta = [W, \bar{\mu}, \sigma]$. Here, columns of W correspond to the vectors that span the principal components. We also find the corresponding projection of the observed data onto the latent dimensions. PPCA projections approach the PCA projections as $\sigma \to 0$. Here, we use expectation maximization (EM) algorithm as described in [TB99] to estimate the model parameters. The EM is an iterative algorithm consists of the expectation step (E-step) where the expected distribution of Z given the model parameters for that iteration and the maximization step (M-step) calculates the model parameters that maximize the complete log likelihood of the data.

1. E-step
   In the E-step we calculate $f(Z|X)$. Using Bayes' rule and substituting from (1), (3) and (4) we get -

$$\begin{aligned}
f(Z|X, \theta) &= \frac{f(X|Z)f(Z)}{f(X)} \\
&= \frac{\mathcal{N}(WZ + \bar{\mu}, \sigma^2 I)\mathcal{N}(\bar{0}, I)}{\mathcal{N}(\bar{\mu}, WW^T + \sigma^2 I)} \\
&= \mathcal{N}(C^{-1}W^T(X - \bar{\mu}), \sigma^2 C^{-1})
\end{aligned}$$

Where, $C = W^T W + \sigma^2 I$. Thus, the latent dimension for the nth data sample,

$$\mathbb{E}[\bar{z}_n] = C^{-1}W^T(\bar{x}_n - \bar{\mu}) \tag{5}$$

$$\mathbb{E}[\bar{z}_n\bar{z}_n^T] = \sigma^2 C^{-1} + \mathbb{E}[\bar{z}_n]\mathbb{E}[\bar{z}_n]^T \tag{6}$$

2. M-step
   In the M-step, we estimate we find the model parameters $\theta = [W, \bar{\mu}, \sigma]$ that maximize the complete data log-likelihood. Consider the complete data log-likelihood -

$$\begin{aligned} L = log(f(X,Z)) &= \sum_{n=1}^{N} log(f(\bar{x}_n, \bar{z}_n)) \\ &= \sum_{n=1}^{N} log(f(\bar{x}_n|\bar{z}_n)) + log(f(\bar{z}_n)) \\ &= \sum_{n=1}^{N}(-\frac{m}{2}log(2\pi\sigma^2) - \frac{1}{2}(\bar{x}_n - W\bar{z}_n - \bar{\mu})^T(\sigma^2 I)^{-1}(\bar{x}_n - W\bar{z}_n - \bar{\mu}) - \frac{k}{2}log(2\pi) - \frac{1}{2}\bar{z}_n^T\bar{z}_n) \end{aligned}$$

Taking expectation on Z over both sides,

$$\begin{aligned} \mathbb{E}[L] = \sum_{n=1}^{N}(-\frac{m}{2}log(2\pi\sigma^2) - \frac{1}{2\sigma^2}((\bar{x}_n - \bar{\mu})^T(\bar{x}_n - \bar{\mu}) - \mathbb{E}[\bar{z}_n^T W^T(\bar{x}_n - \bar{\mu})] \\ - \mathbb{E}[(\bar{x}_n - \bar{\mu})^T W\bar{z}_n] + \mathbb{E}[\bar{z}_n^T W^T W\bar{z}_n]) - \frac{k}{2}log(2\pi) - \frac{1}{2}\mathbb{E}[\bar{z}_n^T\bar{z}_n]) \end{aligned}$$

$$\begin{aligned} \mathbb{E}[L] = \sum_{n=1}^{N}(-\frac{m}{2}log(2\pi\sigma^2) - \frac{1}{2\sigma^2}((\bar{x}_n - \bar{\mu})^T(\bar{x}_n - \bar{\mu}) - \mathbb{E}[\bar{z}_n]^T W^T(\bar{x}_n - \bar{\mu}) \\ - (\bar{x}_n - \bar{\mu})^T W\mathbb{E}[\bar{z}_n] + Tr(W^T W\mathbb{E}[\bar{z}_n^T\bar{z}_n])) - \frac{k}{2}log(2\pi) - \frac{1}{2}\mathbb{E}[\bar{z}_n^T\bar{z}_n]) \end{aligned} \tag{7}$$

To estimate W, we find the maxima of (7) by differentiating w.r.t. W.

$$\frac{d\mathbb{E}[L]}{dW} = \sum_{n=1}^{N}(-\frac{1}{2\sigma^2}(-2(\bar{x}_n - \bar{\mu})^T\mathbb{E}[\bar{z}_n] + 2W\mathbb{E}[\bar{z}_n^T\bar{z}_n])) = 0$$

$$W(\sum_{n=1}^{N}\mathbb{E}[\bar{z}_n^T\bar{z}_n]) = (\sum_{n=1}^{N}(\bar{x}_n - \bar{\mu})^T\mathbb{E}[\bar{z}_n])$$

Thus, the updated W is

$$\hat{W} = (\sum_{n=1}^{N}(\bar{x}_n - \bar{\mu})^T\mathbb{E}[\bar{z}_n])(\sum_{n=1}^{N}\mathbb{E}[\bar{z}_n^T\bar{z}_n])^{-1} \tag{8}$$

Similarly, To estimate σ^2, we find the maxima of (7) by differentiating w.r.t. σ^2.

$$\begin{aligned} \frac{d\mathbb{E}[L]}{d\sigma^2} = \sum_{n=1}^{N}(-\frac{m}{2\sigma^2} + \frac{1}{2\sigma^4}((\bar{x}_n - \bar{\mu})^T(\bar{x}_n - \bar{\mu}) - \mathbb{E}[\bar{z}_n]^T W^T(\bar{x}_n - \bar{\mu}) \\ - (\bar{x}_n - \bar{\mu})^T W\mathbb{E}[\bar{z}_n] + Tr(W^T W\mathbb{E}[\bar{z}_n^T\bar{z}_n])) = 0 \end{aligned}$$

Thus, updated σ^2 is

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{mN}\sum_{n=1}^{N}((\bar{x}_n-\bar{\mu})^T(\bar{x}_n-\bar{\mu}) - \mathbb{E}[\bar{z}_n]^T W^T(\bar{x}_n-\bar{\mu}) \\ &\quad - (\bar{x}_n-\bar{\mu})^T W\mathbb{E}[\bar{z}_n] + Tr(W^T W\mathbb{E}[\bar{z}_n^T\bar{z}_n])) \\ &= \frac{1}{mN}(Tr(\sum_{n=1}^{N}(\bar{x}_n-\bar{\mu})(\bar{x}_n-\bar{\mu})^T) - Tr(\sum_{n=1}^{N}(\bar{x}_n-\bar{\mu})\mathbb{E}[\bar{z}_n]^T W^T) \\ &\quad - Tr(\sum_{n=1}^{N} W\mathbb{E}[\bar{z}_n](\bar{x}_n-\bar{\mu})^T) + Tr(W\sum_{n=1}^{N}\mathbb{E}[\bar{z}_n^T\bar{z}_n]W^T))\end{aligned}$$

Substituting the $\hat{W}$, we get,

$$\hat{\sigma}^2 = \frac{1}{mN}Tr(\sum_{n=1}^{N}(\bar{x}_n-\bar{\mu})(\bar{x}_n-\bar{\mu})^T) - \hat{W}\sum_{n=1}^{N}\mathbb{E}[\bar{z}_n](\bar{x}_n-\bar{\mu})^T) \tag{9}$$

To estimate $\bar{\mu}$,

$$\begin{aligned}\frac{d\mathbb{E}[L]}{d\bar{\mu}} &= \sum_{n=1}^{N}(-\frac{1}{2\sigma^2}(2(\bar{x}_n-\bar{\mu})) = 0 \\ \bar{\mu} &= \frac{1}{N}\sum_{n=1}^{N}\bar{x}_n \qquad (10)\end{aligned}$$

The ML estimate of $\bar{\mu}$ is the sample mean of the observed variable and is not updated in the EM algorithm. Thus, the initial model parameters are assigned randomly and then the EM algorithm is run and values of W and σ^2 are updated each iteration until convergence.

3 Convergence of EM

This contents of this section are derived from the following references [TB99; Row97; DLR77; Zan67; Rao+73; Wu83]. Let $\mathcal{X}$ and $\mathcal{Z}$ be two spaces, where we observe X and the low dimensional latent space Z is not observed. Let the density functions of Z, X, (X, Z) and $X|Z$ be respectively $g(Z|\theta)$, $p(X|\theta)$, $k(X, Z|\theta)$, and $f(X|Z, \theta)$ for some parameters $\theta \in \Omega$. The log likelihood function is -

$$\begin{aligned} L(\theta') &= log(p(X|\theta')) \\ &= \mathbb{E}[log(k(X, Z|\theta'))|Z, \theta] - \mathbb{E}[log(f(X|Z, \theta))|Z, \theta] \\ &= Q(\theta', \theta) - H(\theta', \theta) \end{aligned}$$

In the i^{th} iteration of the EM algorithm,

1. E-step: Determine the $Q(\theta, \theta_i)$ function. In the PPCA above, this involved finding $f(Z|X)$ which was further used in the formula for $\mathbb{E}[log(f(Z, X))]$, which is $Q(\theta, \theta_i)$ in this general case.
2. M-step: Find θ_{i+1} such that it maximizes $Q(\theta, \theta_i)$

Definition 3.1. Given two sets A and B, a point-to-set map M defined on A with range in B assigns each $a \in A$ to a subset $M(a)$ in B. It can be written as $M : A \to 2^B$.

For the EM algorithm, we define the M-step as a point-to-set map M such that when $\theta_i \to \theta_{i+1} \in M(\theta_i)$. By definition of M-step we have,

$$Q(\theta', \theta) \geq Q(\theta, \theta), \text{ for all } \theta' \in M(\theta) \tag{11}$$

Lemma 3.2. For any pair (θ', θ), $H(\theta', \theta) \leq H(\theta, \theta)$ with equality if and only if $f(X|Z, \theta') = f(X|Z, \theta)$.
Proof. Using eq. 1e.6.6 from [Rao+73], we have if $\int_S (f - g)du \geq 0$ then $\int_S (log(f) - log(g))du \geq 0$ with equality only when $f = g$ (a.s)

$$\begin{gathered} Q(\theta', \theta) \geq Q(\theta, \theta) \\ \mathbb{E}[log(k(X, Z|\theta'))|Z, \theta] \geq \mathbb{E}[log(k(X, Z|\theta))|Z, \theta] \\ \mathbb{E}[k(X, Z|\theta')|Z, \theta] \geq \mathbb{E}[k(X, Z|\theta)|Z, \theta] \end{gathered}$$

We know that $\mathbb{E}[p(Z|\theta')|Z, \theta]$ is constant and $k(X, Z) = f(X/Z)p(Z)$, thus using equation eq. 1e.6.6,

$$\begin{gathered} \mathbb{E}[f(X|Z, \theta')|Z, \theta] \leq \mathbb{E}[f(X|Z, \theta)|Z, \theta] \\ \mathbb{E}[log(f(X|Z, \theta'))|Z, \theta] \leq \mathbb{E}[log(f(X|Z, \theta))|Z, \theta] \\ H(\theta', \theta) \leq H(\theta, \theta) \end{gathered}$$

Thus from eq. 1e.6.6, $H(\theta', \theta) \leq H(\theta, \theta)$ with equality if and only if $\mathbb{E}[log(f(X|Z, \theta'))|Z, \theta] = \mathbb{E}[log(f(X|Z, \theta))|Z, \theta]$, or $f(X|Z, \theta') = f(X|Z, \theta)$.

Lemma 3.3. For every EM algorithm, $L(M(\theta)) \geq L(\theta)$ for all $\theta \in \Omega$ and equality holds only when $Q(\theta', \theta) = Q(\theta, \theta)$ and $f(X|Z, \theta') = f(X|Z, \theta)$.
Proof. Consider, $L(M(\theta)) - L(\theta) = \{Q(M(\theta)|\theta) - Q(\theta|\theta)\} - \{H(M(\theta)|\theta) - H(\theta|\theta)\}$ By definition of EM M-step (11), $Q(M(\theta)|\theta) \geq Q(\theta|\theta)$.
By Lemma 3.2, $H(M(\theta)|\theta) \leq H(\theta|\theta)$ with equality only when $f(X|Z, \theta') = f(X|Z, \theta)$.
Thus, $L(M(\theta)) \geq L(\theta)$ with equality only when $Q(\theta', \theta) = Q(\theta, \theta)$ and $f(X|Z, \theta') = f(X|Z, \theta)$.

For the EM algorithm, we make the following assumptions as described in [Wu83].

1. The parameter space $\Omega \subset \mathbb{R}^r$, where r is the number of parameters.

2. $\Omega_{\theta_0} = \{\theta \in \Omega : L(\theta) \geq L(\theta_0)\}$ for all $L(\theta_0) > -\infty$.

3. L is continuous and differentiable in interior of Ω.

From these assumptions, $(L(\theta_i))_{i \geq 0}$ is bounded above for any $\theta_0 \in \Omega$.

Theorem 3.4 (Zangwill's global convergence theorem). Let M be a point-to-set map from Ω to Ω, for a given initial point $\theta_0 \in \Omega$, it generates a sequence $\{\theta_i\}_{i=1}^{\infty}$ such that $\theta_{i+1} \in M(\theta_i)$.
Let $T \subset \Omega$ be a solution set and if

1. $\{\theta_i\}_{i=1}^{\infty} \subset S$, where $S \subset \Omega$ is compact set.

2. there is a continuous, increasing and bounded function L on Ω such that

   (a) if $\theta \notin T$, then $L(\theta') \geq L(\theta)$ for all $\theta' \in M(\theta)$

   (b) if $\theta \in T$, then $L(\theta') \geq L(\theta)$ for all $\theta' \in M(\theta)$

3. the map M is closed at all points of $X - T$

Then, the limit of any convergent subsequence of $\{\theta_i\}_{i=1}^{\infty}$ is a solution.
Proof. Let $\hat{\theta}$ be the limit of the sequence $\{\theta_i\}_{i=1}^{\infty}$.
Then there exists a subsequence, $\{\theta_{i_j}\}_{j=1}^{\infty}$ such that $\theta_{i_j} \to \hat{\theta}$ as $j \to \infty$.
Since, the function L is continuous, $L(\theta_{i_j}) \to L(\hat{\theta})$ as $j \to \infty$.
From 2. of the problem statement, we know that L is monotonically increasing on the sequence $\{\theta_i\}_{i=1}^{\infty}$ because $\theta_{i+1} \in M(\theta_i)$.
Thus, $L(\hat{\theta}) - L(\theta_i) \geq 0$, for all i.
Since, $L(\theta_{i_j}) \to L(\hat{\theta})$, for an $\epsilon \geq 0$ and j_0,
$L(\hat{\theta}) - L(\theta_{i_j}) \leq \epsilon$ for $j \geq j_0$.
Consider, $L(\hat{\theta}) - L(\theta_i) = L(\hat{\theta}) - L(\theta_{i_{j_0}}) + L(\theta_{i_{j_0}}) - L(\theta_i) < \epsilon$ for all $i \geq j_0$.
This can be rewritten as $\sum_i i_\epsilon |L(\hat{\theta}) - L(\theta_i)| \leq \infty$ almost surely.
Thus, $L(\theta_i) \to L(\hat{\theta})$ as $i \to \infty$.

Next, we need to show that $\hat{\theta} \in T$.
Proof by contradiction, assume that $\hat{\theta} \notin T$.
Consider a subsequence $\{\theta_{i_j+1}\}_{j=1}^{\infty}$ that converges to $\theta_{i_j+1} \to \bar{\theta}$ as $j \to \infty$.
Based on the assumption, $\hat{\theta} \notin T$, then $\bar{\theta} \in M(\hat{\theta})$.
But, since $L(\theta_i) \to L(\hat{\theta})$, then $L(\bar{\theta}) = L(\hat{\theta})$ which contradicts 2(b).

Corollary 3.5. Let M be the M-step point-to-set map of the EM algorithm.

1. $\{\theta_i\}_{i=1}^{\infty} \subset S$, where $S \subset \Omega$ is compact set if $Q(\theta'|\theta)$ is continuous [Wu83].

2. Let the solution set T be the set of all local maxima.

3. From Lemma 3.3, $L(\theta') \geq L(\theta)$ for all $\theta' \notin T$.

Thus, if $Q(\theta'|\theta)$ is continuous, then the EM algorithm converges to a local maxima almost surely. Furthermore, [TB99] have shown that the global maxima is the only stable local extremum for PPCA. Thus, the EM algorithm for PPCA is stable irrespective of the initial model parameters.

4 Open questions and research directions

In this section, we discuss further extensions of PPCA using the EM algorithm.

4.1 Gaussian process factor analysis with custom kernels

PPCA is a simple generative model that finds latent dimensions of the raw data. PPCA has a noise model, where noise covariance is modelled as $\mathbb{E}[\epsilon\epsilon^T] = \sigma^2 I$, i.e., the model assumes equal noise variation in each observation channel. This can be improved with a more complex noise model - $\mathbb{E}[\epsilon\epsilon^T] = R$, where R is a diagonal matrix and channel can have different noise variance. This new model is called factor analysis (FA) and is also widely used [Har76; Rum88].

Further extensions of FA are also possible. One such extention is the Gaussian Process Factor Analysis (GPFA). The model equations of GPFA are as follows:

$$f(Z) = \mathcal{N}(\bar{0}, K)$$

$$f(X|Z) = \mathcal{N}(WZ + \bar{\mu}, R)$$

where, R is the noise covariance matrix as described in FA. K is the prior covariance matrix for the latent variable Z. Other terminologies are same as PPCA. Yu et. al., used an auto-covariance matrix K to model the dynamics of time-series data and obtain latent dimensions varying at different time-scales [Yu+08]. This provides a powerful tool to model the spatial and temporal aspects of a time-series data and research involving feasibility of novel kernels can be performed. E.g. kernels designed to capture latent dimensions that have maximum variance in certain frequency bands can be envisaged.

4.2 Constrained PPCA

Constrained Principal Component Analysis (CPCA) is supervised dimensionality reduction technique that combines multivariate multiple regression and PCA into a unified framework [TH01]. CPCA aims to find latent dimensions not based on the variance of the raw data (as done by PCA), but based on the variance that is predictable from a predictor matrix. CPCA can be mainly split into steps, multivariate multiple regression is used to separate the total variance in the dependent variable (X) into variance predictable from the predictor variables (G) and variance that is not. This procedure splits the dependent variable matrix (X) into two matrices: one composed of variance predictable by the predictor variables(G), or regression-based predicted scores (GC), and the residual variance or error scores (E). This is represented as follows:

$$X = GC + E$$

where, X = matrix of dependent variables, G = matrix of predictor variables, C = (G'G)-1G'Z is a matrix of regression coefficients, GC = regression-based predicted score matrix, and E = residual scores (variance in Z not explained by predictor variables in G).

This followed by PCA only on the predicted scores (GC) to extract latent dimensions of interest. This two step process can be combined into one generative model whose parameters can be estimated using EM algorithm. Although, complex extensions of PPCA can be very useful, but they come with their own set of challenges - most of the complex models do not converge to a global maxima and only converge to a local maxima. The convergence of these complex models can be studied in more detail.

A Exercises

1. For $X : \Omega \to \mathbb{R}$ and $h : \mathbb{R} \to [a, b]$, show that $\mathbb{P}(h(X) \geq c) \geq \dfrac{b - \mathbb{E}[h(X)]}{b - c}$.
   Consider, $b - h(X) \leq (b - c)1_{h(X)>c}$
   Taking expectation on both sides,
   $b - \mathbb{E}[h(X)] \leq (b - c)\mathbb{P}(h(X) \geq c)$
   Therefore, $\mathbb{P}(h(X) \geq c) \geq \dfrac{b - \mathbb{E}[h(X)]}{b - c}$.

2. For $X : \Omega \to \mathbb{R}$ and $h : \mathbb{R} \to [a, b]$, show that $\mathbb{P}(h(X) \geq c) \leq \dfrac{\mathbb{E}[h(X)] - a}{c - a}$.
   Consider, $h(X) - a \geq (c - a)1_{h(X)>c}$
   Taking expectation on both sides,
   $\mathbb{E}[h(X)] - a \geq (c - a)\mathbb{P}(h(X) \geq c)$
   Therefore, $\mathbb{P}(h(X) \geq c) \leq \dfrac{\mathbb{E}[h(X)] - a}{c - a}$.

References

[DLR77] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1 (1977), pp. 1–22.

[Har76] H. H. Harman. *Modern factor analysis.* University of Chicago press, 1976.

[Moo96] T. K. Moon. "The expectation-maximization algorithm". In: *IEEE Signal processing magazine* 13.6 (1996), pp. 47–60.

[Rao+73] C. R. Rao et al. *Linear statistical inference and its applications.* Vol. 2. Wiley New York, 1973.

[Row97] S. Roweis. "EM algorithms for PCA and SPCA". In: *Advances in neural information processing systems* 10 (1997), pp. 626–632.

[Rum88] R. J. Rummel. *Applied factor analysis.* Northwestern University Press, 1988.

[TH01] Y. Takane and M. A. Hunter. "Constrained principal component analysis: a comprehensive theory". In: *Applicable Algebra in Engineering, Communication and Computing* 12.5 (2001), pp. 391–419.

[TB99] M. E. Tipping and C. M. Bishop. "Probabilistic principal component analysis". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61.3 (1999), pp. 611–622.

[Wu83] C. J. Wu. "On the convergence properties of the EM algorithm". In: *The Annals of statistics* (1983), pp. 95–103.

[Yu+08] B. M. Yu et al. "Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity". In: *Advances in neural information processing systems* 21 (2008), pp. 1881–1888.

[Zan67] W. I. Zangwill. "Non-linear programming via penalty functions". In: *Management science* 13.5 (1967), pp. 344–358.