Factor Analysis, Probabilistic PCA, and Variational Inference

Statistical Machine Learning (ENGG*6600*08)

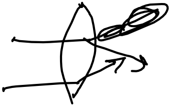
School of Engineering,
University of Guelph, ON, Canada

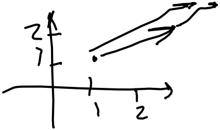
Course Instructor: Benyamin Ghojogh

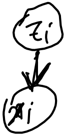
Fall 2023

Variational Inference

Variational Inference







$\to P(x_i ; \theta)$

$\to P(x_i \mid \theta)$

- Consider a dataset $\{\boldsymbol{x}_i\}_{i=1}^n$. Assume that every data point $\boldsymbol{x}_i \in \mathbb{R}^d$ is generated from a latent variable $\boldsymbol{z}_i \in \mathbb{R}^p$. This latent variable has a prior distribution $\mathbb{P}(\boldsymbol{z}_i)$. According to Bayes' rule, we have:

likelihood   prior

posterior

$$\mathbb{P}(\boldsymbol{z}_i \mid \boldsymbol{x}_i) = \frac{\mathbb{P}(\boldsymbol{x}_i \mid \boldsymbol{z}_i)\, \mathbb{P}(\boldsymbol{z}_i)}{\mathbb{P}(\boldsymbol{x}_i)}. \tag{1}$$

$\to P(x_i , z_i)$

$\to P(z_i , x_i)$

- Let $\mathbb{P}(\boldsymbol{z}_i)$ be an arbitrary distribution denoted by $q(\boldsymbol{z}_i)$. Suppose the parameter of conditional distribution of $\boldsymbol{z}_i$ on $\boldsymbol{x}_i$ is denoted by $\boldsymbol{\theta}$; hence, $\mathbb{P}(\boldsymbol{z}_i \mid \boldsymbol{x}_i) = \mathbb{P}(\boldsymbol{z}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})$. Therefore, we can say:

prior

x	y	$x \wedge y$
0	0	0
0	1	0
1	0	0
1	1	1

$$* \quad \boxed{\mathbb{P}(\boldsymbol{z}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}) = \boxed{\frac{\mathbb{P}(\boldsymbol{x}_i \mid \boldsymbol{z}_i, \boldsymbol{\theta})\, \mathbb{P}(\boldsymbol{z}_i \mid \boldsymbol{\theta})}{\mathbb{P}(\boldsymbol{x}_i \mid \boldsymbol{\theta})}}} \tag{2}$$

$4 = 2 \times 2$ Kant

{ synthetic
{ analytic

* { a posteriori ←
{ a priori ←

Variational Inference

$\underbrace{E(X)}\ \underbrace{E(X^2)}\ \underbrace{E(X^3)} \cdots$

$z_i \to x_i$

- Consider the Kullback-Leibler (KL) divergence [1] between the prior probability of the latent variable and the posterior of the latent variable:

$$
\begin{aligned}
\mathrm{KL}\big(q(\boldsymbol{z}_i)\,\|\,\mathbb{P}(\boldsymbol{z}_i\,|\,\boldsymbol{x}_i,\boldsymbol{\theta})\big) &\overset{(a)}{=} \int q(\boldsymbol{z}_i)\log\big(\frac{q(\boldsymbol{z}_i)}{\mathbb{P}(\boldsymbol{z}_i\,|\,\boldsymbol{x}_i,\boldsymbol{\theta})}\big)\,d\boldsymbol{z}_i \\
&= \int q(\boldsymbol{z}_i)\big(\log(q(\boldsymbol{z}_i)) - \log(\mathbb{P}(\boldsymbol{z}_i\,|\,\boldsymbol{x}_i,\boldsymbol{\theta}))\big)\,d\boldsymbol{z}_i \\
&\overset{(2)}{=} \int q(\boldsymbol{z}_i)\big(\log(q(\boldsymbol{z}_i)) - \log(\mathbb{P}(\boldsymbol{x}_i\,|\,\boldsymbol{z}_i,\boldsymbol{\theta})) - \log(\mathbb{P}(\boldsymbol{z}_i\,|\,\boldsymbol{\theta})) + \log(\mathbb{P}(\boldsymbol{x}_i\,|\,\boldsymbol{\theta}))\big)\,d\boldsymbol{z}_i \\
&\overset{(b)}{=} \log(\mathbb{P}(\boldsymbol{x}_i\,|\,\boldsymbol{\theta})) + \int q(\boldsymbol{z}_i)\big(\log(q(\boldsymbol{z}_i)) - \log(\mathbb{P}(\boldsymbol{x}_i\,|\,\boldsymbol{z}_i,\boldsymbol{\theta})) - \log(\mathbb{P}(\boldsymbol{z}_i\,|\,\boldsymbol{\theta}))\big)\,d\boldsymbol{z}_i \\
&= \log(\mathbb{P}(\boldsymbol{x}_i\,|\,\boldsymbol{\theta})) + \int q(\boldsymbol{z}_i)\log\big(\frac{q(\boldsymbol{z}_i)}{\mathbb{P}(\boldsymbol{x}_i\,|\,\boldsymbol{z}_i,\boldsymbol{\theta})\mathbb{P}(\boldsymbol{z}_i\,|\,\boldsymbol{\theta})}\big)\,d\boldsymbol{z}_i \\
&= \log(\mathbb{P}(\boldsymbol{x}_i\,|\,\boldsymbol{\theta})) + \int q(\boldsymbol{z}_i)\log\big(\frac{q(\boldsymbol{z}_i)}{\mathbb{P}(\boldsymbol{x}_i,\boldsymbol{z}_i\,|\,\boldsymbol{\theta})}\big)\,d\boldsymbol{z}_i \\
&= \log(\mathbb{P}(\boldsymbol{x}_i\,|\,\boldsymbol{\theta})) + \mathrm{KL}\big(q(\boldsymbol{z}_i)\,\|\,\mathbb{P}(\boldsymbol{x}_i,\boldsymbol{z}_i\,|\,\boldsymbol{\theta})\big),
\end{aligned}
$$

$\underbrace{P(x_i|z_i)\,P(z_i)}_{P(x_i, z_i)}$

where (a) is for definition of KL divergence and (b) is because $\log(\mathbb{P}(\boldsymbol{x}_i\,|\,\boldsymbol{\theta}))$ is independent of $\boldsymbol{z}_i$ and comes out of integral and $\int d\boldsymbol{z}_i = 1$.

- Hence:

$$
\log(\mathbb{P}(\boldsymbol{x}_i\,|\,\boldsymbol{\theta})) = \mathrm{KL}\big(q(\boldsymbol{z}_i)\,\|\,\mathbb{P}(\boldsymbol{z}_i\,|\,\boldsymbol{x}_i,\boldsymbol{\theta})\big) - \mathrm{KL}\big(q(\boldsymbol{z}_i)\,\|\,\mathbb{P}(\boldsymbol{x}_i,\boldsymbol{z}_i\,|\,\boldsymbol{\theta})\big). \tag{3}
$$

Variational Inference

- We found:

$$\log(\mathbb{P}(\boldsymbol{x}_i \mid \boldsymbol{\theta})) = \mathrm{KL}\big(q(\boldsymbol{z}_i) \,\|\, \mathbb{P}(\boldsymbol{z}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})\big) - \mathrm{KL}\big(q(\boldsymbol{z}_i) \,\|\, \mathbb{P}(\boldsymbol{x}_i, \boldsymbol{z}_i \mid \boldsymbol{\theta})\big).$$

- We define the **Evidence Lower Bound (ELBO)** as:

$$\mathcal{L}(q, \boldsymbol{\theta}) := -\mathrm{KL}\big(q(\boldsymbol{z}_i) \,\|\, \mathbb{P}(\boldsymbol{x}_i, \boldsymbol{z}_i \mid \boldsymbol{\theta})\big). \tag{4}$$

  So:

$$\log(\mathbb{P}(\boldsymbol{x}_i \mid \boldsymbol{\theta})) = \mathrm{KL}\big(q(\boldsymbol{z}_i) \,\|\, \mathbb{P}(\boldsymbol{z}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})\big) + \mathcal{L}(q, \boldsymbol{\theta}).$$

- Therefore:

$$\mathcal{L}(q, \boldsymbol{\theta}) = \log(\mathbb{P}(\boldsymbol{x}_i \mid \boldsymbol{\theta})) - \underbrace{\mathrm{KL}\big(q(\boldsymbol{z}_i) \,\|\, \mathbb{P}(\boldsymbol{z}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})\big)}_{\geq 0}. \tag{5}$$

- As the second term is negative with its minus, the ELBO is a lower bound on the log likelihood of data:

$$\boxed{\mathcal{L}(q, \boldsymbol{\theta}) \leq \log(\mathbb{P}(\boldsymbol{x}_i \mid \boldsymbol{\theta})).} \tag{6}$$

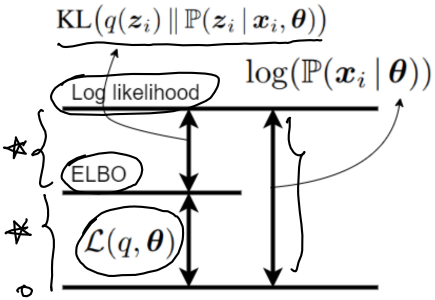
  The likelihood $\mathbb{P}(\boldsymbol{x}_i \mid \boldsymbol{\theta})$ is also referred to as the **evidence**.
- Note that this lower bound gets tight when:

$$\mathcal{L}(q, \boldsymbol{\theta}) \approx \log(\mathbb{P}(\boldsymbol{x}_i \mid \boldsymbol{\theta})) \implies 0 \leq \mathrm{KL}\big(q(\boldsymbol{z}_i) \,\|\, \mathbb{P}(\boldsymbol{z}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})\big) \overset{\text{set}}{=} 0$$
$$\implies q(\boldsymbol{z}_i) = \mathbb{P}(\boldsymbol{z}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}). \tag{7}$$

Variational Inference

- We found:

$$\log(\mathbb{P}(\boldsymbol{x}_i \,|\, \boldsymbol{\theta})) = \text{KL}\big(q(\boldsymbol{z}_i) \,\|\, \mathbb{P}(\boldsymbol{z}_i \,|\, \boldsymbol{x}_i, \boldsymbol{\theta})\big) + \mathcal{L}(q, \boldsymbol{\theta}).$$



Expectation Maximization in Variational Inference

- According to MLE, we want to maximize the log-likelihood of data. According to Eq. (6):

$$\underbrace{\mathcal{L}(q, \boldsymbol{\theta}) \leq \log(\mathbb{P}(\boldsymbol{x}_i \,|\, \boldsymbol{\theta}))}, \quad *$$

maximizing the ELBO will also maximize the log-likelihood.

- The Eq. (6) holds for any prior distribution q. We want to find the best distribution to maximize the lower bound.
- Hence, EM for variational inference is performed iteratively as:

$$\text{E-step:} \quad q^{(t)} := \arg\max_{q} \; \underbrace{\mathcal{L}(q, \boldsymbol{\theta}^{(t-1)})}, \tag{8}$$

$$\text{M-step:} \quad \boldsymbol{\theta}^{(t)} := \arg\max_{\boldsymbol{\theta}} \; \underbrace{\mathcal{L}(q^{(t)}, \boldsymbol{\theta})}, \tag{9}$$

where t denotes the iteration index.

Expectation Maximization in Variational Inference

- **E-step in EM for Variational Inference**: The E-step is:

$$\max_q \mathcal{L}(q, \boldsymbol{\theta}^{(t-1)}) \overset{(5)}{=} \max_q \log(\mathbb{P}(\boldsymbol{x}_i \mid \boldsymbol{\theta}^{(t-1)})) + \max_q \big(-\text{KL}\big(q(\boldsymbol{z}_i) \,\|\, \mathbb{P}(\boldsymbol{z}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(t-1)})\big)\big)$$
$$= \max_q \log(\mathbb{P}(\boldsymbol{x}_i \mid \boldsymbol{\theta}^{(t-1)})) + \min_q \text{KL}\big(q(\boldsymbol{z}_i) \,\|\, \mathbb{P}(\boldsymbol{z}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(t-1)})\big).$$

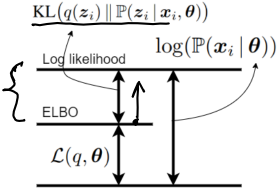
- The second term is always non-negative; hence, its minimum is zero:

$$\text{KL}\big(q(\boldsymbol{z}_i) \,\|\, \mathbb{P}(\boldsymbol{z}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(t-1)})\big) \overset{\text{set}}{=} 0 \implies q(\boldsymbol{z}_i) = \mathbb{P}(\boldsymbol{z}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(t-1)}),$$

which was already found in Eq. (7). Thus, the E-step assigns:

$$\boxed{q^{(t)}(\boldsymbol{z}_i) \leftarrow \mathbb{P}(\boldsymbol{z}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(t-1)}).} \tag{10}$$

- In other words, in the figure, it pushes the middle line toward the above line by maximizing the ELBO.



Expectation Maximization in Variational Inference

- **M-step in EM for Variational Inference:** The M-step is:

$$
\begin{aligned}
\max_{\boldsymbol{\theta}} \mathcal{L}(q^{(t)}, \boldsymbol{\theta}) &\overset{(4)}{=} \max_{\boldsymbol{\theta}} \big(- \mathrm{KL}\big(q^{(t)}(\boldsymbol{z}_i) \,\|\, \mathbb{P}(\boldsymbol{x}_i, \boldsymbol{z}_i \,|\, \boldsymbol{\theta})\big)\big) \\
&\overset{(a)}{=} \max_{\boldsymbol{\theta}} \Big[- \int q^{(t)}(\boldsymbol{z}_i) \log\Big(\frac{q^{(t)}(\boldsymbol{z}_i)}{\mathbb{P}(\boldsymbol{x}_i, \boldsymbol{z}_i \,|\, \boldsymbol{\theta})}\Big) \, d\boldsymbol{z}_i \Big] \\
&= \max_{\boldsymbol{\theta}} \int q^{(t)}(\boldsymbol{z}_i) \log(\mathbb{P}(\boldsymbol{x}_i, \boldsymbol{z}_i \,|\, \boldsymbol{\theta})) \, d\boldsymbol{z}_i - \max_{\boldsymbol{\theta}} \int q^{(t)}(\boldsymbol{z}_i) \log(q^{(t)}(\boldsymbol{z}_i)) \, d\boldsymbol{z}_i,
\end{aligned}
$$

where (a) is for definition of KL divergence.

- The second term is constant w.r.t. $\boldsymbol{\theta}$. Hence:

$$
\begin{aligned}
\max_{\boldsymbol{\theta}} \mathcal{L}(q^{(t)}, \boldsymbol{\theta}) &= \max_{\boldsymbol{\theta}} \int q^{(t)}(\boldsymbol{z}_i) \log(\mathbb{P}(\boldsymbol{x}_i, \boldsymbol{z}_i \,|\, \boldsymbol{\theta})) \, d\boldsymbol{z}_i \\
&\overset{(a)}{=} \max_{\boldsymbol{\theta}} \mathbb{E}_{\sim q^{(t)}(\boldsymbol{z}_i)} \big[\log \mathbb{P}(\boldsymbol{x}_i, \boldsymbol{z}_i \,|\, \boldsymbol{\theta}) \big],
\end{aligned}
$$

where (a) is because of definition of expectation. Thus, the M-step assigns:

$$
\boldsymbol{\theta}^{(t)} \leftarrow \arg\max_{\boldsymbol{\theta}} \; \mathbb{E}_{\sim q^{(t)}(\boldsymbol{z}_i)} \big[\log \mathbb{P}(\boldsymbol{x}_i, \boldsymbol{z}_i \,|\, \boldsymbol{\theta}) \big]. \tag{11}
$$

$$\max_{\theta} \; f(\theta) + g(\lambda)$$

$$\max_{\theta} \; f(\theta) \longrightarrow \cdots \longrightarrow \theta^*$$

$$\underset{\theta}{\operatorname{argmax}} \; f(\theta) + g(\lambda) \longrightarrow \theta^*$$

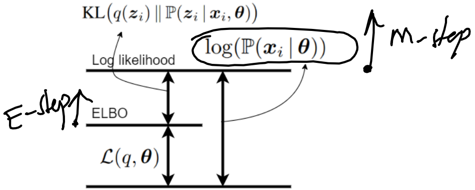
$$f(\theta^*) + \underline{g(\lambda)}$$

Expectation Maximization in Variational Inference

- We found:

$$\boldsymbol{\theta}^{(t)} \leftarrow \arg\max_{\boldsymbol{\theta}} \; \mathbb{E}_{\sim q^{(t)}(\boldsymbol{z}_i)}\big[\log \mathbb{P}(\boldsymbol{x}_i, \boldsymbol{z}_i \,|\, \boldsymbol{\theta})\big].$$

- In other words, in the figure, it pushes the above line higher.



- The E-step and M-step together somehow play a **game** where the E-step tries to reach the middle line (or the ELBO) to the log-likelihood and the M-step tries to increase the above line (or the log-likelihood). This procedure is done repeatedly so the two steps help each other improve to higher values.
- To summarize, the EM in variational inference is:

$$q^{(t)}(\boldsymbol{z}_i) \leftarrow \mathbb{P}(\boldsymbol{z}_i \,|\, \boldsymbol{x}_i, \boldsymbol{\theta}^{(t-1)}), \tag{12}$$

$$\boldsymbol{\theta}^{(t)} \leftarrow \arg\max_{\boldsymbol{\theta}} \; \mathbb{E}_{\sim q^{(t)}(\boldsymbol{z}_i)}\big[\log \mathbb{P}(\boldsymbol{x}_i, \boldsymbol{z}_i \,|\, \boldsymbol{\theta})\big]. \tag{13}$$

Expectation Maximization in Variational Inference

- It is noteworthy that, in variational inference, sometimes, the parameter $\boldsymbol{\theta}$ is absorbed into the latent variable $\boldsymbol{z}_i$.
- According to the chain rule, we have:

$$\mathbb{P}(\boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{\theta}) = \mathbb{P}(\boldsymbol{x}_i \,|\, \boldsymbol{z}_i, \boldsymbol{\theta})\, \mathbb{P}(\boldsymbol{z}_i \,|\, \boldsymbol{\theta})\, \mathbb{P}(\boldsymbol{\theta}).$$

- Considering the term $\mathbb{P}(\boldsymbol{z}_i \,|\, \boldsymbol{\theta})\, \mathbb{P}(\boldsymbol{\theta})$ as one probability term, we have:

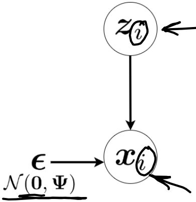
$$\mathbb{P}(\boldsymbol{x}_i, \boldsymbol{z}_i) = \mathbb{P}(\boldsymbol{x}_i \,|\, \boldsymbol{z}_i)\, \mathbb{P}(\boldsymbol{z}_i),$$

where the parameter $\boldsymbol{\theta}$ disappears because of absorption.

Factor Analysis

Factor Analysis

- Factor analysis [2, 3, 4, 5] is one of the simplest and most fundamental **generative** models.
- Factor analysis assumes that every data point $\boldsymbol{x}_i \in \mathbb{R}^d$ is generated from a latent variable $\boldsymbol{z}_i \in \mathbb{R}^p$. The **latent variable** is also referred to as the **latent factor**; hence, the name of factor analysis comes from the fact that it analyzes the latent factors.
- In factor analysis, we assume that the data point $\boldsymbol{x}_i$ is obtained through the following steps: (1) by **linear projection** of the p-dimensional $\boldsymbol{z}_i$ onto a d-dimensional space by projection matrix $\boldsymbol{\Lambda} \in \mathbb{R}^{d \times p}$, then (2) applying some **linear translation**, and finally (3) **adding a Gaussian noise** $\boldsymbol{\epsilon} \in \mathbb{R}^d$ with covariance matrix $\boldsymbol{\Psi} \in \mathbb{R}^{d \times d}$.
- Note that as the noises in different dimensions are **independent**, the covariance matrix $\boldsymbol{\Psi}$ is diagonal.
- Factor analysis can be illustrated as a graphical model [6] where the visible data variable is conditioned on the latent variable and the noise random variable.



Factor Analysis

- For simplicity, the prior distribution of the latent variable can be assumed to be a multivariate Gaussian distribution:

$$\mathbb{P}(\boldsymbol{z}_i) = \underline{\mathcal{N}(\boldsymbol{z}_i \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)} = \frac{1}{\sqrt{(2\pi)^p |\boldsymbol{\Sigma}_0|}} \exp\Big(-\frac{(\boldsymbol{z}_i - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{z}_i - \boldsymbol{\mu}_0)}{2}\Big), \tag{14}$$

where $\boldsymbol{\mu}_0 \in \mathbb{R}^p$ and $\boldsymbol{\Sigma}_0 \in \mathbb{R}^{p \times p}$ are the mean and the covariance matrix of $\boldsymbol{z}_i$ and $|.|$ is the determinant of matrix.

- $\boldsymbol{x}_i$ is obtained through (1) the linear projection of $\boldsymbol{z}_i$ by $\boldsymbol{\Lambda} \in \mathbb{R}^{d \times p}$, (2) applying some linear translation, and (3) adding a Gaussian noise $\boldsymbol{\epsilon} \in \mathbb{R}^d$ with covariance $\boldsymbol{\Psi} \in \mathbb{R}^{d \times d}$.
- Hence, the data point $\boldsymbol{x}_i$ has a **conditional multivariate Gaussian distribution given the latent variable**; its conditional likelihood is:

$$\underline{\mathbb{P}(\boldsymbol{x}_i \mid \boldsymbol{z}_i)} = \mathbb{P}(\boldsymbol{x}_i \mid \boldsymbol{z}_i, \boldsymbol{\Lambda}, \boldsymbol{\mu}, \boldsymbol{\Psi}) = \underline{\mathcal{N}(\boldsymbol{\Lambda}\boldsymbol{z}_i + \boldsymbol{\mu}, \boldsymbol{\Psi})}, \tag{15}$$

where $\boldsymbol{\mu}$, which is the translation vector, is the mean of data $\{\boldsymbol{x}_i\}_{i=1}^n$:

$$\mathbb{R}^d \ni \boldsymbol{\mu} := \frac{1}{n} \sum_{i=1}^n \boldsymbol{x}_i. \tag{16}$$

Factor Analysis

- The marginal distribution of $\boldsymbol{x}_i$ is:

$P(x_i, z_i)$

$x \mapsto \Lambda x$

$\mu \mapsto \Lambda \mu$

$\Sigma \mapsto \Lambda \Sigma \Lambda^\top$

$$\mathbb{P}(\boldsymbol{x}_i) = \int \mathbb{P}(\boldsymbol{x}_i \,|\, \boldsymbol{z}_i)\,\mathbb{P}(\boldsymbol{z}_i)\,d\boldsymbol{z}_i \implies$$

$$\mathbb{P}(\boldsymbol{x}_i \,|\, \boldsymbol{\Lambda}, \boldsymbol{\mu}, \boldsymbol{\Psi}) = \int \mathbb{P}(\boldsymbol{x}_i \,|\, \boldsymbol{z}_i, \boldsymbol{\Lambda}, \boldsymbol{\mu}, \boldsymbol{\Psi})\,\mathbb{P}(\boldsymbol{z}_i \,|\, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)\,d\boldsymbol{z}_i$$

$$\overset{(a)}{=} \mathcal{N}(\boldsymbol{\Lambda}\boldsymbol{\mu}_0 + \boldsymbol{\mu}, \boldsymbol{\Psi} + \boldsymbol{\Lambda}\boldsymbol{\Sigma}_0\boldsymbol{\Lambda}^\top) \qquad \Lambda \Sigma_0 \Lambda^\top = \underbrace{\Lambda \Sigma_0^{1/2}}_{\widehat{\Lambda}} \underbrace{\Sigma_0^{1/2} \Lambda^\top}_{\widehat{\Lambda}^\top} \tag{17}$$

$$= \mathcal{N}(\widehat{\boldsymbol{\mu}}, \boldsymbol{\Psi} + \widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{\Lambda}}^\top), \tag{18}$$

where $\mathbb{R}^d \ni \widehat{\boldsymbol{\mu}} := \boldsymbol{\Lambda}\boldsymbol{\mu}_0 + \boldsymbol{\mu}$, $\mathbb{R}^{d\times d} \ni \widehat{\boldsymbol{\Lambda}} := \boldsymbol{\Lambda}\boldsymbol{\Sigma}_0^{(1/2)}$, and (a) is because mean is linear and variance is quadratic so the mean and variance of projection are applied linearly and quadratically, respectively.

- As the mean $\widehat{\boldsymbol{\mu}}$ and covariance $\widehat{\boldsymbol{\Lambda}}$ are needed to be learned, we can absorb $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ into $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ and assume that $\boldsymbol{\mu}_0 = \boldsymbol{0}$ and $\boldsymbol{\Sigma}_0 = \boldsymbol{I}$.
- In summary, factor analysis assumes every data point $\boldsymbol{x}_i \in \mathbb{R}^d$ is obtained by projecting a latent variable $\boldsymbol{z}_i \in \mathbb{R}^p$ onto a d-dimensional space by projection matrix $\boldsymbol{\Lambda} \in \mathbb{R}^{d\times p}$ and translating it by $\boldsymbol{\mu} \in \mathbb{R}^d$ and finally adding some Gaussian noise $\boldsymbol{\epsilon} \in \mathbb{R}^d$ (whose dimensions are independent) as:

$U^\top z_i = \Lambda z_i$, $\quad U \to \Lambda$

$$\boldsymbol{x}_i := \boldsymbol{\Lambda}\boldsymbol{z}_i + \boldsymbol{\mu} + \boldsymbol{\epsilon}, \tag{19}$$

$$\mathbb{P}(\boldsymbol{z}_i) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \tag{20}$$

$$\mathbb{P}(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi}). \tag{21}$$

$E(x_i) = \underbrace{E(\Lambda z_i)}_{0} + \underbrace{E(\mu)}_{\mu} + \underbrace{E(\varepsilon)}_{0}$

Factor Analysis

- The joint distribution of $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ is:

$$\boldsymbol{y}_i := \begin{bmatrix} \boldsymbol{x}_i \\ \boldsymbol{z}_i \end{bmatrix} \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y). \tag{22}$$

- The expectation of $\boldsymbol{x}_i$ is:

$$\mathbb{E}[\boldsymbol{x}_i] \overset{(19)}{=} \mathbb{E}[\boldsymbol{\Lambda} \boldsymbol{z}_i + \boldsymbol{\mu} + \boldsymbol{\epsilon}] = \boldsymbol{\Lambda} \overbrace{\mathbb{E}[\boldsymbol{z}_i]}^{0} + \boldsymbol{\mu} + \overbrace{\mathbb{E}[\boldsymbol{\epsilon}]}^{0} \overset{(a)}{=} \boldsymbol{\mu}, \tag{23}$$

where (a) is because of Eqs. (20) and (21).

- Hence:

$$\boldsymbol{\mu}_y := \begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_z \end{bmatrix} \overset{(a)}{=} \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{0} \end{bmatrix}, \tag{24}$$

where (a) is because of Eqs. (20) and (23).

Factor Analysis

- Lemma:

Lemma

Consider two random variables $\boldsymbol{x}_i \in \mathbb{R}^d$ and $\boldsymbol{z}_i \in \mathbb{R}^p$ and let $\boldsymbol{y}_i := [\boldsymbol{x}_i^\top, \boldsymbol{z}_i^\top]^\top \in \mathbb{R}^{d+p}$. Assume that $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ are jointly multivariate Gaussian; hence, the variable $\boldsymbol{y}_i$ has a multivariate Gaussian distribution, i.e., $\boldsymbol{y}_i \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$. The mean and covariance can be decomposed as:

$$\boldsymbol{\mu}_y = [\boldsymbol{\mu}^\top, \boldsymbol{\mu}_0^\top]^\top \in \mathbb{R}^{d+p}, \tag{25}$$

$$\boldsymbol{\Sigma}_y = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \in \mathbb{R}^{(d+p)\times(d+p)}, \tag{26}$$

where $\boldsymbol{\mu} \in \mathbb{R}^d$, $\boldsymbol{\mu}_0 \in \mathbb{R}^p$, $\boldsymbol{\Sigma}_{11} \in \mathbb{R}^{d\times d}$, $\boldsymbol{\Sigma}_{22} \in \mathbb{R}^{p\times p}$, $\boldsymbol{\Sigma}_{12} \in \mathbb{R}^{d\times p}$, and $\boldsymbol{\Sigma}_{21} = \boldsymbol{\Sigma}_{12}^\top \in \mathbb{R}^{p\times d}$.

Factor Analysis

- Lemma [7]:

Lemma

$$\mathbb{R}^d \ni \boldsymbol{\mu}_{x|z} := \boldsymbol{\mu} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{z}_i - \boldsymbol{\mu}_0), \tag{27}$$

$$\mathbb{R}^{d\times d} \ni \boldsymbol{\Sigma}_{x|z} := \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}, \tag{28}$$

and likewise for $\boldsymbol{z}_i|\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_{z|x}, \boldsymbol{\Sigma}_{z|x})$*:*

$$\mathbb{R}^p \ni \boldsymbol{\mu}_{z|x} := \boldsymbol{\mu}_0 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}), \tag{29}$$

$$\mathbb{R}^{p\times p} \ni \boldsymbol{\Sigma}_{z|x} := \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}. \tag{30}$$

Factor Analysis

$\mathbb{E}(z_i z_i^\top) = \mathbb{E}[(z_i - \mu_z)(z_i - \mu_z)^\top]$

- According to Eq. (20), we have $\boldsymbol{\Sigma}_{22} = \boldsymbol{\Sigma}_z = \boldsymbol{I}$. According to Eq. (19), we have:

$$\begin{aligned}
\boldsymbol{\Sigma}_{11} = \boldsymbol{\Sigma}_x &= \mathbb{E}[(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top] \\
&= \mathbb{E}[(\boldsymbol{\Lambda z}_i + \boldsymbol{\mu} + \boldsymbol{\epsilon} - \boldsymbol{\mu})(\boldsymbol{\Lambda z}_i + \boldsymbol{\mu} + \boldsymbol{\epsilon} - \boldsymbol{\mu})^\top] \\
&= \mathbb{E}[\boldsymbol{\Lambda z}_i \boldsymbol{z}_i^\top \boldsymbol{\Lambda}^\top + \boldsymbol{\epsilon z}_i^\top \boldsymbol{\Lambda}^\top + \boldsymbol{\Lambda z}_i \boldsymbol{\epsilon}^\top + \boldsymbol{\epsilon\epsilon}^\top] \\
&= \boldsymbol{\Lambda}\mathbb{E}[\boldsymbol{z}_i \boldsymbol{z}_i^\top]\boldsymbol{\Lambda}^\top + \mathbb{E}[\boldsymbol{\epsilon}]\mathbb{E}[\boldsymbol{z}_i]^\top \boldsymbol{\Lambda}^\top + \boldsymbol{\Lambda}\mathbb{E}[\boldsymbol{z}_i]\mathbb{E}[\boldsymbol{\epsilon}]^\top + \mathbb{E}[\boldsymbol{\epsilon\epsilon}^\top] \\
&\overset{(a)}{=} \boldsymbol{\Lambda I \Lambda}^\top + \boldsymbol{0} + \boldsymbol{0} + \boldsymbol{\Psi} = \boldsymbol{\Lambda\Lambda}^\top + \boldsymbol{\Psi},
\end{aligned} \tag{31}$$

$\mathbb{E}((\varepsilon - \mu_\varepsilon)(\varepsilon - \mu_\varepsilon)^\top)$

where (a) is because of Eqs. (20) and (21).

- Moreover, we have:

$$\begin{aligned}
\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{xz} &= \mathbb{E}[(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{z}_i - \boldsymbol{\mu}_0)^\top] \\
&\overset{(a)}{=} \mathbb{E}[(\boldsymbol{\Lambda z}_i + \boldsymbol{\mu} + \boldsymbol{\epsilon} - \boldsymbol{\mu})(\boldsymbol{z}_i - \boldsymbol{0})^\top] \\
&\overset{(b)}{=} \boldsymbol{\Lambda}\mathbb{E}[\boldsymbol{z}_i \boldsymbol{z}_i^\top] + \mathbb{E}[\boldsymbol{\epsilon}]\mathbb{E}[\boldsymbol{z}_i^\top] = \boldsymbol{\Lambda I} + (\boldsymbol{00}^\top) = \boldsymbol{\Lambda},
\end{aligned} \tag{32}$$

where (a) is because of Eqs. (19) and (20) and (b) is because $\boldsymbol{z}_i$ and $\boldsymbol{\epsilon}$ are independent.

- We also have $\boldsymbol{\Sigma}_{21} = \boldsymbol{\Sigma}_{12}^\top = \boldsymbol{\Lambda}^\top$. Therefore:

$$\begin{bmatrix} \boldsymbol{x}_i \\ \boldsymbol{z}_i \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Lambda\Lambda}^\top + \boldsymbol{\Psi} & \boldsymbol{\Lambda} \\ \boldsymbol{\Lambda}^\top & \boldsymbol{I} \end{bmatrix} \right). \tag{33}$$

Factor Analysis

- Hence, the marginal distribution of data point $\boldsymbol{x}_i$ is:

$$\star \quad \mathbb{P}(\boldsymbol{x}_i) = \mathbb{P}(\boldsymbol{x}_i \,|\, \boldsymbol{\Lambda}, \boldsymbol{\mu}, \boldsymbol{\Psi}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}). \tag{34}$$

According to Eqs. (29) and (30) [Lemma], the posterior or the conditional distribution of latent variable given data is:

$$\begin{aligned} q(\boldsymbol{z}_i) \overset{(12)}{=} \mathbb{P}(\boldsymbol{z}_i \,|\, \boldsymbol{x}_i) &= \mathbb{P}(\boldsymbol{z}_i \,|\, \boldsymbol{x}_i, \boldsymbol{\Lambda}, \boldsymbol{\mu}, \boldsymbol{\Psi}) \\ &= \mathcal{N}(\boldsymbol{\mu}_{z|x}, \boldsymbol{\Sigma}_{z|x}), \end{aligned} \tag{35}$$

where:

$$\star \begin{cases} \mathbb{R}^p \ni \boldsymbol{\mu}_{z|x} := \boldsymbol{\Lambda}^\top(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi})^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}), & (36) \\ \mathbb{R}^{p\times p} \ni \boldsymbol{\Sigma}_{z|x} := \boldsymbol{I} - \boldsymbol{\Lambda}^\top(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi})^{-1}\boldsymbol{\Lambda}. & (37) \end{cases}$$

- Recall that the conditional distribution of data given the latent variable, i.e. $\mathbb{P}(\boldsymbol{x}_i \,|\, \boldsymbol{z}_i)$, was introduced in Eq. (15):

$$\star\star \quad \mathbb{P}(\boldsymbol{x}_i \,|\, \boldsymbol{z}_i) = \mathbb{P}(\boldsymbol{x}_i \,|\, \boldsymbol{z}_i, \boldsymbol{\Lambda}, \boldsymbol{\mu}, \boldsymbol{\Psi}) = \mathcal{N}(\boldsymbol{\Lambda}\boldsymbol{z}_i + \boldsymbol{\mu}, \boldsymbol{\Psi}).$$

If data $\{\boldsymbol{x}_i\}_{i=1}^n$ are centered, i.e. $\boldsymbol{\mu} = \boldsymbol{0}$, the marginal of data, Eq. (34), and the likelihood of data, Eq. (15), become:

$$\mathbb{P}(\boldsymbol{x}_i \,|\, \boldsymbol{\Lambda}, \boldsymbol{\Psi}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi} + \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top), \tag{38}$$

$$\mathbb{P}(\boldsymbol{x}_i \,|\, \boldsymbol{z}_i, \boldsymbol{\Lambda}, \boldsymbol{\Psi}) = \mathcal{N}(\boldsymbol{\Lambda}\boldsymbol{z}_i, \boldsymbol{\Psi}), \tag{39}$$

respectively. In some works, people center the data as a pre-processing to factor analysis.

Factor Analysis

- We can find the parameters $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ using Expectation Maximization.
- See our tutorial "Factor analysis, probabilistic principal component analysis, variational inference, and variational autoencoder: Tutorial and survey" [8] for the details of EM steps in factor analysis.

Probabilistic Principal Component Analysis

Probabilistic Principal Component Analysis

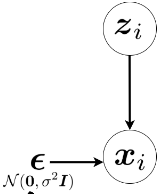
- **Probabilistic PCA (PPCA)** (1997-1999) [9, 10] is a **special case of factor analysis** where the variance of noise is equal in all dimensions of data space with covariance between dimensions, i.e.:

$$\boldsymbol{\Psi} = \sigma^2 \boldsymbol{I}. \tag{40}$$

- In other words, PPCA considers an **isotropic noise** in its formulation. Therefore, Eq. (21) is simplified to:

$$\mathbb{P}(\boldsymbol{\epsilon}) = \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}). \tag{41}$$

- Because of having zero covariance of noise between different dimensions, PPCA assumes that the data points are **independent** of each other given latent variables.
- PPCA can be illustrated as a graphical model, where the visible data variable is conditioned on the latent variable and the isotropic noise random variable.



Probabilistic Principal Component Analysis

- As PPCA is a special case of factor analysis, it also is solved using EM. Similar to factor analysis, it can be solved **iteratively using EM** [9].
- However, one can also find a **closed-form solution to its EM approach** [10]. Hence, by restricting the noise covariance to be isotropic, its **solution becomes simpler and closed-form**.
- We can find the parameters $\mathbf{\Lambda}$ and σ using Expectation Maximization.
- See our tutorial "Factor analysis, probabilistic principal component analysis, variational inference, and variational autoencoder: Tutorial and survey" [8] for the details of EM steps in PPCA.

Acknowledgment

- Some slides are based on our tutorial paper: "Factor analysis, probabilistic principal component analysis, variational inference, and variational autoencoder: Tutorial and survey" [8]
- Some slides of this slide deck are inspired by teachings of deep learning course at the Carnegie Mellon University (you can see their YouTube channel).
- Factor analysis in sklearn: `https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.FactorAnalysis.html`

References

[1] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[2] B. Fruchter, *Introduction to factor analysis*.
Van Nostrand, 1954.

[3] R. B. Cattell, "A biometrics invited paper. factor analysis: An introduction to essentials i. the purpose and underlying models," *Biometrics*, vol. 21, no. 1, pp. 190–215, 1965.

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

[5] D. Child, *The essentials of factor analysis*.
Cassell Educational, 1990.

[6] Z. Ghahramani and G. E. Hinton, "The EM algorithm for mixtures of factor analyzers," tech. rep., Technical Report CRG-TR-96-1, University of Toronto, 1996.

[7] A. Ng, "CS229 lecture notes for factor analysis," tech. rep., Stanford University, 2018.

[8] B. Ghojogh, A. Ghodsi, F. Karray, and M. Crowley, "Factor analysis, probabilistic principal component analysis, variational inference, and variational autoencoder: Tutorial and survey," *arXiv preprint arXiv:2101.00734*, 2021.

[9] S. Roweis, "EM algorithms for PCA and SPCA," *Advances in neural information processing systems*, vol. 10, pp. 626–632, 1997.

References (cont.)

[10] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.