

Gaussian Linear Latent Variable Models

François Ged

Laboratoire de Mathématiques de Bretagne Atlantique
UMR 6205, CNRS – Université de Brest –

Besançon meeting on Probability, Ecology and Evolution 2026



I. Gaussian Linear Latent Variable Models

1. Notation:

We work with continuous densities on subsets of $\mathbb{R}^d$.

For a joint density $p_{(X,Y)}$ of RV (X, Y) , we write:

- $p(x, y) = p_{(X,Y)}(x, y)$;
- $p(x) = p_X(x)$ and $p(y) = p_Y(y)$;
- $p(x | y) = p_{X|Y}(x | y) = \frac{p(x,y)}{p(y)}$;
- ...

I. Gaussian Linear Latent Variable Models

1. Notation:

We work with continuous densities on subsets of $\mathbb{R}^d$.

For a joint density $p_{(X,Y)}$ of RV (X, Y) , we write:

- $p(x, y) = p_{(X,Y)}(x, y)$;
- $p(x) = p_X(x)$ and $p(y) = p_Y(y)$;
- $p(x | y) = p_{X|Y}(x | y) = \frac{p(x,y)}{p(y)}$;
- ...

Similarly when working with a dataset $(x^i)_{i \in [n]} \stackrel{\text{i.i.d.}}{\sim} p_X$, $[n] := \{1, \dots, n\}$, we write

- $p((x^i)_{i \in [n]}) = \bigotimes_{i \in [n]} p_X(x^i)$.

We make all necessary assumptions for quantities to be well defined a.e.

I. Gaussian Linear Latent Variable Models

1. Notation:

We work with continuous densities on subsets of $\mathbb{R}^d$.

For a joint density $p_{(X,Y)}$ of RV (X, Y) , we write:

- $p(x, y) = p_{(X,Y)}(x, y)$;
- $p(x) = p_X(x)$ and $p(y) = p_Y(y)$;
- $p(x | y) = p_{X|Y}(x | y) = \frac{p(x,y)}{p(y)}$;
- ...

Similarly when working with a dataset $(x^i)_{i \in [n]} \stackrel{\text{i.i.d.}}{\sim} p_X$, $[n] := \{1, \dots, n\}$, we write

- $p((x^i)_{i \in [n]}) = \bigotimes_{i \in [n]} p_X(x^i)$.

We make all necessary assumptions for quantities to be well defined a.e.

Henceforth, if p is parametrised by **fixed** parameters θ , we write p_θ and stop using subscript for specifying the random variables (like in p_X).

I. Gaussian Linear Latent Variable Models

2. Gaussian Models

2.1. The Generative Model

Let $p, k \in \mathbb{N}$ with $k \leq p$. Let $\mathbf{W} \in \mathbb{R}^{p \times k}$. Let $\Psi \in \mathbb{R}^{p \times p}$ be symmetric and strictly positive definite (denoted by $\Psi \succ 0$). Let $\mu \in \mathbb{R}^p$.

Define the random variable X in $\mathbb{R}^p$ by

$$\begin{aligned} X &= \mu + \mathbf{W}Z + \epsilon, & \mathbf{W} &\in \mathbb{R}^{p \times k}, \\ Z &\stackrel{d}{\sim} \mathcal{N}(0, \mathbf{I}_k), & \epsilon &\stackrel{d}{\sim} \mathcal{N}(0, \Psi), \quad Z \perp \epsilon. \end{aligned}$$

Define $\Sigma := \mathbf{W}\mathbf{W}^\top + \Psi \in \mathbb{R}^{p \times p}$ (symmetric and $\Sigma \succ 0$).

I. Gaussian Linear Latent Variable Models

$$X = \mu + \mathbf{W}Z + \epsilon, \quad \mathbf{W} \in \mathbb{R}^{p \times k},$$
$$Z \stackrel{d}{\sim} \mathcal{N}(0, \mathbf{I}_k), \quad \epsilon \stackrel{d}{\sim} \mathcal{N}(0, \mathbf{\Psi}), \quad Z \perp \epsilon.$$

Define $\mathbf{\Sigma} := \mathbf{W}\mathbf{W}^\top + \mathbf{\Psi} \in \mathbb{R}^{p \times p}$ (symmetric and $\mathbf{\Sigma} \succ 0$).

Lemma 1

It holds that

$$X \stackrel{d}{\sim} \mathcal{N}(\mu, \mathbf{\Sigma}), \quad (\text{marginal})$$

$$\begin{pmatrix} Z \\ X \end{pmatrix} \stackrel{d}{\sim} \mathcal{N}\left(\begin{pmatrix} 0 \\ \mu \end{pmatrix}, \begin{pmatrix} \mathbf{I}_k & \mathbf{W}^\top \\ \mathbf{W} & \mathbf{\Sigma} \end{pmatrix}\right), \quad (\text{joint})$$

$$Z \mid X = x \stackrel{d}{\sim} \mathcal{N}(\mathbf{M}^{-1}\mathbf{W}^\top\mathbf{\Psi}^{-1}(x - \mu), \mathbf{M}^{-1}), \quad (\text{posterior})$$

where $\mathbf{M} := \mathbf{I}_k + \mathbf{W}^\top\mathbf{\Psi}^{-1}\mathbf{W}$.

I. Gaussian Linear Latent Variable Models

2.2. The Log-Likelihood

Let $(x^i)_{i \in [n]} \subset \mathbb{R}^p$ be a fixed dataset. (Assumed i.i.d. under $p_{\mu, \mathbf{W}, \Psi}$.) Let $\mathbf{X} := (x^1, \dots, x^n)^\top \subset \mathbb{R}^{n \times p}$. The log-likelihood map (also called *log-evidence*) is defined by

$$\begin{aligned}\mathcal{L}(\mathbf{X}, \mu, \mathbf{W}, \Psi) &:= \ln p_{\mu, \mathbf{W}, \Psi}(\mathbf{X}) \\ &= \sum_{i \in [n]} \ln p_{\mu, \mathbf{W}, \Psi}(x^i).\end{aligned}$$

I. Gaussian Linear Latent Variable Models

2.2. The Log-Likelihood

Let $(x^i)_{i \in [n]} \subset \mathbb{R}^p$ be a fixed dataset. (Assumed i.i.d. under $p_{\mu, \mathbf{W}, \Psi}$.) Let $\mathbf{X} := (x^1, \dots, x^n)^\top \subset \mathbb{R}^{n \times p}$. The log-likelihood map (also called *log-evidence*) is defined by

$$\begin{aligned}\mathcal{L}(\mathbf{X}, \mu, \mathbf{W}, \Psi) &:= \ln p_{\mu, \mathbf{W}, \Psi}(\mathbf{X}) \\ &= \sum_{i \in [n]} \ln p_{\mu, \mathbf{W}, \Psi}(x^i).\end{aligned}$$

Define $\bar{x} := n^{-1} \sum x^i$, $\mathbf{S} := n^{-1} \sum_{i \in [n]} (x^i - \mu)(x^i - \mu)^\top$.

Theorem 2

It holds that

$$\mathcal{L}(\mathbf{X}, \mu, \mathbf{W}, \Psi) = -\frac{n}{2} \left(p \ln(2\pi) + \ln \det(\Sigma) + \text{tr}(\Sigma^{-1} \mathbf{S}) + (\bar{x} - \mu)^\top \Sigma^{-1} (\bar{x} - \mu) \right).$$

Proof: Lemma 1 and cyclic property of the trace. $\square$.

From now on, w.l.o.g. we assume that $\bar{x} = 0$.

I. Gaussian Linear Latent Variable Models

We say that $\mathbf{W} \in \mathbb{R}^{p \times k}$, $0 \prec \Psi \in \mathbb{R}^{p \times p}$ symmetric form a **Maximum Likelihood Estimator** iff

$$(\mathbf{W}, \Psi) \in \operatorname{argmax}_{(\hat{\mathbf{W}}, \hat{\Psi})} \mathcal{L}(\mathbf{X}, \hat{\mathbf{W}}, \hat{\Psi}).$$

I. Gaussian Linear Latent Variable Models

We say that $\mathbf{W} \in \mathbb{R}^{p \times k}$, $0 \prec \Psi \in \mathbb{R}^{p \times p}$ symmetric form a **Maximum Likelihood Estimator** iff

$$(\mathbf{W}, \Psi) \in \operatorname{argmax}_{(\hat{\mathbf{W}}, \hat{\Psi})} \mathcal{L}(\mathbf{X}, \hat{\mathbf{W}}, \hat{\Psi}).$$

Theorem 3

For valid $(\mathbf{W}, \Psi)$, $\nabla_{\mathbf{W}} \mathcal{L}(\mathbf{X}, \mathbf{W}, \Psi) = 0 \iff \mathbf{S} \Sigma^{-1} \mathbf{W} = \mathbf{W}$.

Hint for the proof: Jacobi's formula for the derivative of the determinant. Then trace and derivative commute, and derivative of inverse matrix. Play with linear algebra.

We now look at two specific examples : **Probabilistic PCA** and **Factor Analysis**.

The PPCA generative model

The PPCA model

The PPCA generative model is defined by

$$\begin{aligned} X &= \mathbf{W}Z + \epsilon, & \mathbf{W} &\in \mathbb{R}^{p \times k}, \sigma > 0, \\ Z &\stackrel{d}{\sim} \mathcal{N}(0, \mathbf{I}_k), & \epsilon &\stackrel{d}{\sim} \mathcal{N}(0, \sigma^2 \mathbf{I}_p), \quad Z \perp \epsilon. \end{aligned}$$

Gaussian latent model with **isotropic** noise $\Psi = \sigma^2 \mathbf{I}_p$.

The PPCA model

The PPCA generative model is defined by

$$\begin{aligned} X &= \mathbf{W}Z + \epsilon, & \mathbf{W} &\in \mathbb{R}^{p \times k}, \sigma > 0, \\ Z &\stackrel{d}{\sim} \mathcal{N}(0, \mathbf{I}_k), & \epsilon &\stackrel{d}{\sim} \mathcal{N}(0, \sigma^2 \mathbf{I}_p), \quad Z \perp \epsilon. \end{aligned}$$

Gaussian latent model with **isotropic** noise $\Psi = \sigma^2 \mathbf{I}_p$.

Write $\Sigma = \mathbf{W}\mathbf{W}^\top + \sigma^2 \mathbf{I}_p$ and $\mathbf{M} = \mathbf{W}^\top \mathbf{W} + \sigma^2 \mathbf{I}_k$. By Lemma 1,

$$\begin{aligned} \begin{pmatrix} Z \\ X \end{pmatrix} &\stackrel{d}{\sim} \mathcal{N}\left(0, \begin{pmatrix} \mathbf{I}_k & \mathbf{W}^\top \\ \mathbf{W} & \Sigma \end{pmatrix}\right), \\ X &\stackrel{d}{\sim} \mathcal{N}(0, \Sigma), \end{aligned}$$

$$Z \mid X = x \stackrel{d}{\sim} \mathcal{N}(\mathbf{M}^{-1} \mathbf{W}^\top x, \sigma^2 \mathbf{M}^{-1}).$$

What does PPCA seek for?

Infinite dataset intuition

Suppose Σ and σ known (i.e. law of X known). Matrices $\mathbf{W}$ and Σ are diagonalised with same basis:

$$\Sigma = \mathbf{U} \bar{\Sigma} \mathbf{U}^\top, \quad \mathbf{W} \mathbf{W}^\top = \mathbf{V} \bar{\mathbf{W}} \mathbf{V}^\top \implies \boxed{\mathbf{V} = \mathbf{U}, \quad \bar{\mathbf{W}} = \bar{\Sigma} - \sigma^2 \mathbf{I}_p.}$$

Infinite dataset intuition

Suppose Σ and σ known (i.e. law of X known). Matrices $\mathbf{W}$ and Σ are diagonalised with same basis:

$$\Sigma = \mathbf{U} \bar{\Sigma} \mathbf{U}^\top, \quad \mathbf{W} \mathbf{W}^\top = \mathbf{V} \bar{\mathbf{W}} \mathbf{V}^\top \implies \boxed{\mathbf{V} = \mathbf{U}, \quad \bar{\mathbf{W}} = \bar{\Sigma} - \sigma^2 \mathbf{I}_p.}$$

Let $(\lambda_i)_{i \in [k]}$ be the ranked singular values of $\mathbf{W}$ and $(s_i)_{i \in [p]}$ the ranked eigenvalues of Σ , then $\lambda_i^2 = s_i - \sigma^2$, and for the loading matrix and the fraction of variance explained, this translates to

$$\mathbf{W} = \mathbf{U}_k (\bar{\Sigma}_k - \sigma^2 \mathbf{I}_k)^{1/2}, \quad \rho = \frac{\sum_{i \in [k]} (s_i - \sigma^2)}{\sum_{\ell \in [p]} s_\ell}, \quad \rho_j = \frac{s_j - \sigma^2}{\sum_{\ell \in [p]} s_\ell}.$$

Infinite dataset intuition

Suppose Σ and σ known (i.e. law of X known). Matrices $\mathbf{W}$ and Σ are diagonalised with same basis:

$$\Sigma = \mathbf{U} \bar{\Sigma} \mathbf{U}^\top, \quad \mathbf{W} \mathbf{W}^\top = \mathbf{V} \bar{\mathbf{W}} \mathbf{V}^\top \implies \boxed{\mathbf{V} = \mathbf{U}, \quad \bar{\mathbf{W}} = \bar{\Sigma} - \sigma^2 \mathbf{I}_p.}$$

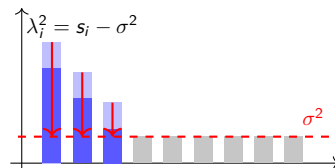
Let $(\lambda_i)_{i \in [k]}$ be the ranked singular values of $\mathbf{W}$ and $(s_i)_{i \in [p]}$ the ranked eigenvalues of Σ , then $\lambda_i^2 = s_i - \sigma^2$, and for the loading matrix and the fraction of variance explained, this translates to

$$\mathbf{W} = \mathbf{U}_k (\bar{\Sigma}_k - \sigma^2 \mathbf{I}_k)^{1/2}, \quad \rho = \frac{\sum_{i \in [k]} (s_i - \sigma^2)}{\sum_{\ell \in [p]} s_\ell}, \quad \rho_j = \frac{s_j - \sigma^2}{\sum_{\ell \in [p]} s_\ell}.$$

Output

PPCA returns the principal direction-principal value pairs $(u_i, \lambda_i^2)_{i \in [k]}$ with $(u_1, \dots, u_k) = \mathbf{U}_k$.

eigenvalues of Σ



Inferring parameters from observations with the log-likelihood.

Log-likelihood and rotation invariance

Let $x^1, \dots, x^n \stackrel{\text{i.i.d.}}{\sim} p_*$, let $\mathbf{X} := (x^1, \dots, x^n)^\top$ and $\mathbf{S} = \frac{1}{n} \sum_i x^i (x^i)^\top$. By Theorem 2

$$\mathcal{L}(\mathbf{X}, \mathbf{W}, \sigma^2 \mathbf{I}_p) = -\frac{n}{2} \left(p \log(2\pi) + \log \det \Sigma + \text{tr}(\Sigma^{-1} \mathbf{S}) \right).$$

Log-likelihood and rotation invariance

Let $x^1, \dots, x^n \stackrel{\text{i.i.d.}}{\sim} p_*$, let $\mathbf{X} := (x^1, \dots, x^n)^\top$ and $\mathbf{S} = \frac{1}{n} \sum_i x^i (x^i)^\top$. By Theorem 2

$$\mathcal{L}(\mathbf{X}, \mathbf{W}, \sigma^2 \mathbf{I}_p) = -\frac{n}{2} \left(p \log(2\pi) + \log \det \Sigma + \text{tr}(\Sigma^{-1} \mathbf{S}) \right).$$

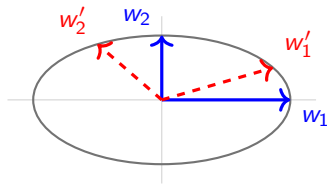
Let $O_k := \{\mathbf{R} \in \mathbb{R}^{k \times k} : \mathbf{R}\mathbf{R}^\top = \mathbf{R}^\top \mathbf{R} = \mathbf{I}_k\}$.

Lemma 4 (Invariance)

$\forall \mathbf{R} \in O_k : \mathcal{L}(\mathbf{X}, \mathbf{W}\mathbf{R}, \sigma^2 \mathbf{I}_p) = \mathcal{L}(\mathbf{X}, \mathbf{W}, \sigma^2 \mathbf{I}_p)$.

Proof:

$$\tilde{\Sigma} := (\mathbf{W}\mathbf{R})(\mathbf{W}\mathbf{R})^\top = \mathbf{W}\mathbf{R}\mathbf{R}^\top \mathbf{W}^\top = \mathbf{W}\mathbf{W}^\top = \Sigma. \square$$



$\mathbf{W}' = \mathbf{W}\mathbf{R}$: same $\Sigma \Rightarrow$ non-identifiable

Equivariance in the data

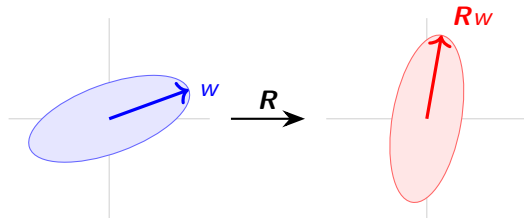
Applying a rotation on the data and the inverse rotation on the loading matrix (in ambient space) also leaves the likelihood unchanged.

Lemma 5 (Equivariance)

$$\forall \mathbf{R} \in O_p : \quad \mathcal{L}(\mathbf{X}\mathbf{R}^\top, \mathbf{W}, \sigma^2 \mathbf{I}_p) = \mathcal{L}(\mathbf{X}, \mathbf{R}\mathbf{W}, \sigma^2 \mathbf{I}_p).$$

Proof:

$$\mathbf{R}\mathbf{X} \stackrel{d}{\sim} \mathcal{N}(0, \mathbf{R}\Sigma\mathbf{R}^\top), \quad \mathbf{R}\Sigma\mathbf{R}^\top = (\mathbf{R}\mathbf{W})(\mathbf{R}\mathbf{W})^\top + \sigma^2 \mathbf{I}_p. \square$$



The MLE

$(u_i, s_i)_{i \in [p]}$: orthonormal eigenpairs of $\mathbf{S}$, $s_1 \geq \dots \geq s_p$; $\mathbf{U}_k = (u_1, \dots, u_k)$, $\bar{\mathbf{S}}_k = \text{diag}(s_1, \dots, s_k)$.

Theorem 6

Suppose $k < \min\{p, n\}$. Then $(\mathbf{W}_{\text{MLE}}, \sigma_{\text{MLE}}^2 \mathbf{I}_p)$ maximises $\mathcal{L}$ iff there exists $\mathbf{R} \in O_k$ such that

$$\mathbf{W}_{\text{MLE}} = \mathbf{U}_k (\bar{\mathbf{S}}_k - \sigma_{\text{MLE}}^2 \mathbf{I}_k)^{1/2} \mathbf{R}, \quad \sigma_{\text{MLE}}^2 = \frac{1}{p-k} \sum_{i=k+1}^p s_i.$$

Moreover, the likelihood of all MLEs is given by

$$\mathcal{L} = -\frac{n}{2} \left(p \ln(2\pi) + \sum_{i \in [k]} \ln s_i + (p-k) \ln \sigma_{\text{MLE}}^2 + p \right).$$

Remark

When $\mathbf{R} = \mathbf{I}_k$, for the MLE it holds that $\lambda_i^2 = \|\mathbf{W}_{\cdot, i}\|^2$.

Sketch of proof

Stationarity (Theorem 3): $\nabla_{\mathbf{W}} \mathcal{L}(\mathbf{W}, \sigma^2 \mathbf{I}_p, \mathbf{X}) = 0 \Leftrightarrow \mathbf{S} \mathbf{\Sigma}^{-1} \mathbf{W} = \mathbf{W}.$

Sketch of proof

Stationarity (Theorem 3): $\nabla_{\mathbf{W}} \mathcal{L}(\mathbf{W}, \sigma^2 \mathbf{I}_p, \mathbf{X}) = 0 \Leftrightarrow \mathbf{S} \mathbf{\Sigma}^{-1} \mathbf{W} = \mathbf{W}$. Write the SVD $\mathbf{W} = \mathbf{A} \overline{\mathbf{W}} \mathbf{B}^\top$, $\overline{\mathbf{W}} = \text{diag}(\lambda_1, \dots, \lambda_k)$, so that $\mathbf{\Sigma} = \mathbf{A} \overline{\mathbf{W}}^2 \mathbf{A}^\top + \sigma^2 \mathbf{I}_p$ and $\mathbf{\Sigma}^{-1} \mathbf{W} = \mathbf{A} (\overline{\mathbf{W}}^2 + \sigma^2 \mathbf{I}_k)^{-1} \overline{\mathbf{W}} \mathbf{B}^\top$.

Sketch of proof

Stationarity (Theorem 3): $\nabla_{\mathbf{W}} \mathcal{L}(\mathbf{W}, \sigma^2 \mathbf{I}_p, \mathbf{X}) = 0 \Leftrightarrow \mathbf{S} \Sigma^{-1} \mathbf{W} = \mathbf{W}$. Write the SVD $\mathbf{W} = \mathbf{A} \overline{\mathbf{W}} \mathbf{B}^\top$, $\overline{\mathbf{W}} = \text{diag}(\lambda_1, \dots, \lambda_k)$, so that $\Sigma = \mathbf{A} \overline{\mathbf{W}}^2 \mathbf{A}^\top + \sigma^2 \mathbf{I}_p$ and $\Sigma^{-1} \mathbf{W} = \mathbf{A} (\overline{\mathbf{W}}^2 + \sigma^2 \mathbf{I}_k)^{-1} \overline{\mathbf{W}} \mathbf{B}^\top$. Then, by stationarity,

$$\mathbf{S} \mathbf{A} (\overline{\mathbf{W}}^2 + \sigma^2 \mathbf{I}_k)^{-1} \overline{\mathbf{W}} \mathbf{B}^\top = \mathbf{A} \overline{\mathbf{W}} \mathbf{B}^\top$$

$$\Leftrightarrow \mathbf{S} \mathbf{A} = \mathbf{A} (\overline{\mathbf{W}}^2 + \sigma^2 \mathbf{I}_k). \quad (\text{suppose } \mathbf{W} \text{ full rank})$$

Sketch of proof

Stationarity (Theorem 3): $\nabla_{\mathbf{W}} \mathcal{L}(\mathbf{W}, \sigma^2 \mathbf{I}_p, \mathbf{X}) = 0 \Leftrightarrow \mathbf{S} \Sigma^{-1} \mathbf{W} = \mathbf{W}$. Write the SVD $\mathbf{W} = \mathbf{A} \overline{\mathbf{W}} \mathbf{B}^\top$, $\overline{\mathbf{W}} = \text{diag}(\lambda_1, \dots, \lambda_k)$, so that $\Sigma = \mathbf{A} \overline{\mathbf{W}}^2 \mathbf{A}^\top + \sigma^2 \mathbf{I}_p$ and $\Sigma^{-1} \mathbf{W} = \mathbf{A} (\overline{\mathbf{W}}^2 + \sigma^2 \mathbf{I}_k)^{-1} \overline{\mathbf{W}} \mathbf{B}^\top$. Then, by stationarity,

$$\mathbf{S} \mathbf{A} (\overline{\mathbf{W}}^2 + \sigma^2 \mathbf{I}_k)^{-1} \overline{\mathbf{W}} \mathbf{B}^\top = \mathbf{A} \overline{\mathbf{W}} \mathbf{B}^\top$$

$$\Leftrightarrow \mathbf{S} \mathbf{A} = \mathbf{A} (\overline{\mathbf{W}}^2 + \sigma^2 \mathbf{I}_k). \quad (\text{suppose } \mathbf{W} \text{ full rank})$$

Hence the k columns (a_i) of $\mathbf{A}$ are eigenvectors of $\mathbf{S}$ such that $\mathbf{S} a_i = (\lambda_i^2 + \sigma^2) a_i = s_{j_i} a_i$ for some reindexing j_i .

Sketch of proof

We showed that the columns (a_i) of $\mathbf{A}$ are eigenvectors of $\mathbf{S}$ such that $\mathbf{S}a_i = (\lambda_i^2 + \sigma^2)a_i$. By Lemma 1, the likelihood at such a stationary point is

$$\mathcal{L}(\mathbf{W}, \Psi, \mathbf{X}) = -\frac{n}{2} (p \ln(2\pi) + \ln \det \Sigma + \text{tr}(\Sigma^{-1} \mathbf{S})) .$$

Sketch of proof

We showed that the columns (a_i) of $\mathbf{A}$ are eigenvectors of $\mathbf{S}$ such that $\mathbf{S}a_i = (\lambda_i^2 + \sigma^2)a_i$. By Lemma 1, the likelihood at such a stationary point is

$$\mathcal{L}(\mathbf{W}, \Psi, \mathbf{X}) = -\frac{n}{2} (p \ln(2\pi) + \ln \det \Sigma + \text{tr}(\Sigma^{-1} \mathbf{S})) .$$

We have $\ln \det \Sigma = \sum_{i \in [k]} \ln(s_{j_i}) + (p - k) \ln(\sigma^2)$.

Sketch of proof

We showed that the columns (a_i) of $\mathbf{A}$ are eigenvectors of $\mathbf{S}$ such that $\mathbf{S}a_i = (\lambda_i^2 + \sigma^2)a_i$. By Lemma 1, the likelihood at such a stationary point is

$$\mathcal{L}(\mathbf{W}, \Psi, \mathbf{X}) = -\frac{n}{2} (p \ln(2\pi) + \ln \det \Sigma + \text{tr}(\Sigma^{-1} \mathbf{S})) .$$

We have $\ln \det \Sigma = \sum_{i \in [k]} \ln(s_{j_i}) + (p - k) \ln(\sigma^2)$. On the other hand, $\text{tr}(\Sigma^{-1} \mathbf{S}) = k + \frac{1}{\sigma^2} \sum_{i=k+1}^p s_{j_i}$.

Sketch of proof

We showed that the columns (a_i) of $\mathbf{A}$ are eigenvectors of $\mathbf{S}$ such that $\mathbf{S}a_i = (\lambda_i^2 + \sigma^2)a_i$. By Lemma 1, the likelihood at such a stationary point is

$$\mathcal{L}(\mathbf{W}, \Psi, \mathbf{X}) = -\frac{n}{2} (p \ln(2\pi) + \ln \det \Sigma + \text{tr}(\Sigma^{-1} \mathbf{S})).$$

We have $\ln \det \Sigma = \sum_{i \in [k]} \ln(s_{j_i}) + (p - k) \ln(\sigma^2)$. On the other hand, $\text{tr}(\Sigma^{-1} \mathbf{S}) = k + \frac{1}{\sigma^2} \sum_{i=k+1}^p s_{j_i}$. We deduce that

$$\mathcal{L}(\mathbf{W}, \sigma^2 \mathbf{I}_p) = -\frac{n}{2} \left(p \ln(2\pi) + \sum_{i \in [k]} \ln s_{j_i} + \frac{1}{\sigma^2} \sum_{i=k+1}^p s_{j_i} + (p - k) \ln \sigma^2 + k \right).$$

Differentiating in σ shows that $\sigma_{\text{MLE}}^2 = \frac{1}{p-k} \sum_{i=k+1}^p s_{j_i}$. Plugging this in the above, one can show that s_{j_i} for $i \in [k]$ are the k largest eigenvalues of $\mathbf{S}$ at an MLE (see Tipping and Bishop '99), which concludes the proof. $\square$

Fraction of variance explained

Comparing PCA and PPCA, the fraction of variance explained reads as

$$\rho^{\text{PCA}} = \frac{\sum_{i \in [k]} s_i}{\sum_{j \in [p]} s_j}, \quad \rho^{\text{PPCA}} = \rho^{\text{PCA}} - \frac{k \sigma_{\text{MLE}}^2}{\sum_{j \in [p]} s_j}.$$

Fraction of variance explained

Comparing PCA and PPCA, the fraction of variance explained reads as

$$\rho^{\text{PCA}} = \frac{\sum_{i \in [k]} s_i}{\sum_{j \in [p]} s_j}, \quad \rho^{\text{PPCA}} = \rho^{\text{PCA}} - \frac{k \sigma_{\text{MLE}}^2}{\sum_{j \in [p]} s_j}.$$

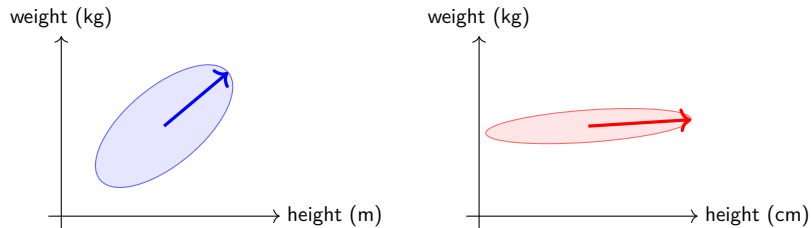
Per factor, choosing $\mathbf{R} = \mathbf{I}_k$ and setting $\mathbf{U}_k = \mathbf{W}_{\text{MLE}} \mathbf{V}(\bar{\mathbf{S}}_k - \sigma_{\text{MLE}}^2 \mathbf{I}_k)^{-1/2}$:

$$(R_i^{\text{PCA}})^2 = \frac{s_i}{\sum_{j \in [p]} s_j}, \quad (R_i^{\text{PPCA}})^2 = \frac{s_i - \sigma_{\text{MLE}}^2}{\sum_{j \in [p]} s_j} \xrightarrow{\sigma \rightarrow 0^+} \text{PCA}.$$

Dependence on the variables' scales.

Problem with PPCA: the scale matters

As for PCA, the PPCA model suffers from strong dependence in the variables' scales:



$$X \mapsto DX : \quad W \mapsto DW, \quad \sigma^2 I_p \mapsto \sigma^2 D^2 \neq \tilde{\sigma}^2 I_p.$$

Cure: let each coordinate have its **own** noise variance.

Factor Analysis: allowing the noise to be diagonal non-isotropic to make the method scale invariant.

The FA model

The generative model of *Factor Analysis* (FA) is the following:

$$\begin{aligned} X &= \mathbf{W}Z + \epsilon, & \mathbf{W} &\in \mathbb{R}^{p \times k}, \quad \mathbf{\Psi} = \text{diag}(\psi_1, \dots, \psi_p) \succ 0, \\ Z &\stackrel{d}{\sim} \mathcal{N}(0, \mathbf{I}_k), & \epsilon &\stackrel{d}{\sim} \mathcal{N}(0, \mathbf{\Psi}), \quad Z \perp \epsilon. \end{aligned}$$

Z_i is called the *i-th factor*, and $\mathbf{W}_{j,i}$ is the *loading* of the *i*-th factor on the *j*-th variable X_j .

The FA model

The generative model of *Factor Analysis* (FA) is the following:

$$\begin{aligned} X &= \mathbf{W}Z + \epsilon, & \mathbf{W} &\in \mathbb{R}^{p \times k}, \quad \mathbf{\Psi} = \text{diag}(\psi_1, \dots, \psi_p) \succ 0, \\ Z &\stackrel{d}{\sim} \mathcal{N}(0, \mathbf{I}_k), & \epsilon &\stackrel{d}{\sim} \mathcal{N}(0, \mathbf{\Psi}), \quad Z \perp \epsilon. \end{aligned}$$

Z_i is called the *i-th factor*, and $\mathbf{W}_{j,i}$ is the *loading* of the *i*-th factor on the *j*-th variable X_j .

Write $\mathbf{\Sigma} = \mathbf{W}\mathbf{W}^\top + \mathbf{\Psi}$ and $\mathbf{M} = \mathbf{I}_k + \mathbf{W}^\top \mathbf{\Psi}^{-1} \mathbf{W}$. Lemma 1 shows that

$$\begin{aligned} \begin{pmatrix} Z \\ X \end{pmatrix} &\stackrel{d}{\sim} \mathcal{N}\left(0, \begin{pmatrix} \mathbf{I}_k & \mathbf{W}^\top \\ \mathbf{W} & \mathbf{\Sigma} \end{pmatrix}\right), \\ X &\stackrel{d}{\sim} \mathcal{N}(0, \mathbf{\Sigma}), \end{aligned}$$

$$Z \mid X = x \stackrel{d}{\sim} \mathcal{N}(\mathbf{W}^\top \mathbf{\Sigma}^{-1} x, \mathbf{I}_k - \mathbf{W}^\top \mathbf{\Sigma}^{-1} \mathbf{W}) = \mathcal{N}(\mathbf{M}^{-1} \mathbf{W}^\top \mathbf{\Psi}^{-1} x, \mathbf{M}^{-1}).$$

What does FA seek for?

Infinite dataset intuition: FA = PPCA on rescaled data

- Σ known and Ψ fixed. Let $\tilde{X} := \Psi^{-1/2}X$ and $\tilde{W} := \Psi^{-1/2}W$, so that

$$\tilde{X} = \tilde{W}Z + \Psi^{-1/2}\epsilon, \quad \Psi^{-1/2}\epsilon \sim \mathcal{N}(0, I_p), \quad \tilde{\Sigma} = \tilde{W}\tilde{W}^\top + I_p :$$

a PPCA model with $\sigma = 1$. As before,

$$\tilde{\Sigma} = \tilde{U}\tilde{\Sigma}\tilde{U}^\top, \quad \tilde{W}\tilde{W}^\top = \tilde{V}\tilde{W}\tilde{V}^\top \quad \Rightarrow \quad \boxed{\tilde{V} = \tilde{U}, \quad \tilde{W} = \tilde{\Sigma} - I_p.}$$

Infinite dataset intuition: FA = PPCA on rescaled data

- Σ known and Ψ fixed. Let $\tilde{X} := \Psi^{-1/2}X$ and $\tilde{W} := \Psi^{-1/2}W$, so that

$$\tilde{X} = \tilde{W}Z + \Psi^{-1/2}\epsilon, \quad \Psi^{-1/2}\epsilon \sim \mathcal{N}(0, I_p), \quad \tilde{\Sigma} = \tilde{W}\tilde{W}^\top + I_p :$$

a PPCA model with $\sigma = 1$. As before,

$$\tilde{\Sigma} = \tilde{U}\tilde{\Sigma}\tilde{U}^\top, \quad \tilde{W}\tilde{W}^\top = \tilde{V}\tilde{W}\tilde{V}^\top \implies \boxed{\tilde{V} = \tilde{U}, \quad \tilde{W} = \tilde{\Sigma} - I_p.}$$

We get $\tilde{W} = U_k(\tilde{\Sigma}_k - I_k)^{1/2}$, and back in the original units, $W = \Psi^{1/2}U_k(\tilde{\Sigma}_k - I_k)^{1/2}$.

- Σ known and W fixed. Ψ fits the individual variances $\Psi = \text{diag}(\Sigma - WW^\top)$.

Communalities and uniquenesses

The variance of the variables splits into the sum of *communality* and *uniqueness*:

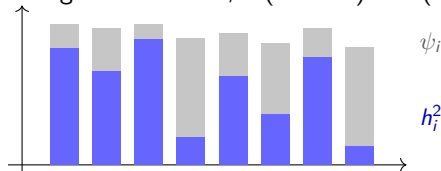
$$\mathbb{E}[X_i^2] = \Sigma_{i,i} = \underbrace{h_i^2}_{\text{communality}} + \underbrace{\psi_i}_{\text{uniqueness}}, \quad h_i^2 := \sum_{j \in [k]} w_{i,j}^2 = (\mathbf{w}\mathbf{w}^\top)_{i,i}.$$

Communalities and uniquenesses

The variance of the variables splits into the sum of *communality* and *uniqueness*:

$$\mathbb{E}[X_i^2] = \Sigma_{i,i} = \underbrace{h_i^2}_{\text{communality}} + \underbrace{\psi_i}_{\text{uniqueness}}, \quad h_i^2 := \sum_{j \in [k]} \mathbf{w}_{i,j}^2 = (\mathbf{w}\mathbf{w}^\top)_{i,i}.$$

The **total shared variance** reads as $\text{tr}(\tilde{\mathbf{W}}\tilde{\mathbf{W}}^\top) = \sum_{i \in [p]} \frac{h_i^2}{\psi_i}$. The form of the PPCA MLE also entails that letting $(d_i)_{i \in [k]}$ be the ranked eigenvalues of $\tilde{\Sigma}$, $\text{tr}(\tilde{\mathbf{W}}\tilde{\mathbf{W}}^\top) = \text{tr}(\tilde{\Sigma}) - p = \sum_{i \in [k]} (d_i - 1)$.



While $\text{tr}(\tilde{\mathbf{W}}\tilde{\mathbf{W}}^\top)$ is rotation invariant, the per-factor fraction of explained variance is **not**, and unlike PPCA there is no canonical rotation.

The FA log-likelihood.

Invariance and equivariance

Recall $\mathcal{L}(\mathbf{X}, \mathbf{W}, \Psi) = -\frac{n}{2} \left(p \log(2\pi) + \log \det \Sigma + \text{tr}(\Sigma^{-1} \mathbf{S}) \right)$.

Lemma 7

For all $\mathbf{W} \in \mathbb{R}^{p \times k}$, $\psi \in (0, \infty)^p$:

❶ $\forall \mathbf{R} \in O_k : \mathcal{L}(\mathbf{X}, \mathbf{WR}, \Psi) = \mathcal{L}(\mathbf{X}, \mathbf{W}, \Psi);$

❷ $\forall \mathbf{A}$ diagonal invertible:

$$\mathcal{L}(\mathbf{XA}, \mathbf{AW}, \mathbf{A}\Psi\mathbf{A}) = \mathcal{L}(\mathbf{X}, \mathbf{W}, \Psi) - n \sum_{i \in [p]} \ln |\mathbf{A}_{ii}|.$$

Proof of (ii): $\tilde{\Sigma} = \mathbf{A}\Sigma\mathbf{A}$, so $\ln \det \tilde{\Sigma} = \ln \det \Sigma + 2 \ln |\det \mathbf{A}|$, and

$$\text{tr}(\tilde{\Sigma}^{-1} \tilde{\mathbf{S}}) = \text{tr}(\mathbf{A}^{-1} \Sigma^{-1} \mathbf{A}^{-1} \mathbf{A} \mathbf{S} \mathbf{A}) = \text{tr}(\Sigma^{-1} \mathbf{S}).$$

$\tilde{\mathbf{W}} = \Psi^{-1/2} \mathbf{W}$ is unchanged $\Rightarrow$ covariance or correlation matrix: same FA.

How to find the FA MLE?

No closed form solution

The stationary point equation of the log-likelihood entails:

- given Ψ : $\mathbf{W} = \Psi^{1/2} \mathbf{U}_k (\mathbf{D}_k - \mathbf{I}_k)^{1/2} \mathbf{R}$ from $\Psi^{-1/2} \mathbf{S} \Psi^{-1/2} = \mathbf{U} \mathbf{D} \mathbf{U}^\top$;
- given $\mathbf{W}$: $\Psi = \text{diag}(\mathbf{S} - \mathbf{W} \mathbf{W}^\top)$.

the system is coupled, and unlike PPCA it admits **no closed-form solution**. Possible existence of local maxima, ill-defined likelihood... → next chapter.

No closed form solution

The stationary point equation of the log-likelihood entails:

- given Ψ : $\mathbf{W} = \Psi^{1/2} \mathbf{U}_k (\mathbf{D}_k - \mathbf{I}_k)^{1/2} \mathbf{R}$ from $\Psi^{-1/2} \mathbf{S} \Psi^{-1/2} = \mathbf{U} \mathbf{D} \mathbf{U}^\top$;
- given $\mathbf{W}$: $\Psi = \text{diag}(\mathbf{S} - \mathbf{W} \mathbf{W}^\top)$.

the system is coupled, and unlike PPCA it admits **no closed-form solution**. Possible existence of local maxima, ill-defined likelihood... → next chapter.

Note: at stationarity, the log-likelihood is known: (admitted)

$$\mathcal{L}(\mathbf{X}, \mathbf{W}, \Psi) = -\frac{n}{2} \left(p \ln(2\pi) + p + \ln \det \Psi + \sum_{i \in [k]} \ln \tilde{s}_i \right),$$

where $\tilde{s}_1 \geq \dots \geq \tilde{s}_p$ are the eigenvalues of $\Psi^{-1/2} \mathbf{S} \Psi^{-1/2}$.

On the interpretability of the factors: the sparsity criterion.

Interpretable factors: varimax

Rotation is free: $\mathbf{W}' = \mathbf{W}\mathbf{R}$, $\mathbf{R} \in O_k$. Choose $\mathbf{R}$ so that each column has **few large entries**. With $\hat{w}_{ij} = \mathbf{W}'_{ij}/h_i$,

$$\bar{w}_j = \frac{1}{p} \sum_i \hat{w}_{ij}^2, \quad v_j = \frac{1}{p} \sum_i (\hat{w}_{ij}^2 - \bar{w}_j)^2, \quad \mathbf{R}^* \in \arg \max_{\mathbf{R} \in O_k} V(\mathbf{R}) := \sum_{j \in [k]} v_j.$$

before, $V = 0.007$

0.64	0.64
0.64	0.50
0.50	0.50
0.64	-0.64
0.64	-0.50
0.50	-0.50

after, $V = 0.490$

0.00	0.90
0.10	0.80
0.00	0.70
0.90	0.00
0.80	0.10
0.70	0.00

$$\mathbf{R}^* = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$$

same Σ , same $\mathcal{L}$,
 same $h_i^2 = (.81, .65, .49, \dots)$;
 SS per column 2.11, 1.79 $\rightarrow$ 1.95, 1.95:
 ordering of factors lost.