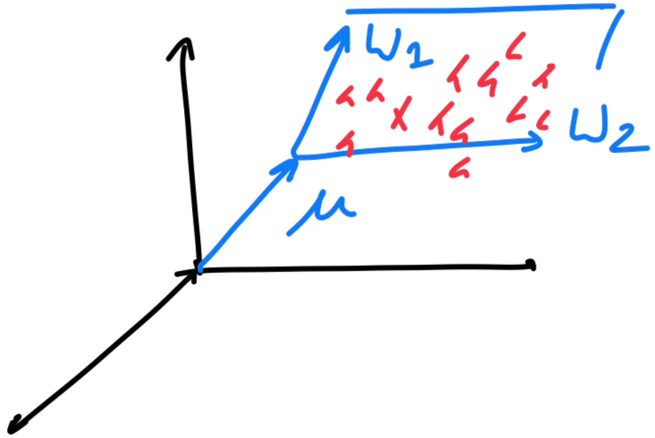
Latent variable Models



Factor Analysis

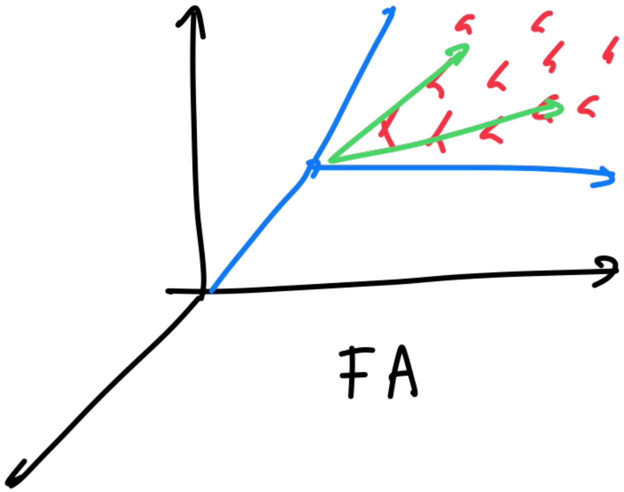
$W = [w_1, w_2]$

$$X = WZ + \mu + \varepsilon$$

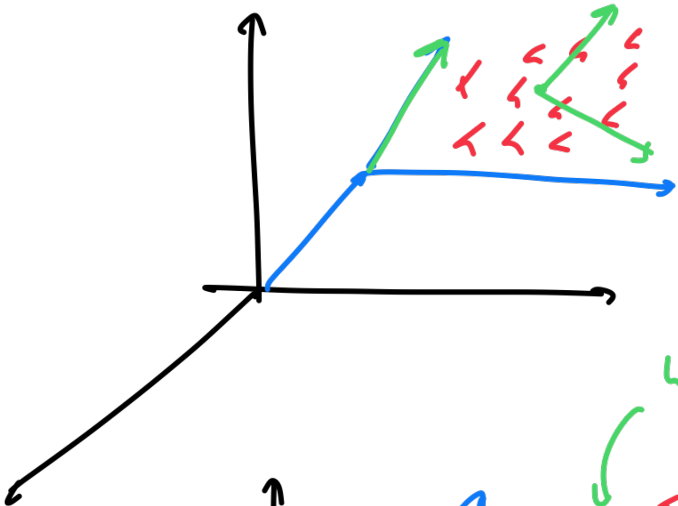
$\varepsilon \sim \mathcal{N}(0, \Psi)$ $\quad$ $\underbrace{Z \sim \mathcal{N}(0, \sigma^2 I)}$

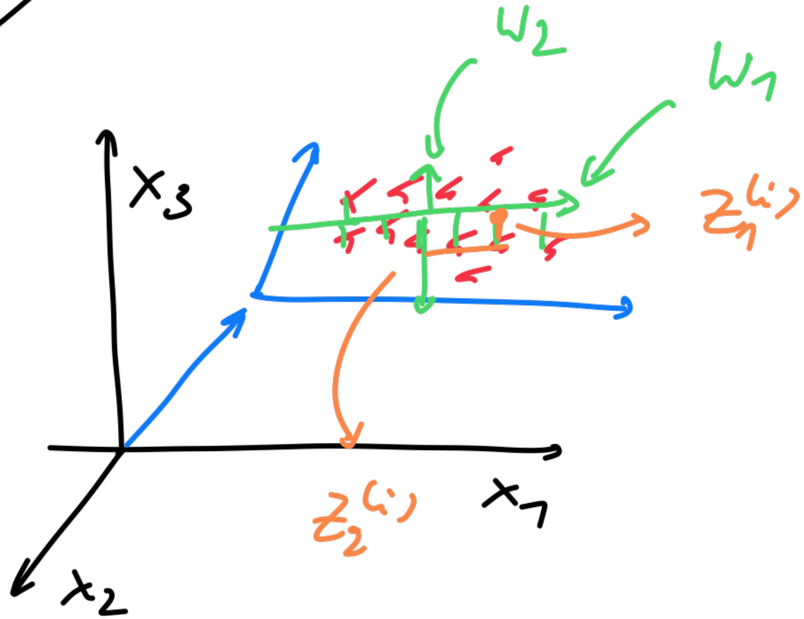
→ Problem non uniqueness of W

→ Several solutions to further "constraint" the data



FA





MLE $\longrightarrow$ PCA

Karhunen Loeve
formulation

To retrieve the W and z from the $\{x^{(i)}\}$ in the orthogonal W setting, we can solve the following problem (known as PCA)

Assuming $x^{(i)}$'s have been centred, we look for the solution of

PCA

$$\min_{W, z} \frac{1}{N} \sum_{i=1}^{N} \| x^{(i)} - W z^{(i)} \|_2^2$$

$x^{(i)} \in \mathbb{R}^D$

$z^{(i)} \in \mathbb{R}^K$

$K \ll D$

$$\min_{W, Z} \frac{1}{N} \| X - WZ \|_F^2$$

$W \in \mathbb{R}^{D \times K}$

→ Solution to the problem can be derived through the eigenvalue decomposition of the empirical covariance

→ Step 1 center the data $\quad x^{(i)} \leftarrow x^{(i)} - \frac{1}{N} \sum_{i=1}^{N} x^{(i)}$

→ Step 2 assemble empirical covariance $\quad \Sigma = \frac{1}{N} \sum_{i=1}^{N} x^{(i)} (x^{(i)})^T$

→ Step 3 → define $W = V_K$

where $\Sigma = V \Lambda V^T$ is the eigenvalue decomposition of the covariance Σ

$$z^{(i)} = V_K^T x^{(i)}$$

To understand the solution to the PCA problem, let us first consider the dimension-1 subspace w_1 and let us look for the representation $z_1^{i} w_1$ that best captures the $x^{(i)}$'s

$$\min_{w_1, z_1} \frac{1}{N} \sum_{i=1}^{N} \| x^{(i)} - z_1^{(i)} w_1 \|_2^2$$

PCA with $K = 1$

$W = \overline{w_1}$

$$= \frac{1}{N} \sum_{i=1}^{N} \left(x^{(i)} - z_1^{(i)} w_1 \right)^T \left(x^{(i)} - z_1^{(i)} w_1 \right)$$

$$\overset{\frown}{J} = \frac{1}{N} \sum_{i=1}^{N} (x^{(i)})^T (x^{(i)}) + (z_1^{(i)})^2 w_1^T w_1 - 2 \underline{z_1^{(i)} w_1^T x^{(i)}} \quad (*)$$

We want to find the minimum of $(*)$ for $w_1^T w_1 = 1$

→ to find the minimum, we set the derivatives to zero

Starting with $z_1^{(i)}$ we get

$$\frac{\partial J}{\partial z_1^{(i)}} = \frac{2}{N} z_1^{(i)} \underbrace{w_1^T w_1}_{= 1} - \frac{2}{N} w_1^T x^{(i)}$$

$$\rightarrow z_1^{(i)} = w_1^T x^{(i)}$$ (assuming w_1 has been normalized)

→ Substituting $z_1^{(i)}$ in J we get

$$\rightarrow J(w_1) = \frac{1}{N} \sum_{i=1}^{N} (x^{(i)})^T (x^{(i)}) + (w_1^T x^{(i)})^2 - 2(w_1^T x^{(i)})^2$$

$$\min_{w_1} J(w_1) = \frac{1}{N} \sum_{i=1}^{N} - \underbrace{(w_1^T x^{(i)})^2}_{(z_1^{(i)})^2} \quad \text{s.t} \quad w_1^T w_1 = 1$$

$$\max_{w_1} J(w_1) = \frac{1}{N} \sum_{i=1}^{N} w_1^T x^{(i)} x^{(i)T} w_1 \quad s.t \quad w_1^T w_1 = 1$$

$$\left. \begin{array}{l} \max\limits_{w_1} \quad w_1^T \underbrace{\frac{1}{N} \sum_{i=1}^{N} x^{(i)} (x^{(i)})^T}_{\hat{\Sigma}} w_1 \\ s.t \quad w_1^T w_1 = 1 \end{array} \right) (*)$$

→ To solve the problem we can consider the Lagrangian

$$\mathcal{L} = \max_{w_1} \quad w_1^T \hat{\Sigma} w_1 - \lambda (w_1^T w_1 - 1)$$

From the KKT conditions we know that the solution of problem (*) has to satisfy $\frac{\partial \mathcal{L}}{\partial w_1} = 0$

$$\underbrace{\hat{\Sigma} w_1 - \lambda_1 w_1 = 0}$$

→ w_1 is an eigenvector (with associated eigenvalue λ_1) of $\hat{\Sigma}$

→ We can proceed in a similar manner with w_2, adding the constraint $\underline{w_2^T w_1 = 0}$

$$\|\vec{v}\|_p = \sqrt[p]{\sum_{j=1}^{D} |v_j|^p} \rightarrow \|\vec{v}\|_2^2 = \sum_{j=1}^{D} |v_j|^2$$

Motivation for PCA formulation

$$p(x|z) \sim \mathcal{N}(\mu + Wz \mid \Psi) \qquad z \sim \mathcal{N}(0, I)$$

$$\Psi = \sigma^2 I$$

$$p(x,z) \propto \exp\left(-\frac{1}{2}(x-\mu-Wz)^T \Psi^{-1}(x-\mu-Wz)\right)$$

$$\exp\left(-\frac{1}{2}z^T z\right)$$

$$\propto \exp\Big(-\frac{1}{2}(x-\mu)^T \Psi^{-1}(x-\mu) - \frac{1}{2}z^T W^T \Psi^{-1} W z$$

$$+ \frac{1}{2} z^T W^T \Psi^{-1}(x-\mu)$$

$$+ \frac{1}{2}(x-\mu)^T \Psi^{-1} W z - \frac{1}{2} z^T z\Big)$$

$$\propto \exp\left(-\frac{1}{2}\begin{bmatrix} x-\mu & z \end{bmatrix}^T \underbrace{\begin{bmatrix} \Psi^{-2} & \Psi^{-2}W \\ W^T\Psi^{-2} & I+W^T\Psi^{-2}W \end{bmatrix}}_{\Sigma^{-1}} \begin{bmatrix} x-\mu \\ z \end{bmatrix}\right)$$

$$\begin{bmatrix} \mu & 0 \end{bmatrix}$$

In order to use an MLE approach we need $p(x)$

Now note that if $[x_1, x_2]$ that follow a Multivariate Gaussian with mean and Covariance $[\mu_1, \mu_2]$ $Cov = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{22} \end{bmatrix}$

then $p(x_1)$ & $p(x_2)$ are both Gaussian with

Marginal distribution for x_2

respective parameters μ_1, Σ_{11} and μ_2 Σ_{22}

$$x \sim \mathcal{N}(\mu \mid \Psi + WW^T)$$

From this if we assume independence of the $x^{(i)}$'s we can write $p(\{x^{(i)}\}_{i=1}^{N})$ and then we take log

$$\log \prod_{i=1}^{N} \exp\left(-\frac{1}{2}(x^{(i)}-\mu)^T(\Psi + WW^T)^{-1}(x-\mu)\right)\frac{1}{(2\pi)^{D/2}|\Psi+WW^T|^{1/2}}$$

$$\sum_{i=1}^{N} -\frac{1}{2}(x^{(i)}-\mu)^T(\Psi+WW^T)^{-1}(x-\mu) + \frac{1}{2}\log|\Psi+WW^T|$$

$$-\frac{1}{2}\sum_{i=1}^{N}\langle (x^{(i)}-\mu)(x^{(i)}-\mu)^T, S\rangle + \frac{1}{2}\log|S^{-1}|$$

$$-\left\langle \sum_{i=1}^{N} (x^{(i)}-\mu)(x^{(i)}-\mu)^T, S \right\rangle + \frac{1}{2}\log|S^{-1}|$$

$\longrightarrow$ compute the gradient, set it to zero

use $\frac{\partial}{\partial A} \log(\det(A)) = (A^{-1})^T$

$$-(WW^T + \Psi) + \sum_{i=1}^{N} (x^{(i)}-\mu)(x^{(i)}-\mu)^T = 0$$

$U \Lambda U^T$

Simplifying the inverse and solving for W gives

$$\longrightarrow W = U_q(\Lambda_q - \sigma^2 I)^{1/2} R$$

$$WW^T + \Psi = \sum_{i=1}^{N} (x^{(i)}-\mu)(x^{(i)}-\mu)^T$$

$$x^{(i)} = \mu + W z^{(i)} \longrightarrow \begin{bmatrix} z^{(1)} & \cdots & z^{(N)} \end{bmatrix}$$

$$\begin{bmatrix} | & & | \\ x^{(1)} & \cdots & x^{(N)} \\ | & & | \end{bmatrix} \approx W Z$$

$$U \Sigma V^T$$

assuming

X is centered

($x^{(i)}$'s are centered)

Independent Component Analysis

$$x^{(i)} = W s^{(i)} \qquad i = 1, \dots, N$$

in PCA $\quad p(z_j^{(i)}) = \prod_{j=1}^{L} \mathcal{N}(z_j ; 0, I)$

in ICA we will consider

$$p(s^{(i)}) = \prod_{j=1}^{L} p(s_j^{(i)})$$

for non Gaussian p

→ To recover the sources and the mixing matrix, we will assume whitened data so that $\mathbb{E}\{x\} = 0$

$\mathbb{E}\{xx^T\} = I$

in particular the whitening implies

sources are independent

$$\mathbb{E}\{xx^T\} = \mathbb{E}\{W s s^T W^T\} = W \, \mathbb{E}\{s s^T\} W^T = WW^T = I$$

→ Mixing matrix is orthogonal.

$$P(X \leq x) = P(f(S) \leq x)$$

$$P(S \leq f^{-1}(x)) = P_S(f^{-1}(x))$$

$$p_x = \frac{d}{dx} P(X \le x) = \frac{d}{dx} P_S(f^{-1}(x))$$

$$= \frac{d}{ds} P_S(f^{-1}(x)) \cdot \frac{ds}{dx}$$

$$= p_S(\underbrace{f^{-1}(x)}_{s}) \cdot \left| \frac{ds}{dx} \right|$$

$X = WS \qquad S = W^{-1}X \longrightarrow$ Jacobian matrix

In the multivariate setting $\quad V = W^{-1}$

$$p_x(x) = p_S(s) \cdot \left| \det\left[\left(\frac{\partial s_i}{\partial x_j} \right)_{ij} \right] \right|$$

$$f = \left(f_1(x_1, x_2),\ f_2(x_1, x_2) \right)$$

$$\rightarrow \text{Jacobian} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} \end{bmatrix} = \left(\frac{\partial f_i}{\partial x_j} \right)_{i,j}$$

$$s = W^{-1} x = f(x)$$

$$V = \begin{bmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{bmatrix}$$

$$s_1 = v_{11} x_1 + v_{12} x_2$$

$$s_2 = v_{21} x_1 + v_{22} x_2$$

$$\left(\frac{\partial s_i}{\partial x_j} \right) = \left(v_{ij} \right) = V$$

W^{-1}

$$p_x(x) = p_s(s) \cdot |\det(V)|$$

We can now use the pdf of x to derive the likelihood function

$$p(\{x^{(i)}\}_{i=1}^{N}) = \prod_{i=1}^{N} p_s(s^{(i)}) \cdot |\det(V)|$$

$$\begin{bmatrix} x_1^{(i)} \\ x_2^{(i)} \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \end{bmatrix}^{(i)}$$

taking the log, we get

$$LLE = \underbrace{\overline{\log|\det(V)|}}_{\substack{\parallel \\ 0}} + \frac{1}{N}\sum_{i=1}^{N} \log(p_s(s^{(i)}))$$

$$\begin{bmatrix} - & v_1^T & - \\ - & v_2^T & - \end{bmatrix}$$

data whitened implies W orthogonal which in turns implies $\det(W) = \pm 1$

$$s^{(i)} = W^{-1} x^{(i)} = V x^{(i)}$$

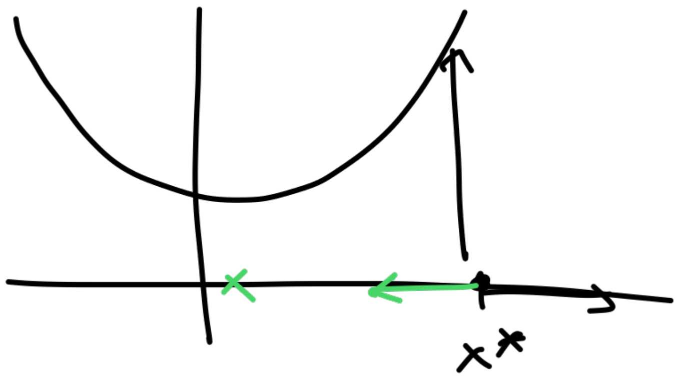
$$s_j^{(i)} = v_j^T x^{(i)}$$

$$\mathrm{NLLE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{L}\log\left(p_{s_j}\left(v_j^T x^{(i)}\right)\right)$$

?

FAST ICA relies on a minimization of the NLLE through a second order iteration of the form

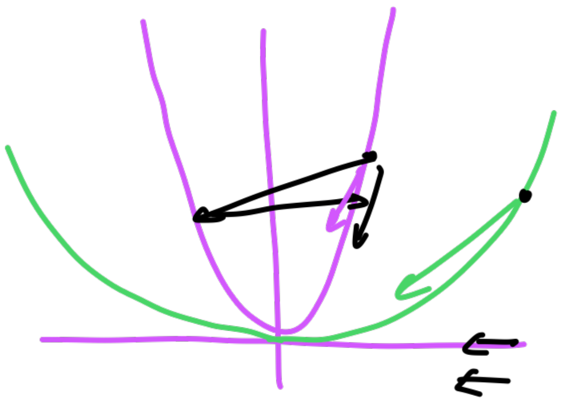
$$v^{(k+1)} \leftarrow v^{(k)} - (\mathrm{Hess}\, L)^{-1}\,\mathrm{grad}\, L \qquad (*)$$



$$x^2 \to \underbrace{2x}$$

$$\underbrace{2}$$

$$v^{(k+1)} \leftarrow v^{(k)} - \frac{\mathrm{grad}}{2}$$



$f_1(x) = 100\,x^2 \longrightarrow v \leftarrow v - (\text{Hess } f)^{-1}\,\text{grad}$

$\qquad\qquad\qquad\qquad \leftarrow v - \frac{1}{100}\,\text{grad}$

$f_2(x) = \frac{1}{100}\,x^2$

$v \leftarrow v - 100\,\text{grad}$

$f(x_1, x_2)$

$$\text{Hess} = \begin{bmatrix} \frac{\partial f}{\partial x_1^2} & \frac{\partial f}{\partial x_1 \partial x_2} \\ \frac{\partial f}{\partial x_2 \partial x_1} & \frac{\partial f}{\partial x_2^2} \end{bmatrix} = \left(\frac{\partial f}{\partial x_i \partial x_j} \right)_{i,j}$$

FAST ICA corresponds to second order iterations of the form (*) applied to the loss corresponding to the negative log likelihood function (NLLE)

Note that we still haven't specified the distribution p_{s_j} of the sources. There are several possible choices:

- We can either rely on super Gaussian distributions such as Laplace. Alternatively classical implementations will prefer smoother function (such as the logistic distribution)

$$\log(p(s)) = -2 \log \cosh\left(\frac{\pi}{2\sqrt{3}} s\right) - \log \frac{4\sqrt{3}}{\pi}$$