- Some properties of MVNs
- Review of EM algo. + GMMs
- Factor analysis model
- Solving the FAM w/ EM
- Comparing FA to PCA

Douglas Rubin

The Factor Analysis Model

Multivariate Normals (Facts about MVNs w/o proofs)

- For a random vector, $X \in \mathbb{R}^n$ that is Gaussian distributed with an invertible covariance matrix, $\Sigma \in \mathbb{R}^{n \times n}$ (w/ Σ symmetric PSD), it's PDF is given by:

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

w/ $\mu \in \mathbb{R}^n$. Note that for a Gaussian random vector, Σ may not be invertible, in which case we say the Gaussian is degenerate (as we will see later). We write this as:

$$x \sim \mathcal{N}(\mu, \Sigma)$$

↪

It is sometimes convenient to write this in block (or partioned form):

$$\begin{matrix} r \\ s \end{matrix}\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{matrix} r \\ s \end{matrix}\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{matrix} r \\ s \end{matrix}\begin{bmatrix} \overbrace{\Sigma_{11}}^{r} & \overbrace{\Sigma_{12}}^{s} \\ \Sigma_{12}^T & \Sigma_{22} \end{matrix}\right),$$

$X_1 \in \mathbb{R}^r, X_2 \in \mathbb{R}^s$

where $r+s=n$. This is just convenient shorthand, eg:

$$X = \begin{matrix} r \\ s \end{matrix}\begin{bmatrix} X^1 \\ \vdots \\ X^r \\ X^{r+1} \\ \vdots \\ X^n \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \text{ and } \Sigma_{11} = \begin{bmatrix} \sigma_{11}^2 & \cdots & \sigma_{1r}^2 \\ \vdots & \ddots & \vdots \\ \sigma_{r1}^2 & \cdots & \sigma_{rr}^2 \end{bmatrix}.$$

$\Sigma_{21} = \Sigma_{12}^T$ since we know Σ must be symmetric.

It is not hard to show that multiplication w/ block vectors/matrices behaves the same way as if you treated them as scalars in normal matrix multiplication.

<u>marginal and conditional dist.</u>

The expression above represents the joint density of X_1, X_2: $P(X_1, X_2) = P(X^1, \ldots, X^r, \ldots, X^n)$. It can be shown that the marginal density of X_1 ($P(X_1) = \int P(X_1, X_2) dX_2$) and X_2 are <u>also</u> normal:

$$X_1 \sim \mathcal{N}(\mu_1, \Sigma_{11}); \quad X_2 \sim \mathcal{N}(\mu_2, \Sigma_{22})$$

→

and the conditional dists ($p(x_1|x_2) = \frac{p(x_1,x_2)}{\int_{x_1 \in \mathbb{R}} p(x_1,x_2)\,dx_1}$)

are normal as well:

$$x_1|x_2 \sim \mathcal{N}\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2),\ \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)$$

$x_2|x_1 \sim$ as above w/ $1 \leftrightarrow 2$.

This is not difficult to show and relies on the matrix form of the completion of squares trick to turn quadratics in the exponential to the familiar gaussian form: $e^{ax^2+bx+c} = e^{a(x-h)^2+k} \propto e^{a(x-h)^2}$.

<u>the sum of indep. normally dist. rvs is also normal.</u>

If $X \sim \mathcal{N}(\mu_x, \Sigma_x)$

$Y \sim \mathcal{N}(\mu_y, \Sigma_y)$

and $Z \equiv X+Y$, then

$$Z \sim \mathcal{N}(\mu_x+\mu_y, \Sigma_x+\Sigma_y)$$

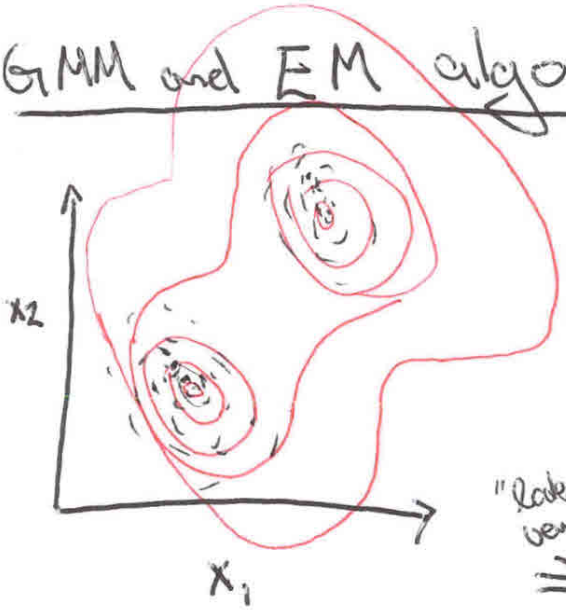
→

- Finally, MVNs rvs under an affine transformation are also normal (affine property):

- let $X \in \mathbb{R}^n$ be $X \sim \mathcal{N}(\mu, \Sigma)$ and $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$,

then $\boxed{AX + b \sim \mathcal{N}(A\mu + b, A\Sigma A^T)}$

GMM and EM algo



model

① Flip an unfair coin to determine class

② draw a rv. from the appropriate Gaussian

"Latent" variable

$\Rightarrow z^{(i)} \sim \text{Bern}(\phi)$ (or multinoull: depending on # of classes)

$x^{(i)} | z^{(i)} = j \sim \mathcal{N}(\mu_j, \Sigma_j)$

Can use for:

1) estimation of density field

2) soft clustering of data, where we obtained soft guesses on cluster assignment through inference!

$$P(z^{(i)} = j \mid x^{(i)})$$

(not Bayesian inference though!)

- For a known (modeled) conditional $p(z^{(i)} | x^{(i)}; \theta)$ and joint density dist. $p(x^{(i)}, z^{(i)}; \theta)$, we may use the EM algo. to find the MLE of θ, θ^*:

EM Algo.

① initialize θ^*

② repeat til convergence {

(e-step) 2a) $Q_i(z^{(i)}) := p(z^{(i)} | x^{(i)}; \theta^*)$

(m-step) 2b) $\theta^* := \underset{\theta}{\text{argmax}} \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}$

}

- note that if $z^{(i)}$ is continuous, then $\sum_{z^{(i)}} \rightarrow \int_{z^{(i)}}$.

- note that we usually will solve for 2b with the standard calculus maximization technique.

Factor Analysis $\mathcal{D} = \{x^{(1)}, \ldots, x^{(m)}\}$; $x^{(i)} \in \mathbb{R}^n$

- Provides:
  1. density estimation (also provides a valid/invertible cov. matrix if needed)
  2. dimensionality reduction (mostly used for this)

I will motivate the problem through density estimation, and then move to feature red.)

Suppose, like GMM we want to fit a MVN, but $m < n$. We know this will be problematic b/c it will be hard to fit a pdf when we have very few data points / unit volume.

- In fact, we can't even construct a MVN (using MLE, in which case the MLE of Σ = sample covariance) ~~if we use the sample covariance as its covariance~~ b/c it is not invertible, which we need for the MVN pdf.

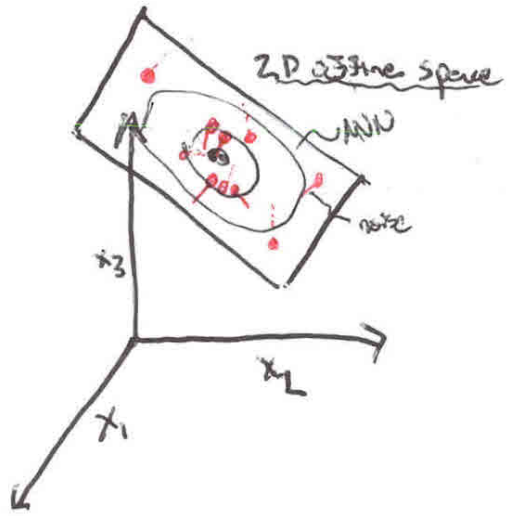
eg: 2 points in $\mathbb{R}^3$

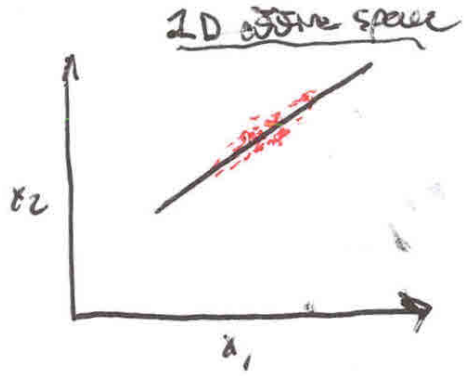
$$\hat{\Sigma} = \frac{1}{m} X^T X = \frac{1}{m} \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} = \frac{1}{m} \begin{bmatrix} 17 & 22 & 27 \\ 22 & 29 & 36 \\ 27 & 36 & 45 \end{bmatrix}$$

but, $-\begin{bmatrix} 17 \\ 22 \\ 27 \end{bmatrix} + 2\begin{bmatrix} 22 \\ 29 \\ 36 \end{bmatrix} = \begin{bmatrix} 27 \\ 36 \\ 45 \end{bmatrix} \Rightarrow$ cols of $\hat{\Sigma}$ not lin. indep. and thus $\hat{\Sigma}$ not invertible.

Solution

- Fit a MVN to much smaller portion of the space, ie to an affine subspace (w/ a degenerate Gaussian), and model deviation from the affine space w/ Gaussian noise ε off that plane. (We can easily fit an affine plane to the data since for an n-dim. plane, the data will go right through the plane).
- In fact, even when $m > n$, if we suspect our data lie approx. on an affine subspace w/ a Gaussian dist we can use this model.



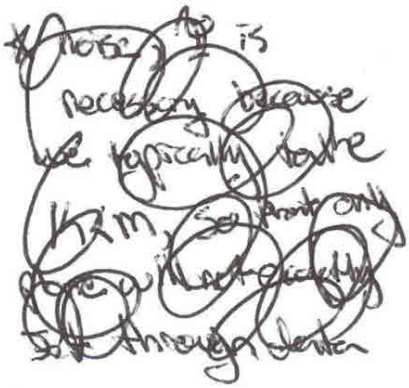


In Math: latent variable

① $z^{(i)} \sim N(0, I_k)$; $z \in \mathbb{R}^k$; $k < n$

② $\mu + \Lambda z^{(i)} \sim N(\mu, \Lambda\Lambda^T)$,
$\mu \in \mathbb{R}^n$; $\Lambda \in \mathbb{R}^{n \times k}$ (by affine property)

③ $x^{(i)} \equiv \mu + \Lambda z^{(i)} + \varepsilon$, w/ ε axis aligned Gaussian noise ($\varepsilon \sim N(0, \Psi)$)
$\Psi \in \mathbb{R}^{n \times n}$; (diagonal)



In words:

① draw a MVN rv in $\mathbb{R}^k$ from a $0, I$ gaussian

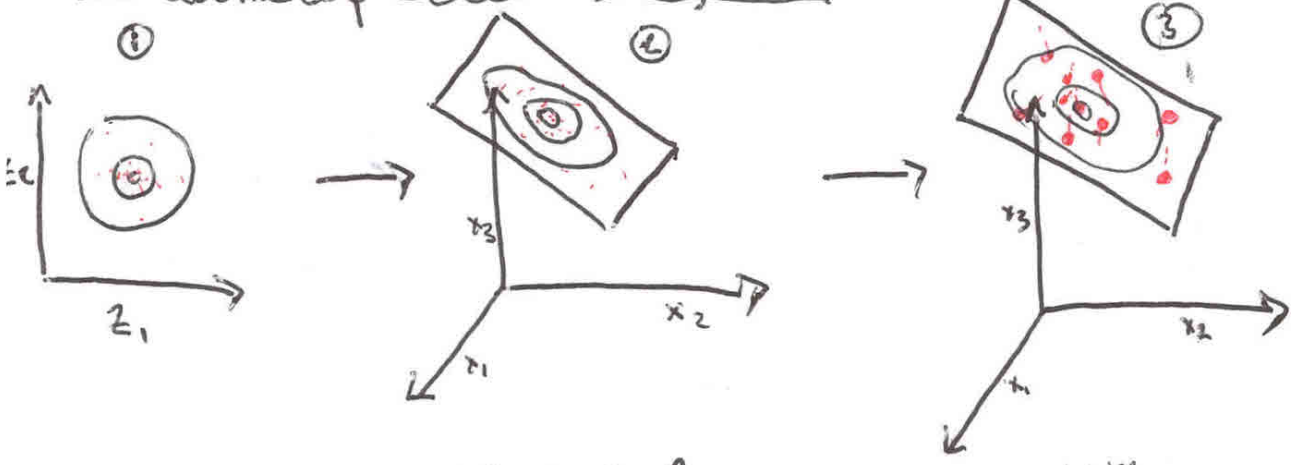
② map that vector to the affine space in $\mathbb{R}^n$ by multiplying by Δ and adding μ (dist. will still be Gaussian on the affine subspace) (by affine property)

③ add axis aligned noise (which can be thought of as measurement noise)

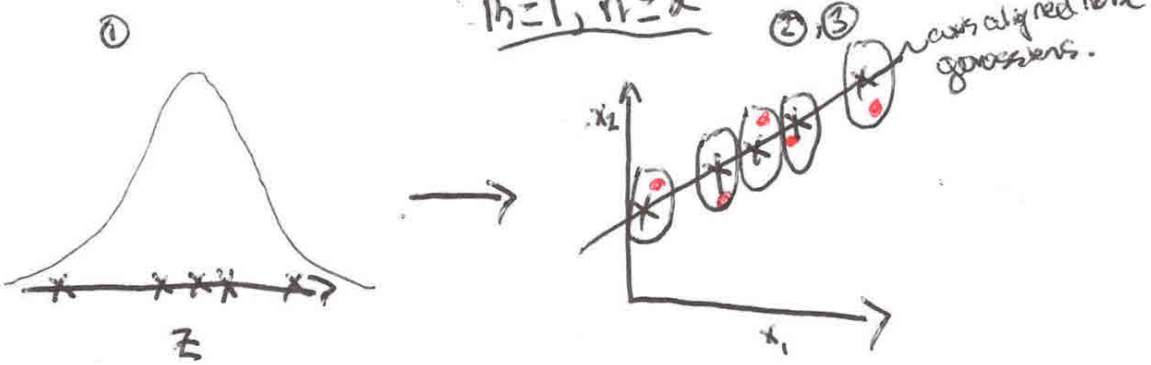
In pictures

affine space

$k=2, n=3$



$k=1, n=2$



→

- we solve for μ, Λ, Ψ using their max. likelihood estimates, and we therefore need to derive a few more dists. to write down ℓ.

- Another way to write this process is that:

$$\begin{cases} z^{(i)} \sim \mathcal{N}(0, I) & \text{(marg. of } z\text{)} \\ x^{(i)} | z^{(i)} \sim \mathcal{N}(\mu + \Lambda z^{(i)}, \Psi) & \text{(cond of } x\text{)} \end{cases}$$

(much like GMM)

Since sum of 2 indep Gaussian rv is Gaussian:

$$x^{(i)} \sim \mathcal{N}(\mu, \Lambda\Lambda^T + \Psi) \quad \text{(marg. of } x\text{)}$$

(this thus shows us that we are modeling the observed density w/ a MVN)

Ψ cov. from ε

covariance from $\mu + \Lambda z$

- joint is also Gaussian

$$\begin{bmatrix} z^{(i)} \\ x^{(i)} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_z \\ \mu_x \end{bmatrix}, \begin{bmatrix} \Sigma_{zz} & \Sigma_{zx} \\ \Sigma_{xz} & \Sigma_{xx} \end{bmatrix} \right)$$

→ thrown from marginals

Since:

$$\begin{aligned} x^{(i)} | z^{(i)} &\sim \mathcal{N}(\mu_x + \Sigma_{xz}\Sigma_{zz}^{-1}(z - \mu_z), \Sigma_{xx} - \Sigma_{xz}\Sigma_{zz}^{-1}\Sigma_{zx}) \\ &\sim \mathcal{N}(\mu + \Sigma_{xz} I^{-1}(z), \Lambda\Lambda^T + \Psi - \Sigma_{xz} I^{-1} \Sigma_{xz}^T) \\ &\sim \mathcal{N}(\mu + \Sigma_{xz} z, \Lambda\Lambda^T + \Psi - \Sigma_{xz}\Sigma_{xz}^T) \end{aligned}$$

$\Rightarrow$ Comparing this dist. to the cond. dist. for X, we see that:

$$\Sigma_{XZ} = \Lambda, \text{ so:}$$

$$\begin{bmatrix} z^{(i)} \\ x^{(i)} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ \mu \end{bmatrix}, \begin{bmatrix} I_k & \Lambda^T \\ \Lambda & \Lambda\Lambda^T + \Psi \end{bmatrix} \right) \quad \text{(joint)}.$$

We may also easily write down the cond. of $z^{(i)}$:

$$z^{(i)} | x^{(i)} \sim \mathcal{N}(\mu_{z^{(i)}|x^{(i)}}, \Sigma_{z^{(i)}|x^{(i)}}) \quad \text{(cond. of } z\text{)}$$

w/ $\mu_{z^{(i)}|x^{(i)}} = \Lambda^T(\Lambda\Lambda^T + \Psi)^{-1}(x^{(i)} - \mu)$

$$\Sigma_{z^{(i)}|x^{(i)}} = I_k - \Lambda^T(\Lambda\Lambda^T + \Psi)^{-1}\Lambda$$

dim. reduction

- recall from the GMM, the latent variables represented the unobserved class labels for each data point.
- to obtain class probabilities for each data point, we used inference by calculating $P(z^{(i)} = j | x^{(i)})$ for each data point, for each class, j.

(we got this for free in the E-step)

- in FA, we posit that there is some unobserved vector (of unobserved features), $z^{(i)}$ (w/ dim $= k < n$) that fundamentally describes each data point (like how the unobserved class in the GMM fundamentally describes each data point).

- we again use inference (note this is not Bayesian inference). Our prior on $z^{(i)}$ is $N(0, I_k)$ and ~~obtain the posterior~~ the posterior, $z^{(i)} | x^{(i)}$ is the MVN w/ mean $\mu_{z^{(i)}|x^{(i)}}$, $\Sigma_{z^{(i)}|x^{(i)}}$. (where now the posteriors are not pdf) We use the values of Λ, μ, Ψ obtained from MLE.

- We have a dist. over the posterior of $z^{(i)}$. To ascribe a single value for this data point, we (as we did in the GMM) therefore use the mode of the posterior:

$$x^{(i)}_{FA} = \mu_{z^{(i)}|x^{(i)}} = \hat{\Lambda}^T (\hat{\Lambda}\hat{\Lambda}^T + \hat{\Psi})^{-1} (x^{(i)} - \hat{\mu}),$$

where the hats denote the MLEs.

Solving the MLE w/ EM

The log-likelihood is:

$$\ell(\mu, \Lambda, \Psi) = \log p(x^{(1)}, \dots, x^{(m)}; \mu, \Lambda, \Psi)$$

$$= \log \prod_{i=1}^{m} p(x^{(i)}; \mu, \Lambda, \Psi)$$

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

$$= -\left\{ \sum_{i=1}^{m} \log\left((2\pi)^{n/2}\right) + \log |\Lambda\Lambda^T + \Psi|^{1/2} \right.$$

$$\left. + \frac{1}{2} (x^{(i)} - \mu)^T (\Lambda\Lambda^T + \Psi)^{-1} (x^{(i)} - \mu) \right\},$$

which is intractable to maximize in closed form w/ calculus techniques. We therefore appeal to the EM algo:

E-step: $Q_i(z^{(i)}) = p(z^{(i)} | x^{(i)}; \mu^*, \Lambda^*, \Psi^*) =$

(we derived this cond. dist. before)

$$\frac{1}{(2\pi)^{k/2} \left| \Sigma_{z^{(i)}|x^{(i)}} \right|^{1/2}} \exp\left\{ -\frac{1}{2} \left(z^{(i)} - \mu_{z^{(i)}|x^{(i)}} \right)^T \Sigma_{z^{(i)}|x^{(i)}}^{-1} \left(z^{(i)} - \mu_{z^{(i)}|x^{(i)}} \right) \right\}$$

w/ $\mu_{z^{(i)}|x^{(i)}} = \Lambda^{*T} (\Lambda^* \Lambda^{*T} + \Psi^*)^{-1} (x^{(i)} - \mu^*)$

$\Sigma_{z^{(i)}|x^{(i)}} = I_k - \Lambda^{*T} (\Lambda^* \Lambda^{*T} + \Psi^*)^{-1} \Lambda^*$

- The M step update is more difficult, but doable.

M-step:

$$\Lambda^*, \mu^*, \Psi^* = \underset{\substack{\Lambda \in \mathbb{R}^{n \times k} \\ \mu \in \mathbb{R}^n \\ \Psi \in \mathbb{R}^{n \times n} \text{ (diag)}}}{\arg\max} \sum_{i=1}^{m} \int_{z^{(i)} \in \mathbb{R}^k} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \mu, \Lambda, \Psi)}{Q_i(z^{(i)})} dz^{(i)}$$

- Expressions for Λ^*, μ^*, Ψ^* can be found in closed form by taking gradients, $\nabla_\Lambda, \nabla_\mu, \nabla_\Psi$, setting the expression to the zero vector/matrix, and solving (using several matrix and trace identities).

- Connection between FA and PCA

PCA: geometric approach
- directly project data down to subspace of $\pm k$ eigenvectors of sample cov.

FA: probabilistic approach
- posit the existence of a k dim. latent variable and perform inference to estimate its value.

- Recall 2 assumptions made by PCA
  1. Principle axes are $\perp$
  2. principle axes are linear

$\Rightarrow$ data distributed as an ellipsoidal blob. I.e., data distr. in $\mathbb{R}^n$ as a MVN (just like the marginal of $x^{(i)}$ in FA).

- In fact if $\Psi = \sigma^2 I$ (isotropic noise) we may solve for $\hat{\Lambda}$ analytically. In limit $\sigma \to 0$ (i.e., data become deterministic), the soln. for Λ for X_{FA} is identical to that of PCA. (called probabilistic PCA)

- I.e., PCA is just the deterministic limit of a special case of FA.