Lecture 23 - scribbles

Tuesday, December 3, 2019 14:31

today's: Gaussian networks
- factor analysis & PCA
- VAE

Gaussian networks

$X \sim N(\mu, \Sigma)$ $\mu \in \mathbb{R}^p$ $\Sigma \in \mathbb{R}^{p \times p}$ $\Sigma \succ 0$

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

$$= \exp\left(-\frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}(x-\mu)(x-\mu)^T\right)\right)$$

$$(x-\mu)(x-\mu)^T = xx^T - \mu x^T - x\mu^T + \mu\mu^T$$

sufficient statistics $T(x) = \begin{pmatrix} x \\ -\frac{1}{2}xx^T \end{pmatrix}$ → canonical parameter

$$\langle \Sigma^{-1}, \tfrac{1}{2}xx^T \rangle + \langle \Sigma^{-1}\mu, x \rangle - \frac{1}{2}\mu^T\Sigma^{-1}\mu$$

$\Lambda \triangleq \Sigma^{-1}$ precision matrix

$\triangleq \eta$

$\mu = \Sigma\eta = \Lambda^{-1}\eta$

canonical parameter $\tilde{\eta}(\Theta) = \begin{pmatrix} \eta \\ \Lambda \end{pmatrix} = \begin{pmatrix} \Sigma^{-1}\mu \\ \Sigma^{-1} \end{pmatrix}$

$$p(x; \eta, \Lambda) = \exp\left(\eta^T x + \langle \Lambda, -\tfrac{1}{2}xx^T \rangle - \underbrace{\left[\frac{1}{2}\eta^T\Lambda^{-1}\eta + \frac{p}{2}\log 2\pi - \frac{1}{2}\log|\Lambda|\right]}_{A(\eta,\Lambda)}\right)$$

$$\Omega = \{(\eta, \Lambda) : \eta \in \mathbb{R}^p, \Lambda \succ 0\} \qquad \Omega = \Lambda^T, \Lambda \in \mathbb{R}^{p \times p}$$

useful exercise: $\nabla_\eta A(\eta, \Lambda) = \mathbb{E}[x] = \mu = \Lambda^{-1}\eta$

$\nabla_\Lambda A(\eta, \Lambda) = \mathbb{E}\left[-\frac{xx^T}{2}\right]$

UGM viewpoint:

$$p(x; \eta, \Lambda) = \exp\left(-\frac{1}{2}\sum_{i,j}\Lambda_{ij}x_i x_j + \sum_i \eta_i x_i - A(\eta, \Lambda)\right)$$

$p \in \mathcal{L}(G)$ where $E \triangleq \{\{i,j\} \text{ s.t. } \Lambda_{ij} \neq 0\}$

Zeros in precision matrix => cond. indep properties (from UGM perspective)

"Gaussian network"

$$p(x) = \prod_{\{i,j\} \in E} \psi_{ij}(x_i, x_j)$$

Quick Schur-complement digression:

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \Sigma_{11}^{-1} + * & -\Sigma_{11}^{-1}\Sigma_{12}M^{-1} \\ -M^{-1}\Sigma_{21}\Sigma_{11}^{-1} & M^{-1} \end{pmatrix}$$

$$M \triangleq \Sigma/\Sigma_{11} \triangleq \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$ "Schur complement of Σ wrt to Σ_{11}"

* use this derive "Woodbury-sherman-Morrison inversion formula"

$$\Sigma/\Sigma_{22} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

property: $|\Sigma| = |\Sigma_{11}|\,|\Sigma/\Sigma_{11}| = |\Sigma_{22}|\,|\Sigma/\Sigma_{22}|$

$$p(\underset{\dim p_1}{x_1}, \underset{p_2}{x_2}) = \frac{1}{\sqrt{(2\pi)^{p_1}|\Sigma_{11}|}} \exp\left(-(x_1-\mu_1)\Sigma_{11}^{-1}(x_1-\mu_1)\right) \cdot \Big\} p(x_1)$$

$$\frac{1}{\sqrt{(2\pi)^{p_2}|\Sigma/\Sigma_{11}|}} \exp\left(-(x_2-\mu_2-b(x_1))^T\left(\Sigma/\Sigma_{11}\right)^{-1}(x_2-\mu_2-b(x_1))\right) \Big\} p(x_2|x_1)$$

where $b(x_1) \triangleq \Sigma_{21}\Sigma_{11}^{-1}(x_1-\mu_1)$

mean parameterization of marginal & conditionals

$$\mu_1^M = \mu_1, \quad \Sigma_1^M = \Sigma_{11}$$ } super simple! } param. for marginal on x_1

$$\mu_{2|1}^{cond} = \mu_2 + b(x_1)$$
$$\Sigma_{2|1}^{cond.} = \Sigma/\Sigma_{11} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$ } for conditional $x_2|x_1$

in canonical param.

$$\Lambda_{2|1}^{cond} = \Lambda_{22}$$ (simple)

$$\eta_{2|1}^{cond.} = \eta_2 - \Lambda_{21}x_1$$

$$\eta_1^m = \eta_1 - \Lambda_{12}\Lambda_{22}^{-1}\eta_2$$ (more complicated)

$$\Lambda_1^m = \Lambda_{11} - \Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21} = \Lambda/\Lambda_{22}$$

x_2 x_1

block $\underbrace{\{i,j\}}_{I}$ | rest

$$\mathrm{cov}(X_I \mid X_{rest}) = \Sigma_{II|rest} = \Lambda_{II}^{-1}$$

$$= \begin{pmatrix} \Lambda_{ii} & \Lambda_{ij} \\ \Lambda_{ji} & \Lambda_{jj} \end{pmatrix}^{-1}$$

if $\Lambda_{ij}=0$ get $\Sigma_{II|rest} = \begin{pmatrix} \Lambda_{ii}^{-1} & 0 \\ 0 & \Lambda_{jj}^{-1} \end{pmatrix}$

$$\Rightarrow \boxed{X_i \perp\!\!\!\perp X_j \mid X_{rest}}$$

15h20

(also true by Markov property of UGM)

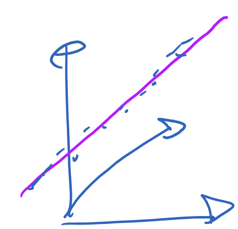
Factor analysis

latent variable model $\quad z \in \mathbb{R}^k \to x \in \mathbb{R}^d \quad$ learn a "latent representation" or dimensionality reduction $k \ll d$

PCA for dimensionality reduction

synthesis view: find k orthonormal vectors in $\mathbb{R}^d$ $w_1, \dots, w_k$

s.t. project x on span$\{w_1,\dots,w_k\}$ is a good approx of x



$$W = \begin{bmatrix} | & & | \\ w_1 & \cdots & w_k \\ | & & | \end{bmatrix} \qquad W^TW = I_k \text{ (by orthonormality)}$$

WW^T?

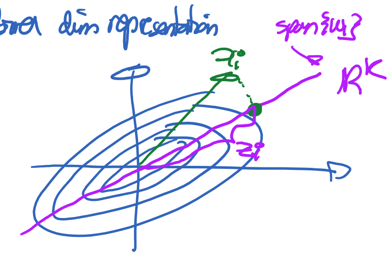
$P_W \triangleq WW^T \qquad P_W^2 = W\underbrace{W^TW}_{I_k}W^T = P_W$

$\hookrightarrow$ orthogonal projection matrix on span$\{w_1,\dots,w_k\} = \mathrm{col}(W)$

$$P_W x = WW^Tx = \begin{pmatrix} w_1 & \cdots & w_k \end{pmatrix}\begin{pmatrix} \langle w_1, x\rangle \\ \vdots \\ \langle w_k, x\rangle \end{pmatrix} = \sum_k w_k \underbrace{\langle w_k, x\rangle}_{(z)_k} = Wz$$

$\hookrightarrow$ get lower dim representation

$z = W^Tx \in \mathbb{R}^k$



PCA $\quad \min_{W \in \mathbb{R}^{d\times k}} \sum_i \|x_i - W\underbrace{W^Tx_i}_{z_i}\|^2$

W is not unique, only col(W)

PCA $\min_{W \in \mathbb{R}^{d\times k},\ W^TW=I_k} \sum_i \|x_i - WW^Tx_i\|^2$ (z_i)

$\text{col}(W) \triangleq$ "principal subspace"

W is not unique, only col(W)

e.g. $\tilde{W} = WR$ where $RR^T = R^TR = I_k$

then $\tilde{W}\tilde{W}^T = WRR^TW^T = WW^T$ ($R^TR = I_k$)

$X = \begin{pmatrix} -x_1^T- \\ \vdots \\ -x_n^T- \end{pmatrix}$ $n\times d$

$\|X^T - WW^TX^T\|_F^2$

$\frac{1}{n}X^TX = \frac{1}{n}\sum_i x_i x_i^T$ ↑ empirical covariance of x when $\sum x_i = 0$ (mean = 0)

$= \|(I - P_W)X^T\|_F^2$

$= \text{tr}\left(X\underbrace{(I-P_W)^T(I-P_W)}_{I-P_W}X^T\right)$

$= \text{tr}(X(I-P_W)X^T)$

$= \text{tr}(X^TX(I-P_W))$

min rec. error ⇔ maximize $\text{tr}(X^TXWW^T) = \sum_k w_k^TX^TXw_k$

"analysis view of PCA"

max sum of empirical variances in new representation

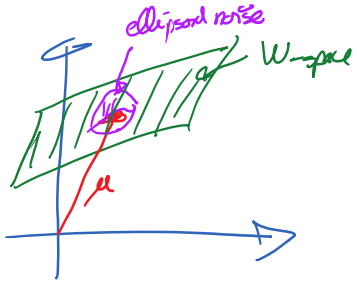
(computation of PCA → top k e-vectors of X^TX)

factor analysis → simplest generative model

$z \sim N(0, I_k)$

$x = Wz + \mu + \varepsilon$ (noise)

$\varepsilon \perp\!\!\!\perp z$, $\varepsilon \sim N(0, D)$ ↑ dxd diagonal matrix



$x|z \sim N(Wz+\mu, D)$

$p(x)$ is Gaussian; $\mathbb{E}[x] = \mathbb{E}[\mathbb{E}[x|z]]$

$\mathbb{E}[Wz+\mu] = 0 + \mu = \mu$

$\text{cov}(X,X) = \text{cov}(Wz+\mu+\varepsilon, Wz+\mu+\varepsilon)$ (indep.)

$= \text{cov}(Wz, Wz) + \text{cov}(\varepsilon,\varepsilon)$

$W \underbrace{\mathrm{cov}(z)}_{I_k} W^T + D$

$= WW^T + D$

equivalent model on $x \sim N(\mu, WW^T + D)$

WW^T: $d \times k$ — low rank covariance piece

D: diagonal $\Rightarrow d$ degrees of freedom

estimate W, D, μ by MLE

$\rightsquigarrow$ do EM (latent variable model)

get $p(z|x) \rightarrow$ Gaussian with mean

$$\mathbb{E}[z|x] = W^T (WW^T + D)^{-1} (x - \mu_x)$$

probabilistic PCA: special case of factor analysis where suppose $D = \sigma^2 I_d$

$$\lim_{\sigma \to 0} W^T (WW^T + \sigma I)^{-1} = W^{\dagger} \quad \text{(pseudo-inverse)}$$

$$= W^T \text{ if } W^T W = I_k$$

show PCA is limit as $\sigma \to 0$ of PPCA

(sidenote: LDA model for text is basically $\theta \sim \mathrm{Dir}(\alpha)$)

$x|\theta \sim \mathrm{Mult}(W\theta, n)$

"discrete version of PPCA"

Kalman filter:

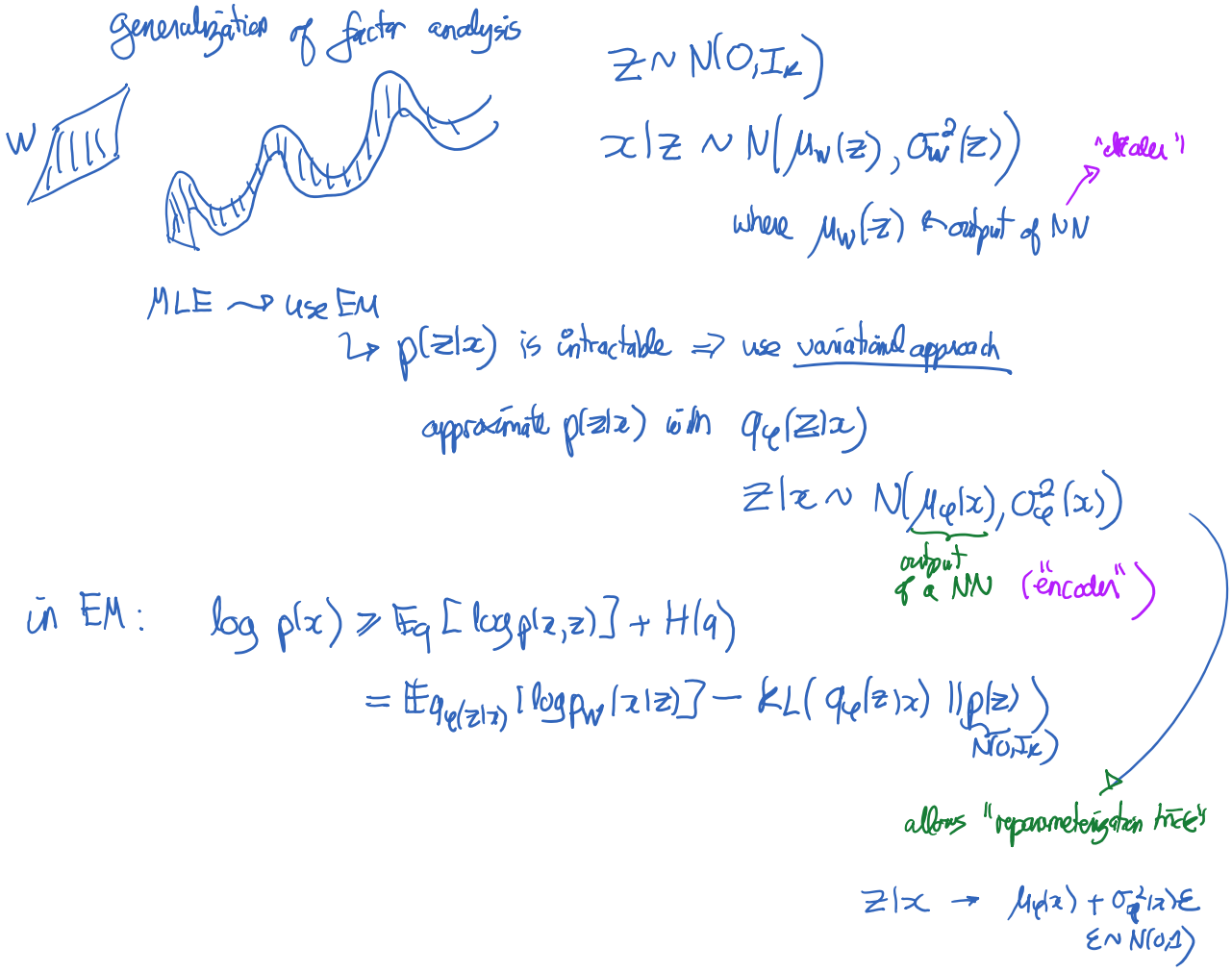
factor analysis

state space model: unroll in time (HMM style)

Kalman filter: $z_t | z_{t-1} \sim N(Az_{t-1}, B)$

$\rightarrow$ doing "sum-product" alg. in HMM $p(z_t | x_{1:t})$ get "Kalman filtering alg."

variational auto-encoder:



- VAE innovations:
  - share parameters phi among data points for their variational approximation q_phi(z|x)
  - re-parameterization trick to only have parameters appear in simple deterministic transformation, stochasticity is all left in N(0,1) (no parameters) => allow simple backpropagation of gradient through expectations
  - for more details, see: Slides on VAE by Aaron Courville - deep learning class Winter 2017

Other skipped parts, for more details:

- see 2016 lecture 17 scribbles for more info on Schur complement & block decomposition of inverse
- see 2016 lecture 18 scribbles for more info on SVD, and also CCA