Lecture 24 - scribbles

Friday, December 1, 2017
12:56

today: · Gaussian networks
· factor analysis & PCA etc.
· VAE

Gaussian networks:

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

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

$$(x-\mu)^T \Sigma^{-1}(x-\mu) = \underbrace{x^T \Sigma^{-1} x} - 2\mu^T \Sigma^{-1} x + \mu^T \Sigma^{-1}\mu$$

$$\operatorname{tr}(x^T \Sigma^{-1} x) = \operatorname{tr}(\Sigma^{-1} x x^T)$$

$$= \langle \Sigma^{-1}, x x^T \rangle \quad -2\langle \underbrace{\Sigma^{-1}\mu}_{\triangleq \eta}, x \rangle$$

bilinear form

precision matrix

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

sufficient statistics: $T(x) = \begin{pmatrix} x \\ -\frac{1}{2} x x^T \end{pmatrix}$

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

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

$$p(x; \mu, \Sigma) = \exp\left(\eta^T x + \langle \Lambda, -\tfrac{1}{2} x x^T \rangle - \left[\tfrac{1}{2}\eta^T \Lambda^{-1} \eta + \tfrac{p}{2}\log 2\pi\right]\right)$$

$-\frac{1}{2}\log|\Lambda|$

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

$A(\eta, \Lambda)$

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

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

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

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

Zeros in precision matrix $\Rightarrow$ cond. indep properties Ⓢ

"Gaussian network"

quick Schur-complement digression

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

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

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

schur-complement of Σ with respect to Σ_{11}

use this to derive "Woodbury-Sherman-Morrison inversion formula"

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

using above,

$$p(x_1, x_2) = \underbrace{\frac{1}{\sqrt{(2\pi)^{p_1}|\Sigma_{11}|}}\exp\left(-(x_1-\mu_1)\Sigma_{11}^{-1}(x_1-\mu_1)\right)}_{p(x_1)} \cdot \underbrace{\frac{1}{\sqrt{(2\pi)^{p_2}|\Sigma/\Sigma_{11}|}}\exp\left(-(x_2-\mu_2-b(x_1))^T(\Sigma/\Sigma_{11})^{-1}(x_2-\mu_2-b(x_1))\right)}_{p(x_2|x_1)}$$

(block of dim. p_1)

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

mean parameterization of marginal and conditionals:

marginal on x_1:
$$\mu_1^m = \mu_1$$
$$\Sigma_1^m = \Sigma_{11}$$

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

---

in canonical parameterization

$$\Lambda_{2|1}^{cond.} = \Lambda_{22} \quad \text{(simple)}$$
$$\eta_{2|1}^{cond} = \eta_2 - \Lambda_{21}x_1$$
$$\eta_1^m = \eta_1 - \Lambda_{12}\Lambda_{22}^{-1}\eta_2$$
$$\Lambda_1^m = \Lambda_{11} - \Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21} = \Lambda/\Lambda_{22}$$

block $\tilde{x}_2$ | rest $\tilde{x}_2$

$\tilde{\Sigma}_{ij}$ ← $\tilde{x}_2$

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

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

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

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

(also true by Markov property of UGM)

Factor analysis:

Latent variable model

$z \in \mathbb{R}^k$

$x \in \mathbb{R}^d$

learning a "latent representation" or dimensionality reduction

PCA for dimensionality reduction:

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

s.t. the projection of x on $\mathrm{span}\{w_1, \ldots, w_k\}$ is a good approximation of x

$$W = \underset{d \times k}{\begin{bmatrix} | & & | \\ w_1 & \cdots & w_k \\ | & & | \end{bmatrix}} \qquad W^T W = I_k \quad \text{(orthonormality)}$$

what about WW^T?

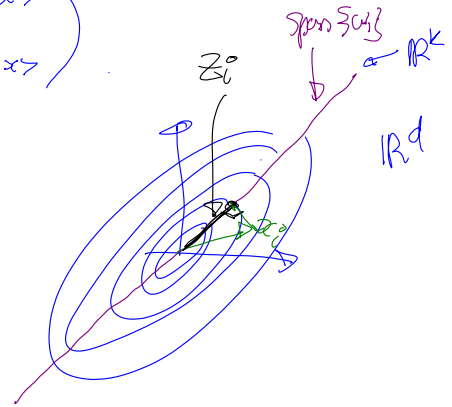
$D \triangleq WW^T$? projection matrix on $\mathrm{span}\{w_1, \ldots, w_k\} = \mathrm{col}(W)$

$P_W \triangleq WW^T$ ← projection matrix on span $\{w_1, \dots, w_k\} = col(W)$

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

$$P_W x = WW^Tx = (w_1 \cdots w_k)\begin{pmatrix}\langle w_1, x\rangle \\ \vdots \\ \langle w_k, x\rangle\end{pmatrix}$$

$$= \sum_k w_k \underbrace{\langle w_k, x\rangle}_{(z)_k}$$



get lower dim. representation $z = W^Tx \in \mathbb{R}^k$ $\quad P_W x = Wz$

PCA $\min_{W \in \mathbb{R}^{d\times k},\ W^TW = I_k} \sum_i \|x_i - \underbrace{WW^Tx_i}_{z_i}\|^2$ ← reconstruction error

"synthesis view"

$col(W) \triangleq$ "principal subspace"

$\tilde{W} = WR$ where R is $k\times k$ rotation $RR^T = R^TR = I_k$

then $\tilde{W}\tilde{W}^T = WRR^TW^T = WW^T$

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

$X^TX = \sum_i x_i x_i^T$

$$\|X^T - WW^TX^T\|_F^2$$
$$= \|(I - P_W)X^T\|_F^2$$
$$= tr\left(X(I-P_W)^T(I-P_W)X^T\right)$$
$$= tr\left(X(I-P_W)X^T\right)$$
$$= tr\left(X^TX(I-P_W)\right)$$

"analysist view of PCA"

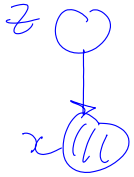
min rec. error $\Leftrightarrow$ maximize $tr(X^TXWW^T) = \sum w_i^T X^T X w_i$

min rec. error $\Leftrightarrow$ maximize $tr(X^T X W W^T) = \sum_k w_k^T X^T X w_k$

maximizing variance in new representation

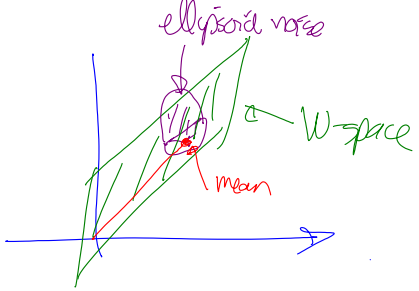
(computation of PCA directions $\rightarrow$ top k e-vectors of $X^T X$)

Factor analysis $\rightarrow$ generative model (latent variable)



$z \sim N(0, I_k)$

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



$\varepsilon \perp\!\!\!\perp z$, $\varepsilon \sim N(0, D)$

$d \times d$ diagonal matrix

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

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

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

$Cov(x, x) = Cov(Wz + \mu + \varepsilon, Wz + \mu + \varepsilon)$ (ε indep.)

$= \underbrace{Cov(Wz, Wz)}_{W^T Cov(z) W} + \underbrace{Cov(\varepsilon, \varepsilon)}_{D}$

$= WW^T + D$

equivalent model for x : $x \sim N(\mu, WW^T + D)$
W: $d \times k$, D: diagonal $\Rightarrow d$ degrees of freedom

get $p(z|x)$:

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

$\mu_x = \mu$ $\quad \Sigma_{xx} = WW^T + D$
$\mu_z = 0$ $\quad \Sigma_{zz} = I_k$

$\text{cov}(x, z) = W$ (exercise)

$$E[z|x] = \mu_z + \Sigma_{zx}\Sigma_{xx}^{-1}(x - \mu_x)$$
$$= 0 + W^T(WW^T + D)^{-1}(x - \mu_x)$$

probabilistic PCA : suppose $D = \sigma^2 I_d$

$\lim_{\sigma \to 0} W^T(WW^T + \sigma^2 I)^{-1} = W^+$ pseudoinverse
$= W^T$ if $W^T W = I_k$

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

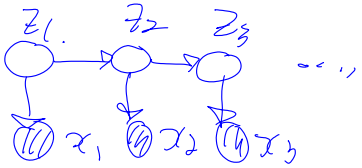
(x) for factor analysis, estimate $\hat{W}, \hat{D}, \hat{\mu}$ using ML
→ use EM → need $p(z|x)$

(sidenote : LDA model for text is basically $\theta \sim \text{Dir}(\alpha)$
$x|\theta \sim \text{Mult}(W\theta, n)$)
"discrete version of PPCA"

note that $\hat{W}$ is not identifiable (can only be identified up to rotation)

Kalman filter:

factor analysis 

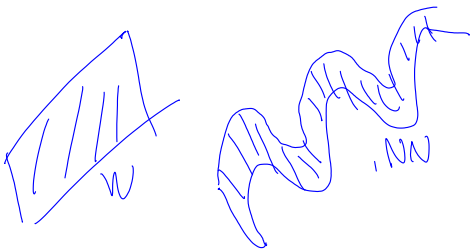
state space model: unroll in time: 

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

Variational auto-encoder:

$z \sim N(0, I_k)$

generalization of factor analysis where $x | z \sim N(\mu_w(z), \sigma_w^2(z))$ "decoder"

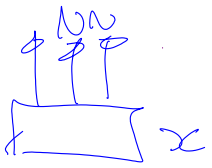


$\mu_w(z)$ ← output of NN

$p(z|x)$ is intractable ⇒ use variational approach

approximate $p(z|x)$ with $q_\varphi(z|x)$ "encoder"

parameters $z|x \sim N(\mu_\varphi(x), \sigma_\varphi^2(x))$

$q_\varphi(z|x)$ 

in EM,

$\log p(x) \geq E_q[\log p(x,z)] + H(q)$

$$\mathbb{E}_{q_\phi(z|x)}[\log p(x|z)] - KL(q_\phi(z|x) \| p(z))$$

- 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) noise variables (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 Old lecture 17 scribbles for more info on Schur complement & block decomposition of inverse
- see Old lecture 18 scribbles for more info on SVD, and also CCA