Lecture 23 - Gaussian networks; factor analysis; VAE

Thursday, November 30, 2023 2:35 PM

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

Gaussian networks

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

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

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

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

put it in exponential family

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

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

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

$\triangleq \eta$

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

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

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

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

useful exercise: check $\nabla_\eta A(\eta, \Lambda) = \mathbb{E}[X] = \mu$

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

UGM viewpoint

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

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

Zeros in precision matrix ⇒ cond. indep. properties

"Gaussian network" $p(x) = \frac{1}{Z}\prod_{\{i,j\}} \psi_{ij}(x_i, x_j) \prod_i \psi_i(x_i)$

quick Schur-complement digression

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

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

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

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

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

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

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

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

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

mean parameterization of marginal on X_1 and conditional $X_2|X_1$

$\mu_1^M = \mu_1$, $\Sigma_1^M = \Sigma_{11}$ } super simple! } param. for marginal on X_1

$\mu_{2|1}^{cond.} = \mu_2 + b(x_1)$

$\Sigma_{2|1}^{cond.} = \Sigma/\Sigma_{11} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$ } param. for cond. $X_2|X_1$

---

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)

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

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

$$\Lambda = \begin{pmatrix} \Lambda_{RR} & \Lambda_{RI} \\ \Lambda_{IR} & \Lambda_{II} \end{pmatrix}$$

$$\mathrm{cov}(X_I|X_{rest}) = \Sigma_{I|rest} = \Lambda_{I|rest}^{-1} = \Lambda_{II}^{-1} = \begin{pmatrix} \Lambda_{ii} & \Lambda_{ij} \\ \Lambda_{ji} & \Lambda_{jj} \end{pmatrix}^{-1}$$

if $\Lambda_{ij} = 0$ $\Lambda_{II} = \begin{pmatrix} \Lambda_{ii} & 0 \\ 0 & \Lambda_{ii} \end{pmatrix}$

if $\Lambda_{ij}=0$ $\quad \Lambda_{I\bar{I}} = \begin{pmatrix} \Lambda_{ii} & 0 \\ 0 & \Lambda_{jj} \end{pmatrix}$

get $\Sigma_{I|rest} = \begin{pmatrix} \Lambda^{-1}_{ii} & 0 \\ 0 & \Lambda^{-1}_{jj} \end{pmatrix}$

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

(also true by Markov property of UGM)

15h28

Factor analysis:

latent variable model

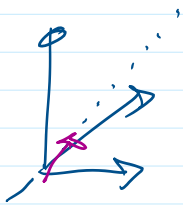
$z \in \mathbb{R}^k$

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

learn "latent representation" in $\mathbb{R}^k$ or dimensionality reduction $k<<d$

PCA for dimensionality reduction

synthetic view: find k orthonormal vectors in $\mathbb{R}^d$ $w_1, \dots, w_k$ s.t. projection of x on $\text{span}\{w_1, \dots, w_k\}$ is a good approx. of x



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

$WW^T ? = I_d$

$P_W \triangleq WW^T$ $\quad P_W^2 = WW^T WW^T = P_W$

↳ orthogonal projection on $\text{span}\{w_1, \dots, w_k\}$

$P_W x = W(W^T x)$

$= \begin{pmatrix} | & & | \\ w_1 & \dots & w_k \\ | & & | \end{pmatrix} \begin{pmatrix} \langle w_1, x \rangle \\ \vdots \\ \langle w_k, x \rangle \end{pmatrix}$

$= \sum_\ell w_\ell \underbrace{\langle w_\ell, x \rangle}_{(z)_\ell} = Wz$

auto-encoder

PCA $\quad \min_{W \in \mathbb{R}^{d \times k}} \sum_{i=1}^n \| x_i - W \underbrace{W^T x_i}_{z_i} \|_2^2$

$W^T W = I_k$ $\quad \text{col}(W) \triangleq$ principal subspace

$z = W^T x$

↳ lower dimensional representation

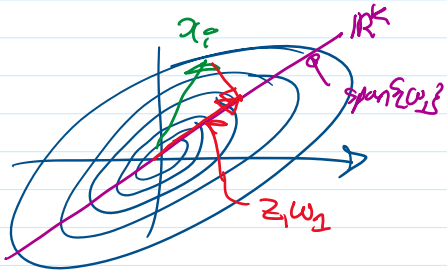
$X = \underset{n \times d}{\begin{pmatrix} - & x_1^T & - \\ & \vdots & \\ - & x_n^T & - \end{pmatrix}}$

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

$= \| (I_d - WW^T) X^T \|_F^2$

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

$\frac{1}{n} X^T X = \frac{1}{n} \sum x_i x_i^T$



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

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

$= tr\left(X (Id - P_W) X^T \right) = tr\left(X^T X (Id - P_W) \right) = tr(cst.) - tr(X^T X P_W)$

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

"analysis view of PCA" — max sum of empirical variances of new representation

(w_k: eigenvector of $(X^T X)$, λ_k)

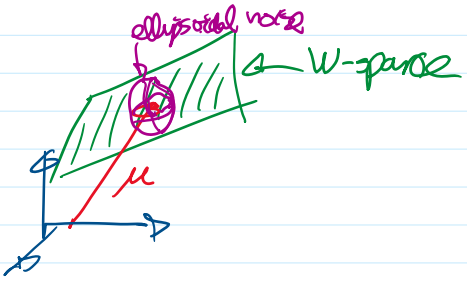
(computation of PCA → top k e-vectors of $X^T X$)

⊛ W is not unique, only $col(W)$

e.g. $\tilde{W} = WR$ where $R^T R = R R^T = I_k$ ($d \times k$)

$\tilde{W}\tilde{W}^T = W R R^T W^T = W I_k W^T$

$\max_{\|a\|=1} a^T \Sigma a = \lambda_{max} \|a\|^2$



Factor analysis → simplest generative model

$z \sim N(0, I_k)$

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

$\varepsilon \perp\!\!\!\perp z$, $\varepsilon \sim N(0, D)$ ($d \times d$ diagonal matrix)

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

$p(x)$ is Gaussian

$\mathbb{E}[x] = \mathbb{E}[\underbrace{\mathbb{E}[x|z]}_{Wz+\mu}] = 0 + \mu = \mu$

$cov(x, x) = cov(Wz + \mu + \varepsilon, Wz + \mu + \varepsilon)$ (z, ε indep)

$= cov(Wz, Wz) + cov(\varepsilon, \varepsilon)$

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

$= WW^T + D$

equivalent model on $x \sim N(\mu, \underbrace{WW^T}_{d \times k} + D)$ ← low rank covariance piece + diagonal → d degrees of freedom

estimate W, D & μ by MLE

↝ do EM (latent variable model)

get $p(z|x)$ → Gaussian with mean

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

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

(μ_x = mean of x)

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

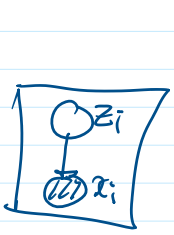
$$\lim_{\sigma \to 0} W^T(WW^T+\sigma^2 I)^{-1} = W^\dagger \leftarrow \text{pseudo-inverse}$$

$$= W^T \quad \text{if } W^TW = I_K$$

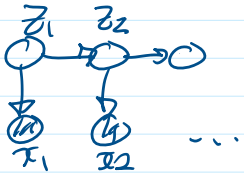
this suggests PCA is limit of PPCA as $\sigma \to 0$

Kalman filter:

factor analysis



Kalman filter



move to state space model: unroll ~~in time~~ (HMM style)

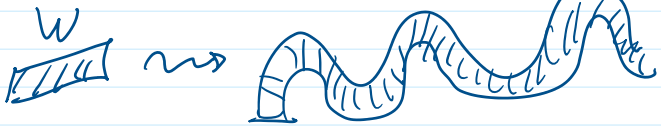
Kalman filter: $z_t | z_{t-1} \sim N(Az_{t-1}, B)$

→ doing "sum-product" alg. in HMM $p(z_t|x_{1:t})$

get "Kalman filter" alg.

Variational auto-encoder

generalization of factor analysis



$$z \sim N(0, I_K)$$

$$x|z \sim N(\mu_W(z), \sigma_W^2(z)) \quad \text{(diagonal noise)}$$

where $\mu_W(z)$ ← output of NN_W "decoder"

MLE → use EM

↳ $p(z|x)$ is intractable

⇒ approximate with variational approach

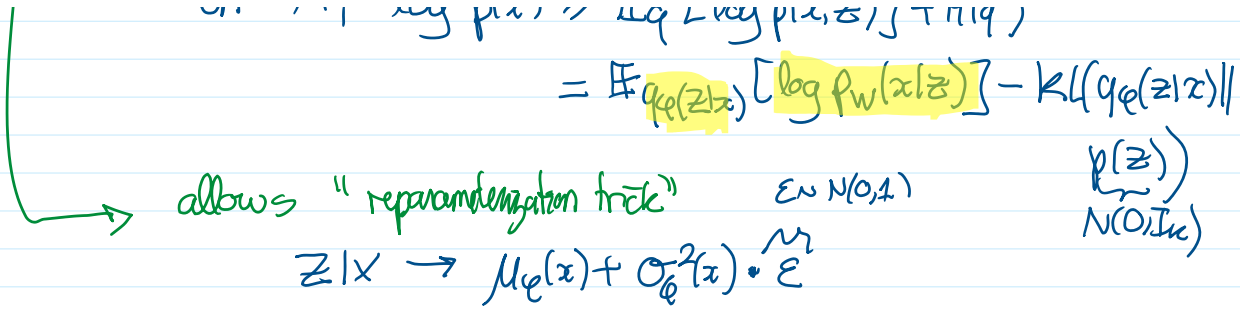
approximate $p(z|x)$ with $q_\phi(z|x)$

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

($\mu_\phi(x)$: output of NN "encoder")

in EM, $\log p(x) \geq \mathbb{E}_q[\log p(x,z)] + H(q)$

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



- 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 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