Lecture 24 - Gaussian networks; factor analysis; VAE

Thursday, December 5, 2024 2:35 PM

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

Gaussian networks

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

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

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

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

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

$\eta \triangleq \Sigma^{-1}\mu$

sufficient statistics $T(x) = \begin{pmatrix} x \\ -\frac{xx^T}{2} \end{pmatrix}$ $\xrightarrow{\text{canonical parameters}}$ $\Lambda \triangleq \Sigma^{-1}$

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

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

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

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

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 $\Rightarrow$ cond. indep property

"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

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

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

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

$$p(x_1, x_2) = \underbrace{\frac{1}{\sqrt{(2\pi)^{d_1}|\Sigma_{11}|}} \exp\left(-\frac{1}{2}(x_1-\mu_1)^T\Sigma_{11}^{-1}(x_1-\mu_1)\right)}_{p(x_1)}$$

(d_1: dim x_1, d_2: dim x_2)

$$\underbrace{\frac{1}{\sqrt{(2\pi)^{d_2}|\Sigma/\Sigma_{11}|}} \exp\left(-\frac{1}{2}(x_2-\mu_2-b(x_1))^T(\Sigma/\Sigma_{11})^{-1}(x_2-\mu_2-b(x_1))\right)}_{p(x_2|x_1)}$$

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 can. 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

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

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

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

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

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

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

(also true by Markov property of UGM)

Factor analysis :

latent variable model

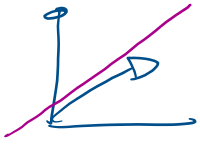
$z \in \mathbb{R}^k$ learn a "latent representation" in $\mathbb{R}^k$

$x \in \mathbb{R}^d$ or dimensionality reduction $k \ll d$

15h28

PCA for dimensionality reduction

synthetic view: find k orthonormal vector 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



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

$WW^T ? \neq I_d$

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

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

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

$= \begin{pmatrix} | & & | \\ w_1 & \cdots & w_k \\ | & & | \end{pmatrix} \begin{pmatrix} \langle w_1, x \rangle \\ \langle w_2, x \rangle \\ \vdots \end{pmatrix}$

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

$z = W^T x$

↳ lower dimensional representation

linear auto-encoder

PCA $\min_{W \in \mathbb{R}^{d \times k}} \sum_i \| x_i - WW^T x_i \|_2^2$ (z_i)

$W^T W = I_k$ $C(W) \triangleq$ principal subspace

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

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

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

$= tr\left(X (I_d - WW^T)(I_d - WW^T) X^T \right)$

$= tr\left(X (I_d - P_W) X^T \right) = tr\left(X^T X (I_d - P_W) \right) = C st. - tr(X^T X P_W)$

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

↑ empirical covariance of x when $\sum x_i = 0$ (mean = 0)

min rec. error $\Leftrightarrow$ maximizing $tr(XX^T WW^T) = \sum_k w_k^T X^T X w_k$

"analysis view" max (leading vector)

(mean = 0)

"analysis view of PCA"

max sum of empirical covariances of new representations

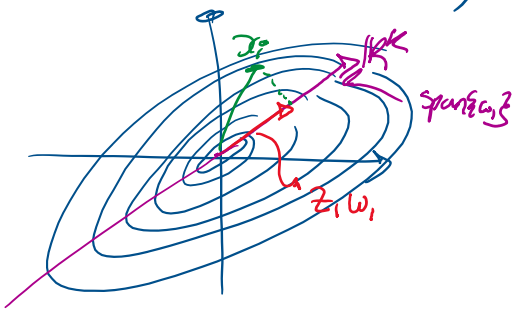
long vector of $(z_{i,k})_{i=1}^{n}$

⊛ W is not unique, only col(W)

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

$\tilde{W}\tilde{W}^T = WRR^TW^T$ ($RR^T = I_k$)

(compute soln of PCA ⇒ top eigenvectors of X^TX)



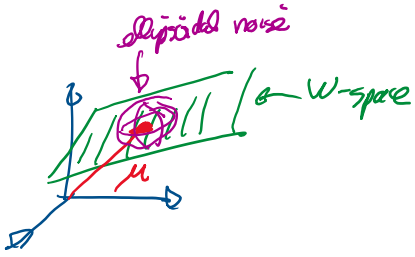
Factor analysis → simplest generative model

$z \sim N(0, I_k)$

$x = Wz + \mu + \varepsilon$

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

diagonal matrix



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

$p(z)$ is Gaussian

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

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

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

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

$= WW^T + D$

equivalent model on $\boxed{x \sim N(\mu, WW^T + D)}$ (W: $d \times k$; D: diagonal → d degrees of freedom)

low rank covariance structure ⊛⊛

estimate W & D & μ by MLE

↝ do EM (latent variable model)

get $p(z|x)$ → Gaussian with mean

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

Probabilistic PCA: special case of factor analysis

where suppose $D = \sigma^2 I$

+

where suppose $\Sigma = \sigma^2 I$

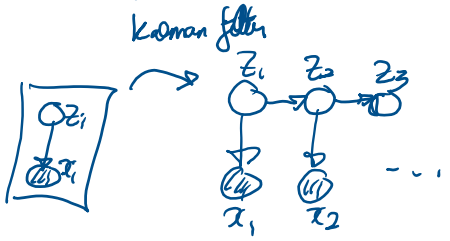
$\lim_{\sigma \to 0} W^T (WW^T + \sigma^2 I)^{-1} = W^+$ ← pseudo inverse

$= W^T$ if $W^T W = I_K$

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

Kalman filter

factor analysis



more to state space model ; unroll 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

$z \sim N(0, I_K)$

diagonal noise

$x | z \sim N(\mu_w(z), \sigma_w^2(z))$

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 $q_\varphi(z|x)$

parameter sharing

"amortized inference"

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

output of NN "encoder"

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

$= E_{q_\varphi(z|x)}[\log p_w(x|z)] - KL(q_\varphi(z|x) \,||\, p(z))$

$N(0,I)$

allows "reparametrization trick"

$\varepsilon \sim N(0, I)$

$z|x \to \mu_\varphi(x) + \sigma_\varphi^2(x) \cdot \varepsilon$

- 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