Lecture 24 - Gaussian networks; factor analysis; VAE

November 30, 2021 4:25 PM

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

Gaussian networks

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

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

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

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

put it in exponential family

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

$$\langle \Sigma^{-1}, -\frac{xx^T}{2} \rangle + \langle \underbrace{\Sigma^{-1}\mu}_{\triangleq \eta}, x \rangle - \frac{1}{2}\mu^T\Sigma^{-1}\mu$$

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

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

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

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

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

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

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

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

quick Schur-complement digression:

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

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

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

$M \triangleq \Sigma/\Sigma_{11} \triangleq \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$ "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. of marginal on x_1

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

$\Sigma_{2|1}^{\text{cond.}} = \Sigma/\Sigma_{11} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$ } param. for cond. $x_2|x_1$

in canonical param

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

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

$\eta_1^M = \eta_1 - \Lambda_{12}\Lambda_{22}^{-1}\eta_2$ (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

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

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

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

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

(also true by Markov property of UGM)

(also true by Markov property of UGM)

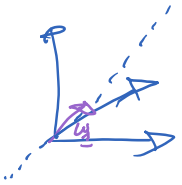
17h34

Factor analysis:

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

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



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

$WW^T ? \neq I_d$

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

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

$P_W x = WW^T x$

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

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

$z = W^T x$

↳ lower dimensional representation

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

col(W) $\triangleq$ principal subspace

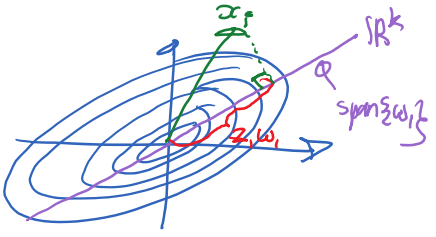
$$X_{n \times d} = \begin{pmatrix} - x_1^T - \\ \vdots \\ - x_n^T - \end{pmatrix}$$

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

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

$= \text{tr}\left(X (I_d - P_W)^T (I_d - P_W) X^T \right)$ $\quad (I_d - P_W)$

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



W is not unique, only col(W)

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

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

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

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

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

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

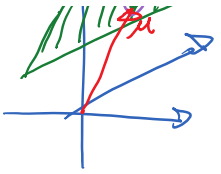
(computation of PCA → top k e-vectors of $X^T X$)

Factor analysis → simplest generative model

$z \sim N(0, I_k)$, noise

$z \sim N(0, I_k)$

$x = Wz + \mu + \varepsilon$ ← noise



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

D: diagonal matrix

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

$p(x)$ is Gaussian

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

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

($Wz+\mu$, ε: indep.)

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

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

$$= WW^T + D$$

equivalent model as $\boxed{x \sim N(\mu, WW^T + D)}$

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

D: diagonal → d degrees of freedom

estimate W & D & μ by MLE

↝ do EM (latent variable model)

get $p(z|x)$ → Gaussian with mean

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

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

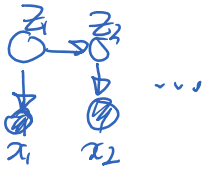
$$\lim_{\sigma \to 0} W^T(WW^T + \sigma^2 I)^{-1} = W^+ \leftarrow \text{pseudo-inverse}$$

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

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

Kalman filter:

factor analysis



more 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 $p(z_t | x$

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

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

"decoder"

where $\mu_w(z)$ ← output of NN

MLE → use EM

↳ $p(z|x)$ is intractable ⇒ approximate with variational approach

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

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

output of a NN

"encoder"

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

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

$N(0, I_K)$

allows "reparameterization trick"

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

$\varepsilon \sim N(0,1)$

- 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