Lecture 25 - Gaussian networks; factor analysis; VAE

Tuesday, December 1, 2020 14:31

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

Gaussian networks

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

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

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

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

put in exponential family

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

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

$\Lambda \triangleq \Sigma^{-1}$ precision matrix $\quad \triangleq \eta \quad \mu = \Sigma\eta = \Lambda^{-1}\eta$

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

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

$$A(\eta, \Lambda)$$

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

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

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

UGM viewpoint:

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

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

zeros in precision matrix ⇒ cond. indep. properties (from UGM perspective)

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

quick Schur-complement digression:

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

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

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

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

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

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

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

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

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

mean parameterization of marginals & conditionals

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

$$\mu_{2|1}^{\text{cond.}} = \mu_2 + b(x_1)$$
$$\Sigma_{2|1}^{\text{cond.}} = \Sigma/\Sigma_{11} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$ } for conditional $X_2 \mid X_1$

---

in canonical param.

$$\Lambda_{2|1}^{\text{cond.}} = \Lambda_{22} \quad \text{(simple!)}$$
$$\eta_{2|1}^{\text{cond.}} = \eta_2 - \Lambda_{21}x_1$$

- - - - - - -

$$\eta_1^m = \eta_1 - \Lambda_{12}\Lambda_{22}^{-1}\eta_2$$
$$\Lambda_1^m = \Lambda_{11} - \Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21} = \Lambda/\Lambda_{22}$$ (more complicated)

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

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

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

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

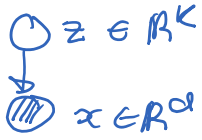
$$\Rightarrow \boxed{X_i \perp\!\!\!\perp X_j \mid X_{rest}}$$

(also true by Markov property of UGM)

15h38

Factor analysis:

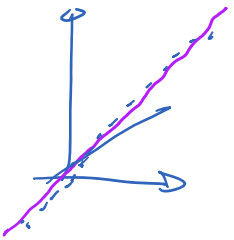
latent variable model



learn "latent representation" 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



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

$$WW^T ? \neq I_d$$

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

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

$$P_W x = WW^T x$$

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

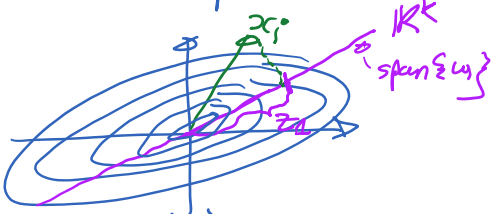
$$= \sum_k w_k \underbrace{\langle w_k, x \rangle}_{(z)_k} = Wz$$

$$\boxed{\text{PCA} \quad \min_{W \in \mathbb{R}^{d \times k},\ W^T W = I_k} \sum_i \| x_i - W \underbrace{W^T x_i}_{z_i} \|_2^2}$$

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

$z = W^T x$

↳ lower dimensional representation



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

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

W is not unique, only $\text{col}(W)$

e.g. ~

$x_n^T -$)

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

W is not unique, only col(W)

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

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

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

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

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

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

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

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

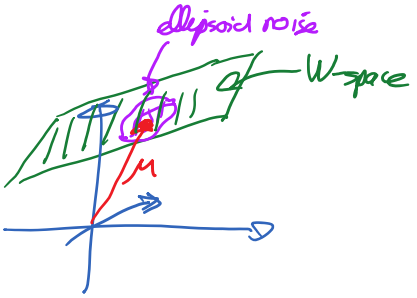
(computation of PCA → top k e-vectors of X^TX)

Factor analysis → simplest generative model

$z \sim N(0, I_k)$

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

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



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

$p(x)$ is Gaussian;

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

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

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

$= W\underbrace{cov(z,z)}_{=I_k}W^T + 0$

$= WW^T + D$

equivalent model as $x \sim N(\mu, \underbrace{WW^T}_{d\times k} + D)$

low rank covariance piece; diagonal $\Rightarrow$ d degrees of freedom

estimate W, D, μ by MLE

↝ 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 $D = \sigma^2 I$

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

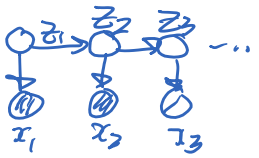
$= W^T$ if $W^TW = I_K$

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

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

Kalman filter

factor analysis 



more: 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

$z \sim N(0, I_K)$

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 ⇒ use variational approach

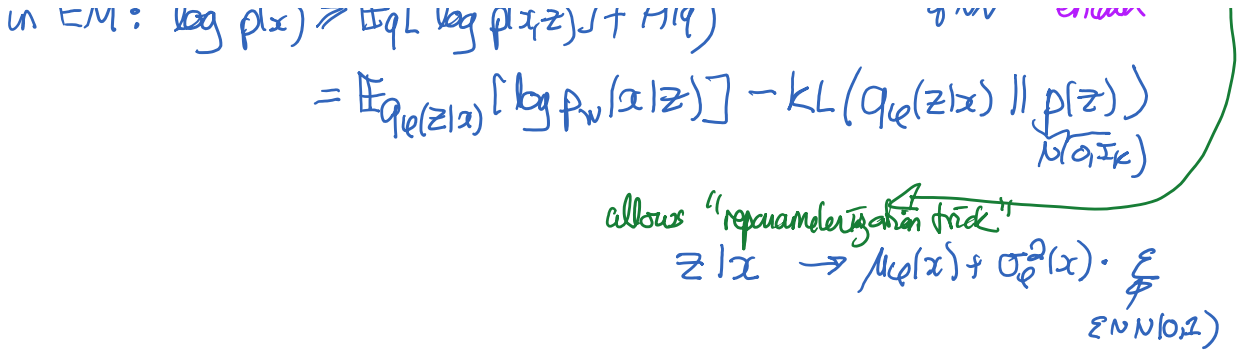
approximate $p(z|x)$ with $q_\varphi(z|x)$

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

output of NN "encoder"

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

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



- 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 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