Lecture 22 - Gaussian networks; factor analysis; VAE

Friday, November 25, 2022 3:01 PM

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

mean field continuation

$\min_{q \in \mathcal{Q}_{MF}} KL(q \| p)$

$\mathcal{Q}_{MF} = \{ q : q(x) = \prod_i q_i(x_i) \}$

• $KL(\cdot \| p)$ is a convex of q (as vector)

but $\mathcal{Q}_{MF}$ is a non-convex constraint set

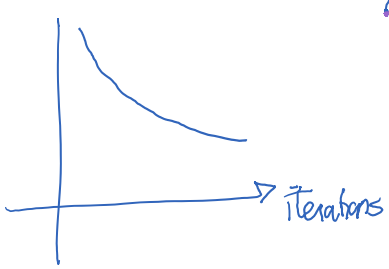
⇒ can get stuck in local minima

e.g. Ising model $\mu_{ij} \overset{?}{=} \mu_i \cdot \mu_j$ non-convex constraint

[see lecture 27 in 2017, for "marginal polytope"

but can monitor progress

$KL(q^{(t)} \| p) + cst$



pros & cons of variational methods vs. Sampling

⊕ optimization based ⇒ often faster to run & easier to debug

⊖ noisy ⇒ harder to debug
mixing problems for chains

⊖ biased estimate

$\mathbb{E}_{q(x)}[f(z)] \neq \mathbb{E}_p[f(z)]$

⊕ unbiased estimate

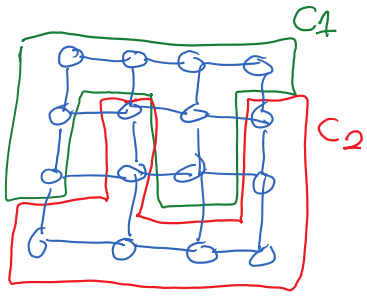
$\mathbb{E}\left[\mathbb{E}_{q_\omega}[f(z)]\right] = \mathbb{E}_p[f(z)]$

$\mathbb{E}$ with respect to random sample

structured mean field:

idea $q(z) = \prod_{j=1}^{K} q_j(z_{C_j})$ where $C_1, \dots, C_K$ is a partition of V

and q_j's are tractable distributions (for example tree UGM)



15h43

Gaussian networks

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

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

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

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

put in exponential family

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

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

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

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

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

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

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

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

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

UGM viewpoint:

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

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

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

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

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

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

$$p(\underbrace{x_1}_{\dim d_1}, \underbrace{x_2}_{\dim d_2}) = \frac{1}{\sqrt{(2\pi)^{d_1}|\Sigma_{11}|}} \exp\left(-\frac{1}{2}(x_1-\mu_1)^T\Sigma_{11}^{-1}(x_1-\mu_1)\right) \cdot \Big\} p(x_1)$$

$d = d_1 + d_2$

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

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

mean parameterization of marginal on x_1 and conditional $x_2|x_1$

$$\left.\begin{aligned}\mu_1^m &= \mu_1 \\ \Sigma_1^m &= \Sigma_{11}\end{aligned}\right\} \text{super simple!} \quad \Big\} \text{param. of marginal on } x_1$$

$$\left.\begin{aligned}\mu_{2|1}^{cond} &= \mu_2 + b(x_1) \\ \Sigma_{2|1}^{cond.} &= \Sigma/\Sigma_{11} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\end{aligned}\right\} \text{param for cond. } x_2|x_1$$

---

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

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

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

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

(more complicated)

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

for example: block $\overbrace{i,j}^{I}$ | rest

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

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

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

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

(also true by Markov of UGM)

Factor analysis

latent variable model $\bigcirc\ z \in \mathbb{R}^k$ learn "latent representation"

latent variable model $z \in \mathbb{R}^k$ → $x \in \mathbb{R}^d$ 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. projecting x on $\text{span}\{w_1, \dots, w_k\}$ is a good approx. of x

$W = \underset{d \times k}{\begin{bmatrix} | & | & & | \\ w_1 & w_2 & \cdots & w_k \\ | & | & & | \end{bmatrix}}$ $\quad W^T W = I_k$ (by orthonormality)

$WW^T ? = I_d$

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

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

$P_W x = WW^T x$

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

$= \sum_{i=1}^{k} w_i \underbrace{\langle w_i, x \rangle}_{(z)_i} = Wz$

$z = W^T x$

lower dimensional representation

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

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

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

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

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

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

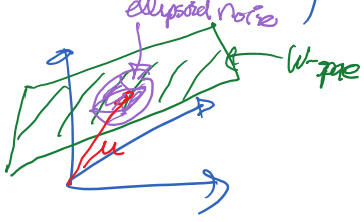
$= \text{tr}\left(X (I_d - P_W) X^T \right) = \text{tr}\left(X^T X (I_d - P_W) \right)$

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

min rec. error $\Leftrightarrow$ 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 eigenvectors of $X^T X$)



factor analysis → simplest generative model

$z \sim N(0, I_k)$

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

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

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