c∗GMΔVÆs—q

Kolb, Yiftach

January 31, 2023

MAX-PLANCK-INSTITUT
FÜR MOLEKULARE GENETIK

c∗GMΔVÆs—q

a Master Thesis in Bioinformatics

Advisor / Reviewer: Professor Martin Vingron
Reviewer: Professor Tim Conrad

Freie Universität Berlin

Topics to cover

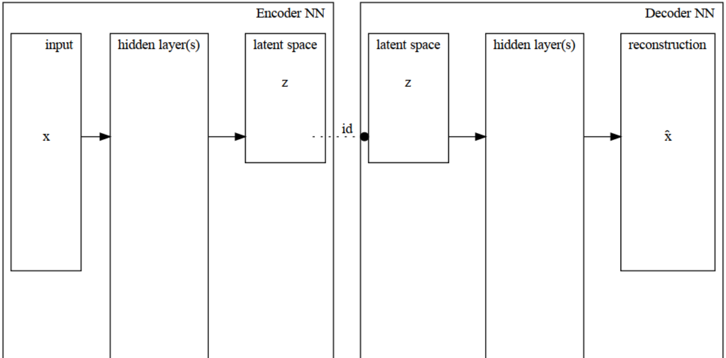
- what was our initial subject of interest
- AEs are non-linear PCA basically
- VAEs
- other animals
- GMVAE and why I derived c$*$GMΔVÆ
- example use of c$*$GMΔVÆ on synthetic conditional-categorical data
- examples on MNIST
- examples on scRNAseq

Autoencoders

$$\Delta \mathrm{abc}abc$$

c∗GMVÆ c∗GMΔVÆ

A "vanilla" autoencoder is a neural networks that "learns" the identity (subject to dimensional restriction).



Autoencoders and PCA

(On centered data[8])

PCA

$$\tilde{\boldsymbol{V}} = \operatorname{argmin}_{\mathbf{W}}\{\|\mathbf{X} - \mathbf{X}\boldsymbol{W}\boldsymbol{W}^T\|_F^2 \quad : \quad \boldsymbol{W} \in \mathbb{R}^{n\times l}, \boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{I}_l\} \tag{1}$$

Linear AE

$$\operatorname{argmin}_{\boldsymbol{E},\boldsymbol{D}}\{\|\mathbf{X} - \mathbf{X}\boldsymbol{E}\boldsymbol{D}\|_F^2 \quad : \quad \boldsymbol{E}, \boldsymbol{D}^{\boldsymbol{T}} \in \mathbb{R}^{n\times l}, \} \tag{2}$$

$$\tilde{\boldsymbol{W}} \in \operatorname{argmin}_{\mathbf{W}}\{\|\mathbf{X} - \mathbf{X}\boldsymbol{W}\boldsymbol{W}^\dagger\|_F^2 \quad : \quad \boldsymbol{W} \in \mathbb{R}^{n\times l}, \} \tag{3}$$

$\operatorname{span}\{\tilde{\boldsymbol{W}}\} = \operatorname{span}\{\tilde{\boldsymbol{V}}\}$

VAEs

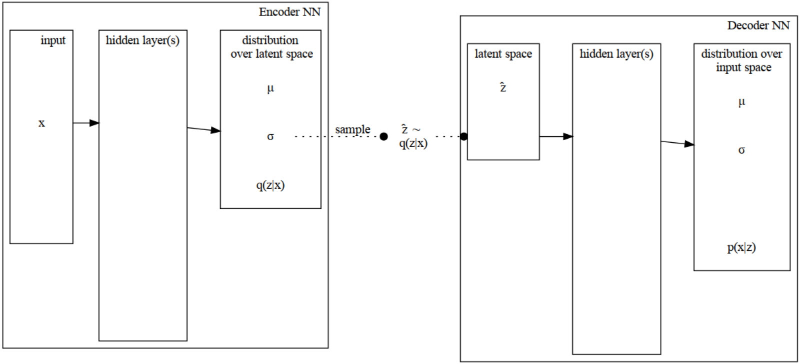


Figure: VAE

VAE: encoding

Instead of deterministic mapping, define distribution.

Define distribution on the laten space ($\mathbf{z}$) by mapping x into the distribution parameters e.g. $\mu(\mathbf{x}), \Sigma(\mathbf{x})$ when we use Gaussian $q(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\mu, \Sigma)$.

VAE: decoding

sample from the latent space $\mathbf{z} \sim \mathcal{N}(\cdot|\mu, \Sigma)$

map $\mathbf{z}$ to a distribution on the input space $p(\mathbf{x}|\mathbf{z})$

VAE: loss function

The *evidence lower bound (ELBO)* with respect to p, q is:

$$-\mathcal{L}(q, p, \mathbf{x}) \triangleq \int \log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z})} dq(\boldsymbol{z}) \tag{4}$$

$$-\mathcal{L}(q, p) \triangleq -\mathcal{L}(q, p, \mathbf{X}) = \frac{1}{N} \sum_{1}^{N} (-\mathcal{L}(q, p, \mathbf{x}_i) \tag{5}$$

$$\approx \mathbf{E}_{\mathbf{x}}[-\mathcal{L}(q, p, \mathbf{x})] \tag{6}$$

We minimize the minus ELBO function:

VAE: log evidence

It can be shown that maximizing the ELBO is equivalent to maximinizing the "log evidence" $\log p(\mathbf{X})$

$$
\begin{aligned}
\frac{1}{N}\log p(\boldsymbol{X}) &= \frac{1}{N}\log\int p(\boldsymbol{X},\boldsymbol{Z})d\boldsymbol{Z} && \text{taking marginal} \\
&= \frac{1}{N}\log\int \frac{p(\boldsymbol{X},\boldsymbol{Z})}{q(\boldsymbol{Z})}q(\boldsymbol{Z})d\boldsymbol{Z} && \text{multiplying by 1 inside} \\
&= \frac{1}{N}\log\int \frac{p(\boldsymbol{X},\boldsymbol{Z})}{q(\boldsymbol{Z})}dq(\boldsymbol{Z}) && \text{definition of } dq(\mathbf{Z}) \\
&\geq \frac{1}{N}\int \log\frac{p(\boldsymbol{X},\boldsymbol{Z})}{q(\boldsymbol{Z})}dq(\boldsymbol{Z}) && \text{Jensen inequality} \\
&= \frac{1}{N}\int\sum_{1}^{N}\log\frac{p(\mathbf{x}_i,\mathbf{z}_i)}{q(\mathbf{z}_i)}dq(\mathbf{z}_i) && \text{using the iid property} \\
&= \frac{1}{N}\sum_{1}^{N}-\mathcal{L}(q,p,\mathbf{x}_i) && \text{definition of } \mathcal{L}(q,p,\mathbf{x}_i) \\
&= -\mathcal{L}(q,p,X) \triangleq -\mathcal{L}(q,p) && \text{again definition of } \mathcal{L}(p,q) \quad \square
\end{aligned}
\tag{7}
$$

VAE: compounding the latent distribution

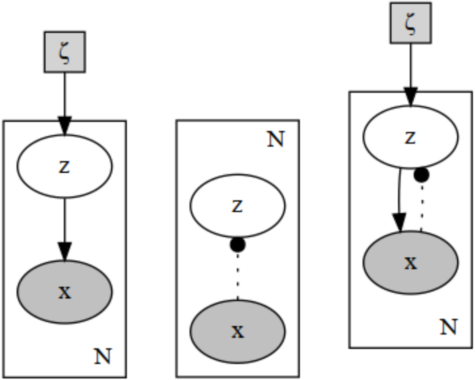
More complicated distributions such as mixture distribution can be modelled by "unpacking" the latent $\mathbf{z}$ and the observed $\mathbf{x}$

1. Define the set of observed random vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_k$, and the set of latent random vectors and stochastic parameters $\mathbf{z}_1, \ldots \mathbf{z}_l$.
2. Specify how to factor the generative model $p(\mathbf{x}_1, \ldots, \mathbf{x}_k | \mathbf{z}_1 \ldots, \mathbf{z}_l)$
3. Specify how to factor the inference model $q(\mathbf{z}_1 \ldots \mathbf{z}_l | \mathbf{x}1, \ldots \mathbf{x}_k)$
4. Choose appropriate priors $p(\mathbf{z}_i)$ and
5. Choose appropriate distribution families for the $\mathbf{x}_i$ and $\mathbf{z}_i$, and choose priors $p(\mathbf{z}_i)$.

VAE: Graphical representation

Evry distribution can be represented by a DAG. Nodes represent random variables (and also priors), and directed arrows represent conditional dependency.

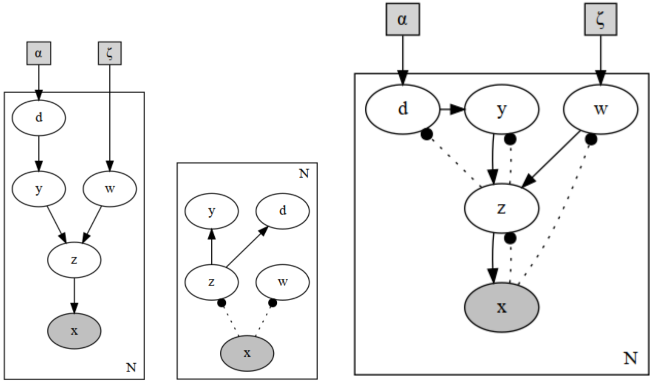
VAE: base case



(a) generative model $p(x, z)$ (b) inference model $q(z|x)$ (c) the combined graphical model

Figure: VAE graphical model

VAE: patholocigacl case



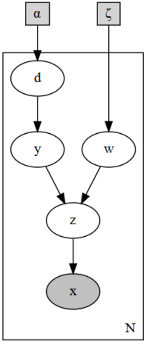
(a) generative model $p(x, z)$

(b) inference model $q(z|x)$

(c) the combined graphical model

Figure: c∗GMΔVÆ graphical model

c∗GMΔVÆ generative model



$$
\begin{aligned}
p(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{w}, \mathbf{d}) &= p(\mathbf{x}|\mathbf{z})p(\mathbf{z}|\mathbf{w}, \mathbf{y})p(\mathbf{y}|\mathbf{d})p(\mathbf{d})p(\mathbf{w}) \\
p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{1}) \\
p(\mathbf{d}) &= \mathrm{Dir}(\mathbf{d}|\alpha) \\
p(\mathbf{y}|\mathbf{d}) &= \mathrm{Cat}(\mathbf{y}|\mathbf{d}) \\
p(\mathbf{z}|\mathbf{w}, \mathbf{y}) &= \mathcal{N}(\mathbf{z}|\mu(\mathbf{w})_{\mathbf{y}}, \sigma(\mathbf{w})_{\mathbf{y}})) \\
p(\mathbf{x}|\mathbf{z}) &= \mathcal{N}(\mathbf{x}|\mu(\mathbf{z}), \sigma(\mathbf{z}))
\end{aligned} \tag{8}
$$

[1] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Vol. 4. 4. Springer, 2006.

[2] Nat Dilokthanakul et al. "Deep unsupervised clustering with gaussian mixture variational autoencoders". In: *arXiv preprint arXiv:1611.02648* (2016).

[3] Diederik P Kingma and Max Welling. "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114* (2013).

[4] Durk P Kingma et al. "Semi-supervised learning with deep generative models". In: *Advances in neural information processing systems* 27 (2014).

[5] Yiftach Kolb. *The "official" c*GMΔVAE project git*. URL: https://github.com/zelhar/mg22.

[6] Yiftach Kolb. *The Github project housing this thesis*. URL: https://github.com/zelhar/mg22.

[7] Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. "scGen predicts single-cell perturbation responses". In: *Nature methods* 16.8 (2019), pp. 715–721.

[8] Elad Plaut. "From principal subspaces to principal components with linear autoencoders". In: *arXiv preprint arXiv:1804.10253* (2018).