ELBO and Variational Autoecoder

Václav Šmídl

June 6, 2022

Recapitulation

- how to sample a general Gaussian using $N(0, 1)$?

Recapitulation

- how to sample a general Gaussian using $N(0, 1)$?
- what optimizes Variational Bayes?

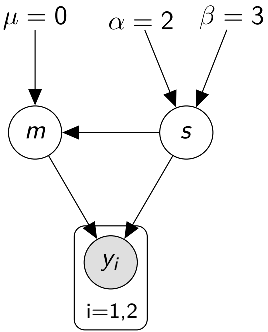
Recapitulation

- how to sample a general Gaussian using $N(0, 1)$?
- what optimizes Variational Bayes?
- generative model of PCA?

Recapitulation

- how to sample a general Gaussian using $N(0,1)$?
- what optimizes Variational Bayes?
- generative model of PCA? linear model $D = AX$,

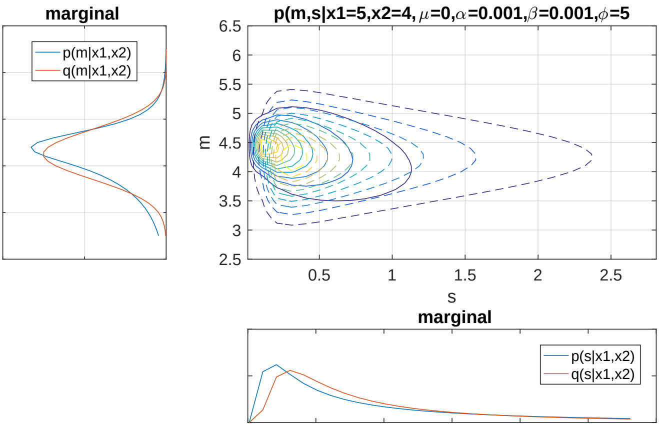
Previous models



$$
\begin{aligned}
p(s) &= iG(\alpha_0, \beta_0) \\
p(m|s) &= \mathcal{N}(\mu, s) \\
p(y_i|m, s) &= \mathcal{N}(m, s)
\end{aligned}
$$

- Observations x_i are sampled from Gaussian with unknown mean and variance.
- We have some prior information about the mean and variance

Approximation via Variational Bayes



s

Divergence minimization

We seek best approximation of intractable distribution $p(x)$ in the chosen class of parametric functions, $q(x|\theta)$, such that

$$\theta^* = \arg\min_\theta D(p, q),$$

where $D(p, q)$ is a statistical divergence.

Divergence minimization

We seek best approximation of intractable distribution $p(x)$ in the chosen class of parametric functions, $q(x|\theta)$, such that

$$\theta^* = \arg\min_\theta D(p, q),$$

where $D(p, q)$ is a statistical divergence.

Different results for different choices of: i) $q(\theta)$, and ii) D.

Divergence minimization

We seek best approximation of intractable distribution $p(x)$ in the chosen class of parametric functions, $q(x|\theta)$, such that

$$\theta^* = \arg\min_{\theta} D(p, q),$$

where $D(p, q)$ is a statistical divergence.

Different results for different choices of: i) $q(\theta)$, and ii) D.

Variational Bayes:

1. conditional independence $q(\theta_1, \theta_2) = q(\theta_1)q(\theta_2)$,
2. (reverse) Kullback-Leibler. $\mathrm{KL}(q||p) \neq \mathrm{KL}(p||q)$

Kullback Leibler divergence

Measure of divergence between two probability densities

$$KL(q||p) = E_q\left(\log \frac{q}{p}\right)$$

Not a metric!

$$E_q\left(\log \frac{q}{p}\right) \neq E_p\left(\log \frac{p}{q}\right)$$

also known as relative/free entropy

$$KL(q||p) = E_q(\log q) - E_q(\log p)$$

with properties:

1. $KL(q||p) \geq 0$,
2. $KL(q||p) = 0, \iff q = p$

Variational Bayes

Is a divergence minimization technique with

$$q^* = \arg\min_q KL(q||p) = \arg\min_q E_q\left(\log \frac{q}{p}\right)$$
$$q(m,s) = q(m|y_1,y_2)q(s|y_1,y_2).$$

which allows *free-form* optimization.

Variational Bayes

Is a divergence minimization technique with

$$q^* = \arg\min_q KL(q||p) = \arg\min_q E_q\left(\log\frac{q}{p}\right)$$
$$q(m, s) = q(m|y_1, y_2)q(s|y_1, y_2).$$

which allows *free-form* optimization.
Result:

$$q(m|y_1, y_2) \propto \exp\left(E_{q(s)}\left[\log p(y_1, y_2, m, s)\right]\right)$$
$$q(s|y_1, y_2) \propto \exp\left(E_{q(m)}\left[\log p(y_1, y_2, m, s)\right]\right)$$

which is a set of implicit functions.

- Proportionality above allows to use $p(y_1, y_2, m, s)$ in place of $p(m, s|y_1, y_2)$
- Variational EM algorithm (E-E algorithm).

E-step: m

Proxy distribution

$$q(m) = \mathrm{E}_{q(s)}(\log p(y_1, y_2, m, s))$$

using

$$p(y_1, y_2, m, s) \propto \frac{1}{s}\frac{1}{s^{\alpha_0+1}} \exp\left(-\frac{1}{2}\frac{(m-y_1)^2}{s} - \frac{1}{2}\frac{(m-y_2)^2}{s} - \frac{1}{2}\frac{(m-\mu)^2}{\phi} - \frac{\beta_0}{s}\right)$$

$$\begin{aligned}
\log p(y_1, y_2, m, s) &\propto (\alpha_0+2)\log s - \frac{1}{2}\frac{(m-y_1)^2}{s} - \frac{1}{2}\frac{(m-y_2)^2}{s} - \frac{1}{2}\frac{(m-\mu)^2}{\phi} - \frac{\beta_0}{s}, \\
&\propto -\frac{1}{2}\mathrm{E}\left(\frac{1}{s}\right)\left[(m-y_1)^2 + (m-y_2)^2\right] - \frac{1}{2}\frac{(m-\mu)^2}{\phi} \\
&\propto -\frac{1}{2}\left[\frac{(m-y_1)^2}{\hat{s}} + \frac{(m-y_2)^2}{\hat{s}}\right] - \frac{1}{2}\frac{(m-\mu)^2}{\phi} \\
q(m) &= \mathcal{N}\left(m; \left(\frac{1}{\phi} + \frac{2}{\hat{s}}\right)^{-1}\left(\frac{\mu}{\phi} + \frac{y_1+y_2}{\hat{s}}\right), \left(\frac{1}{\phi} + \frac{2}{\hat{s}}\right)^{-1}\right)
\end{aligned}$$

E-step: s

Proxy distribution

$$q(s) = \mathrm{E}_{q(m)}(\log p(y_1, y_2, m, s))$$

using

$$p(y_1, y_2, m, s) \propto \frac{1}{s}\frac{1}{s^{\alpha_0+1}} \exp\left(-\frac{1}{2}\frac{(m-y_1)^2}{s} - \frac{1}{2}\frac{(m-y_2)^2}{s} - \frac{1}{2}\frac{(m-\mu)^2}{\phi} - \frac{\beta_0}{s}\right)$$

$$\log p(y_1, y_2, m, s) = -(\alpha_0+2)\log s - \frac{1}{2}\frac{(m-y_1)^2}{s} - \frac{1}{2}\frac{(m-y_2)^2}{s} - \frac{1}{2}\frac{(m-\mu)^2}{\phi} - \frac{\beta_0}{s},$$

$$\mathrm{E}_{q(m)}(\log p(y_1, \cdot)) = -(\alpha_0+2)\log s - \frac{1}{2s}\mathrm{E}_{q(m)}\left[(m-y_1)^2 + (m-y_2)^2\right] - \frac{\beta_0}{s}$$

$$= -(\alpha_0+2)\log s - \frac{1}{2s}\mathrm{E}_{q(m)}\left[(m^2 - 2my_1 + y_2^2 + m^2 - 2y_2m + y_2^2\right]$$

$$q(s) = i\mathcal{G}(\alpha, \beta),$$

$$\alpha = \alpha_0 + 1,$$

$$\beta = 0.5\mathrm{E}(m^2) - \mathrm{E}(m)(y_1+y_2) + 0.5(y_1^2 + y_2^2) + \beta_0$$

Toy: Variational Bayes

Factors:

$$q(m) = \mathcal{N}\left(m; \left(\frac{1}{\phi} + \frac{2}{\hat{s}}\right)^{-1} \left(\frac{\mu}{\phi} + \frac{y_1 + y_2}{\hat{s}}\right), \left(\frac{1}{\phi} + \frac{2}{\hat{s}}\right)^{-1}\right)$$

$$q(s) = i\mathcal{G}\left(\alpha_0 + 1, \mathrm{E}(m^2) - \mathrm{E}(m)(y_1 + y_2) + 0.5(y_1^2 + y_2^2) + \beta_0\right)$$

with

$$\hat{s} = \frac{\mathrm{E}(m^2) - \mathrm{E}(m)(y_1 + y_2) + 0.5(y_1^2 + y_2^2) + \beta_0}{\alpha_0 + 1},$$

$$\mathrm{E}(m) = \left(\frac{1}{\phi} + \frac{2}{\hat{s}}\right)^{-1} \left(\frac{\mu}{\phi} + \frac{y_1 + y_2}{\hat{s}}\right),$$

$$\mathrm{E}(m^2) = \mathrm{E}(m)^2 + \left(\frac{1}{\phi} + \frac{2}{\hat{s}}\right)^{-1},$$

which needs to be (Iterated).

Direct optimization of KL

Is a divergence minimization technique with

$$\begin{aligned} q^* = \arg\min_q \mathrm{KL}(q||p) &= \arg\min_q \mathrm{E}_q \left(\log \frac{q}{p}\right) \\ q(m,s) &= q(m|y_1,y_2)q(s|y_1,y_2). \\ q(m|y_1,y_2) &= \mathcal{N}(\mu_m,\sigma_m) \\ q(s|y_1,y_2) &= i\mathcal{G}(\alpha_s,\beta_s) \end{aligned}$$

Solving task:

$$\mu_m^*,\sigma_m^*,\alpha_s^*,\beta_s^* = \arg\min_{\mu_m,\sigma_m,\alpha_s,\beta_s} \mathrm{E}_q\left(\log\frac{q}{p}\right),$$

using general purpose black-box optimizer (SGD).

Loss function

$$
\begin{aligned}
\mathrm{KL}(q||p) &= \mathrm{E}_{q_m q_s}\left(\log \frac{q(m|x)q(s|x)}{p(m,s|x)}\right) \\
&= \mathrm{E}_{q_m q_s}\left(\log q(m|x) + \log q(s|x) - \log p(x,m,s) + \log p(x)\right), \\
&= -H(q(m|x)) - H(q(s|x)) - \mathrm{E}(\log p(x,m,s)) + \log p(x)
\end{aligned}
$$

Where

$$
\begin{aligned}
&\mathrm{E}_{q_m q_s}\left(\log p(x,m,s)\right) = \\
&\mathrm{E}_{q_m q_s}\left[-(\alpha_0+2)\log s - \frac{1}{2}\frac{(m-y_1)^2}{s} - \frac{1}{2}\frac{(m-y_2)^2}{s} - \frac{1}{2}\frac{(m-\mu)^2}{\phi} - \frac{\beta_0}{s}\right] = \\
&= -(\alpha_0+2)\mathrm{E}_{q_s}(\log s) - \mathrm{E}_{q_s}\left(\frac{\beta_0}{s}\right) \\
&\quad - \mathrm{E}_{q_s}\left(\frac{1}{2s}\right)\mathrm{E}_{q(m)}\left[(m^2 - 2my_1 + y_1^2 + m^2 - 2y_2 m + y_2^2\right] \\
&\quad - \frac{1}{2\phi}\mathrm{E}_{q(m)}\left[m^2 - 2m\mu + \mu^2\right]
\end{aligned}
$$

Loss function II

Using moments

$$\mathrm{E}_{q_m}(m) = \mu_m \qquad \mathrm{E}_{q_m}(m^2) = \mu_m^2 + \sigma_m^2,$$
$$\mathrm{E}_{q_s}(\frac{1}{s}) = \frac{\alpha_s}{\beta_s}, \qquad \mathrm{E}_{q_s}(\log s) = \log\beta - \psi(\alpha)$$

we obtain all components of KL:

$$\begin{aligned}
\mathrm{E}_{q_m q_s}(\log p(x, m, s)) =& -(\alpha_0 + 2)(\log\beta_s - \psi(\alpha_s)) - \frac{\beta_0\alpha_s}{\beta_s} \\
& - \frac{\alpha_s}{\beta_s}\left[\mu_m^2 + \sigma_m^2 - \mu_m(y_1 + y_2) + \frac{y_1^2 + y_2^2}{2}\right] \\
& - \frac{1}{2\phi}\left[\mu_m^2 + \sigma_m^2 - 2\mu_m\mu + \mu^2\right] \\
H(q(m|x)) =& \frac{1}{2}\log(2\pi e\sigma_m^2) \\
H(q(s|x)) =& \alpha + \log(\beta\Gamma(\alpha)) - (\alpha + 1)\psi(\alpha)
\end{aligned}$$

KL and ELBO

Bayes rule for “data” x and “parameters” z:

$$p(z|x) = \frac{p(x,z)}{p(x)}, \quad p(x) = \int p(x|z)p(z)dz$$

Approximation

$$p(z|x) \approx q(z|x)$$

Divergence:

$$\begin{aligned}
\mathrm{KL}(q||p) &= \mathrm{E}_q\left(\log \frac{q(z|x)}{p(z|x)}\right) = \mathrm{E}_q\left(\log \frac{q(z|x)p(x)}{p(x,z)}\right) \\
&= \mathrm{E}_q\left(\log q(z|x) + \log p(x) - \log p(x,z)\right), \\
&= -H(q(z|x)) - \mathrm{E}_q(\log p(x,z)) + \log p(x)
\end{aligned}$$

The normalization $p(x)$ can either:

- neglected for minimization of the KL (constant).
- lower bounded (evidence lower bound):

$$\begin{aligned}
\log p(x) &= H(q(z|x)) + \mathrm{E}_q(\log p(x,z)) + \mathrm{KL}(q||p) \\
&\geq H(q(z|x)) + \mathrm{E}_q(\log p(x,z)) = \mathcal{L}(q(z|x))
\end{aligned}$$

KL minimization

Variational Bayes:

- Free-form: $q(m,s) \approx q(s)q(m)$ where forms of $q()$ are identified
- mean field Variational Bayes

ELBO:

- Direct: $q(m,s)$ is chosen by designer.
- conditionally independent $q(m)q(s)$ (same results)
- they can be **conditioned** $q(m|s)q(s)$
- often slower optimization (tuning SGD)
- allows **non-linear** transformations

Reparametrization trick

Variational Bayes requires knowledge of the moments $q_m = \mathcal{N}(\mu_m, \sigma_m)$

$$\mathrm{E}_{q_m}(m) = \mu_m \qquad\qquad \mathrm{E}_{q_m}(m^2) = \mu_m^2 + \sigma_m^2,$$

in general we may encounter moments, e.g. $\mathrm{E}(\log m^2)$.

Reparametrization trick

Variational Bayes requires knowledge of the moments $q_m = \mathcal{N}(\mu_m, \sigma_m)$

$$\mathrm{E}_{q_m}(m) = \mu_m \qquad\qquad \mathrm{E}_{q_m}(m^2) = \mu_m^2 + \sigma_m^2,$$

in general we may encounter moments, e.g. $\mathrm{E}(\log m^2)$.
Approach: Monte Carlo

$$\mathrm{E}_{q(m)}(f(m)) \approx \tfrac{1}{N} \textstyle\sum_{i=1}^N f(m^{(i)}), \;\; m^{(i)} \sim q(m)$$

Reparametrization trick

Variational Bayes requires knowledge of the moments $q_m = \mathcal{N}(\mu_m, \sigma_m)$

$$\mathrm{E}_{q_m}(m) = \mu_m \qquad\qquad \mathrm{E}_{q_m}(m^2) = \mu_m^2 + \sigma_m^2,$$

in general we may encounter moments, e.g. $\mathrm{E}(\log m^2)$.
Approach: Monte Carlo

$$\mathrm{E}_{q(m)}(f(m)) \approx \tfrac{1}{N}\textstyle\sum_{i=1}^{N} f(m^{(i)}), \quad m^{(i)} \sim q(m)$$

For $q(m) = \mathcal{N}(\mu_m, \sigma_m)$ we can approximate, $m^{(i)} = \mu_m + \sigma_m e^{(i)}$

$$\begin{aligned}\mathrm{E}_{q(m)}(f(m)) &\approx \tfrac{1}{N}\textstyle\sum_{i=1}^{N} f(m^{(i)}) = \tfrac{1}{N}\textstyle\sum_{i=1}^{N} f(\mu_m + \sigma_m e^{(i)}) \\ &= \mathrm{E}_{p(e)}(f(\mu_m + \sigma_m e))\end{aligned}$$

- Exact for large N (GD).
- **Unbiased** estimate for low N, even $N = 1$ (SGD).
- Variance reduction (decreasing learning rate, iterative averaging)

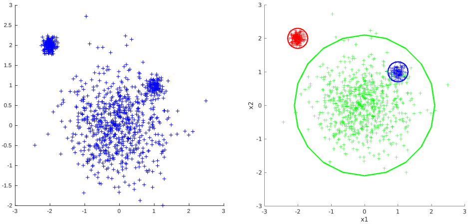
Density estimation

- We have: samples $X = \{x_1, \dots x_n\}$
- We seek: probability density $p(x)$ from which they were generated
  - ideally such that we can generate artificial samples

Density estimation

- We have: samples $X = \{x_1, \ldots x_n\}$
- We seek: probability density $p(x)$ from which they were generated
  - ideally such that we can generate artificial samples

Examples (Mixture):

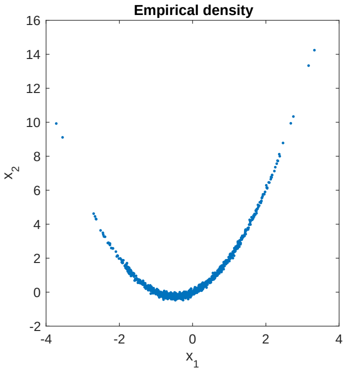


Density model is a mixture with parameters $\alpha_1 \ldots, \alpha_K$, $\mu_1, \ldots, \mu_K$ and $\Sigma_1, \ldots, \Sigma_K$.

- maximum likelihood estimate by gradient descent
- EM algorithm (coordinate descent on negative ELBO)

Generative model

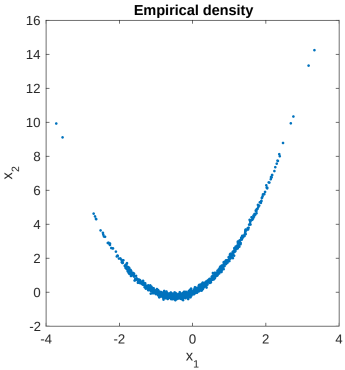
Mixture model?



Generative model

Mixture model?
True generative model:

$$z \sim \mathcal{N}(0,1),$$
$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} z \\ z + z^2 \end{bmatrix} + 0.1e$$



Generative model

Mixture model?
True generative model:

$$z \sim \mathcal{N}(0,1),$$
$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} z \\ z + z^2 \end{bmatrix} + 0.1e$$

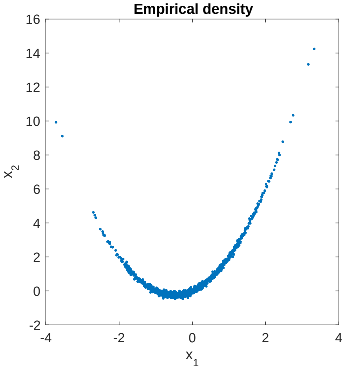
Class of models:

$$z \sim \mathcal{N}(0, I)$$
$$x \sim \mathcal{N}(f_\theta(z), \sigma I),$$

Marginal

$$p(x|\theta) = \int p(x|z,\theta)p(z)dz$$

Find θ. Maximum (marginal) likelihood.



Generative models

Flow models:

$$x = f_\theta(z), \quad p_z(z) = \mathcal{N}(0, I)$$

using transformation of coordinates

$$p(x) = p_z(f_\theta^{-1}(x)) \left| \frac{\partial f_\theta^{-1}(x)}{\partial x} \right| .$$

- only invertible transformations f
- equal dimension of z and x
- exact optimization of θ

Autoencoders:

$$x = f_\theta(z) + \sigma e, \quad p_z(z) = \mathcal{N}(0, I)$$
$$p(e) = \mathcal{N}(0, I)$$

with

$$p(x) = \int p(x|z) p(z) dz.$$

- z can have arbitrary dimension
- no need for inversion of f,
- no analytical solution of $p(x)$

Variational Autoencoder [Kingma, Welling, 2014]

Generative model

$$x = f_\theta(z) + \sigma e, \qquad p_z(z) = \mathcal{N}(0, I), \qquad p(e) = \mathcal{N}(0, I)$$

with

$$p(x) = \int p(x|z)p(z)dz \approx \mathcal{L}(q(z|x)).$$

With choices:

$$q(z|x) = \mathcal{N}(\mu_\psi(x), \mathrm{diag}(\sigma_\psi(x)^2)),$$

- $f_\theta(z)$ is a neural network
- $\mu_\psi(x), \sigma_\psi(x)$, is a neural network

Variational Autoencoder [Kingma, Welling, 2014]

Elaborate:

$$p(x|\theta,\psi) \approx \mathsf{E}_{q(z|x)}\left(\log\frac{p(z,x)}{q(z|x)}\right) = \mathsf{E}_{q(z|x)}\left(\log\frac{p(x|z)p(z)}{q(z|x)}\right)$$
$$= \mathsf{E}_{q(z|x)}\left(\log p(x|z)\right) + \mathsf{E}_{q(z|x)}\left(\log\frac{p(z)}{q(z|x)}\right)$$
$$= \mathsf{E}_{q(z|x)}\left(\log p(x|z)\right) + \mathsf{KL}\left(q(z|x)||p(z)\right)$$

Recall that $x = f_\theta(z) + \sigma e, p(x|z) = \mathcal{N}(f_\theta(z), I)$

$$\mathsf{E}_{q(z|x)}\left(\log p(x|z)\right) = -(x - f(z))^2/\sigma^2$$
$$\mathsf{KL}\left(q(z|x)||p(z)\right) = \sum(-2\log(\sigma_\psi(x)) + \sigma_\psi(x)^2 + \mu_\psi(x)^2)$$

yielding optimization (with reparametrization trick)

$$\theta^*, \psi^* = \arg\min_{\theta,\psi}\sum_{i\in\mathcal{I}}\left((x_i - f_\theta(\mu_\psi(x_i) + \sigma_\psi(x_i)e_i)^2\right) + \mathsf{KL}$$

which is known as autoencoder structure in NN. (for $\sigma_\psi = 0$, KL=0)

Linear model

We have analyzed such model

$$x = Az + e,$$

with analytical solution $p(x) = \mathcal{N}(0, AA^T + \sigma I)$. We can also compute

$$p(z|x) =$$

Linear model

We have analyzed such model

$$x = Az + e,$$

with analytical solution $p(x) = \mathcal{N}(0, AA^T + \sigma I)$. We can also compute

$$p(z|x) = \mathcal{N}((A^T A)^{-1} A^T x, (A^T A)^{-1})$$

In ELBO, need to choose approximation:

$$q(z|x) = \mathcal{N}(Bx, \operatorname{diag}(\beta^2)).$$

Then VAE is:

$$A^*, B^*, \beta^* = \arg\min_{A,B,\beta} \sum_{i \in \mathcal{I}} \Big[\sigma^{-1} \left|\left| x_i - A(Bx_i + \beta \circ e_i) \right|\right|^2 \\ + \textstyle\sum_j \left(-2\log(\beta_j) + \beta_j^2 + (\underline{B}_j x_i)^2 \right) \Big]$$

Explicit density of VAE (reduced dimensions)

Approximation of:

$$p(x) = \int p(x|z)p(z)dz \approx \mathcal{L}(q(z|x)).$$

- Simple approximation:

$$p(x|z = g(x)) = \int p(x|z)\delta(z - g(x))dz$$

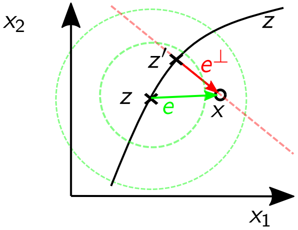
- No noise

$$p(x) \approx p_z(g(x)) \left| \frac{\partial g(x)}{\partial x} \right|$$

- Orthogonal approximation

$$x = f(z') + e'$$

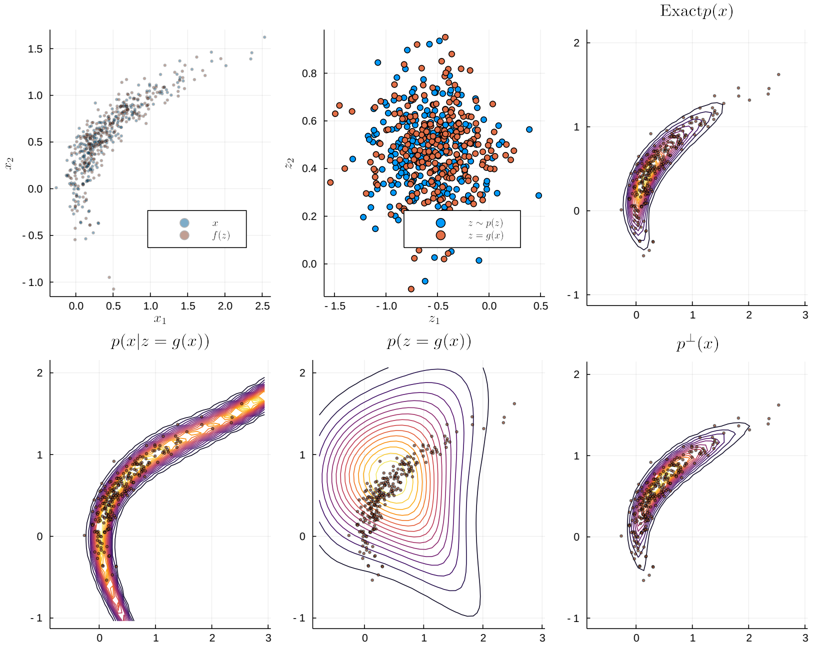
where $f(z')$ and e' are orthogonal.



Then

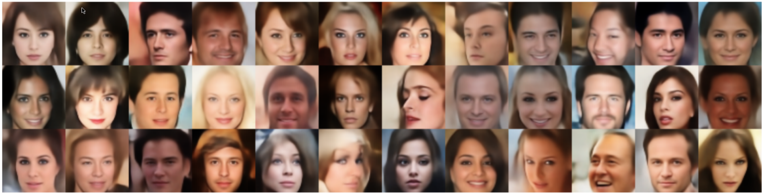
$$p(x) \approx \mathcal{N}(z'|0, \mathbf{I}) \left| \frac{\partial f^{-1}(z')}{\partial z'} \right| \mathcal{N}(x - f(z'), \sigma^2 \mathbf{I}).$$

Results



Variational Autoencoder

- Scales well with dimension (unlike GMM)
- Many extensions
  - two stage vae [Dai&Wipf,2019]: $x = f_z(z), \; z = f_w(w)$
  - First stage ($x = f_z(z)$) only reduces dimension $\dim(z) < \dim(x)$
  - Second stage ($z = f_w(w)$) models distribution $p(w) = \mathcal{N}(0, I)$
- Allows to generate artificial samples of complex distributions



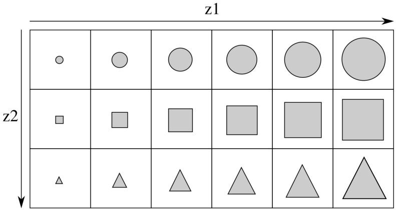
- Competes with GANs (without probabilistic density).

Disentanglement

Original problem of linear source separation is rotation

$$X = AZ = (AT)(T^{-1}Z),$$

Sometimes we want the sources to have a meaning (Bart-Lisa).



Typically we assume that we have a partial information (observation) u such that

$$u = h(z)$$

and we want to learn the state variables that correspond to the meaning.

[Mita, Filippone, Michi, 2020] Learning Optimal Conditional Priors For Disentangled Representations.