Training LDA with Variational EM Algorithm

Hieu Quoc Nguyen

01/28/2021

Contents

References: 2

Overview 3

LDA Model Details 3

E-step: θ and Z 5

M-step: Φ 7

Summary 7

Construct EM Algorithm 8

Prediction 9

Extension 9

Appendix 10

References:

This note is written based primarily on the following sources:

0. D. Blei, Andrew Ng, Jordan M. Latent Dirichlet Allocation. Journal of Machine Learning Research. 2003.
1. Reed C. Latent Dirichlet Allocation: Towards a Deeper Understanding. 2012.
2. Coursera: Bayesian Methods in Machine Learning. National Research University - Higher School of Economics. 2020.
3. Chenouri S. Stat440/840: Computational Inference. University of Waterloo. 2019.
4. Lysy M. Stat946: Advanced Computational Inference. University of Waterloo. 2019.
5. Struthers C. Stat450/850: Estimation and Hypothesis Testing. University of Waterloo. 2018.
6. Bishop C. Pattern Recognition and Machine Learning. Information Science and Statistics. 2006.
7. Wolf W. Deriving Expectation-Maximization. Blog. 2018.
8. Serrano L. Latent Dirichlet Distribution. Youtube. 2020.

Overview

LDA is one of the most important topic models in practice. On a high level, it provides a generative model that describes how the documents in a dataset are generated; i.e. how words are sampled from multiple topics to construct a document. This generative process follows a `bag of words (BOW)` assumption. Hence, the order in which the word occurs is not taken into account.

The core of topic modeling is to analyze unlabeled text data, discover the unknown number of topics and topic distribution in a unsupervised way. This is achieved by making use of statistical inference.

Some definitions before we move on:

- A dataset contains a set of D documents
- `Document`: a probability distribution over the topics
- `Topic`: a probability distribution over the words

The purpose of this note is to *open the box* to explore the mathematical details that enable LDA's effective statistical inference as well as optimization. The most 2 popular approaches to go about model parameter estimation in LDA are `Gibb Sampling` and `Variational E-M` algorithm. This note focuses on deriving model parameter estimation using `Variational E-M` algorithm.

LDA Model Details

The **primary statistical optimization problem** for LDA is to use E-M algorithm to compute the values of model parameter α, Φ such that the likelihood of the observed data $p(\mathbf{W} \mid \alpha, \Phi)$ is maximized. Unfortunately, the exact inference is NP hard. Hence, alternatively we will make use of **Jensen's inequality** to derive the variational lower bound $\mathcal{L}(q, \alpha, \Phi)$, then variational EM algorithm will be used to optimize it.

Let's derive the lower bound of the likelihood function function:

$$
\begin{aligned}
log(p(\mathbf{W} \mid \alpha, \Phi)) &= log \sum_{\mathbf{Z},\theta} p(\mathbf{W}, Z, \theta \mid \alpha, \Phi) \quad \text{(Marginalize over latent variables)} \\
&= log \sum_{\mathbf{Z},\theta} \frac{q(Z)q(\theta)}{q(Z)q(\theta)} p(W, Z, \theta \mid \alpha, \Phi) \quad \text{(multiply by 1)} \\
&= log \sum_{\mathbf{Z},\theta} q(Z)q(\theta) \frac{p(W, Z, \theta \mid \Phi)}{q(Z)q(\theta)} \\
&= log\, \mathbb{E}_{q(\mathbf{Z})q(\theta)} [\frac{p(W, Z, \theta \mid \alpha, \Phi)}{q(Z)q(\theta)}] \\
&\geq \mathcal{L}(q, \alpha, \Phi) \quad \text{(Jensen's inequality)} \\
&= \mathbb{E}_{q(\theta)q(\mathbf{Z})} log[\frac{p(W, Z, \theta \mid \alpha, \Phi)}{q(Z)q(\theta)}] \\
\Rightarrow log(p(\mathbf{W} \mid \alpha, \Phi)) &= \mathbb{E}_{q(\theta)q(\mathbf{Z})} log[\frac{p(W, Z, \theta \mid \alpha, \Phi)}{q(Z)q(\theta)}] + C
\end{aligned}
$$

There is a very neat proof for the following:

$$C = log(p(\mathbf{W} \,|\, \alpha, \Phi)) - \mathbb{E}_{q(\theta)q(\mathbf{Z})} log[\frac{p(\mathbf{W}, \mathbf{Z}, \theta \,|\, \alpha, \Phi)}{q(\mathbf{Z})q(\theta)}]$$

$$= \mathcal{KL}(q(\theta)q(\mathbf{Z}) \,||\, p(\theta, \mathbf{Z}|\mathbf{W}))$$

Hence , we now have:

$$log(p(\mathbf{W} \,|\, \alpha, \Phi)) = \mathbb{E}_{q(\theta)q(\mathbf{Z})} log[\frac{p(\mathbf{W}, \mathbf{Z}, \theta \,|\, \alpha, \Phi)}{q(\mathbf{Z})q(\theta)}] + \mathcal{KL}(q(\theta)q(\mathbf{Z}) \,||\, p(\theta, \mathbf{Z}|\mathbf{W}))$$

$$= \mathcal{L}(q, \alpha, \Phi) + \mathcal{KL}(q(\theta)q(\mathbf{Z}) \,||\, p(\theta, \mathbf{Z}|\mathbf{W}))$$

Where:

- $q(\mathbf{Z})$ and $q(\theta)$ are some known lower bound distributions over the latent variables, i.e. **variational Gaussian distribution**

We can see that maximizing $log(p(\mathbf{W} \,|\, \alpha, \Phi))$ is equivalent of minimizing $\mathcal{KL}(q(\theta)q(z) \,||\, p(\theta, z|w))$

Next, let's consider the posterior distribution of the latent variables $\mathbf{Z}, \theta$ **given** the observed data $\mathbf{W}$ and the model parameter α, Φ. Note that this is the **core statistical inference** problem in LDA.

$$p(\theta, \mathbf{Z} \,|\, \mathbf{W}; \alpha, \Phi) \propto p(\theta, \mathbf{Z}, \mathbf{W} \,|\, \alpha, \Phi) = \prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{n=1}^{N_d} \prod_{t=1}^{T} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn}; \phi_{tw}) \qquad \mathbf{(0)}$$

Where:

- $\mathbf{W}$: the data, a set of documents, each of which has N_d words
- $\mathbf{Z}$: latent variable which represents the topic of each word
- $\mathbf{\Phi}$: a T x N_d word probability matrix for each topic (row) and each word (column)
- θ: latent variable which represents the distribution over the topics for each document
- θ_d: the probability distribution over topics for document d
- w_{dn}: the nth word in document d; $w_{dn} \in \{1, ..., N_d\}$
- z_{dn}: topic of the nth word in document d; $z_{dn} \in \{1, ..., T\}$
- $p(\theta_d) \sim Dir(\theta_d|\alpha) = \frac{1}{C(\alpha)} \prod_{t=1}^{T} \theta_{dt}^{\alpha_t - 1}$
- $p(z_{dn}|\theta_d) = \theta_{dz_{dn}}$
- $p(w_{dn}|z_{dn}) = \Phi_{z_{dn}w_{dn}}$:
  Where: $\phi_{tw} \geq 0$ and $\sum\limits_{w} \phi_{tw} = 1$

Note that the latent variables Z, θ are not observable and being uncovering in the documents of texts via LDA model.

The further derivation of log likelihood function of the mixtures:

$$p(\mathbf{W}, \theta, \mathbf{Z} \,|\, \alpha, \Phi) = \prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{n=1}^{N_d} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn};\phi)$$

$$\iff p(\mathbf{W}, \theta, \mathbf{Z} \,|\, \alpha, \Phi) = \prod_{d=1}^{D} \{\frac{1}{C(\alpha)} \prod_{t=1}^{T} \theta_{dt}^{\alpha_t - 1}\} \prod_{n=1}^{N_d} \{\prod_{t=1}^{T} [z_{dn} = t] \theta_{dt} \phi_{tw_{dn}}\} \quad \text{(Substitution)}$$

$$\Rightarrow \mathit{log}(p(\mathbf{W}, \theta, \mathbf{Z} \,|\, \alpha, \Phi)) = \sum_{d=1}^{D} [\sum_{t=1}^{T} (\alpha_t - 1) \mathit{log}(\theta_{dt}) + \sum_{n=1}^{N_d} \sum_{t=1}^{T} [z_{dn} = t](\mathit{log}\theta_{dt} + \mathit{log}\phi_{tw_{dn}})] + \mathit{const} \quad \text{(Take log both side)}$$

E-step: θ and Z

Objective: $\mathcal{KL}(q(\mathbf{Z})q(\theta) \,||\, p(\theta, \mathbf{Z}|\mathbf{W})) \to \min_{q(\theta)q(\mathbf{Z})}$

Find $q(\theta)$

Consider the joint log likelihood function:

$$\mathit{log}\, p(\mathbf{W}, \theta, \mathbf{Z} \,|\, \alpha, \Phi) = \sum_{d=1}^{D} [\sum_{t=1}^{T} (\alpha_t - 1) \mathit{log}(\theta_{dt}) + \sum_{n=1}^{N_d} \sum_{t=1}^{T} [z_{dn} = t](\mathit{log}\theta_{dt} + \mathit{log}\phi_{tw_{dn}})] + \mathit{const} \qquad \mathbf{(1)}$$

Note that, w.r.t $q(\theta)$, $\sum_{n=1}^{N_d} \sum_{t=1}^{T} [z_{dn} = t] \mathit{log}\phi_{tw_{dn}}$ in **(1)** is a constant so we can ignore it the below derivation of $\mathit{log}[q(\theta)]$.

$$\mathit{log}[q(\theta)] = \mathbb{E}_{q(\mathbf{Z})} \mathit{log}\, [p(Z, \theta, \mathbf{W})] + C_1$$

$$= \mathbb{E}_{q(\mathbf{Z})} \sum_{d=1}^{D} [\sum_{t=1}^{T} (\alpha_t - 1) \mathit{log}(\theta_{dt}) + \sum_{n=1}^{N_d} \sum_{t=1}^{T} [z_{dn} = t] \mathit{log}\theta_{dt}] + C_1$$

$$= \sum_{d=1}^{D} [\sum_{t=1}^{T} (\alpha_t - 1) \mathit{log}(\theta_{dt}) + \sum_{n=1}^{N_d} \sum_{t=1}^{T} \mathbb{E}_{q(\mathbf{z_{dn}})} \{[z_{dn} = t]\} \mathit{log}\theta_{dt}] + C_1$$

Let $\gamma_{dn} = \mathbb{E}_{q(\mathbf{z_{dn}})} \{[z_{dn} = t]\} \Rightarrow q(z_{dn} = t) = \gamma_{dn}^{t}$

$$= \sum_{d=1}^{D} \sum_{t=1}^{T} [(\alpha_t - 1) + \sum_{n=1}^{N_d} \gamma_{dn}]\, \mathit{log}\theta_{dt} + C_1$$

Now take the *exp* both side to get $q(\theta)$:

$$q(\theta) \propto \prod_{d=1}^{D}\prod_{t=1}^{T}\theta_{dt}^{[(\alpha_t-1)+\sum_{n=1}^{N_d}\gamma_{dn}]}$$

$$= \prod_{d=1}^{D}\{\prod_{t=1}^{T}\theta_{dt}^{[(\alpha_t+\sum_{n=1}^{N_d}\gamma_{dn})-1]}\}$$

$$\Rightarrow q(\theta) = \prod_{d=1}^{D} q(\theta_d), \text{where } q(\theta_d) \sim \mathbf{Dir}(\theta_d|\alpha_t + \sum_{n=1}^{N_d}\gamma_{dn}) \qquad \mathbf{(2)}$$

Note that the update rules for $q(\theta)$ depends on γ_{dn} which depends on z_{dn}. So to compete the E-step we will also need to find updating rule for $q(\mathbf{Z})$

Find $q(\mathbf{Z})$

Similarly, W.r.t $q(\mathbf{Z})$ $\sum_{t=1}^{T}(\alpha_t-1)log(\theta_{dt})$ in **(1)** is a constant so we can ignore it the below derivation.

$$log[q(\mathbf{Z})] = \mathbb{E}_{q(\theta)}log\,[p(Z,\theta,\mathbf{W})] + C_2$$

$$= \mathbb{E}_{q(\theta)}\sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{t=1}^{T}[z_{dn}=t](log\theta_{dt}+log\phi_{tw_{dn}}) + C_2 \quad \text{(Substitution from (1))}$$

$$= \sum_{d=1}^{D}\sum_{n=1}^{N_d}\{\sum_{t=1}^{T}[z_{dn}=t](\mathbb{E}_{q(\theta)}[log\theta_{dt}]+log\phi_{tw_{dn}})\} + C_2; \quad (z_{dn} \text{ sums up to 1 over T and } \{*\} \text{ is dist over } z_{dn})$$

Take *exp* both sides we have the product over independent distribution as following:

$$q(\mathbf{Z}) \propto \prod_{d=1}^{D}\prod_{n=1}^{N_d} q(z_{dn})$$

Where:

$$q(z_{dn}=t) \propto exp\{\mathbb{E}_{q(\theta)}[log\theta_{dt}] + log\phi_{tw_{dn}}\}$$

$$= \frac{\phi_{tw_{dn}}exp(\mathbb{E}_{q(\theta_{dt})}[log\theta_{dt}])}{\text{sum over all possible values of t}}$$

$$= \frac{\phi_{tw_{dn}}exp(\mathbb{E}_{q(\theta_{dt})}[log\theta_{dt}])}{\sum_{t'=1}^{T}\phi_{t'w_{dn}}exp(\mathbb{E}_{q(\theta_{dt'})}[log\theta_{dt'}])}$$

$$= \gamma_{dn}^{t}, \quad (\text{since } \gamma_{dn} = \mathbb{E}_{q(\mathbf{z_{dn}})}\{[z_{dn}=t]\} \Rightarrow q(z_{dn}=t) = \gamma_{dn}^{t}) \qquad \mathbf{(3)}$$

Now that we know the update procedure for $q(\mathbf{Z})$, from here we also know the value of γ_{dn}, we can iterate through a loop to update the values for $\theta, \mathbf{Z}$.

M-step: Φ

Objective: $\mathbb{E}_{q(\theta)q(z)}[log(p(W,Z,\theta))] \rightarrow \max\limits_{q(\phi)}$ such that:

- since ϕ_{tw} is a probability: $\phi_{tw} \geq 0\,, \forall t, w$. This is already been satisfied because it is under the log function as we see above.
- The probability distribution should sum up to 1 over vocal size V: $\sum_{w=1}^{V} \phi_{tw} = 1\,, \forall t$

As we already know, optimizing an objective function given a set of constraints, **lagrange multiplier** is the way to go.

$$L = \mathbb{E}_{q(\theta)q(z)}[log(p(W,Z,\theta))] + \sum_{t=1}^{T} \lambda_t (\sum_{w} \phi_{tw} - 1)$$

$$= \mathbb{E}_{q(\theta)q(z)} \sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{t=1}^{T} [z_{dn} = t] log \phi_{tw_{dn}} + \sum_{t=1}^{T} \lambda_t (\sum_{w} \phi_{tw} - 1) \quad \text{(Substitution)}$$

$$= \sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{t=1}^{T} (\mathbb{E}_{q(\theta)}[z_{dn} = t]) log \phi_{tw_{dn}} + \sum_{t=1}^{T} \lambda_t (\sum_{w} \phi_{tw} - 1)$$

$$= \sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{t=1}^{T} \gamma_{dn}^{t} log \phi_{tw_{dn}} + \sum_{t=1}^{T} \lambda_t (\sum_{w} \phi_{tw} - 1)$$

Taking derivative w.r.t ϕ_{tw} then set it to 0 to solve for ϕ_{tw}, we get:

$$\begin{aligned} \phi_{tw} &= \frac{\sum_{d,n} \gamma_{dn}^{t} [w_{dn} = w]}{\text{sum over all possible words in our data } \mathbf{W}} \\ &= \frac{\sum_{d,n} \gamma_{dn}^{t} [w_{dn} = w]}{\sum_{w',d,n} \gamma_{dn}^{t} [w_{dn} = w']} \end{aligned} \qquad \textbf{(4)}$$

Summary

- E-Step: keep Φ fixed, we iteratively update θ and $\mathbf{Z}$ until converges using $\mathcal{KL}(q(\mathbf{Z})q(\theta)\,||\,p(\theta, \mathbf{Z}|\mathbf{W}))$
- M-step: keep θ and $\mathbf{Z}$ fixed, we iteratively update Φ by maximizing $\mathbb{E}_{q(\theta)q(z)}[log(p(W,Z,\theta))]$

Construct EM Algorithm

Input

- T: number of topics
- D: number of documents with corpus size N_d
- V: the size of the overall text corpus

Output

- Model params: $\theta, \mathbf{Z}, \mathbf{\Phi}$

E-step:

Init:

- $N_d \sim Poi(\epsilon)$
- ϕ_{tw}^0 := 1/T, $\forall t \in T, w \in N_d$
- $\alpha_t^0 + \sum_{n=1}^{N_d} \gamma_{dn}^0 := C/T + N_d/K$, $\forall t \in T$, where C is a constant. Choose $C = 50$
- $\gamma_{tw} := 0$, $\forall t \in T, w \in V$
- *loglikelihood* $l_0 := 0$
- Loop starts at index $i := 0$

Updating Loop

- for d = 1 to D:
  - for n = 1 to N_d:
    - for t = 1 to T:

$$q(\theta^{i+1}) = \prod_{d=1}^{D} q(\theta_d), \text{where } q(\theta_d) \sim \mathbf{Dir}(\theta_d|\alpha_t^i + \sum_{n=1}^{N_d} \gamma_{dn}^i)$$

$$q(z_{dn}^{i+1} = t) = \gamma_{dn}^{t;i+1} = \frac{\phi_{tw_{dn}} exp(\mathbb{E}_{q(\theta_{dt}^i)}[log\theta_{dt}^i])}{\sum_{t'=1}^{T} \phi_{t'w_{dn}} exp(\mathbb{E}_{q(\theta_{dt'}^i)}[log\theta_{dt'}^i])}$$

until $\theta, \mathbf{Z}$ converges to $\theta^*, \mathbf{Z}^*$

M-step:

- for d = 1 to D:
  - for t = 1 to T:
    - for w = 1 to V:

$$\phi_{tw}^{i+1} = \frac{\sum_{d,n} \gamma_{dn}^{t;i}[w_{dn} = w]}{\sum_{w',d,n} \gamma_{dn}^{t;i}[w_{dn} = w']}$$

Update *loglikelihood* $l_{new} := l_{old} + \mathbb{E}_{q(\theta^*)q(\mathbf{Z}^*)} log[p(W, \mathbf{Z}^*, \theta^* \mid \phi_{tw}^*)]$ Until ϕ_{tw} converges to ϕ_{tw}^*

if *loglikelihood* l_{new} converges: return $\theta^*, \mathbf{Z}^*, \phi^*$

else: back to `E-step`

Prediction

Objective: $\mathcal{KL}(q(\theta_{d^*})q(\mathbf{Z}_{d^*}) \,||\, p(\theta_{d^*}, \mathbf{Z}_{d^*}|W; \alpha, \Phi)) \rightarrow \min_{q(\theta_{d^*})q(z_{d^*})}$

For the new document, using the model parameters α, Φ that we found through E-M we wanna predict:

- the new value for $\mathbf{Z}$, i.e. assign a topic for each words
- the new value for θ, i.e. the global topic distribution of the document

Extension

1. Treat Φ as a random variable following $DIR(\beta)$

This time we consider Φ as a random variable which follows $DIR(\beta)$. Now the joing probability distribution becomes:

$$p(\theta, \mathbf{Z}, \mathbf{W} \mid \alpha, \Phi) = \prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{n=1}^{N_d} \prod_{t=1}^{T} p(z_{dn}|\theta_d) p(\phi_{tw}|\beta) p(w_{dn}|z_{dn}, \phi_{tw})$$

2. Topic Correlation using Logistic Normal Distribution

https://www.ic.unicamp.br/~tachard/docs/corrlda.pdf

3. Dynamic Topic Modeling

https://dl.acm.org/doi/10.1145/1143844.1143859

Appendix

I. Bayes' Theorem

Let θ be model's parameter and X be the observed data, the posterior probability distribution of θ **given** the observed data X is defined as follow:

$$P(\theta \,|\, X) = \frac{\text{joint probability distribution}}{\text{evidence}}$$

$$= \frac{\text{(likelihood)(prior)}}{\text{evidence}}$$

$$= \frac{P(X, \theta)}{P(X)}$$

$$= \frac{P(X|\theta)P(\theta)}{P(X)}$$

General `Chain Rule`:

$$P(X, Y) = P(X|Y)P(Y) = P(Y|X)P(X)$$

$$P(X, Y, Z) = P(X|Y, Z)P(Y|Z)P(Z)$$

$$P(X_1, ..., X_N) = \prod_{i=1}^{N} P(X_i|X1, ..., X_{i-1})$$

II. General EM Algorithm

Generally speaking, E-M algorithm is used to solve parameter estimation problems when some data is missing/not observed (`latent variable`). This is the main distinction between EM vs traditional MLE approach. This includes approximating the maximum of a likelihood function (`MLE`) or the maximum of a posterior (`MAP`).

Consider:

- $X = \{x_1, ..., x_n\}$
- $Z|X \sim p_\theta$, $\theta \in \Theta$
- p_θ belongs to the exponential family (i.e. Gaussian, Beta, Dirichlet, Exponential, Gamma, etc.), which is the **primary** assumption for EM algo to work well (idk why this is so).
- Main objective: $\hat{\theta}_{EM} \in \, argmax_\theta \, p_\theta(X)$

Iterative algo:

- the loop: for i in range(0,I)
- init $\theta_0 \in \Theta$

- `E-step`: $Q(\theta, \theta_i) = E_{\theta_i}[log\, p_\theta(X, Z|\theta)]$
- `M-step`: $\theta_{i+1} \in argmax_\theta Q(\theta, \theta_i)$

Pros:

- likelihood function is guaranteed to increase from one iteration to another

Cons:

- Not guarantee to reach global maximum
- Computational expensive so convergence rate is slow
- works well only when p_θ belong to exponential family

III. Jensen's Inequality

If $f(x)$ is concave (i.e. $f(\alpha x + (1-\alpha)y) \geq \alpha f(x) + (1-\alpha)f(y)$), then for any random variable X with corresponding probability density function (pdf) $f(X)$:

$$f(\mathbb{E}[X]) \geq \mathbb{E}[f(X)]$$

IV. Kullback Leibler Divergence ($\mathcal{KL}$ Divergence)

KL divergence is used to measure the difference between 2 probability distributions. Note that this is not computed based on parameter wise difference.

Example:

- $\mathcal{KL}[N(0,1)\,||\,N(1,1)] = 0.5$ while their parameter wise difference is 1
- $\mathcal{KL}[N(0,100)\,||\,N(1,100)] = 0.005$ while their parameter wise difference is 1

V. Dirichlet Distribution

$$\theta \sim \mathbf{Dir}(\theta \,|\, \alpha)$$

$$p(\theta \,|\, \alpha) = \frac{1}{C(\alpha)} \prod_{t=1}^{T} \theta_t^{\alpha_t - 1}$$

Where:

- $\theta_t \geq 0$ and $\sum_T \theta_t = 1$
- Model param $\alpha_t > 0$

Statistics:

- Let $\alpha_0 = \sum_T \alpha_t$ where $t \in \{1, 2, ..., T\}$
- $\mathbb{E}[\theta_t] = \frac{\alpha_t}{\alpha_0}$
- $Cov(\theta_t, \theta_j) = \frac{\alpha_t \alpha_0 [i=j] - \alpha_t \alpha_j}{\alpha_0^2(\alpha_0 + 1)}$

VI. Variational Inference

It is used to approximate the posterior distribution:

1. Select a famlily of distributions Q: $Q = \{q \,|\, q(Z) = \prod_{i=1}^{d} q_i(Z_i)\}$; i.e:

$$Q \sim N(\mu, \begin{pmatrix} \sigma_1^2 & 0 & 0 & \dots & 0 & 0 \\ 0 & \sigma_2^2 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \dots & \sigma_{d-1}^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_d^2 \end{pmatrix}$$

2. Next we will try to approximate the full posterior $p(\mathbf{Z})$ with some variational distribution $q(\mathbf{Z})$:

$$\mathcal{KL}[q(Z) \,||\, p(Z)] \to min_{q \in Q}$$

Then apply `Coordinate descend`:

$$\mathcal{KL}[q(Z) \,||\, p(Z)] \to min_{q_1}$$

$$\mathcal{KL}[q(Z) \,||\, p(Z)] \to min_{q_2}$$

$$\dots$$

$$\mathcal{KL}[q(Z) \,||\, p(Z)] \to min_{q_d}$$

Example:

$$p(Z_1, Z_2) \approx q(Z_1)q(Z_2)$$

$$p(Z_1, Z_2) \sim N(0, \Sigma)$$

$$q(Z_1)q(Z_2) \sim N(0, \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix})$$