L3: Variational inference

Caterina De Bacco

May 2, 2025

1. Introduction

In the previous lecture we learned about the standard stochastic block model and its main assumptions. We saw how it can be used to learn hidden community memberships using maximum likelihood. This gives point estimates, i.e. one particular partition. Here we investigate a method to learn the whole posterior distribution, to capture uncertainty about the learned estimates.

2. Variational Inference: the idea

Variational Inference (VI) is an inference approach that approximates probability distributions through optimization.

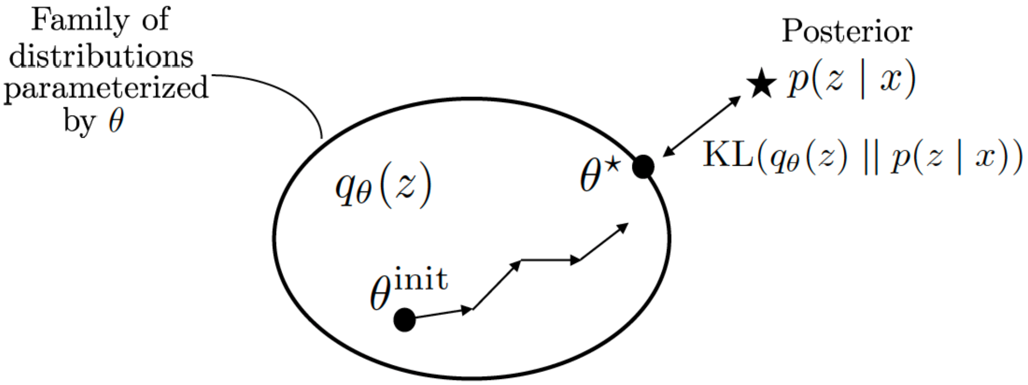
Family of distributions parameterized by θ
Posterior
$\star\ p(z \mid x)$
$\mathrm{KL}(q_\theta(z) \,||\, p(z \mid x))$
$q_\theta(z)$
$\theta^\star$
θ^{init}

FIGURE 1. The idea behind Variational Inference. Figure courtesy of Francisco J. R. Ruiz.

In the field of Bayesian statistics, one is interested in deriving the posterior distribution of model's parameters. Often though, the posterior is not analytically tractable, although the likelihood might be (recall previous lectures).

Consider a joint density of latent variables $\mathbf{z} = (z_1, \dots, z_m)$ and data $\mathbf{x} = (x_1, \dots, x_n)$:

$$p(\mathbf{z}, \mathbf{x}) = p(\mathbf{x}|\mathbf{z})\, p(\mathbf{z}) \quad . \tag{1}$$

Using a model with latent variables is a common approach in many inference models, as one can use these variables to govern the data distribution. Inference in Bayesian modeling consists in posing a prior for these variables $p(\mathbf{z})$ and extracting the posterior $p(\mathbf{z}|\mathbf{x})$, given a likelihood $p(\mathbf{x}|\mathbf{z})$.

The idea behind Variational Inference, is to posit a family of *tractable* distributions $\mathcal{D}$ over the latent variables and find one element of this set that is closest to the untractable posterior. Closeness is measured by Kullback-Leibler (KL) divergence:

$$q^*(\mathbf{z}) = \underset{q_\theta(\mathbf{z})\in\mathcal{D}}{\arg\min}\ KL\left(q_\theta(\mathbf{z})||p(\mathbf{z}|\mathbf{x})\right) \quad . \tag{2}$$

The distribution $q_\theta(\mathbf{z})$ is called *variational distribution*. Fig. 1 gives a sketch of this idea.

REMARK 1. *Other traditional models for inference that go under the Monte Carlo family are based on sampling, rather than optimization.*

REMARK 2. *Other types of cost functions to be optimized could also be considered. Here we focus only on $KL(q||p)$.*

3. The problem

Objective: estimate the posterior $p(\mathbf{z}|\mathbf{x})$ *given* the data.

We have that:

$$p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{x},\mathbf{z})}{p(\mathbf{x})} \quad , \tag{3}$$

where $p(\mathbf{x}) = \int p(\mathbf{x},\mathbf{z})d\mathbf{z}$ is called the *evidence*. This integral is often unavailable in closed form, but this is needed in order to calculate the posterior. This is why inference in these cases is hard.

3.1. Example: Bayesian Mixture of Gaussians (GMM).

Consider a mixture of unit-variance univariate Gaussians. The data $X = \{x_i\}_{i=1}^N$ is a set of one-dimensional variables $x_i \in \mathbb{R}$. The model works as follow: we assume that each data point x_i is extracted from one of the Gaussians, but we do not know which one. We thus need to introduce a cluster assignment latent variable c_i that tells which Gaussian x_i was drawn from. The $c_i = \{1,\dots,K\}$ is assumed to be categorical, but is encoded in an indicator K-vector, with all zeros except one equal to 1 entry corresponding to the cluster. There are K mixture components, one for each Gaussian, with mean $\mu = (\mu_1,\dots,\mu_K)$. In fig. 2 we plot an example of such dataset.

The means in turn are also random variables, drawn from a prior $p(\mu_k)$ which we assume to be Gaussian $\mathcal{N}(0,\sigma^2)$; σ^2 is an hyper-parameter.

Formally, the model is:

$$\mu_k \sim \mathcal{N}(0,\sigma^2) \qquad k = 1,\dots,K \tag{4}$$

$$c_i \sim \text{Categorical}\left(1/K,\dots,1/K\right) \qquad i = 1,\dots,n \tag{5}$$

$$x_i|c_i,\mu \sim \mathcal{N}(c_i^T\mu,1) \qquad i = 1,\dots,n \quad . \tag{6}$$

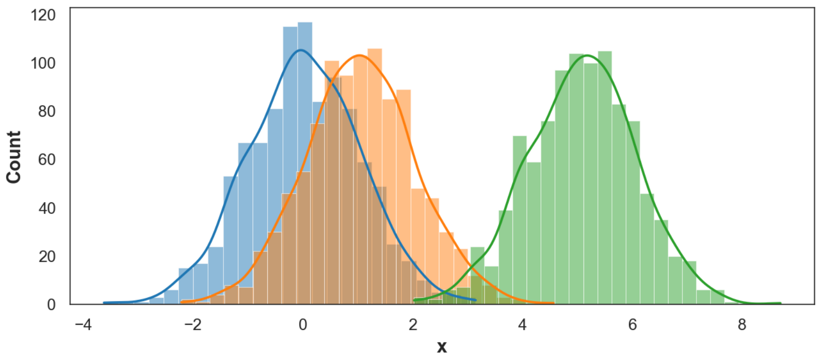

FIGURE 2. Example of a GMM dataset with 3 clusters and their Gaussians components. Here the means of the Gaussians are $\mu_1 = 0\,,\mu_2 = 1\,,\mu_3 = 5$.

REMARK 3. *Notice that each variable is drawn from one single Gaussian, and not from a mixture of Gaussians! The whole set of variables is a mixture of Gaussian.*

The joint density of data and parameters is then:

$$p(\mu, \mathbf{c}, \mathbf{x}) = p(\mu) \prod_{i=1}^{n} p(c_i)\; p(x_i|c_i, \mu) \quad . \tag{7}$$

The latent variables in this case are $\mathbf{z} = (\mu, \mathbf{c})$. The evidence is then:

$$p(\mathbf{x}) = \int p(\mu) \prod_{i=1}^{n} \sum_{c_i} p(c_i)\; p(x_i|c_i, \mu)\, d\mu \quad . \tag{8}$$

Question: Can you calculate that integral?

Unfortunately no. This is because the integrand does contain a separate factor for each μ_k, thus preventing the reduction onto one-dimensional integrals. The sum over the cluster assignments runs over K^n configurations, i.e. exponential in K, which is only doable for very small values of n.

4. The Evidence Lower Bound (ELBO).

Recall that the goal of VI is to optimize (we omit the explicit dependence on θ):

$$q^*(\mathbf{z}) = \arg\min_{q(\mathbf{z})\in\mathcal{D}} KL\left(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})\right) \quad . \tag{9}$$

However, this objective is not computable because we have to compute $\log p(\mathbf{x})$, which is not feasible as we saw before. To see this, let's unpack the KL divergence:

$$\begin{aligned} KL\left(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})\right) &= \mathbb{E}_q\left[\log q(\mathbf{z})\right] - \mathbb{E}_q\left[\log p(\mathbf{z}|\mathbf{x})\right] & (10)\\ &= \mathbb{E}_q\left[\log q(\mathbf{z})\right] - \mathbb{E}_q\left[\log p(\mathbf{z},\mathbf{x})\right] + \log p(\mathbf{x}) \quad , & (11) \end{aligned}$$

which requires calculating $\log p(\mathbf{x})$ as said before.
To skip computing the exact KL, we propose an *alternative* optimization objective.

Question: How do we choose this?

We pick an objective equivalent to the same KL as before up to a constant (in q):

$$\begin{aligned} \text{ELBO}(q) &:= \mathbb{E}_q[\log p(\mathbf{z},\mathbf{x})] - \mathbb{E}_q[\log q(\mathbf{z})] && (12) \\ &= \log p(\mathbf{x}) - KL\left(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})\right) \quad . && (13) \end{aligned}$$

REMARK 4. *Maximizing* ELBO(q) *is equivalent to minimizing* $KL\left(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})\right)$.

4.1. Properties of the ELBO

If we rewrite the ELBO unpacking the first term yields:

$$\begin{aligned} \text{ELBO}(q) &= \mathbb{E}_q[\log p(\mathbf{x}|\mathbf{z})] + \mathbb{E}_q[\log p(\mathbf{z})] - \mathbb{E}_q[\log q(\mathbf{z})] && (14) \\ &= \mathbb{E}_q[\log p(\mathbf{x}|\mathbf{z})] - KL\left(q(\mathbf{z})||p(\mathbf{z})\right) \quad . && (15) \end{aligned}$$

In other words, the ELBO is the sum of the expected log likelihood of the data and (minus) the KL divergence between the variational distribution $q(\mathbf{z})$ and the prior $p(\mathbf{z})$.

Property 1: maximizing the ELBO is then encouraging a combination of increasing the expected log likelihood, i.e. finding a $q(\mathbf{z})$ that explains the data, and choosing $q(\mathbf{z})$ close to the prior.

Because $KL\left(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})\right) \geq 0$, we have that:

$$\begin{aligned} \log p(\mathbf{x}) &= \text{ELBO}(q) + KL\left(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})\right) && (16) \\ &\geq \text{ELBO}(q) \quad . && (17) \end{aligned}$$

Property 2: the ELBO is a lower bound to the evidence (this is the reason of its name).

4.2. The Mean-Field variational family

One relevant aspect of Variational Inference, is that we can pick a variational family at our choice. Of course, this does not guarantee that we will make a good choice. We expect that the more complex the q the better the approximation, at the cost of losing the tractability.

If we want to go for making things analytically tractable, which includes for instance being able to compute $\mathbb{E}_q[\cdot]$, then the best choice is to consider a fully factorized family. This is the Mean-Field variational family:

$$q(\mathbf{z}) = \prod_{j=1}^{m} q_j(z_j) \quad . \tag{18}$$

REMARK 5. *Each latent variable z_j is governed by its own variational factor and these are all independent.*

REMARK 6. *Notice that the variational distribution $q(\mathbf{z})$ does not depend on the data* $\mathbf{x}$. *The dependence on the data is taken care of inside the* ELBO, *through the expected log likelihood term.*

REMARK 7. *The subindex j of $q_j(\cdot)$ is there so that one can specify a different type of distribution for each latent variable. For instance, for binary latent variables $q_j(\cdot)$ can be Bernoulli, for continuous numbers $q_j(\cdot)$ can be Gaussian, and so on.*

5. Coordinate Ascent Mean-Field Variational Inference (CAVI)

So far we have seen the theory being VI and a class of variational family, the Mean-Field distribution. We haven't said anything yet about how to solve the optimization problem. Here we show how to tackle this task by describing the CAVI algorithm.

Idea: iteratively optimize each single variational factor holding the others fixed, until we reach a local maximum of the ELBO.

We will use the following result. Consider z_j. The *complete conditional* $p(z_j|\mathbf{z}_{\setminus j}, \mathbf{x})$ of z_j is a conditional density given all the other latent variables and data.
Fact: the *optimal* $q_j(z_j)$ is proportional to the exponentiated expected log of the complete conditional of z_j:

$$q_j^*(z_j) \propto \exp\left\{\mathbb{E}_{\setminus j}\left[\log p(z_j|\mathbf{z}_{\setminus j}, \mathbf{x})\right]\right\} \quad , \tag{19}$$

where $\mathbb{E}_{\setminus j}[\cdot]$ is over $\prod_{l \neq j} q_l(z_l)$ (recall that $q_l(z_l)$ are currently being held fixed).
An equivalent result is:

$$q_j^*(z_j) \propto \exp\left\{\mathbb{E}_{\setminus j}\left[\log p(z_j, \mathbf{z}_{\setminus j}, \mathbf{x})\right]\right\} \quad , \tag{20}$$

where we considered instead the log of the joint distribution.
Proof. Let's rewrite the ELBO by isolating a variational factor $q_j(z_j)$ absorbing into a constant terms that do not depend on it, and use the Mean-Field approximation:

$$\begin{aligned} \text{ELBO}(q_j) &= \mathbb{E}_j\left[\mathbb{E}_{\setminus j}\left[\log p(z_j, \mathbf{z}_{\setminus j}, \mathbf{x})\right]\right] - \mathbb{E}_j\left[\log q_j(z_j)\right] + const & (21) \\ &= -KL(q_j||q_j^*) + const \quad . & (22) \end{aligned}$$

Thus we maximize the ELBO when we minimize the $KL(q_j||q_j^*)$; this is minimized when $q_j(z_j) \equiv q_j^*(z_j)$.

5.1. Example: Bayesian Mixture of Gaussians (continued).

Let's go back to the GMM and apply what just learned. The MF family for the latent parameters μ_k and c_i is:

$$q(\boldsymbol{\mu}, \mathbf{c}) = \prod_{k=1}^{K} q(\mu_k; m_k, s_k^2) \prod_{i=1}^{n} q(c_i; \rho_i) \tag{23}$$

where we choose:

$$\begin{aligned} q(\mu_k; m_k, s_k^2) &= \mathcal{N}(m_k, s_k^2) & (24) \\ q(c_i; \rho_i) &= \text{Categorical}(\rho_i) \quad , & (25) \end{aligned}$$

and ρ_i is a K-dimensional vector.

REMARK 8. *We chose Gaussian and Categorical distributions. In principle we could have chosen other types. However, the choice impacts the goodness of the approximation. For this case, Gaussian and Categorical are indeed the optimal choice for the Mean-Field family.*

5.1.1. Variational update of the cluster assignment c_i

Using equation (20) yields:

$$\begin{align}
q^*(c_i;\rho_i) &\propto \exp\left\{\mathbb{E}_{\setminus c_i}\left[\log\left(p(c_i)\prod_{l\neq i} p(c_l)\,p(\boldsymbol{\mu})\,p(\mathbf{x}|c_i,\mathbf{c}_{\setminus i},\boldsymbol{\mu})\right)\right]\right\} \tag{26}\\
&= \exp\left\{\log p(c_i) + \mathbb{E}_{\setminus c_i}\left[\sum_{l\neq i}\log p(c_l) + \log p(\boldsymbol{\mu})\right] + \mathbb{E}_{\setminus c_i}\left[\log p(\mathbf{x}|c_i,\mathbf{c}_{\setminus i},\boldsymbol{\mu})\right]\right\} \nonumber\\
&= \exp\left\{\log p(c_i) + \sum_{l\neq i}\mathbb{E}_{q(c_l)}\left[\log p(c_l)\right] + \sum_k \mathbb{E}_{q(\mu_k)}\left[\log p(\mu_k)\right] + \mathbb{E}_{\setminus c_i}\left[\sum_i \log p(x_i|c_i,\boldsymbol{\mu})\right]\right\} \tag{27}
\end{align}$$

The second and third terms are constant in c_i, so they can be neglected. The likelihood term can be further unpacked by keeping only terms containing c_i. Recall that using c_i as an indicator we have:

$$p(x_i|c_i,\boldsymbol{\mu}) = \prod_{k=1}^{K} p(x_i|\mu_k)^{c_{ik}} \quad . \tag{28}$$

Substituting into the previous formula yields:

$$\begin{align}
\mathbb{E}_{\setminus c_i}\left[\log p(x_i|c_i,\boldsymbol{\mu})\right] &= \mathbb{E}_{q(\boldsymbol{\mu})}\left[\log p(x_i|c_i,\boldsymbol{\mu})\right] = \sum_k \mathbb{E}_{q(\mu_k)}\left[c_{ik}\log p(x_i|\mu_k)\right] \tag{29}\\
&= \sum_k c_{ik}\,\mathbb{E}_{q(\mu_k)}\left[\log p(x_i|\mu_k)\right] \tag{30}\\
&= \sum_k c_{ik}\,\mathbb{E}_{q(\mu_k)}\left[-(x_i-\mu_k)^2/2\right] + const \tag{31}\\
&= -\sum_k c_{ik}x_i^2/2 - \sum_k c_{ik}\mathbb{E}_{q(\mu_k)}\left[\mu_k^2/2\right] + x_i\sum_k c_{ik}\mathbb{E}_{q(\mu_k)}\left[\mu_k\right] + const \tag{32}\\
&= x_i\sum_k c_{ik}\mathbb{E}_{q(\mu_k)}\left[\mu_k\right] - \frac{1}{2}\sum_k c_{ik}\mathbb{E}_{q(\mu_k)}\left[\mu_k^2\right] + const \tag{33}\\
&= x_i\sum_k c_{ik}m_k - \frac{1}{2}\sum_k c_{ik}\left(s_k^2 + m_k^2\right) + const \quad . \tag{34}
\end{align}$$

We can finally substitute into Eq. (27):

$$\begin{align}
q^*(c_i;\rho_i) = \prod_k \rho_{ik}^{*\,c_{ik}} &\propto \exp\left\{\log p(c_i) + x_i\sum_k c_{ik}m_k - \frac{1}{2}\sum_k c_{ik}\left(s_k^2 + m_k^2\right)\right\} \tag{35}\\
&= p(c_i)\,\exp\left\{x_i\sum_k c_{ik}m_k - \frac{1}{2}\sum_k c_{ik}\left(s_k^2 + m_k^2\right)\right\} \tag{36}\\
&\propto \prod_k \exp\left\{\left[x_i m_k - \frac{1}{2}\left(s_k^2 + m_k^2\right)\right]c_{ik}\right\} \quad . \tag{37}
\end{align}$$

Which means that the optimal parameter for the categorical variational distribution is:

$$\rho_{ik}^* \propto \exp\left[x_i m_k - \frac{1}{2}\left(s_k^2 + m_k^2\right)\right] \quad . \tag{38}$$

REMARK 9. *These are a function of both data and the variational parameters of the mixture component, but not of the other* ρ_j^*.

5.1.2. Variational update of the mixture components' means μ_k.

We can repeat similar calculations for the variational distributions $q(\mu_k; m_k, s_k^2)$. Again, use equation (20) to get:

$$q(\mu_k) \propto \exp\left\{\log p(\mu_k) + \sum_i \mathbb{E}_{\setminus \mu_k}\left[\log p(x_i|c_i, \mu)\right]\right\} \quad . \tag{39}$$

Using again Eq. (28) and keeping only terms containing μ_k yields:

$$\begin{aligned}
\log q(\mu_k) &= \log p(\mu_k) + \sum_i \mathbb{E}_{q(c_i)}\left[c_{ik}\right] \log p(x_i|\mu_k) + const & (40)\\
&= -\frac{\mu_k^2}{2\sigma^2} - \sum_i \rho_{ik} \frac{(x_i - \mu_k)^2}{2} + const & (41)\\
&= \mu_k \sum_i \rho_{ik} x_i - \frac{\mu_k^2}{2}\left(\frac{1}{\sigma^2} + \sum_i \rho_{ik}\right) + const \quad . & (42)
\end{aligned}$$

For those of you familiar with exponential families, this means that $q(\mu_k)$ is a member of it with sufficient statistics $\left\{\mu_k, \mu_k^2\right\}$ and natural parameters $\left\{\sum_i \rho_{ik} x_i, -\frac{1}{2}\left(\frac{1}{\sigma^2} + \sum_i \rho_{ik}\right)\right\}$. This means that it is a Gaussian distribution. For a Gaussian of mean μ and variance σ^2, the natural parameters are generally $\boldsymbol{\eta} \equiv \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)$. Doing the mapping, we obtain:

$$\begin{aligned}
m_k &= \frac{\sum_i \rho_{ik} x_i}{\frac{1}{\sigma^2} + \sum_i \rho_{ik}} & (43)\\
s_k^2 &= \frac{1}{\frac{1}{\sigma^2} + \sum_i \rho_{ik}} \quad . & (44)
\end{aligned}$$

REMARK 10. *To derive these results, we never used the assumption that $q(\mu_k)$ is a Gaussian. This was coming as a result looking at the shape in terms of exponential families. Thus, for this case, the Gaussian variational distribution is actually the optimal one for the mixing components (as anticipated before).*

REMARK 11. *You can obtain the same results by taking a different approach. Which one?*

REMARK 12. *In general, the complete conditional can be a complicated distribution. However, in many cases this is a member of the exponential family. If this is the case, then Eq. (19) simplifies.*

The algorithm then works as in Algorithm 1.

By applying this algorithm for the example of fig. 2 you can see how it performs in fig. 3.

5.2. VI part A: summary

- VI is an efficient approach for posterior inference
- It is based on an optimization of KL divergence via ELBO
- Mean-field variational family and CAVI updates allow for efficient implementations
- GMM is an example application that has closed-form updates

Further reference for this lecture is Blei *et al.* (2017) for extensive discussions about Variational Inference.

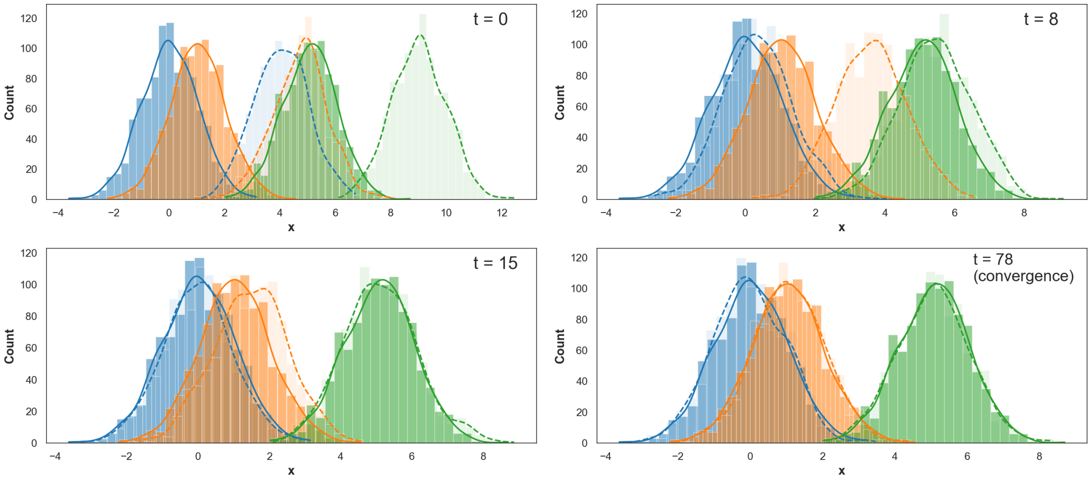

FIGURE 3. Result of CAVI on GMM of fig. 2. The plots are at 4 different iteration steps: at the beginning, at two changing points of the ELBO and at convergence. Dashed lines are corresponding to the variational estimates, regular lines are the exact Gaussians (those used to generate the data).

References

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, Journal of the American statistical Association **112**, 859 (2017).

Algorithm 1: CAVI for a Gaussian mixture model

```
Input: Data $\mathbf{x}$, number of components $K$, prior variance of component means $\sigma^2$
Output: Variational densities $q(\mu_k; m_k, s_k^2)$ (Gaussian) and $q(c_i; \rho_i)$ ($K$-categorical)
Initialize: Variational parameters $\mathbf{m} = m_{1:K}$, $\mathbf{s}^2 = s^2_{1:K}$, and $\boldsymbol{\varphi} = \rho_{1:n}$
while the ELBO has not converged do
    for $i \in \{1, \dots, n\}$ do
        Set $\rho_{ik} \propto \exp\left[x_i m_k - \frac{1}{2}\left(s_k^2 + m_k^2\right)\right]$
    end
    for $k \in \{1, \dots, K\}$ do
        Set $m_k \longleftarrow \dfrac{\sum_i \rho_{ik} x_i}{\frac{1}{\sigma^2} + \sum_i \rho_{ik}}$
        Set $s_k^2 \longleftarrow \dfrac{1}{\frac{1}{\sigma^2} + \sum_i \rho_{ik}}$
    end
    Compute ELBO($\mathbf{m}, \mathbf{s}^2, \boldsymbol{\varphi}$)
end
return $q(\mathbf{m}, \mathbf{s}^2, \boldsymbol{\varphi})$
```