Signal Processing For Data Science (05107-26001)

Final Project

Omri Berman - 305113458

July 19, 2021

Abstract

My final project is based on the paper *Variational Inference: A Review for Statisticians* [2] by Blei et al. The project contains a detailed summary of the paper, as well as an implementation of an extension to 2-dimensions of the CAVI for MOG presented by the paper. In my implementation, I encounter an unexpected behavior, which led me to investigate a direction not addressed by the paper. I concluded the work with some possible follow-up research directions.

Contents

1 Summary **3**
1.1 Problem Formulation . . . 3
1.1.1 Evidence Lower Bound (ELBO) . . . 3
1.2 Mean-Field Variational Family . . . 4
1.3 Coordinate Ascent Variational Inference (CAVI) . . . 4
1.4 Exponential Families . . . 5

2 Bayesian Mixture of Gaussians **6**
2.1 Problem Formulation . . . 6
2.2 CAVI for MOG . . . 7
2.3 My Implementation . . . 7
2.3.1 Concept . . . 7
2.3.2 Results . . . 8
2.3.3 σ^2 Hyper-Parameter Role . . . 11

3 Future Research **13**

4 Appendix **14**

4.1 Simplifications of Update Rules for CAVI of GMM 14

4.1.1 φ_i 14

4.1.2 m_k, s_k^2 14

4.1.3 ELBO 15

1 Summary

The density estimation problem is of particular interest in the field of Bayesian statistics, which utilizes the estimated posterior distribution of latent variables to perform inference. Unfortunately, in most cases, these posterior distribution are very complex, and rarely fit a parametric model, making it hard to estimate them. One method for estimating such complex probability densities is Variational inference (VI) - a statistical-machine-learning method that uses optimization to approximate the complex densities. VI is an alternative method for the commonly used Markov-Chain-Monte-Carlo (MCMC) method, which performs the same density estimation task, but utilizes sampling rather than optimization. While MCMC is much better studied and satisfies some nice asymptotic properties, VI tends to perform faster and scales better for large amounts of data. In some cases, the VI could outperform MCMC even for small data-sets, e.g. in mixture models, where applying a Gibbs-sampling based MCMC tends to result in mode-collapse.

1.1 Problem Formulation

In principle, VI posits a family of prior densities, and optimizes for the most similar member of that family to the distribution of interest, where similarity is measured by the Kullback-Liebler divergence:

$$\hat{q}(z) = \underset{q(z)\in\mathcal{D}}{argmin}\, \mathrm{KL}(q(z)||p(z|x)) \tag{1}$$

Where x is our sampled data, z the latent variable of interest, and $\mathcal{D}$ the parametric family of distributions to which our candidates belong. Clearly, the complexity of the optimization is governed by the complexity of the family of the distributions, as well as the amount and dimensions of the sampled data. Hence, a key consideration in the optimization process is to choose $\mathcal{D}$ that will be sufficiently rich to have candidates close enough to the posterior, yet simple enough for the optimization to be efficient and successful.

As the family of distributions $\mathcal{D}$ in which we look for candidates is parametric, equation (1) is really a problem of parameters estimation:

$$\hat{\theta} = \underset{\theta\in\mathcal{D}(\theta)}{argmin}\, \mathrm{KL}(q(z;\theta)||p(z|x)) \tag{2}$$

Where θ are the parameters by which $\mathcal{D}$ is characterized.

The posterior distribution of interest $p(z|x)$ can be factorized by:

$$p(z|x) = \frac{p(x,z)}{p(x)} = \frac{p(x,z)}{\int p(x,z)dz} \tag{3}$$

Where $p(x,z)$ is the joint distribution of the data and latent variable of interest, and $p(x) = \int p(x,z)dz$ is the marginal of the observations, otherwise known as the evidence. In many cases, the evidence either doesn't have a closed form or requires non-polynomial time to compute, which is the reason Bayesian inference is such a difficult task.

1.1.1 Evidence Lower Bound (ELBO)

The optimization objective introduced in (1) tries to minimize the KL divergence between the optimal member of the suggested family of prior distributions and the posterior. Expanding the term of the KL divergence, we get:

$$\begin{aligned}\mathrm{KL}(q(z)||p(z|x)) &= \mathbb{E}[log(q(z))] - \mathbb{E}[log(p(z|x))] \\ &= \mathbb{E}[log(q(z))] - \mathbb{E}[log(p(z,x))] + \mathbb{E}[log(p(x))] \\ &= \mathbb{E}[log(q(z))] - \mathbb{E}[log(p(z,x))] + log(p(x)) \geq 0\end{aligned} \tag{4}$$

Where the last transition follows since the KL divergence is non-negative, and the transition before that since the expectations are taken w.r.t $q(z)$, where $q(x)$ serves as a constant. isolating the log-evidence, we get:

$$log(p(x)) \geq \mathbb{E}[log(p(z,x))] - \mathbb{E}[log(q(z))] \triangleq ELBO(q) \tag{5}$$

We now note that $\mathrm{KL}(q(z)||p(z|x)) = -ELBO(q) + log(p(x))$, so returning to (1):

$$\begin{aligned}\hat{q}(z) &= \underset{q(z)\in\mathcal{D}}{argmin}(-ELBO(q) + log(p(x))) = \underset{q(z)\in\mathcal{D}}{argmin}(-ELBO(q)) \\ &= \underset{q(z)\in\mathcal{D}}{argmax}(ELBO(q))\end{aligned} \tag{6}$$

Where the 2nd transition is since the evidence is independent on $q(z)$. As we can now notice, VI is equivalent to maximizing the ELBO, which circumvents the need to calculate the evidence, which we established as the bottleneck of the posterior's calculation.

The ELBO can also be factorized in another informative way:

$$\begin{aligned}ELBO(q) &= \mathbb{E}[log(p(z,x))] - \mathbb{E}[log(q(z))] \\ &= \mathbb{E}[log(p(x|z))] + \mathbb{E}[log(p(z))] - \mathbb{E}[log(q(z))] \\ &= \mathbb{E}[log(p(x|z))] - \mathrm{KL}(q(z)||p(z))\end{aligned} \tag{7}$$

Hence, maximizing the ELBO is equivalent to a simultaneous maximiziation of the likelihood and a minimization of the divergence, which is obtained as the proposed distribution $q(z)$ approaches the prior over the latent variable $p(z)$. This clearly reveals the trade-off between likelihood and prior, which is common in the Bayesian setup.

1.2 Mean-Field Variational Family

A special (and common) case of the proposed family densities that try to approximate the posterior is the mean-field variational family. This family is characterized by the latent variables being mutually independent and governed by distinct variational factors, that is: $q(z) = \prod_{j=1}^{m} q_j(z_j)$. In the VI setup, the variational factors $q_j(\cdot)$ are chosen to maximize the ELBO.

1.3 Coordinate Ascent Variational Inference (CAVI)

CAVI is the name of the algorithm that's being used to solve the ELBO maximization under the mean-field variationl family assumption. It performs the optimization iteratively in a pareto-fashion, i.e. optimizes each factor of the variational density while keeping the others fixed. The CAVI algorithm is as follows:

```
Algorithm 1: CAVI
  Data: dataset $x$, model $p(x, z)$
  Result: Approximate variational density $q(z) = \prod_{j=1}^{m} q_j(z_j)$
  Initialize $q_j(z_j)$;
  while ELBO hasn't converged do
      for $j \in [1, ..., m]$ do
          $q_j(z_j) \leftarrow exp\{\mathbb{E}_{-j}[log(p(z_j|z_{-j}, x))]\}$
      end
      Compute $ELBO(q) = \mathbb{E}[log(p(z, x))] - \mathbb{E}[log(q(z))]$
  end
```

As ELBO is generally non-convex, the CAVI algorithm is only guaranteed to converge to a local optimum, and it is clearly sensitive to the initialization.

A typical con of the VI method is an underestimation of the posterior's variance, which is inherent from the formulation of the optimization formulation

1.4 Exponential Families

A particular class of interest in regard to the variational inference problem is the exponential family. It is special since it makes the CAVI algorithm easier to derive, and the inference scales better to massive datasets. If the complete-conditional of every latent variable z_i is exponential, that is:

$$p(z_j|_{z-j,x} = h(z_j)exp(\eta_j(z_{-j}, x)^T z_j - a(\eta_j(z_{-j}, x)))$$

and assuming a mean-field variational inference as described in 1.2, we get:

$$q(z_j) \propto exp(\mathbb{E}[p(z_j|z_{-j}, x)]) \propto h(z_j)exp(\mathbb{E}[\eta_j(z_{-j}, x)]^T z_j) \tag{8}$$

Having this parametric form of the variational factors, the variational parameters update reduces to

$$v_j = \mathbb{E}[\eta_j(z_{-j}, x)] \tag{9}$$

Which significantly simplifies the derivation of the CAVI algorithm.

A special case of an exponential family are conditionally conjugate models, where some latent variables are global, which apply to the entire dataset, and others are local, that is they are relevant to specific samples. This case significantly simplifies the CAVI and turns each iteration to an alternating update between the local paramters and the global parameters.

Despite having nicer scaling properties, even exponential family conditional-conjugate models fail to handle massive datasets. To tackle this, an alternative gradient-based optimization was developed, which is called stochastic variation inference (SVI). SVI samples a single data-sample in each iteration, and uses it to update both the local and the global variational parameters, according to the natural gradient of the ELBO. By utilizing a single data-sample every time, it avoids going over the entire data-set on each iteration, which saves a huge amount of resources when the dataset of interest is massive.

2 Bayesian Mixture of Gaussians

2.1 Problem Formulation

Considering the case of mixture of K Gaussians, all of a unit-variance, where the means of the Gaussians μ_k are latent variables drawn themselves from a zero-mean Gaussian distribution with some variance σ^2, which in our case is a nuisance parameter. The setup can hence be formulated by the following:

$$\begin{aligned} \mu_k &\sim \mathcal{N}(0, \sigma^2) \\ c_i &\sim Unif_{[1,K]} \\ x_i | c_i, \mu &\sim \mathcal{N}(c_i^T \mu, 1) \end{aligned} \tag{10}$$

Where c_i is a one-hot K-long indicator vector of the Gaussian from which the i'th sample was drawn. In this setup, the joint distribution is

$$p(\mu, c, x) = p(\mu) \prod_{i=1}^{n} p(c_i) p(x_i | c_i, \mu) \tag{11}$$

And hence the evidence in this case is:

$$p(x) = \int p(\mu) \prod_{i=1}^{n} \sum_{c_i} p(c_i) p(x_i | c_i, \mu) d\mu = \sum_{c} p(c) \int p(\mu) \prod_{i=1}^{n} p(x_i | c_i, \mu) d\mu \tag{12}$$

Where in both cases, the time complexity of this evidence calculation is in the order of K^n, i.e. exponential in the number of samples.

As for the family of approximation densities $\mathcal{D}$, in the case it is a mean-field variational familiy, it's members have the form:

$$q(\mu, c) = \prod_{k=1}^{K} q(\mu_k; m_k, s_k^2) \prod_{i=1}^{n} q(c_i; \varphi_i) \tag{13}$$

Where $q(\mu_k; m_k, s_k^2)$ is a Gaussian distribution on the kth mixture component's mean μ_k, who's mean is m_k and variance s_k^2 and $q(c_i; \varphi_i)$ is a categorical distribution on the 1-hot vector c_i with a discrete distribution vector φ_i.

2.2 CAVI for MOG

In the particular setup of the mixture of Gaussians, the CAVI algorithm can be implemented by:

Algorithm 2: CAVI

```
**Data:** dataset $\{x_i\}_{i=1}^n$, number of centers K, prior variance of centers means $\sigma^2$
**Result:** Variational densities - K Gaussian $q(\mu_k; m_k, s_k^2)$ and n categorical $q(c_i; \varphi_i)$
initialize $q_j(z_j)$;
while ELBO hasn't converged do
    for $i \in [1, ..., n]$ do
        $\varphi_{ik} \leftarrow exp\{x_i \mathbb{E}[\mu_k; m_k, s_k^2] - \frac{1}{2}\mathbb{E}[\mu_k^2; m_k, s_k^2]\}$
    end
    for $k \in [1, ..., K]$ do
        $m_k \leftarrow \frac{\sum_i \varphi_{ik} x_i}{1/\sigma^2 + \sum_i \varphi_{ik}}$
        $s_k^2 \leftarrow \frac{1}{1/\sigma^2 + \sum_i \varphi_{ik}}$
    end
    Compute $ELBO(m, s^2, \varphi)$
end
Empirical studies conducted by the authors
```

2.3 My Implementation

2.3.1 Concept

In the project's scope, I decided to implement the CAVI algorithm presented in the paper, and extend it to a 2-dimensional MOG setup, where:

$$\begin{aligned} \mu_k &\sim \mathcal{N}(0, \sigma^2 I_{2\times 2}) \\ c_i &\sim Unif_{[1,K]} \\ x_i | c_i, \mu &\sim \mathcal{N}(c_i^T \mu, I_{2\times 2}) \end{aligned} \tag{14}$$

In the experiments described below, I used K=5 classes, n=1000 data samples, and a hyper parameter = 5 as demonstrated in figure 1. The number of data-samples was picked to be large enough to result in a balanced number of points when sampling uniformly over the K classes, yet small enough to allow a quick CAVI convergence. The value of the hyper-parameter σ^2 was selected so that the centers are statistically close enough to allow for non-trivial convergences of the CAVI, yet far enough to avoid mode-collapse in the typical case. This hyper-paremter will be further discussed later on.

To inspect the performance of the implemented algorithm, I focused on 3 different metrics as the CAVI progressed towards convergence:

- $ELBO(m, s^2, \varphi)$
- The variational densities of the latent means ($q(\mu_k; m_k, s_k^2)$)
- The posterior-predictive distribution approximation ($p(x_{new}|x) \approx \frac{1}{K}\sum_1^k p(x_{new}; m_k)$)

In addition, the CAVI was repeatedly initialized and run over the same dataset, to test the non-convexity of the ELBO described in the paper, according to which the CAVI should converge to different values, due to different initial conditions.

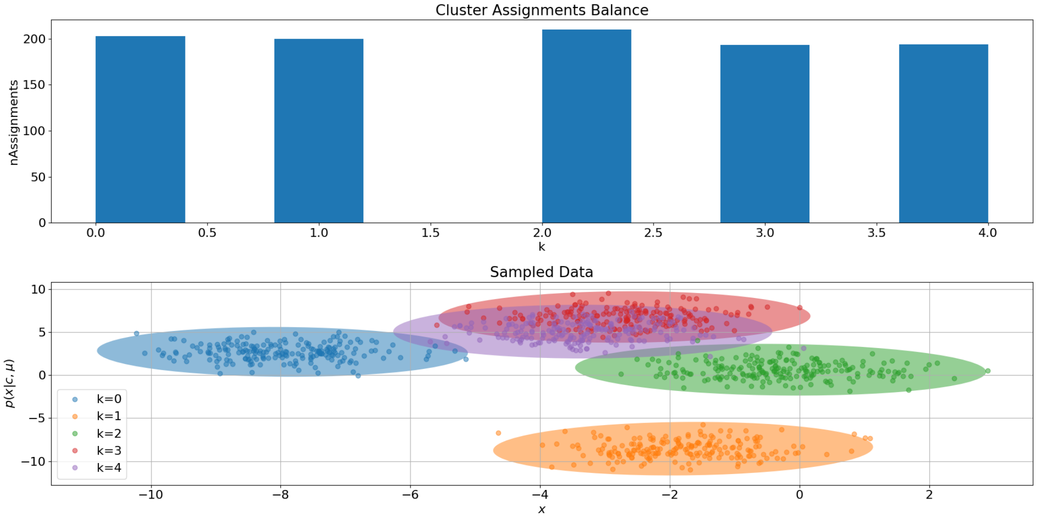
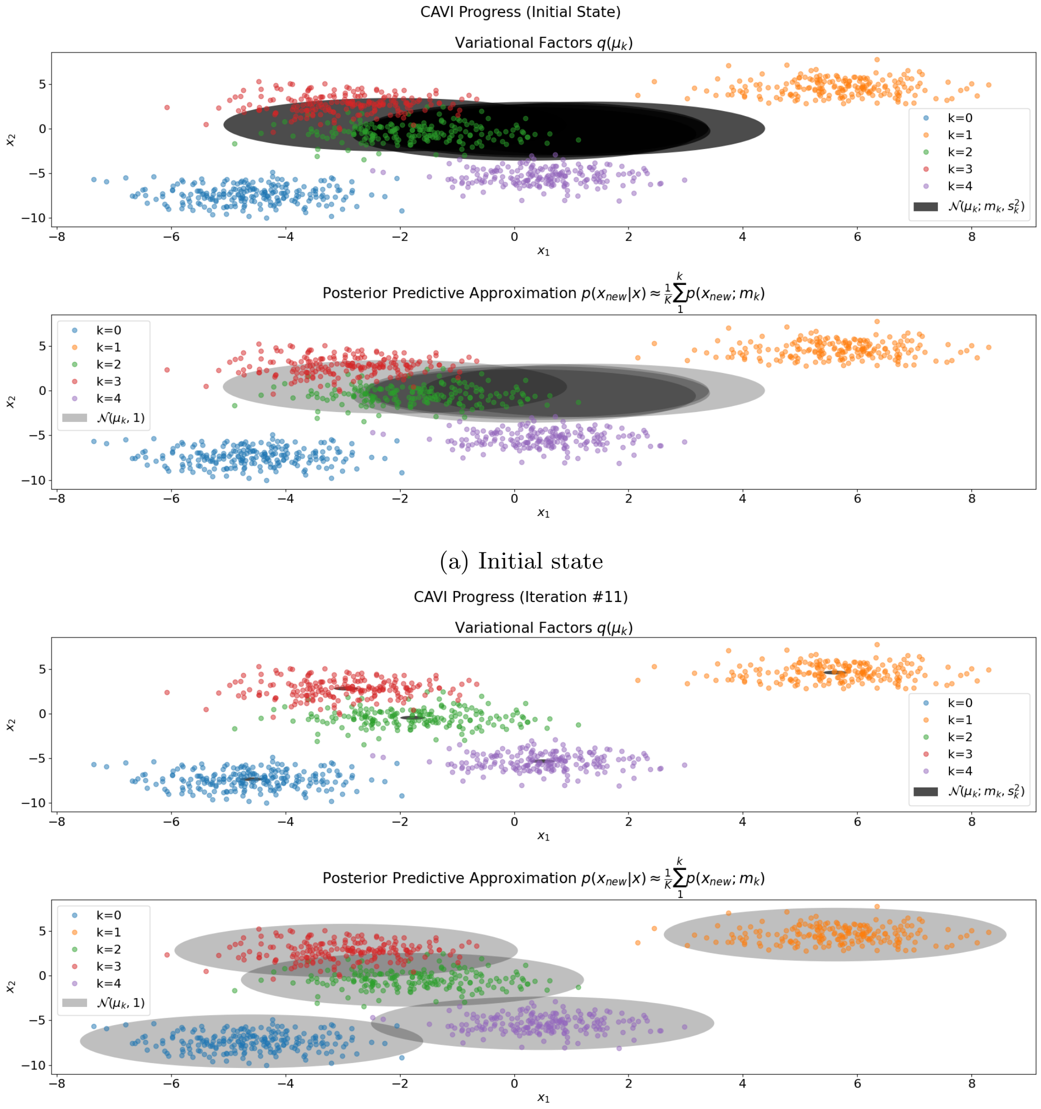


Figure (1) An example for a batch of 1000 data-points, sampled according to the MOG distribution, as presented in the paper. The top histogram indicates that samples are roughly uniformly distributed between the different means.

2.3.2 Results

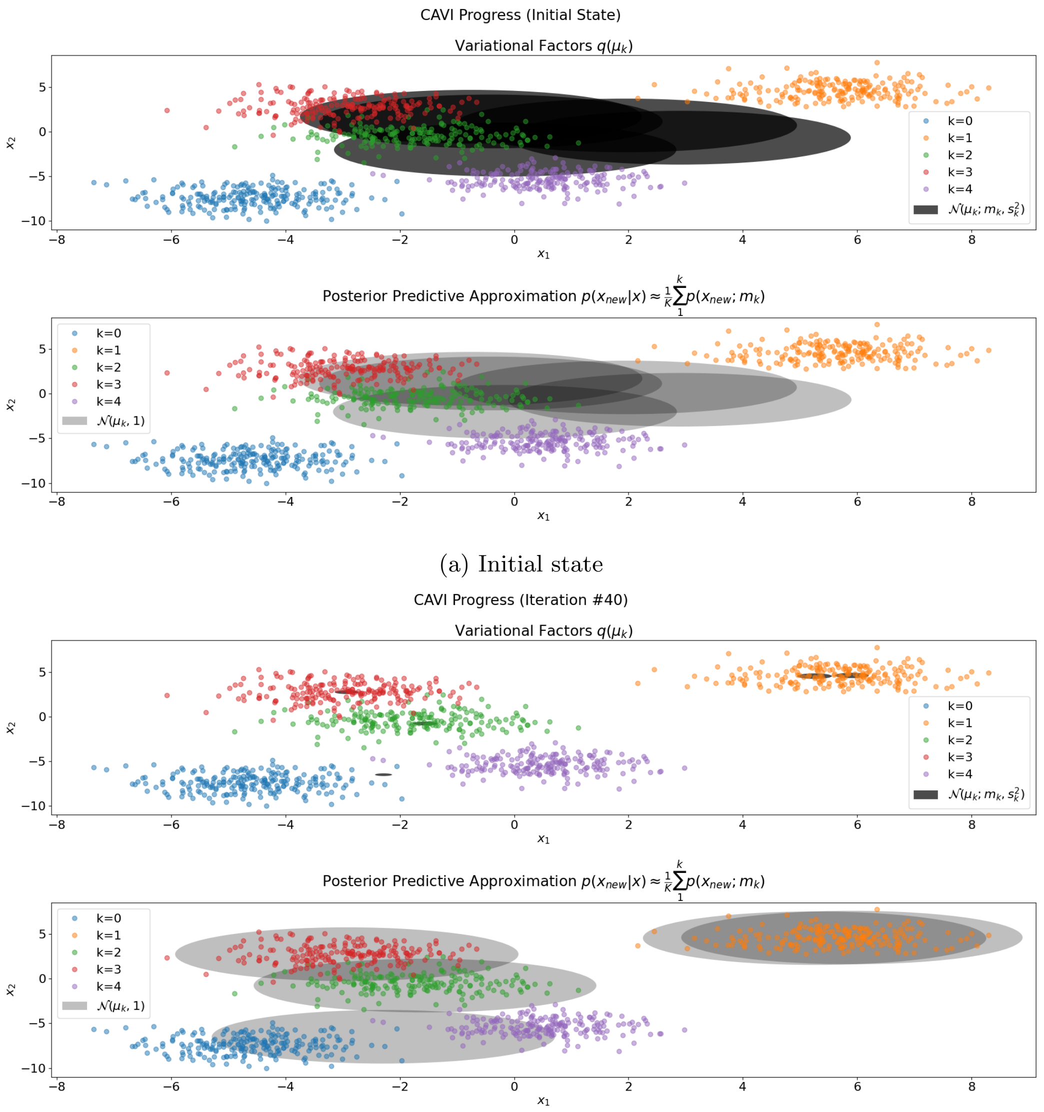
In figure 2, an example of a successful CAVI convergence is depicted, whereas figure 3 demonstrates an unsuccessful convergence, in the sense that the means of the learned variational density aren't as good a fit for the data-samples as in the successful case. After running CAVI with the same dataset for 10 consecutive times, the sensitivity of the convergence to the initialization became even clearer, as reflected by the ELBOs convergence plots in figure 4. These observations are in agreement with the claims made in the paper regarding the non-convexity of the ELBO, and the tendency to converge to different local maxima, depending on the initial conditions.



(a) Initial state

(b) Final state

Figure (2) The variational density and posterior-predictive of a successful CAVI convergence.



(a) Initial state

(b) Final state

Figure (3) The variational density and posterior-predictive of an unsuccessful CAVI convergence. There's clearly a mode-collapse for the "orange" class latent mean, which results in under-fitting of the "blue" and "purple" latent means by a single latent mean that resides in the middle between them.

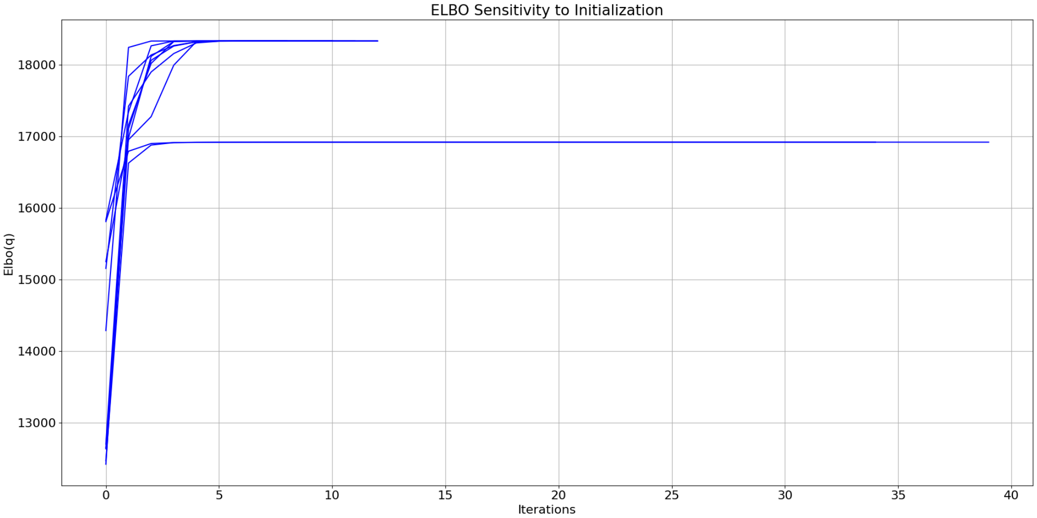
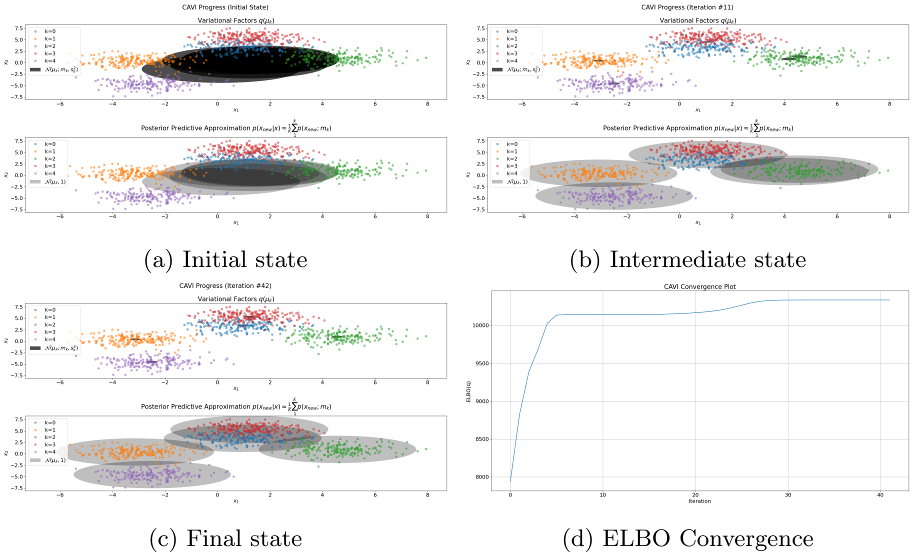


Figure (4) The ELBO obtained for 10 consecutive CAVI runs. The convergence to 2 different local maxima is apparent, reflecting the CAVI's sensitivity to initialization.

2.3.3 σ^2 Hyper-Parameter Role

While seemingly in agreement with the paper, some factor of interest was missing. In the paper, an additional matter of interest were the "elbows" of the ELBO. In the "Emperical study" section of the paper, the authors demonstrated a pattern of elbows appearing in the ELBO convergence trend, suggesting that the CAVI iteratively reaches better plateaus. While the problem described in the paper is essentially different then the one I'm trying to solve (the paper treats the covariance of the data's conditional density as an additional latent variable, making the problem much more complex), I found it odd that I witness no elbows whatsoever. After some thought, I decided to inspect whether the hyper-parameter σ^2 used to generate the GMM samples means μ might have an impact on the discussed phenomenon. This made sense to me, as a small enough σ^2 will result in the sampled data clusters being close enough, in which case the CAVI will initially assign several μ_ks to explain 1 data cluster (and hence also one other μ_k to explain several different clusters). Nonetheless, if the samples of the different clusters are sufficiently separable, the CAVI could slowly figure out that driving the wrongfully assigned μ_k apart (and towards the center of mass of the correct clusters) results in a higher ELBO value.

With that hypothesis in mind, I repeated the test with $\sigma = 3$ (instead of 5), and noticed that indeed, elbows began developing for some of the CAVI initializations. An example of a successful convergence with an elbow in the ELBO convergence profile is depicted in figure 5. A 10 repetition MC exhibits elbows obtained arbitrarily at different iterations, and is depicted in figure 6.



(a) Initial state

(b) Intermediate state

(c) Final state

(d) ELBO Convergence

Figure (5) CAVI convergence with elbows, after setting $\sigma = 3$. The intermediate state displayed corresponds to the plateau observed in the 11th iteration of the CAVI (also referred to as "intermediate state"). As witnessed, 2 latent variational means are attracted to the green center of mass, while 1 mean is fitted to the border between the blue and the red clusters. These means are later driven apart, such that each mean is aligned with the center of a different cluster, which is reflected by a higher ELBO value.

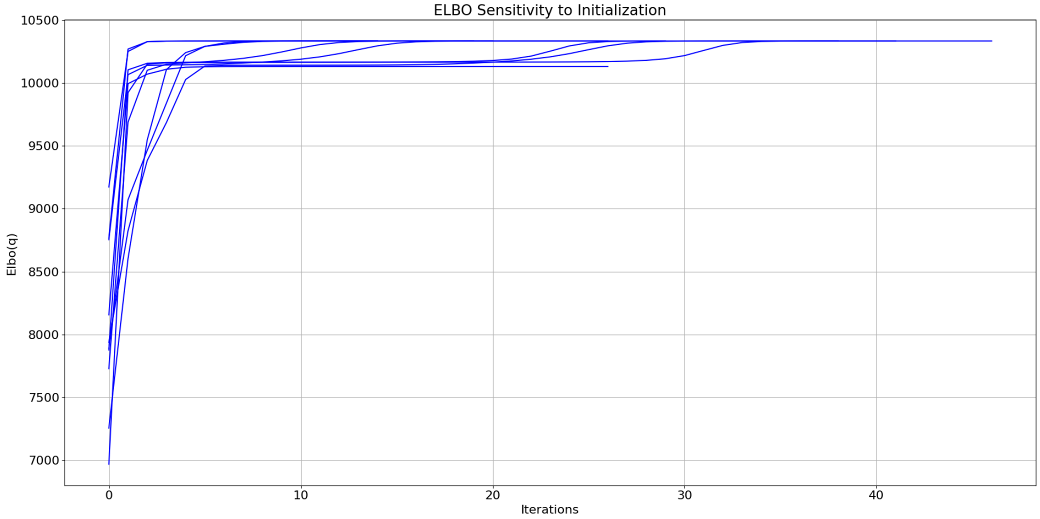


Figure (6) Elbows exhibiting elbows at different iterations of the CAVI algorithm.

3 Future Research

Going over the paper and considering the results obtained by implementing the CAVI, there are several directions I consider interesting to follow in future research:

- **Hyper-parameters sensitivity analysis**: As demonstrated in the previous section, a small change in the variance of the MOG means generation (σ^2) was sufficient to have a significant impact over the CAVI's convergence dynamic. This hyper-parameter is one of several parameters that can potentially have an interesting impact over the algorithm performance, amongst which are the number of clusters K, the number of samples n, and the prior over the MOG means variance $\hat{\sigma}^2$ explicitly used by the CAVI algorithm. While the paper presents and discusses the problem of MOG in depth, it doesn't refer to the impact of the different hyper-parameters over the CAVI algorithm whatsoever. Hence, I find it interesting to preform a comprehensive sensitivity analysis for the CAVI performance to the different hyper-parameters. Such an analysis could include a theoretical analysis to provide the expected impact of each parameter over the performance, followed by an empirical study, which can be divided to pareto-analyses of each hyper-parameter individually, and a combination of hyper-parameters alternations.

- **Distribution similarity metric examination**: The field of variational-inference as described in the paper uses the well-known KL-divergence as the cost-function to be optimized by the CAVI algorithm. While proven successful as a proxy for distribution similarity, there are additional metrics that could also be used to indicate that similarity, and perhaps yield better solutions in terms of posterior approximations. One such metric is the earth-mover distance (also known as Wasserstein distance), that has undergone a resurrection in the past few years as it proved efficient in the field of generative adversarial network. In the original paper by Arjovsky et al. [1], the Wasserstein distance is used as an adversarial loss function (instead of the traditional KL-divergence), and yields superior results. Hence, It would be interesting to inspect the change in the VI formulation and performance when using the Wasserstein distance as a metric to optimize instead of KL divergence.

4 Appendix

4.1 Simplifications of Update Rules for CAVI of GMM

4.1.1 φ_i

$$\begin{aligned}
q(c_i;\varphi_i) &\propto exp\{\mathbb{E}[logp(c_i, c_{-i}, \mu, x)]\} \\
&\propto exp\{\mathbb{E}[logp(c_i) + logp(c_{-i}) + logp(\mu) + \sum_{j=1}^{n} logp(x_j|c_j, \mu)]\} \\
&\propto exp\{\mathbb{E}[logp(c_i) + p(x_i|c_i, \mu)]\} = exp\{\mathbb{E}[\frac{1}{K} + p(x_i|c_i^T \cdot \mu)]\} \\
&\propto exp\{-\frac{1}{2}\mathbb{E}[x_i^2 - 2x_i\mu_{c_i} + \mu_{c_i}^2]\} \\
&\propto exp\{x_i m_{c_i} - \frac{1}{2}(s_{c_i}^2 + \mu_{c_i}^2)\} \\
\Rightarrow \varphi_i &\leftarrow exp\{x_i m_{c_i} - \frac{1}{2}(s_{c_i}^2 + \mu_{c_i}^2)\}
\end{aligned} \tag{15}$$

4.1.2 m_k, s_k^2

$$\begin{aligned}
q(\mu_k; m_k, s_k^2) &\propto exp\{\mathbb{E}[logp(c, \mu, x)]\} \\
&\propto exp\{\mathbb{E}[logp(\mu_k) + \sum_{j=1}^{n} logp(x_j|c_j, \mu)]\} \\
&\propto exp\{-\frac{\mu_k^2}{2\sigma^2} - \frac{1}{2}\sum_{j=1}^{n}\varphi_{jk}(x_j^2 - 2x_j\mu_k + \mu_k^2)]\} \\
&\propto exp\{-\frac{\mu_k^2}{2}(\frac{1}{\sigma^2} + \sum_{j=1}^{n}\varphi_{jk}) + \sum_{j=1}^{n}\varphi_{jk}x_j\mu_k]\} \\
&\propto exp\{-\frac{1}{2}(\frac{1}{\sigma^2} + \sum_{j=1}^{n}\varphi_{jk}) \cdot (\mu_k - \frac{\sum_{j=1}^{n}\varphi_{jk}x_j}{\frac{1}{\sigma^2} + \sum_{j=1}^{n}\varphi_{jk}})^2 \\
&\propto \mathcal{N}(\mu_k; \frac{\sum_{j=1}^{n}\varphi_{jk}x_j}{\frac{1}{\sigma^2} + \sum_{j=1}^{n}\varphi_{jk}}, (\frac{1}{\sigma^2} + \sum_{j=1}^{n}\varphi_{jk})^{-1}) \\
\Rightarrow m_k &\leftarrow \frac{\sum_{j=1}^{n}\varphi_{jk}x_j}{\frac{1}{\sigma^2} + \sum_{j=1}^{n}\varphi_{jk}} \\
s_k^2 &\leftarrow (\frac{1}{\sigma^2} + \sum_{j=1}^{n}\varphi_{jk})^{-1}
\end{aligned} \tag{16}$$

4.1.3 ELBO

$$\begin{aligned} ELBO(m, s^2, \phi) = &\sum_{k=1}^{K} \mathbb{E}[\log p(\mu_k; m_k, s_k^2)] - \mathbb{E}[\log q(\mu_k; m_k, s_k^2)] \\ &+ \sum_{j=1}^{n} \mathbb{E}[\log p(c_j; \varphi_j)] - \mathbb{E}[\log q(c_j; \varphi_j)] + \mathbb{E}[\log p(x_j; \varphi_j, m, s^2)] \end{aligned} \tag{17}$$

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. "Wasserstein Generative Adversarial Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, June 2017, pp. 214–223. URL: http://proceedings.mlr.press/v70/arjovsky17a.html.

[2] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. "Variational Inference: A Review for Statisticians". In: *Journal of the American Statistical Association* 112.518 (Apr. 2017), pp. 859–877. ISSN: 1537-274X. DOI: 10.1080/01621459.2017.1285773. URL: http://dx.doi.org/10.1080/01621459.2017.1285773.