Workshop at the 7th Symposium on Advances in Approximate Bayesian Inference (non-archival), 2025 1–10

Revisiting Influence Functions for Latent Variable Models using Variational Bayes

Dharmesh Tailor D.V.TAILOR@UVA.NL
University of Amsterdam

Mohammad Emtiyaz Khan EMTIYAZ.KHAN@RIKEN.JP
RIKEN Center for Advanced Intelligence Project

Eric Nalisnick NALISNICK@JHU.EDU
Johns Hopkins University

Abstract

Quantifying a model's sensitivity to data is a key tool for model criticism and interpretability. Influence functions are the de-facto method for estimating such quantities. Latent variable models are ubiquitous in modern ML (*e.g.* mixture of experts, deep generative models), but estimating the influence of individual data points can be challenging due to the rigid structure between observed and latent variables. In previous work, Zhu and Lee (2001) proposed to take a Newton-step on a surrogate function inspired by the Expectation-Maximization (EM) algorithm. This exploits the model's structure to decouple the effect of perturbations to the data such that the influence on different parameters can be measured separately. We present a generalization of this approach from the lens of Variational Bayes that does not have the restrictions of EM and can be used in a wide-variety of settings.

1. Introduction

Sensitivity and perturbation have a long history in model diagnostics (*i.e.* pioneering works in the 70s by Cook and others) to detect *influential* examples that lead to the largest changes in the model when removed. This was originally proposed for simple models with analytic solutions such as linear regression (Cook, 1977) or principal components analysis (PCA) (Critchley, 1985) for which the effect can be evaluated exactly.

In recent years, there is increased interest in developing sensitivity-based techniques for deep learning, with use cases such as diagnosing model behaviour (Koh and Liang, 2017), detecting memorized examples (Feldman and Zhang, 2020), attributing training data to test-time predictions (Koh and Liang, 2017; Yeh et al., 2018), constructing validation-free generalization measures (Beirami et al., 2017). In contrast to linear regression or PCA, the effect of perturbation to the data can no longer be evaluated exactly. Various techniques have been proposed to get around this, some more *global* in their nature such as training many models on different subsamples of the data (Feldman and Zhang, 2020; Jiang et al., 2021; Ilyas et al., 2022), or hybrid approaches that track gradients during training (Hara et al., 2019; Pruthi et al., 2020). In this work, we consider a popular *local* sensitivity-based technique called *influence function* (Cook and Weisberg, 1980; Koh and Liang, 2017), that constructs a second-order Taylor expansion to the loss leading to a Newton-step. Although originally proposed for supervised problems that are framed as minimizing an empirical risk, influence function has been demonstrated on generative models (Terashita et al., 2021;

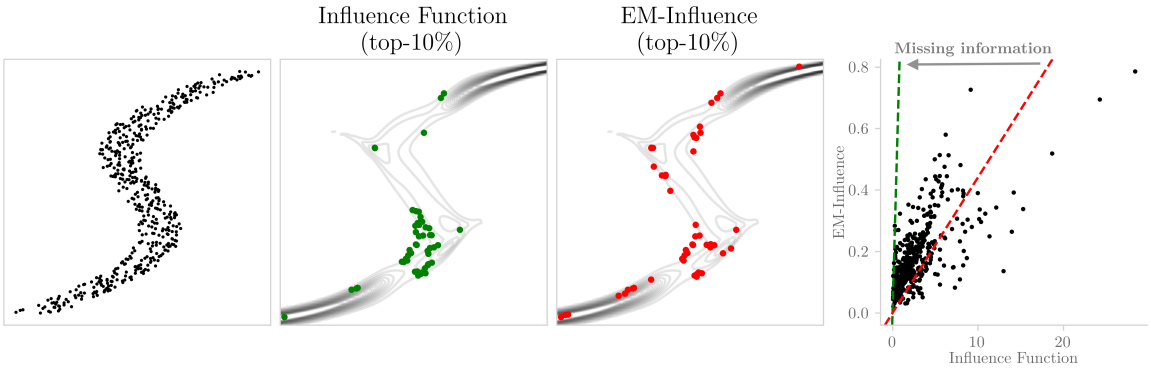

Figure 1: We demonstrate influence function vs EM-influence for a mixture density network trained on a toy dataset (left plot) from Bishop (1994). Whilst both approaches are constructed from a single Newton step, influence function uses the conditional mixture model NLL whereas EM-influence uses its lower bound. Many of the top influential examples can be found in both (middle plots). By incorporating the *missing information* in EM-influence, one can recover influence function (right plot).

Kong and Chaudhuri, 2021; Georgiev et al., 2023), matrix factorization (Cheng et al., 2019) and decentralized settings with bi-level structure (Terashita and Hara, 2022; Zhu et al., 2025). However, in these cases there is often little thought to the original construction which may not be appropriate in certain cases. In particular, here we are interested in models with latent variables. Such models arise for instance from simply combining the outputs of different submodel (*e.g.* state-of-the-art models such as GPT-4 are suggested to be a mixture of experts), or multi-level hierarchies as encountered in federated or meta learning. Naively using influence function in these cases can be challenging. In the mixture model case, the Hessian computation required by the Newton step grows with the number of components where each component could be a neural network with millions of parameters.

We would like to develop sensitivity-based techniques that explicitly take into account latent variables. Previous work (Zhu and Lee, 2001) in the setting of classical latent variable models addressed some of the challenges by taking inspiration from the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). This used the so-called "complete-data" log-likelihood instead which directly takes into account the latent variables. They showed this reduces the computational effort involved with little effect on the accuracy of the influence estimates. However, EM is restrictive in that it requires the posterior over latent variables to be available in closed-form. In this work, we propose a generalization by considering the optimality condition of the Variational Bayes (VB) objective. This allows us to leverage the model's structure, where present, in particular its conditional independence assumptions, to enable efficient influence estimation. This can be seen as a consequence of the factorization constraints in VB that lead to a *decoupling* behaviour where the effect of perturbation in the data can be measured across parameters separately.

2. Preliminaries of Influence Functions

Influence function is most often used in (deep) supervised learning for models trained by minimizing an empirical risk, $\bar{\ell}(\boldsymbol{\theta}) = \sum_i \ell_i(\boldsymbol{\theta}) + \mathcal{R}(\boldsymbol{\theta})$, comprising a sum over N per-example

loss terms $\ell_i(\boldsymbol{\theta})$ with model parameters $\boldsymbol{\theta} \in \mathbb{R}^P$ and a regularizer $\mathcal{R}(\boldsymbol{\theta})$. Given a dataset $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_N\}$, these terms can be identified with a valid likelihood function, $\ell_i(\boldsymbol{\theta}) = -\log p(\mathcal{D}_i|\boldsymbol{\theta})$, decomposed over the N i.i.d. examples and a prior $R(\boldsymbol{\theta}) = -\log p(\boldsymbol{\theta})$. This gives a probabilistic interpretation of the minima $\boldsymbol{\theta}_*$ as the maximum-a-posteriori (MAP) estimate of a Bayesian model. This perspective is not at all restrictive, for instance the cross-entropy loss used for training classifiers is essentially equivalent to a categorical likelihood, and the widely-used regularizer $R(\boldsymbol{\theta}) = \frac{1}{2}\delta \|\boldsymbol{\theta}\|^2$ (*i.e.* weight-decay) corresponds to a zero-mean isotropic Gaussian. This viewpoint is essential for when we consider latent variable models.

In this work, we consider a counterfactual perspective for model diagnosis, seeking to answer the question: *what if a specific example was removed from $\mathcal{D}$, how much would the model change?* By repeating this procedure across all examples in $\mathcal{D}$, we can identify those that are most impactful on the model, referred to as *influential examples*. A naive approach would be to repeat the training process on $\mathcal{D}^{\backslash i}$, that is the dataset without the i^{th} example, however this is infeasible for large N.

Instead, influence functions approach this by estimating the influence from $\boldsymbol{\theta}_*$ that is a model resulting from a *single* training run, via *local* perturbative methods. Let us first denote the objective without the i^{th} example as $\bar{\ell}_{\backslash i}(\boldsymbol{\theta}) = \bar{\ell}(\boldsymbol{\theta}) - \ell_i(\boldsymbol{\theta})$. Then we take a single step of Newton's method from $\boldsymbol{\theta}_*$ on this perturbed objective, $\hat{\boldsymbol{\theta}}_*^{\backslash i} \leftarrow \boldsymbol{\theta}_* - (\nabla^2\bar{\ell}_{\backslash i}(\boldsymbol{\theta}_*))^{-1}\nabla\bar{\ell}_{\backslash i}(\boldsymbol{\theta}_*)$. Using the stationarity condition, $\nabla\bar{\ell}(\boldsymbol{\theta}_*) = 0$, we have $\nabla\bar{\ell}_{\backslash i}(\boldsymbol{\theta}_*) = -\nabla\ell_i(\boldsymbol{\theta}_*)$. It is common to then approximate $\nabla^2\bar{\ell}_{\backslash i}(\boldsymbol{\theta}_*) \approx \nabla^2\bar{\ell}(\boldsymbol{\theta}_*)$ as this reduces the computational burden of influence estimation when repeated for multiple examples. This gives rise to the canonical expression of influence function,

$$\hat{\boldsymbol{\theta}}_*^{\backslash i} - \boldsymbol{\theta}_* = \mathbf{H}_*^{-1}\nabla\ell_i(\boldsymbol{\theta}_*), \qquad \ell_i(\hat{\boldsymbol{\theta}}_*^{\backslash i}) - \ell(\boldsymbol{\theta}_*) \approx \nabla\ell_i(\boldsymbol{\theta}_*)^\top\mathbf{H}_*^{-1}\nabla\ell_i(\boldsymbol{\theta}_*) \tag{1}$$

where $\mathbf{H}_* = \nabla^2\bar{\ell}(\boldsymbol{\theta}_*)$ is the Hessian. On the right side, we show the deviation in the loss, sometimes referred to as Cook's Distance (Cook and Weisberg, 1982), which follows from a 1^{st}-order Taylor expansion to $\ell_i(\boldsymbol{\theta})$ followed by plugging in the left equation. This can be easily extended to measuring the influence on groups of examples.

3. Challenges of Influence Functions for Models with Latent Variables

Let us consider a simple extension of the Bayesian model introduced earlier, in particular introducing a set of latent variables z_i local to each data point $\mathcal{D}_i$. These variables capture the underlying structure or hidden factors that explain the observed data. A popular model of this kind is the mixture model for which the loss terms are given by,

$$\ell_i(\boldsymbol{\theta}) = -\log\left(\sum_{k=1}^{K} p(\mathcal{D}_i, z_i = k|\boldsymbol{\theta})\right) = -\log\left(\sum_{k=1}^{K} \pi_k p_k(\mathcal{D}_i|\mathbf{w}_k)\right) \tag{2}$$

where K is the number of components, π_k are the mixing proportions that satisfy $0 \leq \pi_k \leq 1$ and $\sum_k \pi_k = 1$, $p_k(\cdot)$ is the density of the k^{th} component with parameters $\mathbf{w}_k$, and $\boldsymbol{\theta} := \{\pi_k, \mathbf{w}_k\}_{k=1}^K$ contains all parameters. We can assume the regularizer is selected such that it corresponds to a valid prior (*e.g.* Dirichlet prior for the mixing proportions).

Naturally, one might consider directly applying influence function as given in Eq. (1) to the mixture model. However, this approach faces significant challenges, specifically computing and inverting the Hessian becomes increasingly expensive as its size scales with the number of components. We can clearly see this by manipulating the following identity:

$$\log p(\mathcal{D}, \boldsymbol{\theta}) = \log p(\mathcal{D}, \mathbf{Z}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) - \log p(\mathbf{Z}|\mathcal{D}, \boldsymbol{\theta}) \tag{3}$$

$$\implies \log p(\mathcal{D}, \boldsymbol{\theta}) = \underbrace{\mathbb{E}_{p(\mathbf{Z}|\mathcal{D},\boldsymbol{\theta}_*)}[\log p(\mathcal{D}, \mathbf{Z}|\boldsymbol{\theta})] + \log p(\boldsymbol{\theta})}_{-\mathcal{L}(\boldsymbol{\theta};\boldsymbol{\theta}_*)} \underbrace{- \mathbb{E}_{p(\mathbf{Z}|\mathcal{D},\boldsymbol{\theta}_*)}[\log p(\mathbf{Z}|\mathcal{D}, \boldsymbol{\theta})]}_{-\mathcal{H}(\boldsymbol{\theta};\boldsymbol{\theta}_*)} \tag{4}$$

$$\implies \nabla^2 \bar{\ell}(\boldsymbol{\theta}) = \nabla^2_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{\theta}_*) - \nabla^2_{\boldsymbol{\theta}} \mathcal{H}(\boldsymbol{\theta}; \boldsymbol{\theta}_*) \tag{5}$$

The first line holds for any value of $\mathbf{Z}$. In the second line, we choose to take the expectation of both sides with respect to $p(\mathbf{Z}|\mathcal{D}, \boldsymbol{\theta}_*)$, the posterior evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}_*$. Then in the last line, we take the Hessian of both sides resulting in a relationship that is referred to as the *Missing Information Principle* (Orchard and Woodbury, 1972). This leads to an interpretation of the Hessian as the "complete information" minus the "missing information". The complete-information matrix can benefit from structure present in the model leading to block-diagonal Hessian structure in the aforementioned mixture case (*i.e.* decoupling of the different components). We can see this by writing the expected regularized "complete-data" log-likelihood,

$$\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{\theta}_*) = -\sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_{ik}^{(*)} \left[\log \pi_k + \log p_k(\mathcal{D}_i|\mathbf{w}_k)\right] - \log p(\boldsymbol{\theta}) \tag{6}$$

where we set the following prior $p(\boldsymbol{\theta}) = p(\boldsymbol{\pi}) \prod_k p(\mathbf{w}_k) = \text{Dir}(\boldsymbol{\pi}; \alpha_0) \prod_k \mathbb{N}(\mathbf{w}_k; \mathbf{0}, \delta_k^{-1}\mathbf{I})$. The posterior probabilities $\gamma_{ik}^{(*)} = p(z_i = k|\mathcal{D}_i, \boldsymbol{\theta}_*)$, often called "responsibilities", are given by,

$$\gamma_{ik}^{(*)} = \frac{\pi_{*k} p_k(\mathcal{D}_i|\mathbf{w}_{*k})}{\sum_{k'=1}^{K} \pi_{*k'} p_{k'}(\mathcal{D}_i|\mathbf{w}_{*k'})} \tag{7}$$

which have the effect of introducing per-example (and per-component) weights in Eq. (6). Comparing Eq. (6) with Eq. (2), we can see that the log and summation are interchanged revealing a decomposition into independent parts governed by distinct parameters sets, namely $\boldsymbol{\pi}$ and $\mathbf{w}_k$ $\forall k = 1, \ldots, K$. This is a consequence of the conditional independence assumptions in the model that are not taken into account when locally approximating the regularized (incomplete-data) log-likelihood by a 2^{nd}-order Taylor expansion, as done in influence function. The missing-information matrix is responsible for the off-diagonal blocks in the Hessian resulting in a large, dense matrix. We can see this by writing the following,

$$\mathcal{H}(\boldsymbol{\theta}; \boldsymbol{\theta}_*) = -\sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_{ik}^{(*)} \log \gamma_{ik} \tag{8}$$

where the responsibilities γ_{ik} depend on the normalization constant. It turns out $-\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ is the auxiliary function constructed and maximized in each iteration of the (Generalized) Expectation-Maximization algorithm where the second argument of $\mathcal{L}$ determines the setting of parameters used to evaluate the posterior over latent variables. As clear from

Eq. (4), this maximizes a lower bound to the (regularized) log-likelihood $-\bar{\ell}(\boldsymbol{\theta}) \geq -\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{\theta}_*)$. A standard result (Dempster et al., 1977) tells us that if $\boldsymbol{\theta}_*$ is the minima of the regularized empirical risk, then $\boldsymbol{\theta}_*$ is a fixed point of $\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{\theta}_*)$.

Influence via Expectation-Maximization Zhu and Lee (2001) demonstrated that in the same way each parameter set of $\boldsymbol{\theta}$ in the M-step correspond to their own independent subproblems, we can also estimate the influence on each parameter set independently. Whilst influence function can be viewed as taking a Newton step from $\boldsymbol{\theta}_*$ on the *perturbed* empirical risk, we can take a Newton step on a perturbed auxiliary function. The perturbation is crafted in a similar fashion but instead the i^{th} example is dropped from the complete-data log-likelihood. We can write the following influence estimate on the parameters of the k^{th} component,

$$\hat{\mathbf{w}}_{*k}^{\setminus i} - \mathbf{w}_{*k} = -\gamma_{ik}^{(*)} \mathbf{H}_{*k}^{-1} \nabla_{\mathbf{w}_k} \log p_k(\mathcal{D}_i | \mathbf{w}_{*k}) \qquad (\forall k = 1, \ldots, K) \tag{9}$$

where $\mathbf{H}_{*k} = -\sum_{j=1}^{N} \gamma_{jk}^{(*)} \nabla_{\mathbf{w}_k}^2 \log p_k(\mathcal{D}_j | \mathbf{w}_{*k}) + \delta_k \mathbf{I}$. In the case of $\boldsymbol{\pi}$, it is not necessary to resort to Newton's method as there exists an analytic form (keeping in mind the constraint $\sum_k \pi_k = 1$):

$$\hat{\pi}_{*k}^{\setminus i} - \pi_{*k} = \frac{\pi_{*k} - \gamma_{ik}^{(*)}}{N - 1 + K\alpha_0 - K} \qquad (\forall k = 1, \ldots, K) \tag{10}$$

In Fig. 1, we show a simulation on a mixture density network (see Appendix A for experimental details). Due to the parameter sharing in this architecture, an influence estimate cannot be derived for each parameter set separately.

4. Influence via Variational Bayes

We have seen that through the lens of EM we can leverage the model's structure for efficient influence estimation. However, there are some limitations to this approach. Firstly, it requires that the posterior over the latent variables is available in closed-form which is not always the case. Secondly, the efficiency of influence estimation is tied to conditional independence assumptions in the model. A natural question to ask is whether this cost can be reduced further through additional constraints.

We propose *influence via variational Bayes*, a versatile approach that enables additional decoupling of influence estimates via factorization constraints in the mean-field variational posterior. This introduces a variational distribution over $\boldsymbol{\theta}$ that we refer to as the global variables in addition to the variational distribution over latent variables $\mathbf{Z}$. As a consequence, this naturally handles cases where the latent posterior is not analytic. Since EM can be viewed as coordinate-ascent on the variational objective (Neal and Hinton, 1998), we will show we can recover the influence estimates in the mixture model setting.

But first, let us revisit the simple extension of the Bayesian model with a set of latent variables z_i local to each data point $\mathcal{D}_i$. We start by assuming the following mean-field variational family $q(\boldsymbol{\theta}, \mathbf{Z}) = q(\boldsymbol{\theta}) \prod_{i=1}^{N} q(z_i)$. This is the optimal form of q arising from (conditional) independency assumptions in the model. The optimal variational distribution satisfies the following set of consistency conditions Bishop (2006):

$$q_*(\boldsymbol{\theta}) \propto \exp\left\{\mathbb{E}_{q_*(\mathbf{Z})}\left[\log p(\mathcal{D}, \boldsymbol{\theta}, \mathbf{Z})\right]\right\}, \quad q_*(z_i) \propto \exp\left\{\mathbb{E}_{q_*(\boldsymbol{\theta})}\left[\log p(\mathcal{D}_i, z_i | \boldsymbol{\theta})\right]\right\} \tag{11}$$

These optimal conditions can be equivalently stated in terms of the optimal *natural* parameters $\boldsymbol{\lambda}_*$ corresponding to q_*. Let us denote the natural parameters of $q(\boldsymbol{\theta})$ and $q(z_i)$ by $\boldsymbol{\lambda}_\theta$ and $\boldsymbol{\lambda}_{z_i}$ respectively, then we have:

$$\boldsymbol{\lambda}_{\theta,*} = \eta_\theta + \sum_{i=1}^{N} \widetilde{\nabla}_{\boldsymbol{\lambda}_\theta} \mathbb{E}_{q_*} \left[\log p(\mathcal{D}_i, z_i|\boldsymbol{\theta})\right], \quad \boldsymbol{\lambda}_{z_i,*} = \widetilde{\nabla}_{\boldsymbol{\lambda}_{z_i}} \mathbb{E}_{q_*} \left[\log p(\mathcal{D}_i, z_i|\boldsymbol{\theta})\right] \tag{12}$$

where η_θ is the natural parameter of prior $p(\boldsymbol{\theta})$ and $\widetilde{\nabla}_{\boldsymbol{\lambda}} \mathbb{E}_q(\cdot) = \mathbf{F}(\boldsymbol{\lambda})^{-1} \nabla_{\boldsymbol{\lambda}} \mathbb{E}_q(\cdot)$ is the natural gradient with respect to $\boldsymbol{\lambda}$ given by premultiplying the gradient with the inverse of the Fisher Information Matrix $\mathbf{F}(\boldsymbol{\lambda})$ of q. Eq. (12) arises by taking the natural gradient of the variational objective. It is easy to show that this recovers Eq. (11). The fixed point for $\boldsymbol{\lambda}_\theta$ in Eq. (12) decomposes into a sum of the prior natural parameters and natural gradients for each local factor. This is akin to *message-passing* where the natural gradients on the local factors (*i.e.* local messages) are aggregated to obtain the global parameters. This can be expressed as inference in a conjugate Bayesian model by multiplying by sufficient statistics $\mathbf{T}(\boldsymbol{\theta})$ followed by exponentiation (Khan and Lin, 2017).

Nickl et al. (2023) exploits this perspective to obtain an estimate of the posterior but without the i^{th} example or more generally without data in some subset $\mathcal{M}$. This uses a property of Bayesian models that data examples can be removed by simply dividing their likelihoods from the posterior (Weiss, 1996). In the case of models without latent variables, Nickl et al. (2023) demonstrated that this perspective recovers influence function. In our case, we have the likelihood approximation $\widetilde{p}_i \propto e^{\langle \widetilde{\boldsymbol{\lambda}}_{i,*}, \mathbf{T}(\boldsymbol{\theta})\rangle}$ corresponding to each local factor where $\widetilde{\boldsymbol{\lambda}}_{i,*}$ is its natural gradient. By dividing $\widetilde{p}_i$ (or multiple such terms in a subset $\mathcal{M}$) from $q_*(\boldsymbol{\theta})$, we obtain the following estimate of the deviation in the global natural parameters:

$$\hat{\boldsymbol{\lambda}}_{\theta,*}^{\setminus \mathcal{M}} - \boldsymbol{\lambda}_{\theta,*} = -\sum_{i \in \mathcal{M}} \widetilde{\nabla}_{\boldsymbol{\lambda}_\theta} \mathbb{E}_{q_*} \left[\log p(\mathcal{D}_i, z_i|\boldsymbol{\theta})\right] \tag{13}$$

We now return to the mixture model previously introduced and show that Eq. (13) can be used to recover the influence estimates in Eqs. (9) and (10). The optimal variational distribution on the global variables $\boldsymbol{\theta}$ is given by, $q(\boldsymbol{\theta}) = q(\boldsymbol{\pi}) \prod_k q(\mathbf{w}_k) = \text{Dir}(\boldsymbol{\pi}; \boldsymbol{\alpha}) \prod_k \mathbb{N}(\mathbf{w}_k; \mathbf{m}_k, \mathbf{S}_k^{-1})$, and $q(z_i) = \prod_k \gamma_{ik}^{z_{ik}}$. Specializing Eq. (13) for $q(\boldsymbol{\pi})$ we obtain,

$$\hat{\boldsymbol{\lambda}}_{\pi,*}^{\setminus i} - \boldsymbol{\lambda}_{\pi,*} = -\boldsymbol{\gamma}_i^{(*)} \implies \hat{\alpha}_{k,*}^{\setminus i} - \alpha_{k,*} = -\gamma_{ik}^{(*)} \; (\forall k = 1, \ldots, K) \tag{14}$$

where $\boldsymbol{\gamma}_i = (\gamma_{i1}, \ldots, \gamma_{ik})$. In this case the natural gradient is easy to evaluate due to the presence of conditionally-conjugate structure (see Khan (2023) for further details). To recover Eq. (10) exactly, we can evaluate the mode of the Dirichlet distributions corresponding to $\hat{\boldsymbol{\lambda}}_{\pi,*}^{\setminus i}$ and $\boldsymbol{\lambda}_{\pi,*}$ and then compute their deviation. Now specializing Eq. (13) for $q(\mathbf{w}_k)$, we obtain:

$$\hat{\boldsymbol{\lambda}}_{w_k,*}^{\setminus i} - \boldsymbol{\lambda}_{w_k,*} = -\gamma_{ik}^{(*)} \widetilde{\nabla}_{\boldsymbol{\lambda}_{w_k}} \mathbb{E}_{q_*} \left[\log p_k(\mathcal{D}_i|\mathbf{w}_k)\right] \tag{15}$$

Using the natural parameter pair $\boldsymbol{\lambda}_{w_k} = (\mathbf{S}_k \mathbf{m}_k, -\frac{1}{2}\mathbf{S}_k)$ and expressing the natural gradient in terms of the gradient and Hessian of $\log p_k(\mathcal{D}_i|\mathbf{w}_k)$ (see (Khan and Rue, 2023), Eqs. 10-11) followed by minor manipulation leads to:

$$\hat{\mathbf{m}}_{k,*}^{\setminus i} - \mathbf{m}_k = -\gamma_{ik}^{(*)} \left(\hat{\mathbf{S}}_{k,*}^{\setminus i}\right)^{-1} \mathbb{E}_{q_*}[\nabla_{\mathbf{w}_k} \log p_k(\mathcal{D}_i|\mathbf{w}_k)] \tag{16}$$

Similar to the derivation of influence function from a Newton-step we can approximate $\hat{\mathbf{S}}_{k,*}^{\setminus i} \approx \hat{\mathbf{S}}_{k,*}$. In addition by approximating expectations using delta method and writing $\mathbf{w}_k = \mathbf{m}_k, \mathbf{H}_k = \mathbf{S}_k$, we recover Eq. (9).

5. Conclusion

In this work, we consider the problem of influence estimation for latent variable models through the lens of Variational Bayes. By leveraging the structure of the variational objective and its optimality conditions, we derived efficient influence estimates that explicitly account for the presence of latent variables. Our approach generalizes prior work based on the Expectation-Maximization algorithm, overcoming its limitations by allowing for models with inexact E-step and supporting additional factorization constraints. We demonstrated that this framework can recover influence estimates for mixture models and provides a principled way to extend influence estimation to more complex settings. In future work, we plan to scale up to larger models such as mixture of experts and consider settings with hierarchical structure such as meta-learning.

References

Ahmad Beirami, Meisam Razaviyayn, Shahin Shahrampour, and Vahid Tarokh. On Optimal Generalizability in Parametric Learning. *Advances in Neural Information Processing Systems*, 2017.

Christopher M Bishop. Mixture Density Networks. Technical report, Aston University, 1994.

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, 2006.

Weiyu Cheng, Yanyan Shen, Linpeng Huang, and Yanmin Zhu. Incorporating Interpretability into Latent Factor Models via Fast Influence Analysis. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019.

R Dennis Cook. Detection of Influential Observation in Linear Regression. *Technometrics*, 1977.

R Dennis Cook and Sanford Weisberg. Characterizations of an Empirical Influence Function for Detecting Influential Cases in Regression. *Technometrics*, 1980.

R Dennis Cook and Sanford Weisberg. *Residuals and Influence in Regression*. Chapman and Hall, 1982.

Frank Critchley. Influence in principal components analysis. *Biometrika*, 1985.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum Likelihood from Incomplete Data via the EM algorithm. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 1977.

Vitaly Feldman and Chiyuan Zhang. What Neural Networks Memorize and Why: Discovering the Long Tail via Influence Estimation. *Advances in Neural Information Processing Systems*, 2020.

Kristian Georgiev, Joshua Vendrow, Hadi Salman, Sung Min Park, and Aleksander Madry. The Journey, Not the Destination: How Data Guides Diffusion Models. *ArXiv e-Prints*, 2023.

Satoshi Hara, Atsushi Nitanda, and Takanori Maehara. Data Cleansing for Models Trained with SGD. *Advances in Neural Information Processing Systems*, 2019.

Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Predicting Predictions from Training Data. In *International Conference on Machine Learning*, 2022.

Ziheng Jiang, Chiyuan Zhang, Kunal Talwar, and Michael C Mozer. Characterizing Structural Regularities of Labeled Data in Overparameterized Models. In *International Conference on Machine Learning*, 2021.

Mohammad Emtiyaz Khan. Variational Bayes Made Easy. In *Fifth Symposium on Advances in Approximate Bayesian Inference*, 2023.

Mohammad Emtiyaz Khan and Wu Lin. Conjugate-Computation Variational Inference: Converting Variational Inference in Non-Conjugate Models to Inferences in Conjugate Models. In *International Conference on Artificial Intelligence and Statistics*, 2017.

Mohammad Emtiyaz Khan and Håvard Rue. The Bayesian Learning Rule. *Journal of Machine Learning Research*, 2023.

Pang Wei Koh and Percy Liang. Understanding Black-Box Predictions via Influence Functions. In *International Conference on Machine Learning*, 2017.

Zhifeng Kong and Kamalika Chaudhuri. Understanding Instance-based Interpretability of Variational Auto-Encoders. *Advances in Neural Information Processing Systems*, 2021.

Radford M Neal and Geoffrey E Hinton. A View of the EM Algorithm that Justifies Incremental, Sparse, and other Variants. In *Learning in Graphical Models*. Springer, 1998.

Peter Nickl, Lu Xu, Dharmesh Tailor, Thomas Möllenhoff, and Mohammad Emtiyaz Khan. The Memory-Perturbation Equation: Understanding Model's Sensitivity to Data. In *Advances in Neural Information Processing Systems*, 2023.

Terence Orchard and Max A Woodbury. A Missing Information Principle: Theory and Applications. In *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*, 1972.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating Training Data Influence by Tracing Gradient Descent. *Advances in Neural Information Processing Systems*, 2020.

Naoyuki Terashita and Satoshi Hara. Decentralized Hyper-Gradient Computation over Time-Varying Directed Networks. *ArXiv e-prints*, 2022.

Naoyuki Terashita, Hiroki Ohashi, Yuichi Nonaka, and Takashi Kanemaru. Influence Estimation for Generative Adversarial Networks. In *International Conference on Learning Representations*, 2021.

Robert Weiss. An Approach to Bayesian Sensitivity Analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 1996.

Chih-Kuan Yeh, Joon Kim, Ian En-Hsu Yen, and Pradeep K Ravikumar. Representer Point Selection for Explaining Deep Neural Networks. *Advances in Neural Information Processing Systems*, 2018.

Hong-Tu Zhu and Sik-Yum Lee. Local influence for incomplete data models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2001.

Tongtian Zhu, Wenhao Li, Can Wang, and Fengxiang He. DICE: Data Influence Cascade in Decentralized Learning. In *International Conference on Learning Representations*, 2025.

Appendix A. Experimental details

The toy dataset is generated in the same way as described in Bishop (1994) with 500 points. The mixture density network (MDN) has a single hidden layer with 10 units and a tanh activation function. There are then separate heads for the mixing proportions, mean and standard deviation. A softmax function ensures the constraints for the mixing proportions are met, and an exponential function is used to ensure positivity for the standard deviation. The number of components is set to 3. The MDN is trained by minimizing the negative log-likelihood of the conditional mixture model penalized by a L2-regularizer. We use the Adam optimizer (default settings) with a learning rate of 10^{-2}, a weight decay of 10^{-3} and train for 1000 epochs (full-batch training).