Statistical Science
2019, Vol. 34, No. 2, 229–233
https://doi.org/10.1214/19-STS710
Main article: https://doi.org/10.1214/18-STS674

Comment: Variational Autoencoders as Empirical Bayes

Yixin Wang, Andrew C. Miller and David M. Blei

We thank Professor Efron for his informative and unifying review of empirical Bayes. In this comment, we discuss the connection between empirical Bayes and the variational autoencoder (VAE), a popular statistical inference framework in the machine learning community. We hope this connection motivates new algorithmic approaches for empirical Bayesians and gives new perspectives on VAEs for machine learners.

EMPIRICAL BAYES AND VAES

The key idea of empirical Bayes is to estimate a prior distribution from data. Consider a model where each observation is independently generated by a different, unobserved random variable. The empirical Bayesian first uses *all observations* to estimate a prior over the latent variables; she then infers these variables using the fitted prior. In this model, each latent variable is associated with only one data point. Yet, through the fitted prior, the empirical Bayesian profits by incorporating information from the entire data set into each inference.

This view of empirical Bayes reminds us of the variational autoencoder (VAE) (Kingma and Welling, 2013), an approach to approximate Bayesian inference for a particular class of latent variable models. A VAE refers to both the user-specified generative model and a strategy for approximate posterior inference. Given a dataset, a VAE simultaneously fits the *forward model* (i.e., the generative model) that describes the data and a function that approximates *Bayesian inversion* for the generative model. This inversion maps a data point to the (approximate) posterior of its associated latent variable and, crucially, it is constructed from the entire data set. Below we show that a VAE approximates one form of empirical Bayes inference: in Efron's language, it performs g-modeling with a particular parametric form of g.

THE POSTERIOR INFERENCE PROBLEM

Consider a data set with n data points $\mathbf{x} = (x_1, \dots, x_n)$. Each data point x_i is independently generated from a function f_β of a latent variable z_i, where β parameterizes the function. With prior p_0 on each z_i, observation i is generated

$$z_i \overset{\text{ind}}{\sim} p_0(z_i), \tag{1}$$

$$x_i \mid z_i \overset{\text{ind}}{\sim} p\big(x_i \mid f_\beta(z_i)\big). \tag{2}$$

Assume the prior $p_0(\cdot)$ and probability kernel $p(\cdot)$ are known; for example, they may both be multivariate Gaussian with identity covariance. The form of the function $f_\beta(\cdot)$ is also known, for example, a neural network, but its parameters β are unknown. This class of generative distributions includes both linear and nonlinear factor models as special cases. The goal is to use the data to estimate the parameters β and infer the posterior of the latent variables $\mathbf{z} = (z_1, \dots, z_n)$.

The posterior is a quotient between a joint density and a marginal density; the latter takes the form of an integral,

$$\begin{aligned} p(\mathbf{z} \mid \mathbf{x}; \beta) &= \prod_{i=1}^{n} p(z_i \mid x_i; \beta) \\ &= \prod_{i=1}^{n} \frac{p_0(z_i) p(x_i \mid f_\beta(z_i))}{\int p_0(z_i) p(x_i \mid f_\beta(z_i))\, \mathrm{d}z_i}. \end{aligned}$$

When the function $f_\beta(\cdot)$ is complicated, such as a neural network, the integral in the denominator is often computationally intractable. Hence the posterior $p(\mathbf{z} \mid \mathbf{x}; \beta)$ is also intractable.

Yixin Wang is Ph.D. student, Department of Statistics, Columbia University, New York, New York 10027, USA (e-mail: yixin.wang@columbia.edu). Andrew C. Miller is Postdoctoral Research Scientist, Data Science Institute, Columbia University, New York, New York 10027, USA (e-mail: am5171@columbia.edu). David M. Blei is Professor, Department of Statistics, Department of Computer Science and Data Science Institute, Columbia University, New York, New York 10027, USA (e-mail: david.blei@columbia.edu).

THE VARIATIONAL AUTOENCODER (VAE)

A VAE is an approach to fit the model parameters β and to approximate the intractable posteriors $p(\mathbf{z} \mid \mathbf{x}; \beta)$. The intractable per-data posterior is approximated with a parametric distribution, for example, a Gaussian, whose parameter is a *function* of the associated data point. This function is learned using the entire data set.

Specifically, a VAE approximates the posterior of $\mathbf{z}$ with variational Bayes (VB) (Jordan et al., 1999, Wainwright et al., 2008, Blei, Kucukelbir and McAuliffe, 2017), which casts posterior inference as an optimization problem. In VB, we first posit a family of distributions $\mathcal{Q}$ on the latent variables $\mathbf{z}$. We then find the member of $\mathcal{Q}$ within this family that is closest to the exact posterior $p(\mathbf{z} \mid \mathbf{x}; \beta)$.

A VAE posits a particular *conditional form* for the family of distributions $\mathcal{Q}^n$ for $\mathbf{z}$,

$$\mathcal{Q}^n = \left\{ q_\phi(\mathbf{z}) = \prod_{i=1}^{n} q_\phi(z_i \mid x_i) : \phi \in \Phi \right\}. \tag{3}$$

The form of $q_\phi(\cdot)$ is known but the parameters ϕ are free in its domain Φ. For example, $q_\phi(z_i \mid x_i) = \mathcal{N}(z_i; h_\phi(x_i), I)$, where $h_\phi(\cdot)$ is a neural network with parameters ϕ. This family is called the *recognition model*.[1] It represents a factorizable joint of the latent variables $\mathbf{z}$; the marginal distribution of each latent variable z_i is a function of its associated data point x_i.

A VAE seeks the member within this family that is closest to the exact posterior in Kullback–Leibler (KL) divergence. Its goal is to optimize the parameter ϕ of $q_\phi(\cdot)$ given β:

$$\phi^*(\beta) = \arg\min_{\phi} \mathrm{KL}\big(q_\phi(\mathbf{z}) \parallel p(\mathbf{z} \mid \mathbf{x}; \beta)\big). \tag{4}$$

This closest member $q_{\phi^*(\beta)}(\mathbf{z})$ is the approximate posterior of $\mathbf{z}$.

However, computing the KL divergence to the exact posterior is typically intractable. In practice, the VAE optimizes an alternative objective, the *evidence lower bound* (ELBO),

$$\phi^*(\beta) = \arg\max_{\phi} \mathrm{ELBO}\big(q_\phi(\mathbf{z}), \mathbf{x}; \beta\big),$$

where

$$\begin{aligned} &\mathrm{ELBO}\big(q_\phi(\mathbf{z}), \mathbf{x}; \beta\big) \\ &\quad \triangleq \mathbb{E}_{q_\phi(\mathbf{z})}\big[\log p(\mathbf{x}, \mathbf{z}; \beta)\big] - \mathbb{E}_{q_\phi(\mathbf{z})}\big[\log q_\phi(\mathbf{z})\big]. \end{aligned} \tag{5}$$

Maximizing the ELBO is equivalent to minimizing the KL divergence because

$$\begin{aligned} &\mathrm{ELBO}\big(q_\phi(\mathbf{z}), \mathbf{x}; \beta\big) \\ &\quad = p(\mathbf{x}) - \mathrm{KL}\big(q_\phi(\mathbf{z}) \parallel p(\mathbf{z} \mid \mathbf{x}; \beta)\big). \end{aligned} \tag{6}$$

The ELBO is easier to maximize (Blei, Kucukelbir and McAuliffe, 2017).

Finally, a VAE estimates the parameters β by optimizing the ELBO over β:

$$\beta^* = \arg\max_{\beta} \mathrm{ELBO}\big(q_{\phi^*(\beta)}(\mathbf{z}), \mathbf{x}; \beta\big).$$

Equivalently, a VAE jointly maximizes the parameters β and the approximate posteriors $q_\phi(\mathrm{z})$:

$$\beta^*, \phi^* = \arg\max_{\beta,\phi} \mathrm{ELBO}\big(q_\phi(\mathbf{z}), \mathbf{x}; \beta\big).$$

THE VAE AS EMPIRICAL BAYES

By rewriting the ELBO objective for a VAE, we find that the VAE approximates the g-modeling approach to empirical Bayes when the number of data points n is large.

Efron described the empirical Bayes setup of g-modeling. Each data point x_i is modeled with a latent variable θ_i and a known probability kernel $p(x \mid \theta)$; further assume that each latent variable θ_i is independently drawn from some hidden prior $g(\theta)$,

$$\theta_i \overset{\text{iid}}{\sim} g(\theta_i),$$
$$x_i \mid \theta_i \sim p(x_i \mid \theta_i).$$

Consider a class $\mathcal{G}$ of prior $g \in \mathcal{G}$, for example, all the Gaussian densities, or even all densities. Empirical Bayes estimates this prior $g(\theta)$ by maximizing the marginal likelihood within this class,

$$\begin{aligned} \hat{g} &= \arg\max_{g \in \mathcal{G}} L\big(\mathbf{x}; g(\boldsymbol{\theta})\big) \\ &= \arg\max_{g \in \mathcal{G}} \prod_{i=1}^{n} \int g(\theta_i) p(x_i \mid \theta_i)\,\mathrm{d}\theta_i. \end{aligned}$$

We emphasize this *hidden* prior $g(\theta)$ is different from the *known* prior $p_0(\cdot)$ we posit in the posterior inference problem (equations (1) and (2)). Empirical Bayes finally infers each latent variable θ_i using this estimate prior $\hat{g}(\theta)$,

$$p(\theta_i \mid x_i) = \frac{\hat{g}(\theta_i) p(x_i \mid \theta_i)}{\int \hat{g}(\theta_i) p(x_i \mid \theta_i)\,\mathrm{d}\theta_i}.$$

Taking $\theta_i = f_\beta(z_i)$ and $p(x_i \mid \theta_i) = p(x_i \mid f_\beta(z_i))$ (i.e., the same probability kernel as in equation (2)), a VAE turns out to approximate empirical Bayes with a particular class of $\mathcal{G}$.

[1] It is also referred to as the *encoder* or the *amortized variational family*.

THEOREM 1 (VAE as empirical Bayes). *Bayesian posterior inference via* VAE *approximates the g-modeling of empirical Bayes up to a constant: As the number of data points $n \to \infty$,*

$$(7)\qquad \frac{1}{n}\mathrm{ELBO}\big(q_\phi(\mathbf{z}), \mathbf{x}; \beta\big) - \left[\frac{1}{n}\log L\big(\mathbf{x}; g_\beta(\boldsymbol{\theta})\big) - h(\phi, \beta)\right] \overset{\text{a.s.}}{\to} 0,$$

where each latent variable θ_i is a function of the latent variable z_i:

$$\theta_i = f_\beta(z_i);$$

the prior $g_\beta(\cdot)$ belongs to a class $\mathcal{G}$ parametrized by β:

$$(8)\qquad g_\beta(\theta) = p_0\big(f_\beta^{-1}(\theta)\big) \cdot \big|\big(f_\beta^{-1}\big)'(\theta)\big|.$$

The term $h(\phi, \beta)$ takes the form

$$(9)\quad h(\phi, \beta) = \mathbb{E}_{x_i}\big[\mathrm{KL}\big(q_\phi(z_i \mid x_i) \,\|\, p(z_i \mid x_i; \beta)\big)\big];$$

this term $h(\phi, \beta)$ decreases to zero as we increase the flexibility of q_ϕ. (The expectation in equation (9) is taken over the population distribution of the data x_i.)

(The proof of Theorem 1 relies on rewriting the ELBO and invoking the strong law of large numbers. The full proof is in the Appendix.)

Theorem 1 implies that Bayesian posterior inference with VAE is essentially performing g-modeling up to a term $h(\phi, \beta)$. It optimizes the prior (i.e., the g function) on the latent variables θ_i. Its prior takes a particular parametric form as in equation (8), which involves the parameters β of the generative model (equations (1) and (2)). Though parametric, the form of this prior can be quite flexible; for instance, the function f_β can be a neural network.

The difference between the VAE objective and the empirical Bayes objective is the term $h(\phi, \beta)$; it is the price a VAE pays for *approximating* the posterior. This term $h(\phi, \beta)$ depends on both the parameters β of the generative model (equations (1) and (2)) and the parameters ϕ of the approximate posteriors $q_\phi(\mathrm{z})$, that is, the recognition model (equation (3)); it decreases to zero as we increase the capacity of the recognition model $q_\phi(\cdot)$ (Wang and Blei, 2018). A VAE more closely approximates empirical Bayes as its recognition model $q_\phi(\cdot)$ becomes more flexible.

Finally, we note this connection to empirical Bayes (Theorem 1) is specific to VAE, where the approximate posterior of each $q_\phi(z_i \mid x_i)$ shares the same parameter ϕ; it does not apply to other VB methods like mean-field variational Bayes.

EMPIRICAL BAYES WITH HIGH-DIMENSIONAL DATA

One of the main advantages of VAEs is computational tractability with high-dimensional data. Can we leverage the connection between VAE and empirical Bayes to facilitate the computation of empirical Bayes?

- *High-dimensional g-modeling with* VAE. Jiang and Zhang (2009) showed excellent empirical performance of nonparametric g-modeling with one-dimensional binary x_i's. However, g-modeling is still computationally prohibitive for high-dimensional x_i and θ_i. The computational bottleneck lies in computing and maximizing the marginal likelihood:

$$(10)\quad \lambda^* = \arg\max_{\lambda} \sum_{i=1}^{n} \log \int g_\lambda(\theta_i) p(x_i \mid \theta_i)\, \mathrm{d}\theta_i,$$

where λ is the parameter of the prior $g_\lambda(\cdot)$. Computing this log marginal likelihood (equation (10)) is often intractable with a complicated $g_\lambda(\cdot)$ and high-dimensional θ_i's.

With high-dimensional θ_i, we can optimize λ^* with a VAE approximation to this integral. In particular, we solve

$$(11)\quad \hat{\lambda}^* = \arg\max_{\lambda}\max_{\phi} \sum_{i=1}^{n} \Big[\log \int g_\lambda(\theta_i) p(x_i \mid \theta_i)\, \mathrm{d}\theta_i - \mathrm{KL}\big(q_\phi(\theta_i \mid x_i) \,\|\, p_\lambda(\theta_i \mid x_i; \lambda)\big)\Big],$$

where $p_\lambda(\theta_i \mid x_i; \lambda) \propto g_\lambda(\theta_i) p(x_i \mid \theta_i)$ is the posterior distribution implied by the $g_\lambda(\cdot)$ prior. The function $q_\phi(\cdot)$ is a probability density function as in the recognition model of VAE (equation (3)). The optimization objective in equation (11) resembles the ELBO objective in equation (6); it is computationally tractable with high-dimensional θ_i (Kingma and Welling, 2013). Moreover, this VAE approximation is exact when the function $q_\phi(\cdot)$ is flexible enough. In particular, we have $\lambda^* = \hat{\lambda}^*$ when $p_\lambda(\theta_i \mid x_i; \lambda^*) = q_{\hat{\phi}}(\theta_i \mid x_i)$ for some $\hat{\phi} \in \Phi$.

- *High-dimensional f-modeling with flexible density estimators.* Tweedie's formula (Efron, 2011) enables f-modeling for certain models, which allows us to infer the mean and variance of the latent variables θ_i's by directly modeling the marginal distribution of the data. In practice, f-modeling has been mostly restricted to one-dimensional data using Lindsey's method—binning the data and modeling the counts in each bin with Poisson regression.

However, binning becomes impossible with high-dimensional data.

For high-dimensional data, we can turn to alternative density estimators for f-modeling. Examples include normalizing flows (Rezende and Mohamed, 2015), bidirectional recurrent neural networks (Schuster and Paliwal, 1997, Berglund et al., 2015), neural autoregressive distribution estimators (Larochelle and Murray, 2011) and generative stochastic networks (Bengio et al., 2014). These density estimators are amenable to high-dimensional observations and use flexible deep neural networks, which approximates the nonparametric nature of Lindsey's method. These density estimators are almost everywhere differentiable. This allows us to easily compute derivatives of the log densities $(\log f(x))'$ for use within Tweedie's formulas (Efron, 2019, equation (24)) and f-modeling.

APPENDIX: PROOF OF THEOREM 1

We rewrite the VAE objective to prove Theorem 1. The VAE is fit through variational EM, where the expectation is taken over $z_i \sim q_\phi(z \mid x)$ and the maximization is over (ϕ, β). We remark that the exact posterior $p(z_i \mid x_i; \beta)$ is a function of the parameter β and only depends on x_i.

The objective of VAE is

$$\frac{1}{n}\mathrm{ELBO}\big(q_\phi(\mathbf{z}), \mathbf{x}; \beta\big) \tag{12}$$

$$\begin{aligned} &= \frac{1}{n}\sum_{i=1}^{n}\Big[\log\int p_0(z_i)p\big(x_i \mid f_\beta(z_i)\big)\,\mathrm{d}z_i \\ &\qquad - \mathrm{KL}\big(q_\phi(z_i \mid x_i) \parallel p(z_i \mid x_i; \beta)\big)\Big] \end{aligned} \tag{13}$$

$$\begin{aligned} &= \frac{1}{n}\sum_{i=1}^{n}\log\int p_0(z_i)p\big(x_i \mid f_\beta(z_i)\big)\,\mathrm{d}z_i \\ &\qquad - \frac{1}{n}\sum_{i=1}^{n}\mathrm{KL}\big(q_\phi(z_i \mid x_i) \parallel p(z_i \mid x_i; \beta)\big) \end{aligned} \tag{14}$$

$$\begin{aligned} &= \frac{1}{n}\sum_{i=1}^{n}\log\int \big[p_0\big(f_\beta^{-1}(\theta_i)\big)\cdot\big|\big(f_\beta^{-1}\big)'(\theta_i)\big| \\ &\qquad \cdot p(x_i \mid \theta_i)\big]\,\mathrm{d}\theta_i \\ &\qquad - \frac{1}{n}\sum_{i=1}^{n}\mathrm{KL}\big(q_\phi(z_i \mid x_i) \parallel p(z_i \mid x_i; \beta)\big). \end{aligned} \tag{15}$$

The first equality writes the ELBO objective as a difference between the log marginal probability and the KL divergence. The second equality collects the KL terms of all the data points. The third equality is due to a change-of-variable step: $\theta_i \triangleq f_\beta(z_i)$.

Next we rewrite the last equation (15). We first rewrite its KL term:

$$g(x_i, \phi, \beta) \triangleq \mathrm{KL}\big(q_\phi(z_i \mid x_i) \parallel p(z_i \mid x_i; \beta)\big). \tag{16}$$

This step is because the KL term is only a function of x_i, ϕ, β; the latent random variable z_i is marginalized out. We then apply the strong law of large numbers to conclude

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n}\mathrm{KL}\big(q_\phi(z_i \mid x_i) \parallel p(z_i \mid x_i; \beta)\big) \\ &\quad \xrightarrow{\text{a.s.}} \mathbb{E}_X\big[g(X_i, \phi, \beta)\big] \triangleq h(\phi, \beta). \end{aligned} \tag{17}$$

This step is because x_i's are assumed i.i.d. The expectation in $\mathbb{E}_X[g(X_i, \phi, \beta)]$ is taken over the population distribution of x. We emphasize that $h(\phi, \beta) \triangleq \mathbb{E}_X[g(X_i, \phi, \beta)] \geq 0$ is only a function of ϕ and β.

Equation (15) and equation (17) then lead to equation (7) of Theorem 1.

Equation (15) shows how the VAE connects to empirical Bayes; the VAE maximizes the marginal likelihood of the following model up to a term $h(\phi, \beta)$:

$$\theta_i \overset{\text{iid}}{\sim} p_0\big(f_\beta^{-1}(\theta_i)\big)\cdot\big|\big(f_\beta^{-1}\big)'(\theta_i)\big|, \tag{18}$$

$$x_i \mid \theta_i \sim p(x_i \mid \theta_i). \tag{19}$$

The prior on θ_i is shared across $i = 1, \ldots, n$. It is optimized because the VAE maximizes ϕ and β.

This connection of variational inference to empirical Bayes is specific to VAEs; it does not apply to mean field variational Bayes. The reason is that equation (17) requires all x_i's share the same parameter ϕ. It does not hold for mean field variational Bayes whose approximating family is

$$\mathcal{Q}^n = \Big\{q_\phi(\mathbf{z}) = \prod_{i=1}^{n} q_{\phi_i}(z_i) : \phi \in \Phi\Big\}. \tag{20}$$

with different ϕ_i's for $i = 1, \ldots, n$. In this case, the latent variable θ_i's will not share the same prior.

REFERENCES

BENGIO, Y., LAUFER, E., ALAIN, G. and YOSINSKI, J. (2014). Deep generative stochastic networks trainable by backprop. In *International Conference on Machine Learning* 226–234.

BERGLUND, M., RAIKO, T., HONKALA, M., KÄRKKÄINEN, L., VETEK, A. and KARHUNEN, J. T. (2015). Bidirectional recurrent neural networks as generative models. In *Advances in Neural Information Processing Systems* 856–864.

BLEI, D. M., KUCUKELBIR, A. and MCAULIFFE, J. D. (2017). Variational inference: A review for statisticians. *J. Amer. Statist. Assoc.* **112** 859–877. MR3671776

EFRON, B. (2011). Tweedie's formula and selection bias. *J. Amer. Statist. Assoc.* **106** 1602–1614. MR2896860

EFRON, B. (2019). Bayes, oracle Bayes, and empirical Bayes. *Statist. Sci.* **34** 177–201.

JIANG, W., ZHANG, C.-H. et al. (2009). General maximum likelihood empirical Bayes estimation of normal means. *Ann. Statist.* **37** 1647–1684. MR2533467

JORDAN, M. I., GHAHRAMANI, Z., JAAKKOLA, T. S. and SAUL, L. K. (1999). An introduction to variational methods for graphical models. *Mach. Learn.* **37** 183–233.

KINGMA, D. P. and WELLING, M. (2013). Auto-encoding variational Bayes. Arxiv preprint. Available at arXiv:1312.6114.

LAROCHELLE, H. and MURRAY, I. (2011). The neural autoregressive distribution estimator. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* 29–37.

REZENDE, D. J. and MOHAMED, S. (2015). Variational inference with normalizing flows. Arxiv preprint. Available at arXiv:1505.05770.

SCHUSTER, M. and PALIWAL, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **45** 2673–2681.

WAINWRIGHT, M. J., JORDAN, M. I. et al. (2008). Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.* **1** 1–305.

WANG, Y. and BLEI, D. M. (2018). Frequentist consistency of variational Bayes. *J. Amer. Statist. Assoc.* 1–15.