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.

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 POSTERIOR INFERENCE PROBLEM

Consider a data set with n data points $\mathbf{x} = (x_1, \ldots, x_n)$. 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 $\mathbf{z} = (z_1, \ldots, z_n)$.

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.