NIASRA
NATIONAL INSTITUTE FOR APPLIED
STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

University of Wollongong, Australia

Working Paper

07-22

Nonparametric Empirical Bayes Prediction

Noel Cressie

National Institute for Applied Statistics Research Australia, University of Wollongong,
Wollongong NSW 2522, Australia Phone: +61 2 42215076, Fax: +61 2 42214998.
Email: karink@uow.edu.au

Nonparametric Empirical Bayes Prediction

Noel Cressie[*]
School of Mathematics and Applied Statistics, University of Wollongong
NSW 2522, Australia
(ncressie@uow.edu.au)

Abstract

This is an invited discussion of the *Journal of the American Statistical Association* 2022 Theory and Methods Invited Paper, "Nonparametric Empirical Bayes Analysis" by Nikolaos Ignatiadis and Stefan Wager. Their statistical model is Bayesian: $z \sim p(\cdot \mid \mu)$ and $\mu \sim G$. The paper concentrates on estimation of the optimal predictor, $\theta_G(z) = E(h(\mu) \mid z)$. Uncertainty quantification is important to a Bayesian; here, the posterior variance and prediction intervals should be of particular interest.

Keywords: Bayesian hierarchical model; loss function; mean-squared prediction error; posterior distribution; MSPE

1 Introduction

Henceforth, the *Journal of the American Statistical Association* 2022 Theory and Methods Invited Paper by Nikolaos Ignatiadis and Stefan Wager, "Nonparametric Empirical Bayes Analysis," will be referred to as IW. My thanks to the editors for the opportunity to discuss it and to the authors for their new look at nonparametric empirical Bayes (NEB). NEB can be traced back to Robbins (1956), who saw that marginal probabilities contain important information on nonparametric prior distributions, which could be used for inference.

Rewriting the model (1) in IW (hereafter IW-(1)), we have parameter μ, datum z, data model $[z \mid \mu]$, and parameter model $[\mu]$ (equivalently written as prior distribution function

[*] This research was partially supported by the Australian Research Council Discovery Project DP190100180.

G). Hence the joint distribution is $[z, \mu] = [z \mid \mu][\mu]$, where for any random quantities A and B, $[A, B]$ represents the joint distribution of A and B, $[A \mid B]$ represents the conditional distribution of A given B, and $[B]$ represents the marginal distribution of B. Robbins (1956) considered a Poisson data model,

$$[z \mid \mu] = \frac{e^{-\mu}\mu^z}{z!} I(z \in \{0, 1, ...\})\,; \mu > 0,$$

where $I(C) = 1$ if the event C is true and $= 0$ otherwise. To make inference on μ based on the datum z, he used a $\delta^*(z)$ that minimizes the unconditional risk:

$$R_G(\delta) \equiv E(L(\delta(z), \mu)), \tag{1}$$

where dependence of the risk on the prior G is emphasized. The expectation in (1) is taken over $[z, \mu]$, and L is the squared-error loss function, $L(\delta(z), \mu) = (\delta(z) - \mu)^2$. Hence, minimizing $R_G(\delta)$ results in

$$\delta^*_G(z) = E(\mu \mid z) \equiv \theta_G(z), \tag{2}$$

where again dependence on G is emphasized, and it is convenient to use interchangeably the notation G and $[\mu]$ for the prior and $\delta^*_G(\cdot)$ and $\theta_G(\cdot)$ for the optimal (i.e., Bayes) predictor. In the same vein as (1), notate the marginal distribution, $[z]$, as:

$$p_G(z) = \int [z \mid \mu][\mu]\mathrm{d}\mu = \int [z \mid \mu]\mathrm{d}G(\mu); z \in \{0, 1, ...\}.$$

Robbins (1956) showed that for $[z \mid \mu]$ Poisson, the Bayes predictor is

$$\delta^*_G(z) = (z + 1)\frac{p_G(z + 1)}{p_G(z)};\ z \in \{0, 1, ...\}.$$

Hence, in this case, the marginal distribution $[z]$ contains the right amount of information on the nonparametric prior G to be able to optimally predict μ when z is observed. Robbins (1956) gave related results for the geometric distribution and the binomial distribution.

Cressie (1982a) derived a more general result (given by (3) and (4) below) and applied it to the Sichel distribution, Fisher's logarithmic series distribution, the Riemann distribution, the Good distribution, the negative binomial distribution (Cressie, 1982b), the Rasch distribution (Cressie and Holland, 1983), and the hypergeometric distribution (Cressie and Seheult, 1985). Suppose $\mathcal{L}$ is a linear functional with domain contained in the space of all real-valued functions. Then under mild regularity conditions, Cressie (1982a) showed that

$$\frac{\mathcal{L}(p_G(z))}{p_G(z)} = E\left(\frac{\mathcal{L}([z \mid \mu])}{[z \mid \mu]} \;\middle|\; z\right), \tag{3}$$

where the expectation is taken with respect to $[\mu \mid z]$.

Define the functional,

$$\mathcal{M}(f(\cdot)) \equiv \frac{\mathcal{L}(f(\cdot))}{f(\cdot)}.$$

Consider the general NEB model given by IW-(1). Then often,

$$\mathcal{M}([z \mid \mu]) = u(z)h(\mu) + v(z); u(\cdot) > 0, \tag{4}$$

where $h(\mu)$ is a known real function of μ. With a slight abuse of notation, we shall replace μ in (1) with $h(\mu)$, and write

$$\delta_G^*(z) = E(h(\mu) \mid z) \equiv \theta_G(z), \tag{5}$$

which is in line with IW-(1). Consequently,

$$\theta_G(z) = E(h(\mu) \mid z) = \{\mathcal{M}(p_G(z)) - v(z)\}/u(z).$$

For example, for the Poisson distribution considered by Robbins (1956), and if $\mathcal{L}(f(z)) = f(z+1); z \in \{0, 1, ...\}$, then $\mathcal{M}([z \mid \mu]) = \mu/(z+1)$. That is, $h(\mu) = \mu$, $u(z) = 1/(z+1)$, $v(z) = 0$, and $\delta_G^*(z) = (z+1)p_G(z+1)/p_G(z); z \in \{0, 1, ...\}$, which is the result that Robbins (1956) obtained.

The general result based on $\mathcal{M}$ has a lot of flexibility, since $\mathcal{L}$ can be any linear functional. The papers by Cressie in the 1980s (cited above), represent an exploration of this approach for many different data models $[z \mid \mu]$ and linear functionals $\mathcal{L}$, particularly for discrete distributions but also for continuous ones (Maritz and Lwin, 1975; Cressie, 1982a).

IW present two approaches to the construction of confidence intervals for μ, namely F-localization and AMARI. AMARI depends on writing the posterior mean as a ratio of two linear functionals of G. Define $a_G(z) \equiv \int h(\mu)[z \mid \mu]\mathrm{d}G(\mu)$; then straightforwardly, $\theta_G(z) = a_G(z)/p_G(z)$, and the null hypothesis, $H_0: \ \theta_G(z) = c$, is equivalent to the null hypothesis, $H_0: \ a_G(z) - c \cdot p_G(z) = 0$. This equivalence is used to obtain AMARI confidence intervals for $\theta_G(z)$ through constructing confidence intervals on the linear functional, $a_G(z) - c p_G(z)$, linear in G. This is an approach that can be taken with any data model $[z \mid \mu]$.

Now, when the identity (4) holds, which it does for many data models,

$$\theta_G(z) = \left\{ \frac{\mathcal{L}(p_G(z))}{p_G(z)} - v(z) \right\} \Big/ u(z),$$

and $H_0: \ \theta_G(z) = c$ is equivalent to

$$H_0: \ \mathcal{L}(p_G(z)) - v(z) p_G(z) - c \cdot u(z) = 0.$$

Notice that the linearity of the functional $\mathcal{L}(\cdot)$ yields a linear functional $\mathcal{L}(p_G(z))$ in G. *Question:* Can this structure be exploited when constructing confidence intervals, through either AMARI or through F-localization?

2 The posterior distribution is key

Section 1 describes how NEB has focussed on the posterior mean, $\theta_G(z) = E(h(\mu) \mid z)$, where the prior $[\mu]$ is of unknown form, represented by a nonparametric distribution function G. The marginal distribution, $[z]$, depends on G, so that independent draws $z_1, ..., z_n$

may be used to infer "something" about G. This is an important departure point for NEB analysis compared to Bayesian analysis; NEB's principal goal has been to estimate $\theta_G(z)$ (via point estimation or interval estimation) from the sample $z_1, ..., z_n$.

A Bayesian analysis is based on Bayes' Theorem, which makes it clear that the whole posterior distribution is needed:

$$[\mu \mid z] = \frac{[z \mid \mu][\mu]}{[z]},$$

not just one moment, $E(h(\mu) \mid z)$. To understand the behavior of a probability distribution, at the very least one should summarise it with a measure of its center and of its variability. For the parameter of interest $h(\mu)$, that would be $E(h(\mu) \mid z)$, and

$$\text{var}(h(\mu) \mid z) = E(h(\mu)^2 \mid z) - (\theta_G(z))^2,$$

so in principle the NEB machinery is there to estimate it as well.

Statistical inference can be broadly classified into estimation, prediction, and attribution (Efron, 2020). Under squared-error loss, the mis-named "Bayes estimator" of $h(\mu)$ in IW, namely $E(h(\mu) \mid z)$, should be re-named the *Bayes predictor*. It is important to keep "prediction" (inference on a random quantity, here $h(\mu)$) separate from "estimation" (inference on a fixed but unknown parameter, for example $\theta_G(z)$).

Uncertainty quantification for prediction is different than it is for estimation, although often they are both based on squared error (or even absolute error): Suppose $\delta(z)$ is a predictor of $h(\mu)$; then the squared prediction error is $(\delta(z) - h(\mu))^2$, and hence the mean-squared prediction error (MSPE) is:

$$E(\delta(z) - h(\mu))^2 = R_G(\delta), \tag{6}$$

the unconditional risk that, when minimized with respect to δ, yields the Bayes predictor, (5). Then the *minimized* MSPE is:

$$E(\delta_G^*(z) - h(\mu))^2 = E\big\{E\big((h(\mu) - \delta_G^*(z))^2 \mid z\big)\big\} = E\{\mathrm{var}(h(\mu) \mid z)\}\,,$$

showing the importance of the posterior variance, $\mathrm{var}(h(\mu) \mid z)$, for predictive inference. Moreover, $\mathrm{var}(h(\mu) \mid z)$ is simply the conditional risk, $E((\delta_G^*(z) - h(\mu))^2 \mid z)$.

Inference for prediction extends to intervals: A $100(1-\alpha)\%$ *prediction interval* $I_\alpha(z)$ for $h(\mu)$ (e.g., $\alpha = 0.10$) has the property that

$$\Pr(h(\mu) \in I_\alpha(z)) = 100(1-\alpha)\%; 0 < \alpha < 1,$$

where $\Pr(C)$, the probability of the event C, is calculated here with respect to $[\mu, z]$. Further, a $100(1-\alpha)\%$ *conditional prediction interval* $I_\alpha^c(z)$ for $h(\mu)$ satisfies,

$$\Pr(h(\mu) \in I_\alpha^c(z) \mid z) = 100(1-\alpha)\%; 0 < \alpha < 1,$$

where $\Pr(C \mid z)$ is calculated with respect to $[\mu \mid z]$. One would expect that conditional prediction intervals are narrower (and hence preferable) than their unconditional counterparts, but this is not always the case. Let $q(z;\alpha)$ denote the α-th quantile of $[\mu \mid z]$ for $0 \leq \alpha \leq 1$. Then a 90% conditional prediction interval is easily obtained as $I_{0.1}^c = (q(z; 0.05), q(z; 0.95))$, although clearly it is not the only one.

This discussion brings out an important feature of empirical Bayesian inference: The posterior distribution, or a selection of its moments and quantiles, are typically not known well enough to compute them. When the prior distribution is parametric, for example can be written as $G(\phi)$ for a finite collection of parameters ϕ, then $[\mu \mid z, \phi]$ and $[z \mid \phi]$ depend on ϕ. Parametric empirical Bayes analysis would estimate ϕ, which is often achieved via maximum-likelihood estimation; that is, the estimator based on $z_1, ..., z_n$ is $\hat{\phi} \equiv \arg\sup_\phi \prod_{i=1}^n [z_i \mid \phi]$. Now, the Bayes predictor is $E(h(\mu) \mid z) = \theta_{G(\phi)}(z)$, which is usually estimated as $\theta_{G(\hat{\phi})}(z)$, following the plug-in approach.

NEB could be imagined similarly, but with an infinite number of parameters to be estimated (e.g., all the moments of G). Estimation based only on a sample from the marginal distribution looks daunting, but IW and my discussion in Section 1 show that quite remarkable inferences on $\theta_G(z)$ are possible. Nonetheless, for the purpose of predictive inference, effort in NEB analysis should be directed towards $\mathrm{var}(h(\mu) \mid z)$ and its estimation.

I shall conclude this section with a proposal for the next major step that could be taken in NEB. Uncertainty quantification for predictive inference is based on the MSPE, which we have seen is $R_G(\theta_G(z)) = E(\theta_G(z) - h(\mu))^2 = E(\mathrm{var}(h(\mu) \mid z))$, where $\theta_G(\cdot)$ is given by (5). Let $\hat{\theta}_G(\cdot)$ denote an estimator of $\theta_G(\cdot)$ depending on $z_1, ..., z_n$. Then,

$$\hat{R}_G(\hat{\theta}_G(z)) \equiv \hat{E}\Big(\hat{\theta}_G(z) - h(\mu)\Big)^2, \tag{7}$$

is an estimate of the MSPE. Notice that there are two plug-ins, one to obtain an estimator of $\theta_G(\cdot)$, and the other to estimate the probability measure that defines the expectation. In the area of geostatistics, Zimmerman and Cressie (1992) studied the bias of (7), namely $E\Big\{\hat{R}_G\Big(\hat{\theta}_G(z)\Big) - R_G(\theta_G(z))\Big\}$, where the expectation is taken over the sample $z_1, ..., z_n$. In small area estimation, Prasad and Rao (1990) gave approximations to $E(\theta_G(z) - h(\mu))^2$ based on large-sample corrections to $\hat{R}_G(\hat{\theta}_G(z))$.

Question: Do the authors think that their inference results for estimating $\theta_G(\cdot)$ can help to address uncertainty quantification results for predicting $h(\mu)$?

3 Other models and other loss functions

There has been a lot of research into empirical Bayes analysis where observations, $z_1, ..., z_n$, are vectors, and μ is a vector of latent effects. For example, in plant breeding (e.g., Harville and Jeske, 1992), small area estimation (e.g., Prasad and Rao, 1990), and geostatistics (e.g., Cressie, 1993), the class of predictors is restricted to be linear in the vector z and unbiased, resulting in the Bayes predictor $\delta^*_G(z)$ being referred to as the *BLUP* (*Best Linear Unbiased Predictor*). Here G, the joint distribution of the vector μ, is often specified *semi-parametrically*, with assumptions made about G's first two moments, $E(\mu)$ and $\mathrm{cov}(\mu)$. Often, the data model $[z \mid \mu]$ is assumed Gaussian, but the flexible family of generalized linear models has been used in predictive inference for discrete and skewed data (e.g., Sengupta and Cressie, 2013). Prediction of μ then requires an estimate of $E(\mu)$ and $\mathrm{cov}(\mu)$, which can be plugged-into the BLUP. The resulting predictor is, not surprisingly, referred to as an *EBLUP* (*Empirical BLUP*).

A number of statistical models are most naturally hierarchical, where there is a sequence of conditional-probability measures that capture the joint distribution of the data (z), the process (y, say), and the parameters (μ). Here, ever-present measurement errors are in the data model $[z \mid y, \mu]$, separated from the latent process y where the substantive science resides. A generic *Bayesian hierarchical model (BHM)* is:

$$\begin{aligned}&\text{Data model: } [z \mid y, \mu]\\ &\text{Process model: } [y \mid \mu]\\ &\text{Parameter model (i.e., the prior): } [\mu],\end{aligned}$$

where z, y, and μ could be high-dimensional random vectors, and the substantive process y might consist of interacting sub-processes. Then $[y \mid \mu]$ can be thought of as a physical-statistical model (e.g., Kuhnert, 2014) that captures within-process and between-process variability as well as scientific uncertainty (e.g., Cressie and Wikle, 2011).

The "unknowns" are y and μ, and by Bayes' Theorem their posterior distribution is,

$$[y, \mu \mid z] = \frac{[z \mid y, \mu][y \mid \mu][\mu]}{[z]}.$$

In NEB, $\mu \sim G$ where G is unknown. Then for a BHM, not only should μ be predicted, but so should y. Assuming (weighted) squared-error loss, the optimal predictors are $\theta_G(z) = E(\mu \mid z)$ and $\eta_G(z) \equiv E(y \mid z) = \int\int y[z \mid y, \mu][y \mid \mu][\mu] \; \mathrm{d}y\mathrm{d}G(\mu)\Big/[z]$, and recall that an alternative notation for $[z]$ is $p_G(z) = \int\int[z \mid y, \mu][y \mid \mu][\mu] \; \mathrm{d}y\mathrm{d}G(\mu) = \int[z \mid \mu][\mu] \; \mathrm{d}G(\mu)$. Then, from (3) in Section 1,

$$\mathcal{M}_1(p_G(z)) = E(\mathcal{M}_1([z \mid \mu]) \mid z),$$

where the expectation is taken with respect to $[\mu \mid z]$, and $\mathcal{M}_1(f) \equiv \mathcal{L}_1(f)/f$ for linear functional $\mathcal{L}_1$. Furthermore,

$$\mathcal{M}_2(p_G(z)) = E(\mathcal{M}_2([z \mid y, \mu] \mid z)),$$

where the expectation is taken with respect to $[y, \mu \mid z]$, and $\mathcal{M}_2(f) \equiv \mathcal{L}_2(f)/f$ for linear functional $\mathcal{L}_2$. This generalises NEB analysis to BHMs and relies on being able to find $\mathcal{L}_1$ and $\mathcal{L}_2$ and hence $\mathcal{M}_1(p_G(\cdot))$ and $\mathcal{M}_2(p_G(\cdot))$ that together yield expressions for $\theta_G(z)$ and $\eta_G(z)$ in terms of the marginal probabilities $p_G(\cdot)$.

Another area of generalization is to replace squared-error loss with other loss functions, leading to other optimal predictors such as the posterior mode, the posterior median and, more generally, posterior quantiles. Quantiles can be derived from asymmetric loss functions (except for the median, whose loss function is symmetric) that are not differentiable at the origin. A differentiable asymmetric loss function is the linex loss function (Zellner, 1986),

$$L(\delta(z), \mu) = e^{\{a(\delta(z)-h(\mu))\}} - a(\delta(z) - h(\mu)) - 1,$$

where $a < 0$ is used when the losses incurred due to under-prediction (i.e., predictor $\delta < h(\mu)$) are greater than the losses incurred due to over-prediction (i.e., predictor $\delta > h(\mu)$). For example, suppose $h(\mu)$ is the peak water level in a river town during a flood;

under-prediction would lead to loss of homes, businesses, and even lives, but the loss of over-prediction would be expenditure on preparations for flooding (e.g., constructing levees, filling sandbags, and so forth). For IW-(1) and the linex loss function, the Bayes predictor is $\delta^*_G(z) = (-1/a)\log\left\{E\left(e^{-ah(\mu)} \mid z\right)\right\}$, which is greater than $\theta_G(z) = E(h(\mu) \mid z)$, for $a < 0$. Regardless of the loss function chosen, the key quantity is the posterior distribution $[\mu \mid z]$, from which prediction and uncertainty quantification follows.

4 Concluding remarks

NEB analysis has largely followed the path set out by Robbins (1956), namely for classes of data models, $[z \mid \mu]$, find expressions for Bayes predictors that depend only on the marginal distribution, and give point estimates of them. Subsequent results have been established for new classes of data models and/or for predicting a known function, $h(\mu)$, of μ. I participated in this stream in the 1980s, with encouragement from Herbert Robbins, although my involvement began several years earlier while a PhD student at Princeton University, when Fred Lord and I worked on producing confidence intervals for the success parameter $0 < \mu < 1$ in the binomial distribution (Lord and Cressie, 1975). Interval estimation has been given much less attention, until now with the paper under discussion.

Concentrating on estimators, estimation variances, and confidence intervals for $\theta_G(z)$ takes a frequentist viewpoint in a Bayesian model. I believe that it is most important to seek the posterior distribution $[\mu \mid z]$. From that, the posterior distribution $[h(\mu) \mid z]$, for known h, is readily available, as is $\theta_G(z) = E(h(\mu) \mid z)$ and $\mathrm{var}(h(\mu) \mid z)$. Since the target quantity $h(\mu)$ is random, the appropriate measure of uncertainty is the mean-squared prediction error, $E(\theta_G(z) - h(\mu))^2$.

References

Cressie, N. (1982a). A useful empirical Bayes identity. *Annals of Statistics 10*, 625–629.

Cressie, N. (1982b). Empirical Bayes estimation for discrete distributions. *South African Journal of Statistics 16*, 25–37.

Cressie, N. (1993). *Statistics for Spatial Data* (rev. ed.). New York, NY: Wiley.

Cressie, N. and P. W. Holland (1983). Characterizing the manifest probabilities of latent trait models. *Psychometrika 48*, 129–141.

Cressie, N. and A. Seheult (1985). Empirical Bayes estimation in sampling inspection. *Biometrika 72*, 451–458.

Cressie, N. and C. K. Wikle (2011). *Statistics for Spatio-Temporal Data* . Hoboken, N.J: Wiley.

Efron, B. (2020). Prediction, estimation, and attribution (with discussion). *Journal of the American Statistical Association 115*, 636–655.

Harville, D. A. and D. R. Jeske (1992). Mean squared error of estimation or prediction under a general linear model. *Journal of the American Statistical Association 87*, 724–731.

Kuhnert, P. (2014). Physical-statistical modelling. *Environmetrics 25*, 201–202.

Lord, F. M. and N. Cressie (1975). An empirical Bayes procedure for finding an interval estimate. *Sankhyā 39*, 1–9.

Maritz, J. S. and T. Lwin (1975). Construction of simple empirical Bayes estimators. *Journal of the Royal Statistical Society, Series B 39*, 421–425.

Prasad, N. G. N. and J. N. K. Rao (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association 85*, 163–171.

Robbins, H. (1956). An empirical Bayes approach to statistics. In *Proceedings of the Third Symposium on Mathematical Statistics and Probability, Vol. 1: Contributions to the Theory of Statistics*, Berkeley, CA: University of California.

Sengupta, A. and N. Cressie (2013). Hierarchical statistical modeling of big spatial datasets using the exponential family of distributions. *Spatial Statistics 4*, 14–44.

Zellner, A. (1986). Bayesian estimation and prediction using asymmetric loss functions. *Journal of the American Statistical Association 81*, 446–451.

Zimmerman, D. and N. Cressie (1992). Mean squared prediction error in the spatial linear model with estimated covariance parameters. *Annals of the Institute of Statistical Mathematics 44*, 27–43.