Math. Appl. **8** (2019), 189–209
DOI: 10.13164/ma.2019.12

ON THE UBIQUITY OF THE BAYESIAN PARADIGM IN STATISTICAL MACHINE LEARNING AND DATA SCIENCE

ERNEST FOKOUÉ

Abstract. This paper seeks to provide a thorough account of the ubiquitous nature of the Bayesian paradigm in modern statistics, data science and artificial intelligence. Once maligned, on the one hand by those who philosophically hated the very idea of subjective probability used in prior specification, and on the other hand because of the intractability of the computations needed for Bayesian estimation and inference, the Bayesian school of thought now permeates and pervades virtually all areas of science, applied science, engineering, social science and even liberal arts, often in unsuspected ways. Thanks in part to the availability of powerful computing resources, but also to the literally unavoidable inherent presence of the quintessential building blocks of the Bayesian paradigm in all walks of life, the Bayesian way of handling statistical learning, estimation and inference is not only mainstream but also becoming the most central approach to learning from the data. This paper explores some of the most relevant elements to help to the reader appreciate the pervading power and presence of the Bayesian paradigm in statistics, artificial intelligence and data science, with an emphasis on how the Gospel according to Reverend Thomas Bayes has turned out to be the truly good news, and in some cases the amazing saving grace, for all who seek to learn statistically from the data.

1. Introduction

One of the quintessential building blocks of any data science activity is the explicit or implicit consideration of a dataset, $\mathcal{D}_n = \{(\mathbf{x}_1, \mathrm{y}_1), (\mathbf{x}_2, \mathrm{y}_2), \cdots, (\mathbf{x}_n, \mathrm{y}_n)\}$ where $(\mathbf{x}_i, \mathrm{y}_i)$ are realizations of $(X_i, Y_i) \overset{iid}{\sim} p_{\mathrm{xy}}(\mathbf{x}, \mathrm{y})$, with $X_i \in \mathcal{X}$ and $Y_i \in \mathcal{Y}$, for $i = 1, \cdots, n$. The probability density function $p_{\mathrm{xy}}(\mathbf{x}, \mathrm{y})$ defined on $\mathcal{X} \times \mathcal{Y}$ governs the stochastic mechanism (typically unknown in practice) assumed to have generated the realized data set. Now, given the dataset $\mathcal{D}_n$, one of the overarching goals of machine learning consists of estimating (learning) the patterns of dependencies/relationships between the input space $\mathcal{X}$ and the output space $\mathcal{Y}$. Learning as meant here essentially consists of constructing (searching for) mathematical mappings $f : \mathcal{X} \longrightarrow \mathcal{Y}$, that formally capture and represent the hypothesized dependencies/relationships between the elements of $\mathcal{X}$ and those of $\mathcal{Y}$. Since it turns out to be untenable in practice to search the whole power space $\mathcal{Y}^{\mathcal{X}}$ of all possible mappings from $\mathcal{X}$ to $\mathcal{Y}$, one typically has to consider a more realistic and more manageable space $\mathcal{H}$, usually with some specific properties like

MSC (2010): primary 62F15; secondary 62F07.
Keywords: Bayes' theorem, prior specification, posterior distribution, learning machine, statistical learning theory, bayesian learning, regularization, Bayes learner, data science.

differentiability, or even just continuity or compactness and so on. One typically ends up with a function space $\mathcal{H}$ where members are parameterized and even have hyperparameters, like

$$\mathcal{H} = \left\{ f : \ f(\mathbf{x}) := f(\mathbf{x}|\boldsymbol{\theta}), \ \boldsymbol{\theta} \equiv \boldsymbol{\theta}_{\boldsymbol{\gamma}} \in \boldsymbol{\Theta}, \boldsymbol{\gamma} \in \boldsymbol{\Gamma} \right\}.$$

One such function space is the space of univariate polynomials of some degree p,

$$\mathcal{H} = \left\{ f : \exists \boldsymbol{\theta} = (\theta_1, \cdots, \theta_p) \in \mathbb{R}^p \mid \forall \mathbf{x} \in [a,b], \ f(\mathbf{x}) = \langle \boldsymbol{\theta}, \mathbf{x} \rangle = \sum_{j=1}^{p} \theta_j \mathbf{x}^j \right\}.$$

An overwhelming number of other function spaces exist, some of which will be mentioned in subsequent paragraphs. However, before we even elaborate any further on function spaces, let's recall that the goal of this paper is to draw the attention of the reader on the ubiquity in data science and statistical machine learning, of the Bayesian paradigm and the Bayesian school of thought. The prime reason for such a prevalence of the Bayesian machinery throughout statistical machine learning, resides in the fact that when a mathematical object like $f(X|\boldsymbol{\theta})$ is under consideration, the Bayesian scientist appropriately treats both X and $\boldsymbol{\theta}$ as random variables by virtue of the fact that they are both unknown, and therefore need to be handled and treated with appropriately (suitably) specified probability distributions. *And as we will see throughout this paper, the mechanism that ends up allowing the scientifically rigorous treatment of the quantities of interest in statistical machine learning turns out to be the famous Bayes' Theorem or Bayes' formula.* In its most generic and canonical form, Bayes' theorem is used to connect the conditional and marginal probabilities of two events.

Theorem 1.1. *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be two events with nonzero probabilities such that* $\Pr(\mathbf{A}) > 0$ *and* $0 < \Pr(\mathbf{B}) < 1$*, then the conditional probability of* $\mathbf{B}$ *given that* $\mathbf{A}$ *has occurred, is given by*

$$\Pr(\mathbf{B}|\mathbf{A}) = \frac{\Pr(\mathbf{B})\Pr(\mathbf{A}|\mathbf{B})}{\Pr(\mathbf{A})} = \frac{\Pr(\mathbf{B})\Pr(\mathbf{A}|\mathbf{B})}{\Pr(\mathbf{B})\Pr(\mathbf{A}|\mathbf{B}) + \Pr(\mathbf{B}^c)\Pr(\mathbf{A}|\mathbf{B}^c)}. \tag{1.1}$$

An extension deals with a collection $\mathbf{B}_1, \cdots, \mathbf{B}_K \in \Omega$ of mutually exclusive events, and their probabilistic relationship with some event $\mathbf{A} \in \Omega$.

Theorem 1.2. *Let* $\mathbf{A} \in \Omega$ *be an event with nonzero probability such that* $\Pr(\mathbf{A}) > 0$*, and consider the collection of mutually exclusive events* $\mathbf{B}_1, \cdots, \mathbf{B}_K \in \Omega$ *such that* $\mathbf{B}_k \cap \mathbf{B}_j = \emptyset$*,* $j \neq k$ *and* $\sum_{k=1}^{K} \Pr(\mathbf{B}_k) = 1$*, i.e.* $\bigcup \mathbf{B}_k = \Omega$*, then the conditional probability of* $\mathbf{B}_k$ *given that* $\mathbf{A}$ *has occurred, is given by*

$$\Pr(\mathbf{B}_j|\mathbf{A}) = \frac{\Pr(\mathbf{B}_j)\Pr(\mathbf{A}|\mathbf{B}_j)}{\Pr(\mathbf{A})} = \frac{\Pr(\mathbf{B}_j)\Pr(\mathbf{A}|\mathbf{B}_j)}{\sum_{k=1}^{K}\Pr(\mathbf{B}_k)\Pr(\mathbf{A}|\mathbf{B}_k)}. \tag{1.2}$$

The central tenant of the Bayesian paradigm is the concept of posterior probability of an event. For instance, $\Pr(\mathbf{B}_j|\mathbf{A})$ in (1.2) is the posterior probability of event $\mathbf{B}_j$ given $\mathbf{A}$, which measures the probability that $\mathbf{B}_j$ will occur, given that $\mathbf{A}$ has occurred. This concept of posterior probability provides a powerful mechanism for formulating, modelling and computing prediction and predictive quantities of all kinds. It is important however to emphasize that prediction here is not forecasting,

nor is it meant in the sense of causation. Prediction here is meant in the sense of dependent arising. In Bayesian parlance, $\Pr(\mathbf{B}_j)$ represents the prior belief in $\mathbf{B}_j$ before the dependent event $\mathbf{A}$ occurs, and in that sense, the posterior $\Pr(\mathbf{B}_j|\mathbf{A})$ updates the belief in $\mathbf{B}_j$ given that $\mathbf{A}$ has occurred. $\Pr(\mathbf{A})$ is referred to as the evidence by many in the Bayesian community, enjoys that appellation most appropriately in settings like Bayesian hypothesis testing where $\Pr(\mathbf{H}_0|\,\texttt{data})$ measures the probability that the null hypothesis is true given the evidence provided by the data. Indeed this concept of evidence is key for a variety of reasons. The seminal and posthumously published work of Reverend Thomas Bayes, of which a recognizable simplified form is given in Equations (1.1) and (1.2), permeates virtually every aspect of scientific analysis involving the doctrine of chance and probability. It is so rich indeed in positively transformative concepts that it won't be an exaggeration to refer to it as the *gospel according to Reverend Thomas Bayes*, judging by the sheer plurality of its applications to literally all areas of statistical and probabilistic modelling. As a matter of fact, both explicitly and implicitly, an overwhelmingly large number of the so-called learning machines in artificial intelligence, statistical machine learning or data science, admit a Bayesian formulation often directly or after simple transformations. The multiplicity of such occurrences leads one to recognize the quasi-centrality of the Bayesian paradigm in science in general. The Bayesian paradigm appears ubiquitous, permeating and pervading every scientific activity involving the doctrine of chance and statistical learning from the data. In an era marked by the resurgence of artificial intelligence and the firm establishment of statistical machine learning as a force to reckon with, along with the meteoric rise to prominence of the emerging field of data science, all of which have to deal with uncertainty at their core, it makes sense the statistics, the natural language (along with sister probability) for dealing with uncertainty, should permeate the very fabric of epistemology, theory, methodology, computation and application. Interestingly, as we will see later, the famous Bayes' theorem (Bayes' rule or Bayes' formula as some call it) stands prominently and firmly at the very core, providing a versatile, rich and powerful paradigm for modelling both the simplest and the most complex of phenomena. From the fundamental algebra of finite sets of events to the estimation of model parameters to infinite dimensional function approximation and estimation, the Bayesian paradigm seems to find a way to emerge (sometimes almost miraculously) as the de-facto flexible modelling framework for formulating and/or solving the task at hand. The goal of this paper is not to preach the Gospel according Reverend Thomas Bayes, not is it aimed at reviewing the sophisticated technical niceties of some seminal Bayesian fundamental results. Instead, our goal is to provide a general bird's eye view of the manifold incarnations of the Bayesian machinery in artificial intelligence, statistical machine learning and data science. The question still remains: *What then are some of the ways in which the Bayesian paradigm and its machinery ubiquitously permeate the landscape of statistical machine learning?* All the subsequent sections will provide detailed and elaborate answers to this question.

2. Bayes' impact in statistical learning theory

As will become clearer in the subsequent paragraphs, statistical machine learning brings with it the need to extend Bayes' theorem from events to random variables, especially with concepts of marginal density, conditional density and conditional expectation. We will see throughout that the Bayesian paradigm provides the perfect mechanism for the most fundamental results in pattern recognition, regression, hypothesis testing, signal detection, parameter estimation, function estimation and statistical learning in general.

2.1. Bayes learner in classification and regression

It turns out that one of the first blessings of the Bayesian gospel comes in the form of a mechanism for setting the gold standard in supervised learning, namely providing the definition and characterization of the theoretical best learning machines in both classification and regression. Specifically, upon choosing the desired loss function $\mathcal{L}(\cdot,\cdot)$, for measuring the discrepancy between the true output Y and the predicted output $f(X)$, the expected loss or theoretical risk or generalization error or true error of any function $f \in \mathcal{Y}^{\mathcal{X}}$ is given by

$$R(f) = \mathbb{E}[\mathcal{L}(Y, f(X))] = \int_{\mathcal{X}\times\mathcal{Y}} \mathcal{L}(y, f(\mathbf{x}))p_{\mathbf{xy}}(\mathbf{x}, y)d\mathbf{x}dy,$$

and can be interpreted as the expected discrepancy between $f(X)$ and Y, and indeed a measure of the predictive strength of f. Ideally, one seeks to find the minimizer $f^\star$ of $R(f)$ over all measurable functions $f \in \mathcal{Y}^{\mathcal{X}}$, specifically,

$$f^\star = \operatorname*{arginf}_{f\in\mathcal{Y}^{\mathcal{X}}} \big\{R(f)\big\} = \operatorname*{arginf}_{f\in\mathcal{Y}^{\mathcal{X}}} \Big\{\mathbb{E}[\mathcal{L}(Y, f(X))]\Big\}.$$

For classification learning under the so-called zero-one loss defined as

$$\mathcal{L}(Y, f(X)) = \mathbb{1}(Y \neq f(X)),$$

the universal best classifier $f^\star$ is appropriately called the Bayes' classifier because it is defined via the posterior probability of class membership. Specifically, $\forall \mathbf{x} \in \mathcal{X}$,

$$f^\star(\mathbf{x}) = \operatorname*{argmax}_{c\in\mathcal{Y}} \{\texttt{Prob}(Y = c|\mathbf{x})\} = \operatorname*{argmax}_{c\in\mathcal{Y}} \left\{\frac{\texttt{Prob}(Y = c)p(\mathbf{x}|c)}{p(\mathbf{x})}\right\}.$$

This result, namely that the *Bayes classifier* achieves the universal (global) minimum (infimum) error over all measurable classifiers, is a fundamental result in pattern recognition and statistical learning. The probability theory for pattern recognition is made up of multiple results featuring learning machines whose performance are compared to the performance of the Bayes' classifier [7, 39]. A similar fundamental statistical learning result exists for regression, namely that under the so-called squared error loss, the universal best function is the conditional expectation of Y given X. Specifically, consider functions $f : \mathbb{R}^p \to \mathbb{R}$, and the squared theoretical risk functional

$$R(f) = \mathbb{E}[(Y - f(X))^2] = \int_{\mathcal{X}\times\mathcal{Y}} (y - f(\mathbf{x}))^2 p_{\mathrm{xy}}(\mathbf{x}, y)d\mathbf{x}dy.$$

Then the best function $f^\star = \underset{f}{\operatorname{arginf}}\,\{R(f)\}$ is given by the conditional expectation of Y given X, so that $\forall \mathbf{x} \in \mathcal{X}$,

$$f^\star(\mathbf{x}) = \mathbb{E}[Y|X = \mathbf{x}] = \int_{\mathcal{Y}} y\, p(y|\mathbf{x})dy.$$

Far more detailed accounts are found in both [11] and [14] that clearly show the paramount importance of the Bayesian paradigm in statistical machine learning. We see that for both regression and classification, the Bayesian paradigm provides the best mechanism, at the very least in theory, and that is indeed very important. It is worth mentioning that for most people, the Bayesian school of thought is typically not introduced through results like the ones we just described, but instead through Bayesian estimation and inference in parametric families of models. It is our view that both the Bayes' classifier and the Bayes regressor are just as valuable members of the Bayes' heritage as are the vastly studied results in both parametric and nonparametric Bayesian estimation, inference and prediction. Although the results described earlier were in their purely theoretical forms, applications abound that are based on those foundational results. Studying pattern recognition and regression with a solid knowledge of both the Bayes' classifier and the Bayes regressor which provide the best in both cases is of vital importance. I deem it necessary here to further clarify the premises of my argumentation in favor of the ubiquity of the Bayesian paradigm: *Most people think of the Bayesian paradigm solely in* sensu stricto *whereby there is a very involved and often complex and sometimes controversial topic of prior specification. Our perception and view of the Bayesian school of thought in the context of statistical machine learning is definitely in* sensu lato*, and encompasses all modelling situations where the posterior density or the posterior probability is part of the modelling mechanism. Our intention is not to trigger an epistemological debate, quite far from it.* Personally, it is my view that in *sensu lato*, all statistical learning methods are offshoots of the Bayesian machinery, in the sense that the Bayes learner under the two most commonly used loss functions is always the optimal, and indeed the gold standard. In that sense, at the very least, most so-called non-Bayesians or anti-Bayesians are inherently Bayesians at their core (unknowingly I might add), at least in the most quintessential sense of those universal optimality results that all learning machines essentially attempt to attain.

2.2. Bayesian tools for assessing binary classifiers

Binary pattern recognition plays a central role in classification learning, and the Bayesian framework once again provides a convenient language for the assessment of any binary classifier $f : \mathcal{X} \to \mathcal{Y}$. Specifically, the so-called True Positive Rate (`TPR`) and False Positive Rate (`FPR`) which are the ingredients for constructing the visually compelling graphical tool known as the Receiver Operating Characteristics (ROC) curve, are both defined using the Bayesian concept of posterior probability. Indeed, the True Positive Rate (`TPR`) of f is given by

$$\texttt{TPR}(f) = \Pr(f(X) = 1|Y = 1) = \frac{\Pr(f(X) = 1 \,\text{and}\, Y = 1)}{\Pr(Y = 1)},$$

and the False Positive Rate (FPR) of f is given by

$$\texttt{FPR}(f) = \Pr(f(X) = 1 | Y = -1) = \frac{\Pr(f(X) = 1 \text{ and } Y = -1)}{\Pr(Y = -1)}.$$

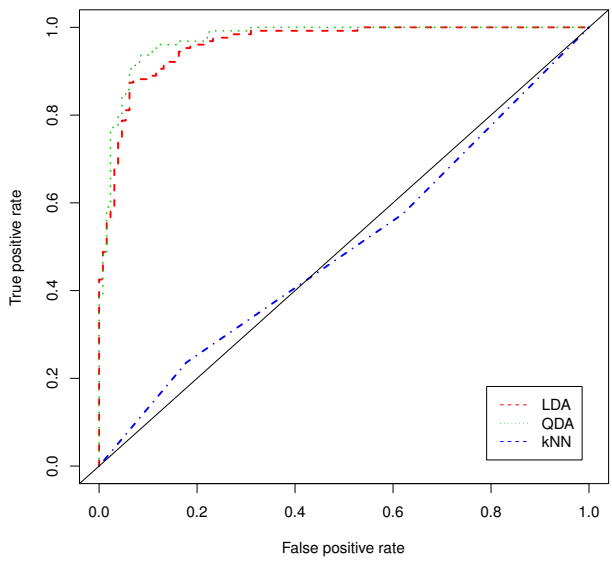

Figure 1. Comparative ROC Curves of three learning machines on the NFL Playoffs prediction data. Gaussian Discriminant Analysis Methods are shown to be far superior to the Nearest Neighbors Learning Machine, judging by their relatively larger area under the curve (AUC).

As one can see on the ROC curve example of Figure (2.2) featuring the performances of different learning machines on the prediction of the playoffs appearance of NFL teams, the visually compelling graphical summary helps compare several learning machines intuitively and readily.

3. Bayes' impact in statistical estimation and inference

The practical construction of the ROC curves mentioned earlier relies heavily on the ability to practically construct the empirical function $\widehat{f}$ in the first place. The history of statistics and machine learning is adorned with pearls of human creativity in the form of a wide variety of approaches to estimating a function f empirically from the data.

3.1. Bayesian parameter estimation

In the case where the function is fully defined and represented by a finite collection of parameters, a wide variety of methods have been developed over the years in statistical estimation and inference. Indeed, the best known setup where the richness of the Bayesian paradigm is practically and more directly revealed, is encountered when we assume that the task of learning the function f, is associated with the estimation of a parameter $\theta \in \Theta \subseteq \mathbb{R}^p$ such that, when treated as

a random variable, the probability density function of θ is given by $p(\theta)$. This is encountered for models involving (a) Parametric density estimation along with elements of prediction; (b) Parametric function estimation along with prediction, when $f(\mathbf{x}) = f(\mathbf{x};\theta)$. In both cases, a key quantity is the posterior density of the parameter θ given the data, namely

$$p(\theta|\mathbf{Y}) = \frac{p(\theta)p(\mathbf{Y}|\theta)}{p(\mathbf{Y})},$$

where $p(\mathbf{Y}|\theta)$ is the likelihood of θ, and $p(\mathbf{Y})$ is the evidence, sometimes referred to as the marginal likelihood of the underlying model. Recall that the likelihood of θ is the joint density of the data vector $\mathbf{Y}$ given the unknown parameter θ, i.e. $\texttt{Likelihood}(\theta) = p(\mathbf{Y}|\theta)$. The maximum likelihood principle is arguably the most commonly used method/approach in statistical analysis because of the central role the likelihood plays in statistical modelling. Now, the maximum likelihood estimator $\widehat{\theta}^{(\texttt{MLE})}$ of θ is given by

$$\widehat{\theta}^{(\texttt{MLE})} = \underset{\theta\in\Theta}{\operatorname{argmax}}\left\{\texttt{Likelihood}(\theta)\right\} = \underset{\theta\in\Theta}{\operatorname{argmax}}\left\{p(\mathbf{Y}|\theta)\right\},$$

while the Bayesian estimator $\widehat{\theta}^{(\texttt{Bayes})}$ of θ under the squared error loss, is

$$\begin{aligned}\widehat{\theta}^{(\texttt{Bayes})} &= \underset{a\in\Theta}{\operatorname{argmin}}\left\{\mathbb{E}[(\theta-a)^2|\mathbf{Y}]\right\}\\ &= \mathbb{E}[\theta|\mathbf{Y}] = \int_{\Theta}\theta p(\theta|\mathbf{Y})d\theta.\end{aligned}$$

A very nice property of both Maximum Likelihood and Bayesian Estimators is that for all continuous functions $g(\cdot)$,

$$\widehat{g(\theta)}^{(\texttt{Bayes})} = \mathbb{E}[g(\theta)|\mathbf{Y}] = \int_{\Theta} g(\theta)p(\theta|\mathbf{Y})d\theta.$$

Now, as far as inference is concerned, the Bayesian paradigm offers: (a) Interval estimation via the construction of credible sets; (b) Bayesian hypothesis testing; (c) Bayesian posterior predictive density specification. All these are seamless offshoots of the quintessential Bayesian learning machinery. For $\alpha \in (0,1)$, the set given by $\texttt{CS}_\alpha(\theta) = [\texttt{LB}_{\alpha,n}(\theta), \texttt{UB}_{\alpha,n}(\theta)]$ such that

$$\int_{\texttt{LB}_{\alpha,n}(\theta)}^{\texttt{UB}_{\alpha,n}(\theta)} p(\theta|\mathbf{Y})d\theta = \alpha,$$

is referred to as the $\alpha\times 100\%$ Bayesian credible set for the unknown parameter θ, which is one of the Bayesian inferential mechanisms for θ, one of the other related mechanisms being Bayesian hypothesis testing. The most common choices are $\texttt{LB}_{\alpha,n}(\theta)$ as the $(1-\alpha)/2$ quantile of $p(\theta|\mathbf{Y})$ and $\texttt{UB}_{\alpha,n}(\theta)$ as the $(1+\alpha)/2$ quantile of $p(\theta|\mathbf{Y})$. It is important to remember that the parameter θ herein studied can be a scale or a vector governing the form of a very complex function. The interpretation of the Bayesian credible set, unlike the arcane and often convoluted interpretation of the Frequentist confidence interval, is straightforward and intuitively clear. Thanks to the Bayesian rigorous and appropriate handling of the unknown random variable θ via the language of probability, one can write

$$\texttt{Prob}\big(\theta \in [\texttt{LB}_{\alpha,n}(\theta), \texttt{UB}_{\alpha,n}(\theta)]\,|\,\mathcal{D}_n\big) = \alpha.$$

Given the data $\mathcal{D}_n$, the probability is α that $\theta \in [\mathtt{LB}_{\alpha,n}(\theta), \mathtt{UB}_{\alpha,n}(\theta)]$, which is uniquely only possible with the Bayesian paradigm, since θ being random is aptly and appropriately spoken of via the powerful language of probability. Note that this intuitive interpretation is not possible in the Frequentist setting. In hypothesis testing, the null and the alternative hypotheses are tested using evidence from the data and theoretical properties inherent in the hypothesized models. The hypotheses are given as:

$$\mathbf{H}_0 : \theta \in \Theta_0 \quad \mathtt{vs} \quad \mathbf{H}_a : \theta \notin \Theta_0.$$

As far as Bayesian hypothesis testing is concerned, the decision about the null hypothesis is conveniently made by measuring the posterior probability of $\mathbf{H}_0$ given the data, which is given by

$$\Pr(\mathbf{H}_0 | \, \mathbf{y}) = \frac{p_Y(\mathbf{y}|\mathbf{H}_0) \Pr(\mathbf{H}_0)}{p_Y(\mathbf{y})}$$

where $\Pr(\mathbf{H}_0) + \Pr(\mathbf{H}_a) = 1$ and $p_Y(\mathbf{y}) = \Pr(\mathbf{H}_0) p_Y(\mathbf{y}|\mathbf{H}_0) + p_Y(\mathbf{y}|\mathbf{H}_a) \Pr(\mathbf{H}_a)$ is the density of the data. It is interesting to note the error of the test is also conveniently defined and calculated as

$$\mathtt{Error} = \Pr(\mathbf{H}_0 \ \mathtt{is\,chosen}|\mathbf{H}_a) \Pr(\mathbf{H}_a) + \Pr(\mathbf{H}_a \ \mathtt{is\,chosen}|\mathbf{H}_0) \Pr(\mathbf{H}_0).$$

Unlike with non-Bayesian inference, the Bayesian paradigm draws conclusions about the test in a clear and intuitive and straightforward based simply on the value of the posterior probability $\Pr(\mathbf{H}_0 | \, Y = \mathbf{y}) \equiv \Pr(\mathbf{H}_0 \, \mathtt{is\,true} | \, \mathtt{Data})$. This Bayesian decision mechanism aligns far better with human intuition than all other mechanisms that are marred is technical niceties and subtleties.

3.2. Bayesian posterior predictive density

Besides providing a mechanism for parameter estimation, the Bayesian paradigm inherently addresses the important need for a predictive density of any new element $\mathrm{y}_{\mathtt{new}} \in \mathcal{Y}$, which is given by

$$p(\mathrm{y}_{\mathtt{new}}|\mathbf{Y}) = \int_{\Theta} p(\mathrm{y}_{\mathtt{new}}|\theta) p(\theta|\mathbf{Y}) d\theta.$$

With the posterior predictive density in place, one can perform a whole host of prediction related activities, including the determination of the quantiles of the posterior predictive density $p(\mathrm{y}_{\mathtt{new}}|\mathbf{Y})$ but also the moments like the expected value $\mathbb{E}(Y_{\mathtt{new}}|\mathbf{Y})$ and the variance $\mathtt{variance}(Y_{\mathtt{new}}|\mathbf{Y})$ which play a central role in the computation of the crucially needed error bars, and prediction bands.

3.3. Maximum a posteriori (MAP) estimator

The so-called Maximum A Posteriori (MAP) Estimator is another type, albeit sometimes deemed inferior, of Bayesian estimator, given by

$$\begin{aligned}
\widehat{\theta}^{(\mathtt{MAP})} &= \underset{\theta \in \Theta}{\operatorname{argmax}} \left\{ p(\theta|\mathbf{Y}) \right\} = \underset{\theta \in \Theta}{\operatorname{argmax}} \left\{ p(\theta) p(\mathbf{Y}|\theta) \right\} \\
&= \underset{\theta \in \Theta}{\operatorname{argmax}} \left\{ \log p(\theta) + \log p(\mathbf{Y}|\theta) \right\} \\
&= \underset{\theta \in \Theta}{\operatorname{argmax}} \left\{ \log \mathtt{Prior}(\theta) + \log \mathtt{Likelihood}(\theta) \right\}.
\end{aligned}$$

Note that $\widehat{\theta}^{(\texttt{MAP})}$ is essentially the mode of the posterior density of θ, which coincide with the expected value of the posterior density if the posterior density is symmetric. Pure statisticians typically tend to avoid the MAP estimator precisely because on the one hand the mode may not even exist or it may not be the "right" estimator.

3.4. Bayesian image denoising and image compression

The MAP estimator has been used a lot in image denoising and image compression. Typically, each image is acquired as an $r \times c$ matrix. For mathematical and computational convenience, each image is represented as a p-dimensional vector of grayscale values, where $p = rc$. The observed image $\mathbf{y} \in \mathbb{R}^p$ is assumed to be a noisy version of the true (noiseless, clean) original image $\mathbf{x} \in \mathbb{R}^p$, where $\texttt{dim}(\mathbf{x}) = \dim(\mathbf{y}) = p = rc$, and $\mathbf{y} = \mathbf{x} + \sigma\mathbf{z}$ where $\sigma > 0$ and $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The goal of image denoising is to recover the clean original image $\mathbf{x}$ from the noisy $\mathbf{x}$, a task the Bayesian machinery accomplishes by specifying a prior density $p(\mathbf{x})$ and then obtaining a Bayesian estimator of $\mathbf{x}$ via Maximum A Posteriori (MAP) estimation

$$\widehat{\mathbf{x}}^{(\texttt{MAP})} = \underset{\mathbf{x}\in\mathcal{X}}{\operatorname{argmax}}\left\{\frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})}\right\}$$

which is typically reduced for mathematical and computational convenience to

$$\widehat{\mathbf{x}}^{(\texttt{MAP})} = \underset{\mathbf{x}\in\mathcal{X}}{\operatorname{argmin}}\left\{-\log p(\mathbf{y}|\mathbf{x}) - \log p(\mathbf{x})\right\}. \tag{3.1}$$

The image denoising mechanism of Equation (3.1) is just the tip of the iceberg as there are thousands upon thousands of research papers, bothe theoretical and applied that have be written featuring several incarnations and variations of that very same formula. Image compression for instance is an even more complex and more frequently encountered extension where it is assumed that $\mathbf{y} = \mathbf{W}\mathbf{x} + \sigma\mathbf{z}$ with $\texttt{dim}(\mathbf{x}) \lll \dim(\mathbf{y})$. Once again, these are vast topics, mentioned here only as evidence of the diverse fields of application of the Bayesian paradigm whose blessings extend far beyond the scope of this paper.

3.5. Bayesian nonnegative matrix factorization

Nonnegative matrix factorization decomposes a $p\times n$ matrix $\mathbf{X} \in \mathbb{R}_+^{p\times n}$ into a product of a loading matrix $\mathbf{W} \in \mathbb{R}_+^{p\times q}$ and feature matrix $\mathbf{H} \in \mathbb{R}_+^{q\times n}$ where $q \lll p$ is the reduced rank. The reconstructed version $\tilde{\mathbf{X}} = \mathbf{W}\mathbf{H}$ of $\mathbf{X}$ is obtained by solving a constrained optimization problem

$$\underset{\mathbf{W}\geq 0,\mathbf{H}\geq 0}{\texttt{argmin}}\left\{\|\mathbf{X} - \mathbf{W}\mathbf{H}\|_{\mathtt{F}}^2\right\},$$

where the Frobenius norm $\|\mathbf{A}\|_{\mathtt{F}}^2$ of $\mathbf{A}$ is given by

$$\|\mathbf{A}\|_{\mathtt{F}}^2 = \sum_{i=1}^{p}\sum_{j=1}^{n}|\mathrm{a_{ij}}|^2 = \sum_{i=1}^{p}\|\mathbf{A}_{i\cdot}\|_2^2 = \sum_{j=1}^{n}\|\mathbf{A}_{\cdot j}\|_2^2 = \operatorname{trace}(\mathbf{A}^\top\mathbf{A}).$$

The `R` package `rNMF` performs nonnegative matrix factorization for feature extraction and image reconstruction. Nonnegative matrix factorization has received tremendous attention from Bayesian leaning researchers, many of whom have typically assumed that the matrix $\mathbf{X} \in \mathbb{R}_+^{p\times n}$ admits the representation $\mathbf{X} = \mathbf{WH}+\sigma\mathbf{Z}$ where $\mathbf{Z} \in \mathbb{R}^{p\times n}$ is a matrix of standard normal random variables. Assuming σ^2 known, the parameter vector in Bayesian NMF is given by $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{H}\}$. Due to the nonnegativity constraint on $\mathbf{W}$ and $\mathbf{H}$, distributions like the exponential and the gamma are used since that are both defined only for positive numbers. All in all, Bayesian NMF uses the likelihood defined by

$$p(\mathbf{X}|\mathbf{W}, \mathbf{H}) = \prod_{i=1}^{n}\prod_{j=1}^{p} \phi(\mathrm{x}_{ij}; (\mathbf{WH})_{ij}, \sigma^2),$$

and proceeds by exploring various aspects of the posterior density

$$p(\mathbf{W}, \mathbf{H}|\mathbf{X}) = \frac{p(\mathbf{X}|\mathbf{W}, \mathbf{H})p(\mathbf{W}, \mathbf{H})}{p(\mathbf{X})}.$$

Many researchers have assumed $p(\mathbf{W}, \mathbf{H}) = p(\mathbf{W})p(\mathbf{H})$, by virtue of the assumed prior independence between $\mathbf{W}$ and $\mathbf{H}$. Some authors have further resorted to the Maximum a Posteriori (MAP) treatment

$$\underset{\mathbf{W}\geq 0, \mathbf{H}\geq 0}{\operatorname{argmin}} \Big\{ -\log p(\mathbf{X}|\mathbf{W}, \mathbf{H}) - \log p(\mathbf{W}) - \log p(\mathbf{H}) \Big\},$$

which is similar in form to the regularized approach used in [4] and [5], namely

$$\underset{\mathbf{W}\geq 0, \mathbf{H}\geq 0}{\operatorname{argmin}} \Big\{ \|\mathbf{X} - \mathbf{WH}\|_{\mathrm{F}}^2 + \alpha \mathtt{Pen}_1(\mathbf{W}) + \beta \mathtt{Pen}_2(\mathbf{H}) \Big\}.$$

Many other unsupervised learning methods have been formulated under the Bayesian paradigm. Among those, Bayesian Factor Analysis, Bayesian Principal Component Analysis, Bayesian Independent Component Analysis, and Bayesian Analysis of Mixtures of Distributions for model based clustering. Interestingly though, all those methods admit a formulation in terms of matrix factorization, hence our choice to feature Bayesian Nonnegative Factorization here without loss of generality as far as most Bayesian approach to unsupervised learning goes.

3.6. Bayesian paradigm as the extension and generalization of others

It turns out that *the Bayesian paradigm is an extension and a generalization of the maximum likelihood principle, an extension that affords greater modelling flexibility, and consequently the capability to solve a wider class of problems. For instance, the maximum likelihood estimator is a special case of the Bayesian estimator.* To see this, note that if the prior density $p(\theta)$ is uniform, i.e. $p(\theta) = c$, then we have

$$\widehat{\theta}^{(\mathtt{MAP})} = \widehat{\theta}^{(\mathtt{Bayes})} = \underset{\theta\in\Theta}{\operatorname{argmax}} \{\log p(\mathbf{Y}|\theta)\} = \underset{\theta\in\Theta}{\operatorname{argmax}} \{\mathtt{Likelihood}(\theta)\} = \widehat{\theta}^{(\mathtt{MLE})}.$$

That the Frequentist (non-Bayesian) result is a special case of the Bayesian counterpart should, in my humble opinion, be a motivation for researchers to give the Bayesian approach another look. This clearly resembles the relationship between Newtonian physics and Einstein's relativity theory, the latter being a generalization and a more complete extension of the former.

3.7. Bayesian inherent shrinkage mechanism

Another powerful property inherent in the Bayesian paradigm is its inherent shrinkage and regularization capability, which turns out to be a powerful remedy that helps circumvent a wide variety of modelling challenges. To gain deeper insights into this regularization and shrinkage property, we consider the Bernoulli experiment, with the parameter $\theta \in (0,1)$ representing the probability of success, and $Y_i \in \{0,1\}$ such that

$$Y_1, Y_2, \cdots, Y_n \overset{iid}{\sim} \texttt{Bernoulli}(\theta).$$

We have

$$p(y_i|\theta) = \theta^{y_i}(1-\theta)^{1-y_i}.$$

Under the conjugacy principle, the conjugate prior for θ is

$$p(\theta|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}.$$

It can be shown that the posterior density of θ is given by

$$p(\theta|D_n) = \frac{\Gamma(a+b+n)}{\Gamma(S_n+a)\Gamma(n-S_n+b)}\theta^{S_n+a-1}(1-\theta)^{n-S_n+b-1}$$

which means that $(\theta|D_n) \sim \texttt{Beta}(a+S_n, b+F_n)$. Now we

$$\widehat{\theta}^{(\texttt{Bayes})} = \mathbb{E}[\theta|\mathbf{Y}] = \int_{\Theta} \theta p(\theta|\mathbf{Y})d\theta = \frac{a+S_n}{a+b+n}.$$

Notice

$$\begin{aligned}\widehat{\theta}^{(\texttt{Bayes})} &= \frac{a+S_n}{a+b+n} = \frac{a+b}{a+b+n}\frac{a}{a+b} + \frac{n}{a+b+n}\frac{S_n/n}{a+b+n} \\ &= w_n\widehat{\theta}_0 + (1-w_n)\widehat{\theta}_{(\texttt{MLE})}.\end{aligned}$$

The Bayesian "point" estimator $\widehat{\theta}^{(\texttt{Bayes})}$ is therefore a convex combination of the prior estimate with the maximum likelihood estimator. Indeed, $\lim_{n\to\infty} w_n = 0$. As a result,

$$\lim_{n\to\infty} \widehat{\theta}^{(\texttt{Bayes})} = \widehat{\theta}^{(\texttt{MLE})},$$

which means that as more data becomes available, the posterior density is dominated by the likelihood, so that the asymptotically the Bayesian estimator coincide with the maximum likelihood estimator. In this sense, the prior is bringing to the estimation (learning) task, items of information that the data from the sampling process does not contain, and this is crucial. As more data becomes available, the information brought by the prior is then overwhelmed by the information richly provided by large amounts of data. This serves as the basis for resorting to the Bayesian paradigm in situations where there isn't enough data to carry the modelling task at hand.

4. Bayes' impact in statistical function estimation

Statistical Function estimation is a very very vast field of statistical science, machine learning and artificial intelligence.

4.1. Bayesian multiple linear regression

It suffices to use the space of p-dimensional multiple linear regression models to help gain insights into some of the ways in which the Bayesian paradigm permeates statistical function estimation.

$$\mathcal{H} = \left\{ f : \exists \boldsymbol{\theta} = (\theta_1, \cdots, \theta_p) \in \mathbb{R}^p, \, | f(\mathbf{x}) = \langle \boldsymbol{\theta}, \mathbf{x} \rangle = \sum_{j=1}^{p} \theta_j \mathrm{x}_j \right\}.$$

To better understand this, we consider multiple linear regression under the Gaussian homoscedastic noise model, $(\mathbf{Y}|\mathbf{X}, \theta, \sigma^2) \sim \mathbf{N}_n(\mathbf{X}\theta, \sigma^2 \mathbf{I}_n)$, for which the likelihood of θ is simply

$$\begin{aligned} \mathtt{L}(\theta|\mathbf{X}, \mathbf{Y}) = p(\mathbf{Y}|\mathbf{X}, \theta, \sigma^2) &= \phi_n(\mathbf{Y}; \mathbf{X}\theta, \sigma^2 \mathbf{I}_n) \\ &= \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp\left(-\frac{1}{2\sigma^2} (\mathbf{Y} - \mathbf{X}\theta)^\top (\mathbf{Y} - \mathbf{X}\theta) \right). \end{aligned}$$

The maximum likelihood estimator $\widehat{\theta}^{(\mathtt{MLE})}$ of θ is the well-known

$$\widehat{\theta}^{(\mathtt{MLE})} = \underset{\theta \in \Theta}{\operatorname{argmax}} \left\{ \mathtt{L}(\theta|\mathbf{X}, \mathbf{Y}) \right\} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}.$$

Combining the fact that $\widehat{\theta}^{(\mathtt{MLE})} \sim N_p(\theta, \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1})$ with the conjugate prior $\theta \sim N_p(\theta_0, \sigma^2 \Lambda_0^{-1})$, the Bayesian estimator $\widehat{\theta}^{(\mathtt{Bayes})}$ of the vector θ of regression coefficients, is given by

$$\widehat{\theta}^{(\mathtt{Bayes})} = \mathbb{E}[\theta|\mathbf{Y}] = (\mathbf{X}^\top \mathbf{X} + \Lambda_0)^{-1} (\mathbf{X}^\top \mathbf{X} \widehat{\theta}^{(\mathtt{MLE})} + \Lambda_0 \theta_0).$$

The famous ridge regression estimator $\widehat{\theta}_\lambda^{(\mathtt{ridge})}$ of θ first proposed by [25] and [24] is shown to be special case of the above Bayesian estimator when $\Lambda_0 = \lambda \mathbf{I}$. Specifically,

$$\begin{aligned} \widehat{\theta}_\lambda^{(\mathtt{ridge})} &= \underset{\theta \in \Theta}{\operatorname{argmin}} \left\{ (\mathbf{Y} - \mathbf{X}\theta)^\top (\mathbf{Y} - \mathbf{X}\theta) + \lambda \theta^\top \theta \right\} \\ &= (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{Y} \\ &= \mathbb{E}[\theta|\mathbf{Y}] = \widehat{\theta}^{(\mathtt{Bayes})}. \end{aligned}$$

4.2. Bayesian learning as regularized learning

It is easy to verify (check) that *the maximum likelihood estimator is a special case of the Bayesian estimator*, in the sense that

$$\lim_{\lambda \to 0} \widehat{\theta}_\lambda^{(\mathtt{ridge})} = \lim_{\lambda \to 0} (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{Y} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} = \widehat{\theta}^{(\mathtt{MLE})}.$$

It also easy to see that the ridge estimator is a shrinkage estimator, with the tendency to shrink all the components of the vector to zero together as λ gets ever larger. Specifically,

$$\lim_{\lambda \to \infty} \widehat{\theta}_\lambda^{(\mathtt{ridge})} = \mathbf{0}.$$

Of great importance to big data analytics is the fact that between the two extremes of zero λ and infinite λ, lies a value of λ that achieves the trade-off between bias and variance, and thereby achieves the smallest cross validation error. The fact that the Bayesian estimator is inherently (by its very design) biased, used to

be a subject of great debates, until numerous findings revealed that unbiasedness while a desired property, is not the be all and end all of statistical estimation and inference, quite far from it. It turns out that most scientific endeavors reveal the fundamental need for a trade-off between bias and variance. For the regression example mentioned earlier, when we (a) either have multicollinearity in the design matrix or (b) the data matrix $\mathbf{X}$ is high dimensional but with a very low sample size ($n \lll p$, underdetermined system), the maximum likelihood estimator is theoretical unbiased but has an ill-conditioned variance matrix that leads to non-existence or non uniqueness or severe instability. Even in case where a numerical solution can realized, the variance is inflated because of the near singularity. The Bayesian approach via ridge regression for instance yields a solution, albeit biased, but with a reduced variance. In fact, in the $n \lll p$ it is impossible to have any solution without a device like the ridge approach. This is the kind of scenarios that make us say that the Bayesian paradigm is a gospel, meaning good news, as it allows workable solution where none appears to exist. Solutions like ridge are nowadays ubiquitous in statistical machine learning and belong to a class of machine learning approaches known as regularization methods, where all the techniques consist of adding constraints to an ill-posed problem to hopefully achieve well-posedness in Hadamard's sense. All the methods in the regularization framework are centered around the regularized empirical risk

$$R_{\texttt{reg}}(f) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(Y_i, f(X_i)) + \lambda \|f\|_{\mathcal{H}}$$

where $\|f\|_{\mathcal{H}}$ is the norm of f in the function space $\mathcal{H}$. For the linear regression learning task for instance, the ridge regularization mentioned earlier has evolved (been developed) alongside the Least Absolute Shrinkage and Selection Operator (LASSO) proposed by [37], which admits a Bayesian formulation using a Laplace prior on θ, but does not yield a closed-form solution like the ridge estimator

$$\widehat{\theta}_{\lambda}^{(\texttt{lasso})} = \underset{\theta \in \Theta}{\operatorname{argmin}} \left\{ (\mathbf{Y} - \mathbf{X}\theta)^{\top}(\mathbf{Y} - \mathbf{X}\theta) + \lambda \|\theta\|_1 \right\}.$$

The well-known greatest strength of the LASSO estimator comes from the fact that it does achieve sparsity and therefore is used for variable selection. Just like the ridge solution, the LASSO, through regularization, is inherently able to yield a solution where the MLE would at best be very unstable. It is interesting to note that the LASSO estimator both shrinks and selects, whereas the ridge estimator simply shrinks while maintaining all the initial variables. Combining ridge and lasso, one gets

$$R_{\texttt{reg}}(f) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(Y_i, f(X_i)) + \lambda \texttt{Penalty}_{\alpha}(\theta),$$

where $\texttt{Penalty}_{\alpha}(\theta) = (1-\alpha)\|\theta\|_1 + \alpha\|\theta\|_2$ is the so-called elastic net penalty. Several implementations exist in R, including [18] and [35].

4.3. Bayesian paradigm in kernel methods

If one were to start conferring awards of excellence to statistical learning machines, many so-called kernel machines would rack up many medals. Despite the

popularity of deep learning these days, there has been a resurgence of the near dominance of kernel machines, partly due to emerging results revealing that kernel machines are just as good if not better than deep neural networks on several benchmarks tasks, with the added benefit for kernel methods that they rest on very solid theoretical foundations.

4.3.1. Bayesian formulation of neural networks. Interestingly both neural networks and kernel learning machines do admit favorable Bayesian formulations that have led to theoretical and practical breakthroughs. Reference [31] gives a very detailed account of Bayesian Neural Networks, and even shows that with a single hidden layer, a neural network with an infinite number of nodes is essentially a Gaussian process which in turn happens to be a kernel learning machine. For clarity, we are in the presence of the space of kernel learning machines if the prediction function $f : \mathcal{X} \longrightarrow \mathbb{R}$ comes from a function space $\mathcal{H}$ where functions are such that $\forall \mathbf{x} \in \mathcal{X}$, the corresponding $f(\mathbf{x})$ is of the form

$$\widehat{f}(\mathbf{x}) = \widehat{f}_n(\mathbf{x}) = \mathsf{g}\left(\sum_{j=1}^{n} \widehat{\mathrm{w}}_j \mathcal{K}(\mathbf{x}, \mathbf{x}_j)\right), \tag{4.1}$$

where $\mathcal{K}(\cdot,\cdot)$, the bivariate function $\mathcal{K} : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ defined on the product $\mathcal{X} \times \mathcal{X}$ is used to measure the similarities between the members/elements of the input space $\mathcal{X}$, and finally $\mathsf{g}(\cdot)$ is what is usually typically referred to as a link function. The two crucial and distinguishing choices when it comes to kernel learning machines are: (a) The choice of the kernel and (b) The way the estimates $\widehat{\mathrm{w}}_j$ of the weights are formulated and obtained. The so-called Gaussian Radial Basis Function (RBF) kernel is arguably the most used in practice, and it is defined as

$$\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2r^2}\right).$$

A more general extension of the Gaussian RBF kernel has been used in practice and has the form

$$\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = \eta \exp\left(-\frac{1}{2}\sum_{k=1}^{p} \rho_k (x_{ik} - x_{jk})^2\right) + \lambda \delta_{ij}.$$

There are tons of other kernels (covariance functions) corresponding to the extremely wide variety of function estimation tasks that arise in classification learning and regression learning.

4.3.2. Gaussian process learning and Bayesian kernel learning. Gaussian process learning machines could well be the most natural members in the family of kernel learning machines, and because of their ubiquity in the resurgence of kernel methods [6, 34], it makes good sense to give a bit more details about them. For each $\mathbf{x}_i$, let $f_i = f(\mathbf{x}_i)$ be the function value yielded by the underlying function f. Let $\boldsymbol{f} = (f_1, f_2, \cdots, f_n)^\top$ be the n-dimensional vector of function values for the points $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$. Now, *the probability distribution of a function $f(\mathbf{x})$ is said to be a Gaussian process if for any selection of points $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$, the density $p(\boldsymbol{f})$ of the corresponding n-dimensional vector of function values is Gaussian.* What we really want is a prior density $p(f)$ over the infinite dimensional function

space $\mathcal{H}$ from where the function f is assumed to originate. Given a new point $\mathbf{x}_{\texttt{new}} \in \mathcal{X}$, it is of interest to study the random variable $f(\mathbf{x}_{\texttt{new}}) = f_{\texttt{new}}$ and the random variable $Y_{\texttt{new}}$. Assume that a priori the vector $\boldsymbol{f}$ follows a multivariate Gaussian distribution with zero mean and covariance matrix $\boldsymbol{K}$, that is

$$\boldsymbol{f} \sim \boldsymbol{N}_n(\mathbf{0}, \boldsymbol{K}).$$

Assume that a priori the vector $\boldsymbol{f}$ follows a multivariate Gaussian distribution with zero mean and covariance matrix $\boldsymbol{K} = (K_{ij})$ with $K_{ij} = \mathsf{cov}(f_i, f_j) = \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)$, that is

$$p(\boldsymbol{f}) = \phi_n(\boldsymbol{f}|\mathbf{0}, \boldsymbol{K}) = \frac{1}{\sqrt{(2\pi)^n|\boldsymbol{K}|}} \exp\left\{-\frac{1}{2}\boldsymbol{f}^\top \boldsymbol{K}^{-1}\boldsymbol{f}\right\}.$$

The posterior obtained leads to an estimated prediction function of the form given in Equation (4.2), with

$$\widehat{\mathrm{w}}_j = [(\boldsymbol{K} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{Y}]_j.$$

A detailed derivation of this result can be readily found in [34]. Figure depicts samples of random functions drawn from a Gaussian process prior with some specific covariance matrix. Gaussian processes, by allowing prior distributions over infinite dimensional function spaces, provided one of the earliest forms of Bayesian nonparametric function estimation.

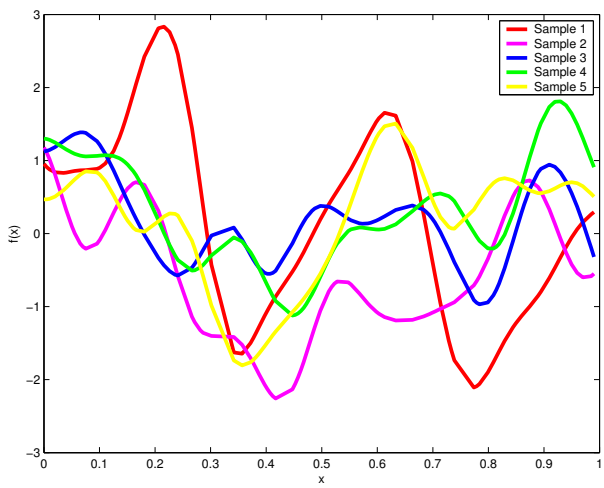

Figure 2. Sample functions drawn from a Gaussian process prior. This ability to define distributions over function spaces without specific parametric forms turns out to be crucial in nonparametric estimation. The samples depicted here are different, although one could make the case for an emerging pattern.

4.4. Bayesian paradigm in ensemble learning methods

When it comes to the popularity of predictive analytics in practical applied data science, random forest and boosting two sister methods in ensemble learning appear to have claimed the two highest places/seats of honor. Their elevated stature

seems quite deserved when one realizes that these two methods have continually and consistently emerged as the most accurate and most precise in most practical settings. Ensemble learning methods are indeed formidable when it comes to optimal prediction maybe in part because of the synergistic nature of the underlying resulting estimating functions. The most generic form of an ensemble learning machine is given by

$$\widehat{f}(\mathbf{x}) = \widehat{f}_n(\mathbf{x}) = \mathsf{g}\left(\sum_{\ell=1}^{L} \widehat{\alpha}_\ell \widehat{h}_\ell(\mathbf{x})\right), \tag{4.2}$$

where $\widehat{h}_\ell(\cdot) \in \mathcal{H}, \ell = 1, \cdots, L$ are the estimators of L base learners from the chosen function space $\mathcal{H}$. The base learners are bonafide mappings from $\mathcal{X}$ to $\mathcal{Y}$, and may only exist in pure algorithmic form. Bagging, Boosting and Random Forest learning machines all admit that very same representation. Interestingly, and in keeping with the spirit of this paper, the Bayesian school of thought has also contributed substantially to ensemble learning. Indeed, one of the Bayesian approaches to ensemble learning is the so-called Bayesian Model Averaging, for which $\widehat{\alpha}_\ell$ is an estimator of α_ℓ, where

$$\alpha_\ell = \Pr(M_\ell|\mathcal{D}_n) = \frac{p(M_\ell)p(\mathcal{D}_n|M_\ell)}{\sum_{\ell'=1}^{L} p(M_{\ell'})p(\mathcal{D}_n|M_{\ell'})}$$

with

$$\Pr(\mathcal{D}_n|M_\ell) = \int_{\boldsymbol{\Theta}} p(\mathcal{D}_n|\theta_\ell, M_\ell)p(\theta_\ell|M_\ell)d\theta_\ell$$

and

$$h_\ell(\mathbf{x}) = \mathbb{E}[Y|\mathbf{x}, \theta_\ell, M_\ell].$$

R packages dealing with ensemble learning include some purely Bayesian ones like library(BART), library(mBART, library(BMA), but also some non-Bayesian ones like library(boosting), library(mboost), library(randomForest), library(ipred), library(ada), and library(caret), library(bagging).

Once again, the goal of this paper is not to explicate the technical niceties of the Bayesian paradigm, but instead draw the reader's attention on its immense modelling potential. As a matter of fact, the above regularization framework based on the elastic net, helps tackle and solve many predictive analytics task in high dimension and low sample size situations as arises with DNA Gene Expression Microarray Data and several other large p small n tasks. It bears repeating that all the above pearls of statistical modelling, while set up in the so-called regularization framework, do inherently admit a Bayesian formulation. Like we said earlier, "To Bayes or Not Bayes?" is no longer the question, but rather "How am I making the most of Bayes?". The key seems to lie in the specification of carefully thought out prior densities that allow one to isolate precisely the kind of solution desired out of the multiplicity of solutions. *Wherever there is statistical learning, especially in settings where there is an ill-posedness challenge, the Bayesian paradigm is forever available as a formidable weapon in the statistical scientist's modelling arsenal.* We see the power of the Bayesian thought directly or indirectly in state

of the art settings such Latent Dirichlet Allocation (LDA) for Topic Modelling [3]. The forever useful Kalman filter, thanks to the latent space has also benefitted heavily from the power of the Bayesian paradigm [20, 21] and [22]. Even before the blessings of affordable computation ushered in the glorious era of the Bayesian thought, Markov Random Fields were being used for the Statistical Analysis of Dirty Pictures [2], already anchoring the palpable power of the Gospel according to Reverend Thomas Bayes. And yes, modern artificial intelligence as also benefitted very immensely from the flexibility that the combination of likelihood with prior affords the statistical scientists, in [30], we see that the explosion of Neural Networks as tools for artificial intelligence and learning was quickly found to have a nice connection to the Bayesian paradigm, and even now works like [17] demonstrate the great appeal of the Bayesian approach for the now very fashionable and in vogue Networks as well. Paper [29] gives a detailed account of Bayesian interpolation and introduces the now popular and widely used concept of automatic relevance determination (ARD). As a matter of fact, [38]'s Sparse Bayesian Learning and the Relevance Vector Machine (RVM) is a nice piece of work inspired by a combination of [29] and [40]. Interestingly, a little after [38], we get [36] exploring Bayesian Methods for Support Vector Machines and more recently [23] with an interesting account of Bayesian Nonlinear Support Vector Machine. The work on Gaussian process regression in [41] and later in [6] with Efficient Algorithms for Bayesian Gaussian Processes, both ushered in a series of contributions in machine learning featuring Bayesian Gaussian processes for regression and classification, later crystalized in [34] which has become one of the main textbook for the use of Bayesian Gaussian processes in machine learning. [6] explore ideas of variational mean field approximations featuring efficient Approaches to Bayesian Gaussian Process Classification. Gaussian Process Priors open the door to a vast universe of nonparametric statistical modelling in the Bayesian framework. This use of prior distributions over function spaces central to Gaussian process learning has recently become mainstream in Bayesian Nonparametric statistical analysis, anchored by the seminal work on the introduction of the Dirirchlet process prior by [8] and [9] which has enriched the statistician and data scientist's modelling arsenal with a formidably powerful weapon in nonparametric statistical analysis, especially allowing prior distributions on function spaces and infinite dimensional spaces in general. The recent years have been marked by what is literally an explosion of methods which are derivatives of or inspired by the seminal work of [8] and [9]. With Bayesian nonparametrics providing extra modelling strength and flexibility to statisticians and data scientists, the vast territory of application of the Bayesian paradigm just keeps on expanding, further justifying our view that the Gospel according to Reverend Thomas Bayes is indeed ubiquitous, pervading and permeating the whole of science. Statistical Machine Learning from its very early days both implicit and explicitly gave a prominent platform and a loud speaking voice to the Bayesian school of thought, Gaussian processes and Dirichlet processes have increased the volume of the loud speaker. In Bayesian computation, the 1990 seminal work of [19] introduce the world to the power of the Gibbs Sampler, and made it possible for Bayesian statisticians to tackle and successfully solve many statistical modelling problems which had eclipsed them

until that milestone. After the [19] paper that launched the Bayesian Computation revolution, software packages like **BUGS** (`Bayesian Inference with the Gibbs Sampler`) began to emerge, making it more and more possible for Bayesians to actually solve interesting and meaningful real life problems. Implementations abound that help practitioners experiment and applied the power of the Gibbs sampler [32]. The statistical software environment R has many packages and an entire view install.view(Bayesian) that contain various functions for Bayesian analyses of all kinds. As a matter of fact, with the development of Bayesian computation which marked the birth of a collection of methods known as Markov Chain Monte Carlo (MCMC) methods, literally every aspect of statistical benefitted from the modelling power of the Bayesian paradigm. The development of Bayesian computation also allowed substantial progress in Bayesian model selection and Bayesian variable selection. Among other contributions, we have Spike and Slab [27] and more recently work on featuring mixtures of g-Priors for Bayesian variable selection [28]. The Bayesian Model Averaging for Linear Regression Models by [33] was later supplemented by a tutorial [26], that further helped put practical BMA on a firm foundation. Later, [1] provided optimal predictive model selection via the so-called Bayesian median model. The Estimation of Atom Prevalence for Optimal Prediction [10] sought to be a flexible and more adaptive counterpart to [1]. As we said earlier, the intention of this paper is far from any attempt to provide an exhaustive technical exploration of the Bayesian paradigm. That would be gargantuan and virtually impossible. Instead, we have sought throughout and hope to have given the reader a visceral sense of the appeal of the Bayesian paradigm as a statistical machine learning tool for data science. We complete by mentioning a few contributions of the Bayesian paradigm to latent variable modelling and kernel regression, with works like [13] which introduces a stable Radial Basis Function Selection via Mixture Modelling of the Sample Path, and [15] that extends it with a fully Bayesian Analysis of the Relevance Vector Machine With Extended Prior. Paper [12] proposes and develops a Bayesian computation of the Intrinsic Structure of Factor Analytic Models, drawing some of its elements from [16] where mixtures of Factor Analysers featuring Bayesian Estimation and Inference by Stochastic Simulation.

5. Conclusion and discussion

In this paper, we have provided a general bird's eye view of the manifold ways in which the Bayesian paradigm has become one of the main tools in the arsenal of all statisticians and data scientists. In our experience and observation, human statistical thought and perception as witnessed in interval estimation and hypothesis testing appears to be inherently and quintessentially Bayesian: indeed when non statisticians are asked to interpret confidence intervals or p-values, most (if not all) say things that are essentially credible sets or posterior probabilities of hypotheses. It does seem that our human statistical thought quintessentially agrees with the Bayesian principle. When it comes to decision making under uncertainty, it is virtually impossible to find a field of study in science or otherwise that has not been heavily and positively by the power of the Bayesian paradigm. Uncertainty is ideally dealt with using the powerful language of probability, hence the

appropriateness of the assignment of property to all unknown quantities, including parameters unfortunately treated by others as fixed. The Bayesian approach is the only way to properly deal with latent variable models. What other way exists to properly model a random variable other than specify or estimate its distribution from the data Penalized Least Squares Estimation turns out to admit a natural Bayesian formulation with penalties capturing a prior belief about distributional aspects of the parameters or function class of interest. In sensu lato and senso stricto, the likelihood principle is a special case (subset) of the Bayesian paradigm. The inherent capacity of the Bayesian paradigm to extend the likelihood principle can be likened to the way in which the relativity theory contributed by Albert Einstein extended, enriched and revolutionized Isaac Newton's fundamental laws of physics. The gospel according to Reverend Thomas Bayes lives on and keeps on gaining more power and transforming more lives through its impact in science, statistical machine learning and data science From an unpublished manuscript, the revolutionary idea of Reverend Thomas Bayes has become one of the most consequential and most pervading transformative concepts in the whole of science and epistemology. Indeed, wherever there is a bona fide likelihood, there is room for Bayes.

References

[1] M. M. Barbieri and J. O. Berger, *Optimal predictive model selection*, The Annals of Statistics **32** (2004), 870–897.

[2] J. Besag, *On the statistical analysis of dirty pictures*, Journal of the Royal Statistical Society. Series B **48** (1986), 259–302.

[3] D. M. Blei, A. Y. Ng and M. I. Jordan, *Latent Dirichlet allocation*, Journal of Machine Learning Research **3** (2003) 993–1022.

[4] M. Corsetti and E. Fokoué, *Nonnegative matrix factorization with Zellner penalty*, Open Journal of Statistics **5** (2015), 777–786.

[5] M. Corsetti and E. Fokoué, *Nonnegative matrix factorization with Toeplitz penalty*, Journal of Informatics and Mathematical Sciences **10** (2018), 201–215.

[6] L. Csató, E. Fokoué, M. Opper, B. Schottky and O. Winther, *Efficient approaches to Gaussian process classification*, in: S. A. Solla, T. K. Leen and K. Müller (eds.), Advances in Neural Information Processing Systems **12** (NIPS 1999), MIT Press, 2000, pp. 251–257.

[7] L. Devroye, L. Györfi and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Stochastic Modelling and Applied Probability, Springer, New York, 1997.

[8] T. S. Ferguson, *A Bayesian analysis of some nonparametric problems*, The Annals of Statistics **1** (1973), 209–230.

[9] T. S. Ferguson, *Prior distributions on spaces of probability measures*, The Annals of Statistics **2** (1974), 615–629.

[10] E. Fokoué, *Estimation of atom prevalence for optimal prediction*, Contemporary Mathematics **443** (2008), 103–129.

[11] E. Fokoué, *Foundational aspects of the theory of statistical function estimation and pattern recognition*, Bulletin of PFUR, Series Mathematics, Information Sciences, Physics, No. 3, 2008, pp. 40–54.

[12] E. Fokoué, *Bayesian computation of the intrinsic structure of factor analytic models*, Journal of Data Science **7** (2009), 285–311.

[13] E. Fokoué, *Stable radial basis function selection via mixture modelling of the sample path*, Journal of Data Science **9** (2011), 345–358.

[14] E. Fokoué, *To Bayes or not to Bayes? That's no longer the question*, Indian Bayesian Society Newsletter **XV** (2018), 2–5.

[15] E. Fokoué, D. Sun and P. Goel, *Fully Bayesian analysis of the relevance vector machine with extended prior*, Statistical Methodology **8** (2011), 83–96.

[16] E. Fokoué and D. M. Titterington, *Mixtures of factor analysers: Bayesian estimation and inference by stochastic simulation*, Machine Learning **50**, 73–94.

[17] M. Fortunato, C. Blundell and O. Vinyals, *Bayesian recurrent neural networks*, CoRR, abs/1704.02798, 2017.

[18] J. Friedman, T. Hastie and R. Tibshirani, *Regularization paths for generalized linear models via coordinate descent*, Journal of Statistical Software **33** (2010), 1–22.

[19] A. E. Gelfand and A. F. M. Smith, *Sampling-based approaches to calculating marginal densities*, Journal of the American Statistical Association **85** (1990), 398–409.

[20] A. J. Haug, *Bayesian estimation for target tracking: part I, general concepts*, Wiley Interdisciplinary Reviews: Computational Statistics **4** (2012), 375–383.

[21] A. J. Haug, *Bayesian estimation for target tracking: part ii, the Gaussian sigma-point Kalman filters*, Wiley Interdisciplinary Reviews: Computational Statistics **4** (2012), 489–497.

[22] A. J. Haug, *Bayesian estimation for target tracking, part iii: Monte Carlo filters*, Wiley Interdisciplinary Reviews: Computational Statistics **4** (2012), 498–512.

[23] R. Henao, X. Yuan and L. Carin, *Bayesian nonlinear support vector machines and discriminative factor modeling*, in: Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence and K. Q. Weinberger (eds.), Advances in Neural Information Processing Systems **27**, Curran Associates, Inc., 2014, pp. 1754–1762.

[24] A. E. Hoerl and R. W. Kennard, *Ridge regression: Applications to nonorthogonal problems*, Technometrics **12** (1970), 69–82.

[25] A. E. Hoerl and R. W. Kennard, *Ridge regression: Biased estimation for nonorthogonal problems*, Technometrics **12** (1970), 55–67.

[26] J. A. Hoeting, D. Madigan, A. E. Raftery and C. T. Volinsky, *Bayesian model averaging: A tutorial*, Statistical Science **14** (1999), 382–417.

[27] H. Ishwaran and J. S. Rao, *Spike and slab variable selection: Frequentist and Bayesian strategies*, The Annals of Statistics **33** (2005), 730–773.

[28] F. Liang, R. Paulo, G. Molina, M. A. Clyde and J. O. Berger, *Mixtures of g priors for Bayesian variable selection*, Journal of the American Statistical Association **103** (2008), 410–423.

[29] D. J. C. MacKay, *Bayesian interpolation*, Neural Computation **4** (1991), 415–447.

[30] R. M. Neal, *Bayesian Learning for Neural Networks*, Springer-Verlag, New York, Secaucus, NJ, USA, 1996.

[31] R. M. Neal, *Priors for infinite networks*, Bayesian Learning for Neural Networks, Lecture Notes in Statistics **118**, Springer, New York, NY, 1996, pp. 29–53.

[32] M. Plummer, *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling*, in: Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), 2003, 10 pp.

[33] A. E. Raftery, D. Madigan and J. A. Hoeting, *Bayesian model averaging for linear regression models*, Journal of the American Statistical Association *92* (1997), 179–191.

[34] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, Adaptive Computation and Machine Learning, MIT Press, Cambridge, MA, USA, 2006.

[35] N. Simon, J. Friedman, T. Hastie and R. Tibshirani, *Regularization paths for Cox's proportional hazards model via coordinate descent*, Journal of Statistical Software **39** (2011), 1–13.

[36] P. Sollich, *Bayesian methods for support vector machines: Evidence and predictive class probabilities*, Machine Learning **46** (2002), 21–52.

[37] R. Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society, Series B **58** (1996), 267–288.

[38] M. E. Tipping, *Sparse Bayesian learning and the relevance vector machine*, Journal of Machine Learning Research **1** (2001), 211–244.

[39] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, 2000.

[40] V. N. Vapnik, S. E. Golowich and A. J. Smola, *Support vector method for function approximation, regression estimation and signal processing*, in: M. I. Jordan, M. C. Mozer and

T. Petsche (eds.), Advances in Neural Information Processing Systems **9**, MIT Press, 1997, pp. 281–287.

[41] C. K. I. Williams and C. E. Rasmussen, *Gaussian processes for regression*, in: D. S. Touretzky, M. C. Mozer and M. E. Hasselmo (eds.), Advances in Neural Information Processing Systems **8**, MIT Press, 1996, pp. 514–520.

Ernest Fokoué, School of Mathematical Sciences, Rochester Institute of Technology, 98 Memorial Drive, Rochester, NY 14623, USA
e-mail: `epfeqa@rit.edu`