mathematics

MDPI

Article

Bayes in Wonderland! Predictive Supervised Classification Inference Hits Unpredictability

Ali Amiryousefi *, Ville Kinnula and Jing Tang

Research Program in Systems Oncology, Faculty of Medicine, University of Helsinki, 00014 Helsinki, Finland; ville.kinnula@helsinki.fi (V.K.); jing.tang@helsinki.fi (J.T.)
* Correspondence: ali.amiryousefi@helsinki.fi

Abstract: The marginal Bayesian predictive classifiers (mBpc), as opposed to the simultaneous Bayesian predictive classifiers (sBpc), handle each data separately and, hence, tacitly assume the independence of the observations. Due to saturation in learning of generative model parameters, the adverse effect of this false assumption on the accuracy of mBpc tends to wear out in the face of an increasing amount of training data, guaranteeing the convergence of these two classifiers under the de Finetti type of exchangeability. This result, however, is far from trivial for the sequences generated under Partition Exchangeability (PE), where even umpteen amount of training data does not rule out the possibility of an unobserved outcome (Wonderland!). We provide a computational scheme that allows the generation of the sequences under PE. Based on that, with controlled increase of the training data, we show the convergence of the sBpc and mBpc. This underlies the use of simpler yet computationally more efficient marginal classifiers instead of simultaneous. We also provide a parameter estimation of the generative model giving rise to the partition exchangeable sequence as well as a testing paradigm for the equality of this parameter across different samples. The package for Bayesian predictive supervised classifications, parameter estimation and hypothesis testing of the Ewens sampling formula generative model is deposited on CRAN as *PEkit* package.

Keywords: partition exchangeability; supervised classification; hypothesis testing; asymptotic statistics; predictive classifiers

MSC: 62C10; 62-08; 62-04

Citation: Amiryousefi, A.; Kinnula, V.; Tang, J. Bayes in Wonderland! Predictive Supervised Classification Inference Hits Unpredictability. *Mathematics* **2022**, *10*, 828. https://doi.org/10.3390/math10050828

Academic Editors: Andrea Prati, Luis Javier García Villalba and Vincent A. Cicirello

Received: 1 February 2022
Accepted: 3 March 2022
Published: 5 March 2022

1. Introduction

Under the broad realm of inductive inference, the goal of the supervised classification is to assign the test objects into a priori defined number of classes learned from the training data [1]. While the choice of a proper inductive operator is often due to an algorithmic search, still one of the most applicable frameworks that optimally handle these scenarios is Bayesian [2]. This is the result of a formulation that with a given prior information and accruing observed data, gradually enhances the precision of the inferred population's parameters [3]. In these cases, with increasing amount of observed data, the accuracy of the predictions regarding the generative models parameters will asymptotically improve. In general, this asymptotic improvement is observed for the behaviour of the classification of maximum likelihood estimates (MLE) [4]. This is the underlying factor for the more specific case where the number of possible outcomes in a given class is predefined and is essentially a closed set of discrete species [5]. The asymptotic learning saturation tailored to the predictive classifiers is due to the fact that the learning of the underlying proportional probabilities assigned to each observing species is getting more accurate based on the law of large numbers [6]. This is, however, not intuitively generalizable to the cases where there is always a positive probability in observing new species. We consider here the general supervised classification case where the sets of species observed for features are

not closed *a priori*, rendering the positive probability of the appearance of new species at any stage. In the limiting form, the mixture of this species sampling sequence is commonly used in Bayesian nonparametric scenarios [7]. In our case, the de Finetti type of exchangeability [8], seems intractable as it mandates the exchangeability under a defined set of species. Nevertheless, under this asymptotic scenarios, one solution is to adhere to a form of partition exchangeability (PE; due to Kingman [9]) that allows the exchangeability of not only *time* indices of observed outcomes but also the *frequency* of the observed species [10]. Next to presentation of the predictive rules under this type of exchangeablity [11] and asymptotic representation of the number of classes for the sampling species sequences [12], our derivation based on the Bayesian classifiers shows that given an infinite amount of data, the simultaneous and marginal predictive classifiers will converge asymptotically. This is congruent with the similar study under the de Finetti exchangeability with multinomial modelling [5]. In general, due to the existence of marginal dependency between the data points, the simultaneous and marginal classifiers are not necessarily equal. On the other hand, their convergence is not intuitive due to the complication posed with a priori unfixed set of observable species. Upon availability of umpteen amount of data, however, the results presented here justifies the replacement of the marginal classifiers with the computationally expensive simultaneous ones.

In the following section, we derive the MLE of the single parameter of the underlying generative model assigned to the data under PE and Lagrangian hypothesis testing scheme regarding this parameter for one-sample and multiple-sample tests. This will be followed by a section on the derivation and computational performance of the marginal Bayesian predictive classifiers (mBpc) and simultaneous Bayesian predictive classifiers (sBpc), and their relative asymptotic behaviour in different classifications settings. We also introduce the algorithms used to implement the classifiers along with classification results on simulated datasets from partition exchangeable sequences. Lastly, some notes regarding the adequacy of each model and its specific use is provided in the last section. We also provide the implementation of all the derived functions presented in this paper as a freely available R package deposited on CRAN as PEkit package (https://cran.r-project.org/web/packages/PEkit/index.html, accessed on 23 December 2021).

2. Partition Exchangeability

Assume that number of species related to our feature is unfixed a priori. Upon availability of the vector of test labels S, under the PE framework, we can deduce the sufficient statistic for each subset of data. To define this statistic, consider the assignment of arbitrary permutation of integers $1, \ldots, |s_c|$ to the items in s_c where $n_c = |s_c|$ is the size of a given class c. Introducing the $I(\ldots)$ indicator function and $n_{cl} = \Sigma_{i \in s_c} I(x_i = l)$ as the frequency of items in class c having value $l \in \chi$ (where χ is the species set $\chi = \{1, 2, ..., r\}$ and $x_i \in \mathbf{x}$ is the ith test item), then in terms of count of the test data in the class c, $\mathbf{x}^{(c)}$ as shown in [13], the sufficient statistic is

$$\rho_{ct} = \sum_{l=1}^{\infty} I(n_{cl} = t), \tag{1}$$

Note that the above is the sufficient statistic for a specific class c. The vector of sufficient statistic $\rho_c = (\rho_{ct})_{t=1}^{n_c}$ indicates a partition of the integer n_c such that ρ_{ct} is the frequency of specific feature values that have been observed only t times in class c of test data. Given the above formulation [14], the random partition is exchangeable if and only if two different sequences having the same vector of sufficient statistics have the same probabilities of occurring. Independent of the classes and according to Kingman's representation theorem [9], the probability distribution of the vector of n observations with sufficient statistic $\boldsymbol{\rho} = (\rho_1, \rho_2, \ldots, \rho_n)$ with $\rho_t = \sum_{l=1}^{\infty} I(n_l = t)$ for $t = 1, 2, \ldots, n$, under partition exchangeability will follow the Poisson–Dirichlet (ψ) (PD) distribution.

This distribution (also known as the Ewens sampling formula [15]) and its comprehensive review, history and its applications are thoroughly discussed in [16] and is presented as;

$$p(\boldsymbol{\rho}|\psi) = \frac{n!}{\psi(\psi+1)\dots(\psi+n-1)} \prod_{t=1}^{n} \left\{ (\frac{\psi}{t})^{\rho_t} \frac{1}{\rho_t!} \right\}, \tag{2}$$

where,

$$\forall\, \psi \in \mathbb{R}^+, \;\; \boldsymbol{\rho} \in \mathfrak{S}_\rho,$$

$$\mathfrak{S}_\rho = \left\{ (\rho_1, \rho_2, \dots \rho_n) \mid \sum_{i=1}^{n} i\rho_i = n, \rho_i \in \mathbb{N}_0, i = 1, 2, \dots n \right\}.$$

Note that the ψ is the only parameter of the PD distribution and is controlling the level of diversity of the observed species such that values close to zero for this parameter is leading to a more degenerative outcomes and bigger values are underlying the higher diversities.

2.1. Parameter Estimation

As one of the main goals of the inductive inference in learning about the underlying parameters of the observed data, we are interested in the MLE of ψ as the dispersal parameter of (2). The MLE of the parameter $\hat{\psi}$ can be derived from a sample as demonstrated in Appendix A by solving the following equation for ψ .

$$\sum_{j=1}^{n} \frac{\psi}{\psi + j - 1} = \sum_{t=1}^{n} \rho_t. \tag{3}$$

The sum of sufficient statistics on the right side of the equation equals the observed number of distinct species in the sample. The left side of the equation equals, as shown by Ewens in [15], the expected number of distinct species observed given parameter ψ. Thus, the MLE of ψ is that number for which the observed number of distinct species equals the expected number of species observed in a sample of size n. There is no closed form solution for this equation for arbitrary value of n, so the MLE $\hat{\psi}$ has to be obtained numerically. As the right side of the equation is a strictly increasing function when $\psi > 0$, a binary search algorithm is adhered to find the root of (3) in PEkit (https://cran.r-project.org/web/packages/PEkit/index.html, accessed on 23 December 2021). For assessing the similarity of the samples under different classes in the supervised classification scenario, an estimate for ψ can be calculated for each class in the training data separately.

2.2. Hypothesis Testing

Once the parameters are estimated based on the MLE, the second question is about the similarity or the degree of their distinctiveness across different samples. Toward this end, we also devised distinguished statistical hypothesis testing frameworks for assessing the similarity of the estimated parameters with either a single value or across samples.

One-sample test . A Lagrange multiplier test as defined in [17] can be used for statistical testing of a hypothesized parameter ψ_0 under the null hypothesis $H_0 : \psi = \psi_0$ for a single sample. The test statistic is constructed as follows:

$$S(\psi_0) = \frac{U(\psi_0)^2}{I(\psi_0)}, \tag{4}$$

where U is the gradient of the log-likelihood $L(\psi)$, and I is the Fisher information of the distribution. Under the null hypothesis the test statistic S follows the χ^2_1-distribution.

In the case of the PD distribution, these quantities become (the proof of is provided in Appendix B).

$$U(\psi_0) = \sum_{i=1}^{n} \left(\frac{\rho_i}{\psi_0} - \frac{1}{\psi_0 + i - 1} \right) \tag{5}$$

$$I(\psi_0) = \sum_{i=1}^{n} \left(\frac{1}{\psi_0(\psi_0 + i - 1)} - \frac{1}{(\psi_0 + i - 1)^2} \right). \tag{6}$$

Multiple-sample test. To infer whether there is a statistically significant difference in the ψ of each sample, we have devised a Likelihood Ratio Test (LRT) [18]. The null hypothesis of the test is that there is no difference in the ψ of each of s samples, $H_0 : \psi_1 = \psi_2 = \ldots = \psi_s$, and consequently the alternative hypothesis is the inequality of at least two ψ. The test statistic Λ is constructed as

$$\Lambda = -2\, log \frac{sup\, \mathcal{L}(\theta_0)}{sup\, \mathcal{L}(\hat{\theta})} \xrightarrow{\mathcal{D}} \chi^2_d, \tag{7}$$

where $\mathcal{L}(\theta_0)$ is the likelihood function of the data given the model under the null-hypothesis, and $\mathcal{L}(\hat{\theta})$ is the unrestricted likelihood of the model. The sup refers to *supremum*, so the likelihood is evaluated at the MLE of the parameters. Λ asymptotically converges in distribution to the χ^2_d, where d equals the difference in the amount of parameters between the models. When testing the ψ of s different samples from the PD distribution with possibly different sample sizes n_s, the model under the null hypothesis has one shared ψ, while the unrestricted model has s different dispersal parameters, $(\psi_1, \psi_2, \ldots, \psi_s)$, so $d = s - 1$. The likelihood $\mathcal{L}$ of multiple independent samples from the PD distribution is a product of the density functions for the partitions ρ of those samples, and under H_1 the MLE of ψ for each of the samples is evaluated as in (3) from each sample independently, as the other samples have no effect on the ψ of a single sample. However, under the null-hypothesis, the samples share a common ψ, the MLE of which according to Section 2.1 would be estimated by solving,

$$\sum_{i=1}^{s} \sum_{j=1}^{n_s} \frac{\psi}{\psi + j - 1} = \sum_{i=1}^{s} \sum_{t=1}^{n_s} \rho_{st}. \tag{8}$$

As the likelihood of multiple independent samples with identical ψ under the null-hypothesis is just a product of their likelihoods, the derivative of the log-likelihood is just a sum of the derivatives of each sample's likelihood. Again, the ψ has to be obtained as the root of the above equation, which can be found with a small modification to the same binary algorithm as the one we use to determine the MLE of a single parameter of the PD distribution. Having found the MLEs of the ψ under null and alternative hypotheses, the likelihood ratio in the test statistic in Equation (7) can then be expressed as,

$$\prod_{j=1}^{s} \mathcal{L}(\boldsymbol{\rho}_j | \hat{\psi}_j) \;/\; \prod_{j=1}^{s} \mathcal{L}(\boldsymbol{\rho}_j | \hat{\psi}_j), \tag{9}$$

where the likelihood function $\mathcal{L}(\boldsymbol{\rho}_j | \hat{\psi}_j)$ is the likelihood function of the PD$(\boldsymbol{\rho}_j | \psi)$ distribution for the partition $\boldsymbol{\rho}$ of the j-th sample. As the restricted model in the numerator can never have a larger likelihood than the unrestricted likelihood in the denominator, and both are positive real numbers, this ratio is bounded between 0 and 1.

3. Supervised Classifiers under PE

Now with the availability of the estimation and testing machinery for the dispersal parameter of the PD distribution, we can tackle the diverse and controlled scenarios of the supervised classification under partition exchangeability. Consider the set of m available training items by M and correspondingly the set of n test items by N. For each item, we

observe only one feature that can take value from species set $\chi = \{1, 2, ..., r\}$. Note that each number in χ is represented with one species such that the first species observed is represented with integer 1, the second species is represented with integer 2, and so on. On the other hand, r is not known a priori, denoting the fact that we are uninformative about all of the species possible in our population. A training item $i \in M$ is characterized by a feature z_i such that, $z_i \in \chi$. Similarly, we have for a test item $i \in N$ the feature x_i such that, $x_i \in \chi$. Collections of the training and test data features are denoted by vectors $\mathbf{z}$ and $\mathbf{x}$, respectively. Furthermore consider that the training data are allocated into k distinct classes and T is a joint labeling of all the training items into these classes. Simultaneous supervised classification will assign labels to all the test data in N in a joint manner. We can consider partitioning of N test elements into k different classes similar to T such that $S = (s_1, \ldots, s_k), s_c \subseteq N, c = 1, ..., k$ be the joint labeling of this partition. The T and S structures indicate a partition of the training and test feature vectors, such that $\mathbf{z}^{(c)}$ and $\mathbf{x}^{(c)}$ represent the subset of training and test items in class $c = 1, ..., k$, respectively. The S denote the space of possible simultaneous classifications for a given N and so $S \in \mathbb{S}$. We also note that the predictive probability of observing a new feature value of species j given a set of prior observations from the PD distribution is

$$p(x_{n+1} = j|\mathbf{n}) = \frac{n_j}{N + \psi} \tag{10}$$

where n_j is the frequency of species j in the observed set $\mathbf{n}$. However, if the value j is of a previously unobserved species, the predictive probability is

$$p(x_{n+1} = j|\mathbf{n}) = \frac{\psi}{N + \psi}. \tag{11}$$

The proof of this arises both from the mechanics of an urn model that [19] showed to generate the Ewens sampling formula, as well as from [20]. The predictive probability of a previously unseen feature value is thus higher for a population with a larger parameter ψ than it is for a population of equal size with a smaller ψ.

mBpc. Now the product predictive distribution for all test data that is assumed to be independent and identically distributed (*i.i.d.*) for the marginal classifier under the partition exchangeability framework becomes:

$$\begin{aligned} p_{mBpc} &= \prod_{c=1}^{k} \prod_{i:S_i \in c} p(x_i = l_i|\mathbf{z}^{(c)}, T^{(c)}, S_i = c) \\ &= \prod_{c=1}^{k} \prod_{i:S_i \in c} \left(\frac{m_{i;cl}}{m_c + \hat{\psi}_c} \right)^{I(m_{i;cl} \neq 0)} \times \\ &\quad \left(\frac{\hat{\psi}_c}{m_c + \hat{\psi}_c} \right)^{I(m_{i;cl} = 0)}. \end{aligned} \tag{12}$$

Note that under a maximum a posteriori rule with training data of equal size for each class, a newly observed feature value that has not been previously seen in the training data for any class, will always be classified to the class with the highest estimated dispersal parameter $\hat{\psi}$. However, each class still has a positive probability of including previously unseen values based on the observed variety of distinct values previously observed within the class.

sBpc. Analogously to the product predictive probability for the case of general exchangeability in [5], the product predictive distribution of the simultaneous classifier under partition exchangeability, is then defined as

$$
\begin{aligned}
p_{sBpc} &= \prod_{c=1}^{k} \prod_{i:S_i \in c} p(x_i = l_i | \mathbf{z}^{(c)}, T^{(c)}, S_i = c) \\
&= \prod_{c=1}^{k} \prod_{i:S_i \in c} \left(\frac{m_{i;cl} + n_{i;cl}}{m_c + n_{i;cl} + \hat{\psi}_c} \right)^{I(m_{i;cl} \neq 0)} \times \\
&\quad \left(\frac{\hat{\psi}_c}{m_c + n_{i;cl} + \hat{\psi}_c} \right)^{I(m_{i;cl}=0)}.
\end{aligned} \tag{13}
$$

An asymptotic relationship between these classifiers is immediately apparent in these predictive probabilities. As the amount of training data in each class m_c increases, the impact of class-wise test data n_c becomes negligible in comparison, and the difference in the predictive probabilities approaches zero asymptotically. As the classifiers are searching for classification structures S that optimize the test data predictive probability given the training data, and the predictive probabilities converge asymptotically, the classifiers are searching for the same optimal labeling.

However, the classifiers handle values unseen in the training data differently. This is the situation where $m_{i;cjl} = 0$ in the predictive probabilities p_{mBpc} and p_{sBpc} The marginal classifier's predictive probability for the test data is always maximized by assigning such a value into the class with the highest $\hat{\psi}$. The simultaneous classifier, however, considers the assignment of other instances of an unseen value as well. This can lead to optimal classification structures where different instances of an unseen value are classified into different classes. Thus, the convergence of the test data predictive probabilities of the marginal and simultaneous classifiers is not certain in the presence of unseen values. In practice, though, as the amount of training data m tends to infinity, the probability of observing new values from the PD(ψ) distribution presented in Equation (11) tends to 0; $\psi/(\psi + m) \rightarrow 0$ as $m \rightarrow \infty$. Unexpected values would be very rare and would have a minimal effect on the classifications made by the different classifiers causing the asymptotic convergence of these classifiers as also proved in [13].

Algorithms for the Predictive Classifiers

Next to the representation of the probabilities of the sBpc and mBpc above, we are interested in computational algorithms for calculating the observed values. Adopting the learning algorithms defined in [5] for obtaining the numerical outcomes of (12) and (13) given observed data, we note that these predictive probabilities normalizing constant is irrelevant for the optimization task. Hence, the computationally more straightforward marginal classifier with the exigency of individual classification of each test data point upon their arrival according to a maximum a posteriori rule would be

$$
\hat{p}_{mBpc} : \underset{c=1,\ldots,k}{argmax}\ p(S_i = c | \mathbf{z}^{(M)}, x_i, T), \tag{14}
$$

which is essentially a choice of a class that maximizes the probability of a given test data to belong. For the simultaneous classification algorithm on the other hand, define classification structure $S_c^{(i)}$ to be an identical classification structure to S with only the item i reclassified to class c. The greedy deterministic algorithm with its maximization algorithm instructions in this scenario is defined as

$$
\hat{p}_{sBpc} : \underset{S \in \mathbb{S}}{argmax}\ p(S | \mathbf{x}^{(N)}, \mathbf{z}^{(M)}, T), \tag{15}
$$

1. Set an initial S_0 with the marginal classifier algorithm $\hat{p}_{mBpc}$.

2. Until S remains unchanged between iteration, do for each test item $i \in N$:

$$\hat{p}_{sBpc} : \underset{S \in \{S_1^{(i)}, \ldots, S_k^{(i)}\}}{argmax} p(S|\mathbf{x}^{(N)}, \mathbf{z}^{(M)}, T). \tag{16}$$

The simultaneous classifier thus is a more elegant construct based on the marginal classifier where for each test item, the potential labelling of other test items is also taken into account.

4. Numerical Illustrations Underlying Convergence

To study the classification performance and convergence of the two different classifiers, we simulated both training and test datasets from the PD distribution with a generative urn model described in [10] called the De Morgan process. We varied the amount of training and test data, the values of the distribution parameter ψ, as well as the amount of classes k. A pool of 4 million data points was created for each ψ. A small sample of size 1000 was used to train the model first. This sample was augmented with more data from the same pool for each ψ up to 2 million (without replacement) to form our two *small* and *big* training data sizes. A test dataset created with the same parameters was kept constant for classification with all training data samples as reflected in Table 1 and Figure 1.

Table 1. The item-wise 0–1 classification error for the marginal and simultaneous classifiers, as well as the 0–1 absolute difference between the predicted labels between the two classifiers (δ).

Training (m)	Test (n)	No. Clusters (k)	Dispersion (ψ)	$\hat{p}_{mBpc}$	$\hat{p}_{sBpc}$	$\lvert\delta\rvert$
2×10^6	2×10^3	2	1, 2	0.491	0.491	0
10^3	2×10^3	3	1, 10, 50	0.3408	0.2823	0.0626
10^5	2×10^3	3	1, 10, 50	0.2768	0.2758	0.0010
10^3	2×10^3	5	$1, 100, 10^3, 5 \times 10^3, 10^4$	0.7535	0.5865	0.167
2×10^6	2×10^3	5	$1, 100, 10^3, 5 \times 10^3, 10^4$	0.434	0.364	0.115

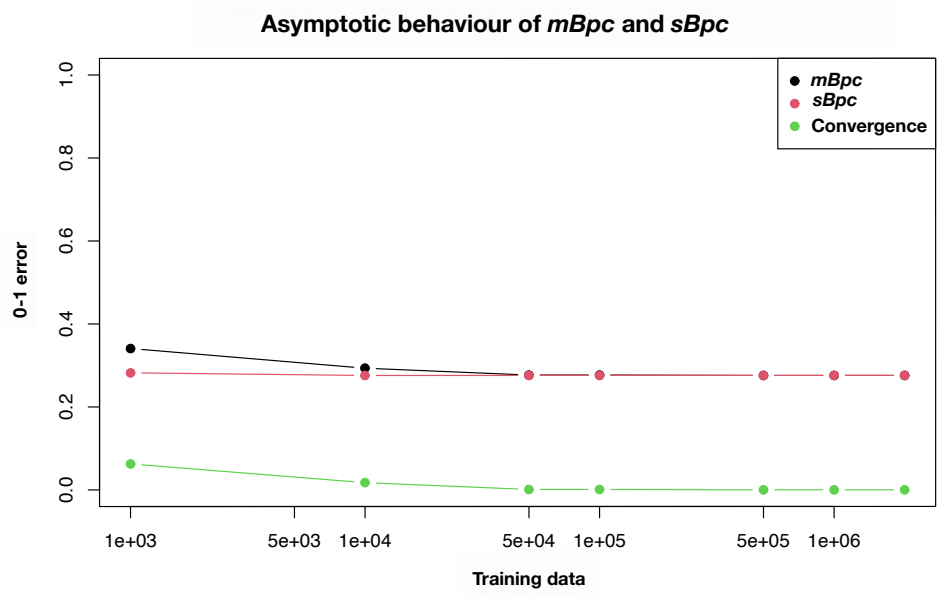

Figure 1. The classification error of the marginal and simultaneous classifiers with datasets from PD distribution with $\psi \in (1, 10, 50)$, as well as the convergence of their labellings. Rows 2 and 3 in Table 1 are included in the figure.

The above results next to our experiments show that the simultaneous classifier performs better than the marginal classifier especially as the number of the classes is higher and training data is small. As the amount of training data increases, the predictions of

the classifiers converge, as the extra information in the test data used by the simultaneous classifier becomes negligible. This behaviour is illustrated in Figure 1. Additionally, the first row of the Table 1 shows that the classifier fails on the binary classification task with datasets created with ψ of 1 and 2. This is because the classes created with ψ of such similar magnitude will be distributed very similarly, tailored both from similarity of these values and their closeness to zero. Thus the classification results are no better than random guessing (accuracy of 0.491). The data is in fact so homogeneous in this case, that the labelings of the two different classifiers already converge with test data of size 1000. On the other hand, the biggest difference underlies the 4th row where, despite a clear distinction between the competing classes ψ, the sBpc is distinctively performing better than mBpc but both predictions are even worse than random guessing. This is due to the small training samples and, as marked, the improvement is considerable for each classifier and also their difference is decreasing. Compared with the three-class groups of the second and third row of the Table 1, we conclude that the number of classes is playing a more pivotal role in accuracy of the classifiers than the distinction between the underlying dispersal parameters. However, this dominating factor is in service to the number of training data as well such that increasing amount of training data could overrule this differentiating cause.

Altogether, the above results support the hypothesis that the marginal and simultaneous classifiers converge in their labelings with enough training data. This justifies the use of the marginal classifier in place of the more accurate simultaneous one with large data. Still, the simultaneous classifier is more accurate with smaller training data, as it benefits from the information in the test data.

5. Discussion

Previously unseen or unanticipated feature values are challenging scenarios in standard Bayesian inductive inference. Under general exchangeability as formalized by de Finetti, upon such an observation, the entire alphabet of anticipated feature values must be retrospectively changed. Due to consideration of the specificity of the type of data being generated under PE, the classifiers introduced here, however, are based on a fully probabilistic framework and, as such, are equipped with a specific formulation that allow the updating of their predictions autonomously. This choice of classifiers are shown to have multiple advantages in population structure studies where there is a need for accommodating the appearance of the unseen species [21]. The classification of previously unseen feature values here is handled through the use of a parameter learned from the training data, instead of, for example, using an uninformative Dirichlet-prior [6]. The superiority in classification accuracy of labeling the test data simultaneously instead of one by one was also shown under partition exchangeability. The assumption that test data is *i.i.d*. is obviously an unrealistic one. The two classifiers considered here only converge in prediction as the amount of training data approaches infinity. This is a special case of the coincidence of the exchangeablity laws in the limiting forms for the mixture of the infinitely divisible distributions under specific conditions [22]. From the computational point of view, however, the simultaneous classifier exhibits the exponentially increasing cost as the amount of test data increases, making it unfeasible to use for large-scale prediction. Additionally, the algorithm presented here is only capable of arriving at local optima, leaving still the chances of being sub-optimal in multimodal scenarios. Further research could be directed at the Gibbs sampler-assisted algorithms presented in [5] to more thoroughly scan the optimization space, although the convergence of such algorithms with large test datasets is also uncertain. An implementation for the supervised classifiers considered could also be extended for the two parameter distribution Poisson–Dirichlet (α, ψ) [23] (Appendix C).

6. Conclusions

We provided an estimation and hypothesis testing scheme for the dispersal parameter of the PD distribution tied with the sequences under PE. These derivations next to the

algorithms for the computational calculation of the simultaneous and marginal Bayesian predictive classifiers and their implementations as a software package PEkit (https://cran.r-project.org/web/packages/PEkit/index.html, accessed on 23 December 2021) in R, would facilitate the smoother application of these classifiers and provides a conductive avenue toward PEs more effective disposition and prevalence in the community. In short, we provided the computational evidence that the marginal and simultaneous predictive classifiers under partition exchangeability converge in their classification assignment with presence of umpteen amount of training data. This asymptotic convergence of the classifiers, justifies the interchangeable use of them and hence, legitimizing forgoing of the sPpc's negligible superiority in accuracy, for the mPpc's huge computational advantage. Of course, under reasonably small amount of training data, this would be more of a preferential interplay, but altogether, these Bayesian classifiers are robust probabilistic judgmental apparatus for supervised classification tasks in the inductive inference domain. Next to these results, we underlie the presented derivations and exclusive tool for PE sequences. Namely, the hypothesis testing paradigm, random generation of the PD distribution, and its numerical MLE formulation, would be of a great use for the experiments underlying species sampling sequences.

Author Contributions: Conceptualization, A.A.; methodology, A.A. and V.K.; software, A.A. and V.K.; validation, A.A.; formal analysis, A.A. and V.K.; investigation, A.A. and V.K.; resources, J.T.; data curation, A.A. and V.K.; writing—original draft preparation, A.A. and V.K.; writing—review and editing, A.A. and J.T.; visualization, A.A. and V.K.; supervision, A.A.; project administration, A.A. and J.T.; funding acquisition, J.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Academy of Finland grant No. 320131.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The package for Bayesian predictive supervised classifications, parameter estimation and hypothesis testing of the Ewens Sampling Formula generative model as well as the synthetic data used in this article is deposited on CRAN as PEkit package (https://cran.r-project.org/web/packages/PEkit/index.html, accessed on 23 December 2021 and https://github.com/AmiryousefiLab/PEkit, accessed on 23 December 2021).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

PE	Partition Exchangeability
mBpc	marginal Bayesian predictive classifiers
sBpc	simultaneous Bayesian predictive classifiers
LRT	Likelihood Ratio Test
MLE	Maximum Likelihood Estimate
PD	Poisson–Dirichlet
i.i.d.	Independent and Identically Distributed

Appendix A. Maximum Likelihood Estimate

Here we provide the technical details for deriving the MLE in Equation (3) as well as the components U and I in the Lagrange Multiplier test described in Equations (5) and (6). The task is to find the first derivative of the logarithm of the Ewens sampling formula and its root in Equation (2), as well as the second derivative needed for the Fisher information.

$$
\begin{aligned}
L(\psi) &= log(n!) + \sum_{i=1}^{n}\{-log(\psi + i - 1) + \\
&\quad \rho_i \, log\, \psi - \rho_i \, log\, \rho_i - log(\rho_i!)\} \\
=> l(\psi) &= \sum_{i=1}^{n}\{-log(\psi + i - 1) \; + \rho_i \, log\, \psi\}
\end{aligned}
$$

$$
\begin{aligned}
l'(\psi) &= \sum_{i=1}^{n}\left(- \frac{1}{\psi + i - 1} + \frac{\rho_i}{\psi}\right) \\
&= U(\psi)
\end{aligned}
$$

The MLE is found by finding the root of the equation:

$$
\begin{aligned}
\sum_{i=1}^{n}\left(- \frac{1}{\hat{\psi} + i - 1} + \frac{\rho_i}{\hat{\psi}}\right) &= 0 \\
\sum_{j=1}^{n} \frac{\hat{\psi}}{\hat{\psi} + j + 1} &= \sum_{t=1}^{n} \rho_t
\end{aligned}
$$

Appendix B. Lagrange Multiplier Test

According to Ewens in [15], the left side of the above equation equals the expected number of unique values observed with this ψ and this sample size n, while the right side is the observed number of unique values:

$$
E[\sum_{t=1}^{n} \rho_t | \psi, n] = \sum_{i=1}^{n} \frac{\hat{\psi}}{\hat{\psi} + i - 1}
$$

This is needed for the Fisher information I, along with the second derivative of $l(\psi)$:

$$
\begin{aligned}
l''(\psi) &= \sum_{i=1}^{n}\left(\frac{1}{(\psi + i - 1)^2} - \frac{\rho_i}{\psi^2}\right) \\
&= \sum_{i=1}^{n} \frac{1}{(\psi + i - 1)^2} - \frac{k}{\sum_{i=1}^{n} \psi^2},
\end{aligned}
$$

where k denotes the observed amount of unique values in the sample. The Fisher information then becomes:

$$
\begin{aligned}
I(\psi) &= -E[l''(k; \psi) | \psi] \\
&= \frac{E[k|\psi]}{\sum_{i=1}^{n} \psi^2} - \sum_{i=1}^{n} \frac{1}{(\psi + i - 1)^2} \\
&= \sum_{i=1}^{n}\left(\frac{1}{\psi(\psi + i - 1)} - \frac{1}{(\psi + i - 1)^2}\right),
\end{aligned}
$$

where the expectation of k is as described from above.

Appendix C. A Note on Two-Parameter PD

A two-parameter formulation of the distribution presented in [23]. The added parameter in PD (α, ψ) is called the discount parameter. The role of the parameters is discussed in length in [16]. In short, the α is defined to fall in the interval $[-1, 1]$. When it is positive, it increases the probability of observing new species in the future proportional to the amount of already discovered species, while decreasing the probability of seeing the newly observed species again. When α is negative, the opposite is true and the number of new

species to be discovered is bounded. The single parameter $PD(\psi)$ is the special case of the two parameter distribution with α set to 0, $PD(0, \psi)$. The two parameter distribution is not considered further here, as the estimation of the parameters becomes a daunting task compared to the simple one parameter formalization.

References

1. Solomonoff, R. A formal theory of inductive inference. *Inf. Ctrl.* **1964**, *7*, 1–22.
2. Falco, I.D.; Cioppa, A.D.; Maisto, D.; Tarantino, E. A genetic programming approach to Solomonoff's probabilistic induction. In *European Conference on Genetic Programming*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 24–35.
3. Hand, D.J.; Yu, K. Idiot's Bayes: Not so stupid after all? *Int. Stat. Rev.* **2001**, *69*, 385.
4. Bryant, P.; Williamson, J.A. Asymptotic behaviour of classification maximum likelihood estimates. *Biometrika* **1978**, *65*, 273–281. [CrossRef]
5. Corer J.; Cui, Y.; Koski, T.; Siren, J. Have I seen you before? Principles of Bayesian predictive classification revisited. *Springer Stat. Comput.* **2011**, *23*, 59–73.
6. Quintana, F.A. A predictive view of Bayesian clustering. *J. Stat. Plan. Inference* **2006**, *136*, 2407–2429. [CrossRef]
7. Bassetti, F.; Ladelli, L. Mixture of Species Sampling Models. *Mathematics* **2021**, *9*, 3127. [CrossRef]
8. Barlow, R.E. Introduction to de Finetti (1937) foresight: Its logical laws, its subjective sources. In *Breakthroughs in Statistics*; Springer: New York, NY, USA, 1992; pp. 127–133.
9. Kingman, J.F.C. Random partitions in population genetics. *Proc. R. Soc. A Math Phys. Eng. Sci.* **1978**, *361*, 1–20.
10. Zabell, S.L. Predicting the unpredictable. *Harv. Bus. Rev.* **1992**, *90*, 205–232. [CrossRef]
11. Hansen, B.; Pitman, J. Prediction rules for exchangeable sequences related to species sampling. *Stat. Probab. Lett.* **2000**, *46*, 251–256. [CrossRef]
12. Bassetti, F.; Ladelli, L. Asymptotic number of clusters for species sampling sequences with non-diffuse base measure. *Stat. Probab.-Lett.* **2020**, *162*, 108749. [CrossRef]
13. Amiryousefi, A. Asymptotic Supervised Predictive Classifiers under Partition Exchangeability. *arXiv* **2021**, arXiv:2101.10950.
14. Kingman, J.F.C. The population structure associated with the Ewens sampling formula. *Theor. Popul. Biol.* **1977**, *11*, 274–283. [CrossRef]
15. Ewens, W. The Sampling Theory of Selectively Neutral Alleles. *Theor. Popul. Biol.* **1972**, *3*, 87–112. [CrossRef]
16. Crane, H. The Ubiquitous Ewens Sampling Formula. *Stat. Sci.* **2016**, *31*, 1–19. [CrossRef]
17. Radhakrishna Rao, C. Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Math. Proc. Camb. Philos. Soc.* **1948**, 44, 50–57. [CrossRef]
18. Neyman, J.; Pearson, E.S. On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. R. Soc. Lond. Ser. A Contain. Pap. Math. Phys. Character* **1933**, *231*, 289–337. [CrossRef]
19. Hoppe, F.M. Polya-like urns and the Ewens sampling formula. *J. Math. Biol.* **1984**, *20*, 91–94. [CrossRef]
20. Karlin, S.; McGregor, J. Addendum to a paper of Ewens. *Theor. Popul. Biol.* **1972**, *3*, 113–116. [CrossRef]
21. Corer J.; Gyllenberg, M.; Koski, T. Random partition models and exchangeability for Bayesian identification of population structure. *Bull. Math. Biol.* **2007**, *69*, 797–815.
22. Fortini, S.; Ladelli, L.; Regazzini, E. A Central Limit Problem for Partially Exchangeable Random Variables. *Theory Probab. Its Appl.* **1997**, *41*, 224–246. [CrossRef]
23. Pitman, J.; Yor, M. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.* **1997**, *25*, 855–900. [CrossRef]