Evaluating Bayesian Models with Posterior Dispersion Indices

Alp Kucukelbir [1] **Yixin Wang** [1] **David M. Blei** [1]

Abstract

Probabilistic modeling is cyclical: we specify a model, infer its posterior, and evaluate its performance. Evaluation drives the cycle, as we revise our model based on how it performs. This requires a metric. Traditionally, predictive accuracy prevails. Yet, predictive accuracy does not tell the whole story. We propose to evaluate a model through posterior dispersion. The idea is to analyze how each datapoint fares in relation to posterior uncertainty around the hidden structure. This highlights datapoints the model struggles to explain and provides complimentary insight to datapoints with low predictive accuracy. We present a family of posterior dispersion indices (PDI) that capture this idea. We show how a PDI identifies patterns of model mismatch in three real data examples: voting preferences, supermarket shopping, and population genetics.

1. Introduction

Probabilistic modeling is a flexible approach to analyzing structured data. Three steps define the approach. First, we specify a model; this captures our structural assumptions about the data. Then, we infer the hidden structure; this means computing (or approximating) the posterior. Last, we evaluate the model; this helps build better models down the road (Blei, 2014).

How do we evaluate models? Decades of reflection have led to deep and varied forays into model checking, comparison, and criticism (Gelman et al., 2013). But a common theme permeates all approaches to model evaluation: the desire to generalize well.

In machine learning, we traditionally use two complementary tools: predictive accuracy and cross-validation. Predictive accuracy is the target evaluation metric. Cross-validation captures a notion of generalization and justifies holding out data. This simple combination has fueled the development of myriad probabilistic models (Bishop, 2006; Murphy, 2012).

Does predictive accuracy tell the whole story? The predictive accuracy of an observation is its per-datapoint likelihood averaged over the posterior. In this sense, predictive accuracy reports a mean value for each datapoint; it ignores how each per-datapoint likelihood *varies* with respect to the posterior.

Main idea. We propose to evaluate probabilistic models through the idea of posterior dispersion, analyzing how each datapoint fares in relation to posterior uncertainty around the hidden structure. To capture this, we propose a family of posterior dispersion indices (PDI). These are per-datapoint quantities, each a variance to mean ratio of the datapoint's likelihood with respect to the posterior. A PDI highlights observations whose likelihoods exhibit the most uncertainty under the posterior.

Consider a model $p(\boldsymbol{x}, \boldsymbol{\theta})$ and the likelihood of a datapoint $p(x_n \mid \boldsymbol{\theta})$. It depends on some hidden structure $\boldsymbol{\theta}$ that we seek to infer. Since $\boldsymbol{\theta}$ is random, we can view the likelihood of each datapoint as a random variable. Predictive accuracy reports the average likelihood of each x_n with respect to the posterior $p(\boldsymbol{\theta} \mid \boldsymbol{x})$. But this ignores how the likelihood changes under the posterior. How can we capture this uncertainty and compare datapoints to each other?

To answer this, we appeal to various forms of dispersion, such as the variance of the likelihood under the posterior. We propose a family of dispersion indices in Section 2.2; they have the following form:

$$\begin{aligned}\text{PDI} &= \frac{\text{variance of likelihood under posterior}}{\text{mean of likelihood under posterior}} \\ &= \frac{\mathbb{V}_{\boldsymbol{\theta}\mid\boldsymbol{x}}[p(x_n \mid \boldsymbol{\theta})]}{\mathbb{E}_{\boldsymbol{\theta}\mid\boldsymbol{x}}[p(x_n \mid \boldsymbol{\theta})]}.\end{aligned}$$

PDIs compliment predictive accuracy. Here is a mental picture. Consider a nuclear power plant where we monitor the temperature of a pool of water. We train a probabilistic model; the posterior represents our uncertainty around some safe temperature, say 80 degrees. Suppose we receive a high measurement t_{high} (Figure 1). Its likelihood varies rapidly

[1]Columbia University, New York City, USA. Correspondence to: Alp Kucukelbir <alp@cs.columbia.edu>.

Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).

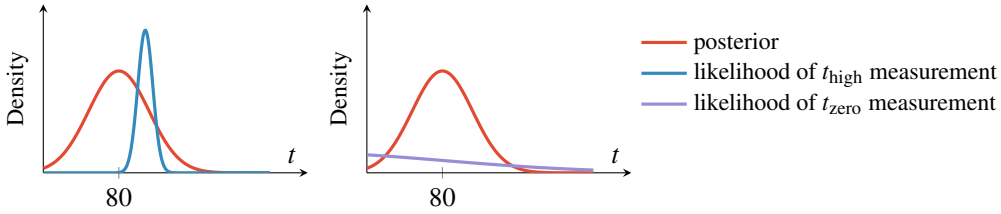

Figure 1. A mental picture of how PDI identify different types of mismatch. While both measurements (t_{high} and t_{zero}) exhibit low predictive accuracy, a PDI differentiates them by also considering how their per-datapoint likelihoods vary under the posterior. See text for more details.

across plausible posterior values for t. This datapoint is reasonably well modeled, but is sensitive to the posterior. Now imagine the thermostat breaks and we receive a zero measurement t_{zero}. This zero datapoint is poorly modeled.

Both datapoints may have similar predictive accuracy values under the model. (See Section 2.3 for how this can occur.) But the high measurement is different than the zero measurement. A PDI differentiates these measurements by considering not only their predictive accuracy scores, but also how their per-datapoint likelihoods vary with respect to the posterior.

Section 3 presents an empirical study of model mismatch in three real-world examples: voting preferences, supermarket shopping, and population genetics. In each case, a PDI provides insight beyond predictive accuracy and highlights potential directions for improvement.

Related work. This paper relates to a constellation of ideas from statistical model criticism.

The first connection is to analysis of variance: PDI bears similarity to ANOVA, which is a frequentist approach to evaluating explanatory variables in linear regression (Davison, 2003). Gelman et al. (1996) cemented the idea of studying predictive accuracy of probabilistic models at the data level; Vehtari et al. (2012) and Betancourt (2015) give up-to-date overviews of these ideas. PDIs add to this body of work by considering the variance of each datapoint in context of its predictive accuracy.

The second connection is to model comparison. Recent research, such as Gelman et al. (2014), Vehtari et al. (2014) and Piironen & Vehtari (2015), explore the relationship between cross validation and information criteria, such as the widely applicable information criterion (WAIC) (Watanabe, 2010; Vehtari et al., 2016). WAIC offers an intuitive connection to cross validation (Vehtari & Lampinen, 2002; Watanabe, 2015); we draw inspiration from it in this paper too. However our focus is on evaluation at the datapoint level, not at the dataset level. In this sense, PDIs and information criteria are complimentary tools.

The third connection is to a body of work from the machine learning community. Gretton et al. (2007) and Chwialkowski et al. (2016) developed effective kernel-based methods for independence and goodness-of-fit tests. Recently, Lloyd & Ghahramani (2015) visualized smooth regions of data space that the model fails to explain. In contrast, we focus directly on the datapoints, which can live in high-dimensional spaces that may be difficult to visualize.

A final connection is to scoring rules. While the literature on scoring rules originally focused on probability forecasting (Winkler, 1996; Dawid, 2006), recent advances draw new connections to decision theory, divergence measures, and information theory (Dawid & Musio, 2014). We discuss how PDIs fit into this picture in the conclusion.

2. Posterior Dispersion Indices

A posterior dispersion index (PDI) highlights datapoints that exhibit the most uncertainty with respect to the hidden structure of a model. Here is the road map for this section. A small case study illustrates how a PDI gives more insight beyond predictive accuracy. Definitions, theory, and another small analysis give further insight; a straightforward algorithm leads into the empirical study.

2.1. 44% outliers?

Hayden (2005) considers the number of days each U.S. president stayed in office; he submits that 44% of presidents may be outliers. Figure 2 plots the data. One-term presidents stay in office for around 1460 days; two-term presidents approximately double that. Yet many presidents deviate from this "two bump" trend.

A reasonable model for such data is a mixture of negative binomial distributions.[1] Consider the parameterization of the negative binomial with mean μ and variance $\mu + \mu^2/\phi$. Posit gamma priors on the (non-negative) latent variables. Set the prior on μ to match the mean and variance of the data (Robbins, 1964). Choose an uninformative prior on ϕ. Three mixtures make sense: two for the typical trends and one for the rest.

[1] A Poisson likelihood is too underdispersed.

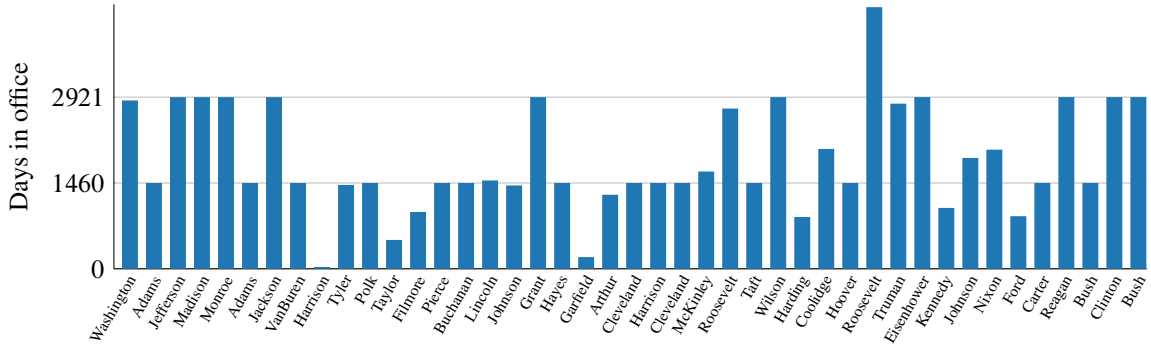

Figure 2. The number of days each U.S. president stayed in office. Typical durations are easy to identify; gray lines indicate one- and two-term stays. Appendix A presents numerical values.

The complete model is

$$
\begin{aligned}
p(\boldsymbol{\pi}) &= \text{Dirichlet}(\boldsymbol{\pi}\,;\, \alpha = (1,1,1)) \\
p(\boldsymbol{\mu}) &= \prod_{k=1}^{3} \text{Gam}(\mu_k\,;\, \text{mean and variance} \\
&\qquad\qquad\qquad\qquad \text{matched to that of data}) \\
p(\boldsymbol{\phi}) &= \prod_{k=1}^{3} \text{Gam}(\phi_k\,;\, a = 1, \beta = 0.01) \\
p(x_n \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\phi}) &= \sum_{k=1}^{3} \pi_k \text{NB}(x_n\,;\, \mu_k, \phi_k).
\end{aligned}
$$

Fitting this model gives posterior mean estimates $\widehat{\boldsymbol{\mu}} = (1461, 2896, \mathbf{1578})$ with corresponding $\widehat{\boldsymbol{\phi}} = (470, 509, \mathbf{1.3})$. The first two clusters describe the two typical term durations, while the third (highlighted in **bold red**) is a dispersed negative binomial that attempts to describe the rest of the data.

We compute a PDI (defined in Section 2.2) and the posterior predictive density for each president $p(x_n \mid \boldsymbol{x})$. Figure 3 compares both metrics and sorts the presidents according to the PDI.

Some presidents are clear outliers: Harrison [31: natural death], Roosevelt [4452: four terms], and Garfield [199: assassinated]. However, there are three presidents with worse predictive accuracy than Harrison: Coolidge, Nixon, and Johnson. A PDI differentiates Harrison from these three because his likelihood is varying rapidly with respect to the dispersed negative binomial cluster.

This PDI also calls attention to McKinley [1655: assassinated] and Arthur [1260: succeeded Garfield], because they are close to the sharp negative binomial cluster at 1460 but not close enough to have good predictive accuracy. They are datapoints whose likelihoods are rapidly changing with respect to a peaked posterior, like the high measurement in the nuclear plant example in the introduction.

This case study suggests that predictive probability does not tell the entire story. Datapoints can exhibit low predictive accuracy in different ways. We now turn to a formal definition of PDIs.

2.2. Definitions

Let $\boldsymbol{x} = \{x_n\}_1^N$ be a dataset with N observations. A probabilistic model has two parts. The first is the likelihood, $p(x_n \mid \boldsymbol{\theta})$. It relates an observation x_n to hidden patterns described by a set latent random variables $\boldsymbol{\theta}$. If the observations are independent and identically distributed, the likelihood of the dataset factorizes as $p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \prod_n p(x_n \mid \boldsymbol{\theta})$.

The second is the prior density, $p(\boldsymbol{\theta})$. It captures the structure we expect from the hidden patterns. Combining the likelihood and the prior gives the joint density $p(\boldsymbol{x}, \boldsymbol{\theta}) = p(\boldsymbol{x} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})$. Conditioning on observed data gives the posterior density, $p(\boldsymbol{\theta} \mid \boldsymbol{x})$.

Treat the likelihood of each datapoint as a function of $\boldsymbol{\theta}$. To evaluate the model, we analyze how each datapoint fares in relation to the posterior density. Consider these expectations and variances with respect to the posterior,

$$
\begin{aligned}
\mu(n) &= \mathbb{E}_{\boldsymbol{\theta}\mid\boldsymbol{x}}[p(x_n \mid \boldsymbol{\theta})] \\
\mu_{\log}(n) &= \mathbb{E}_{\boldsymbol{\theta}\mid\boldsymbol{x}}[\log p(x_n \mid \boldsymbol{\theta})] \\
\sigma^2(n) &= \mathbb{V}_{\boldsymbol{\theta}\mid\boldsymbol{x}}[p(x_n \mid \boldsymbol{\theta})] \\
\sigma^2_{\log}(n) &= \mathbb{V}_{\boldsymbol{\theta}\mid\boldsymbol{x}}[\log p(x_n \mid \boldsymbol{\theta})].
\end{aligned}
\tag{1}
$$

Each includes the likelihood in a slightly different fashion. The first expectation is a familiar object: $\mu(n)$ is the posterior predictive density.

A PDI is a ratio of these variances to expectations. Taking the ratio calibrates this quantity for each datapoint. Recall the mental picture from the introduction. The variance of

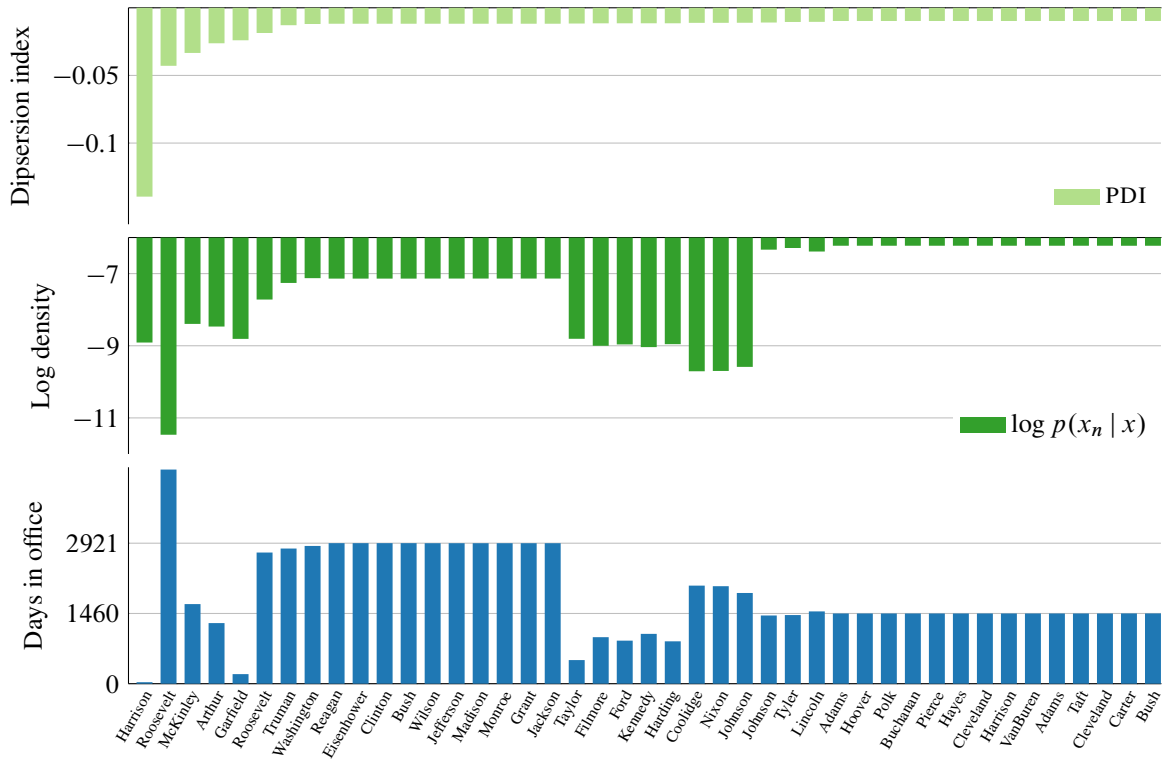

Figure 3. PDI and log predictive accuracy of each president under a mixture of three negative binomials model. Presidents sorted by PDI. The closer to zero, the better. (Code in supplement.)

the likelihood under the posterior highlights potential model mismatch; dividing by the mean calibrates this spread to its predictive accuracy.

Calibration puts all datapoints on a common scale. Imagine a binary classification problem where each datapoint y_n lives in $\{0, 1\}$. The variances of the zero measurements may be numerically quite different than the one measurements; considering the mean renders these values comparable.

Related ratios also appear in classical statistics under a variety of forms, such as indices of dispersion (Hoel, 1943), coefficients of variation (Koopmans et al., 1964), or the Fano factor (Fano, 1947). They all quantify dispersion of samples from a random process. PDIs extend these ideas by connecting to the posterior density of a probabilistic model.

In this paper, we study a particular PDI, called the widely applicable posterior dispersion index (WAPDI),

$$\text{WAPDI}(n) = \frac{\sigma^2_{\log}(n)}{\log \mu(n)}.$$

Its form and name comes from the widely applicable information criterion $\text{WAIC} = -\frac{1}{N}\sum_n \log \mu(n) + \sigma^2_{\log}(n)$. WAIC measures generalization error; it asymptotically equates to leave-one-one cross validation (Watanabe, 2010; 2015). WAPDI has two advantages; both are practically motivated. First, we hope the reader is computing an estimate of generalization error. Gelman et al. (2014) suggests WAIC because it is easy to compute and designed for common machine learning models (Watanabe, 2010). Computing WAIC gives WAPDI for free. Second, the variance is a second-order moment calculation; using the log likelihood gives numerical stability to the computation. (More on computation in Section 2.4.)

WAPDI compares the variance of the log likelihood to the log posterior predictive. This gives insight into *how* the likelihood of a datapoint fares under the posterior distribution of the hidden patterns. We now study this in more detail.

2.3. Intuition: not all predictive probabilities are created equal

The posterior predictive density is an expectation, $\mathbb{E}_{\boldsymbol{\theta}\mid\boldsymbol{x}}[p(x_{\text{new}} \mid \boldsymbol{\theta})] = \int p(x_{\text{new}} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \boldsymbol{x})\, \mathrm{d}\boldsymbol{\theta}$. Expectations are integrals: areas under a curve. Different likelihood and posterior combinations can lead to similar integrals.

A toy model illustrates this. Consider a gamma likelihood with fixed shape, and place a gamma prior on the rate. The

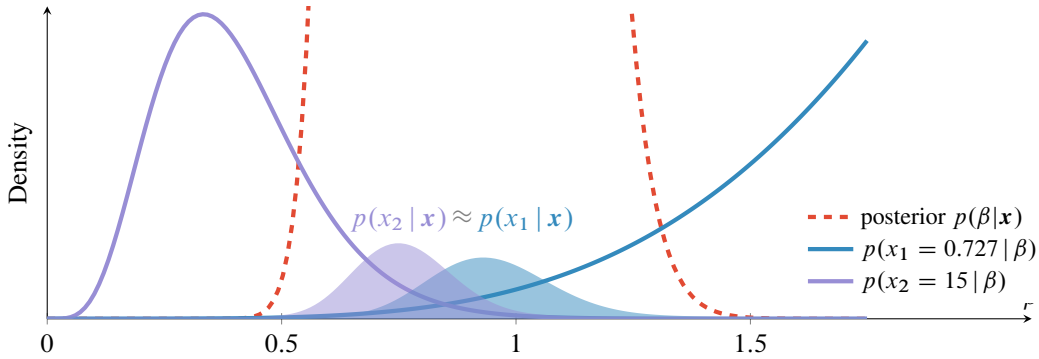

Figure 4. Not all predictive probabilities are created equal. The translucent curves are the two likelihoods multiplied by the posterior (cropped). The posterior predictives $p(x_1 \mid \boldsymbol{x})$ and $p(x_2 \mid \boldsymbol{x})$ for each datapoint is the area under the curve. While both datapoints have the same predictive accuracy, the likelihood for x_2 has higher variance under the posterior; x_2 is more sensitive to the spread of the posterior than x_1. WAPDI captures this effect. (Code in supplement.)

model is

$$p(\beta) = \text{Gam}(\beta\,;\, a_0 = 1, b_0 = 1),$$
$$p(\boldsymbol{x} \mid \beta) = \prod_{n=1}^{N} \text{Gam}(x_n\,;\, a = 5, \beta),$$

which gives the posterior $p(\beta \mid \boldsymbol{x}) = \text{Gam}(\beta\,;\, a = a_0 + 5N, b = b_0 + \sum_n x_n)$.

Now simulate a dataset of size $N = 10$ with $\beta = 1$; the data have mean $a/\beta = 5$. Now consider an outlier at 15. We can find another x value with essentially the same predictive accuracy

$$\log p(x_1 = 0.727 \mid \boldsymbol{x}) = -5.633433,$$
$$\log p(x_2 = 15 \mid \boldsymbol{x}) = -5.633428.$$

Yet their WAPDI values differ by an order of magnitude

$$\text{WAPDI}(x_1 = 0.727) = -0.067,$$
$$\text{WAPDI}(x_2 = 15) = -0.229.$$

In this case, WAPDI highlights $x_2 = 15$ as a more severe outlier than $x_1 = 0.727$, even though they have the same predictive accuracy. What does that mean? Figure 4 depicts the difference.

The following lemma explains how WAPDI measures this effect. (Proof in Appendix B.)

Lemma 1 *If* $\log p(x_n \mid \boldsymbol{\theta})$ *is at least twice differentiable and the posterior* $p(\boldsymbol{\theta} \mid \boldsymbol{x})$ *has finite first and second moments, then a second-order Taylor approximation gives*

$$\text{WAPDI}(n) \approx \frac{\left(\log p'(x_n \mid \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{x}}[\boldsymbol{\theta}])\right)^2 \mathbb{V}_{\boldsymbol{\theta}|\boldsymbol{x}}[\boldsymbol{\theta}]}{\log \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{x}}[p(x_n \mid \boldsymbol{\theta})]}. \quad (2)$$

Corollary 1 WAPDI *highlights datapoints whose likelihood is rapidly changing at the posterior mean estimate of the latent variables.* ($\mathbb{V}_{\boldsymbol{\theta}|\boldsymbol{x}}[\boldsymbol{\theta}]$ *is constant across* n.)

Looking back at Figure 4, the likelihood $p(x_2 = 15 \mid \beta)$ indeed changes rapidly under the posterior. WAPDI reports the ratio of this rate-of-change to the area under the curve. In this specific example, only the numerator matters, since the denominator is effectively the same for both datapoints.

Corollary 2 *Equation (2) is zero if and only if the posterior mean coincides with the maximum likelihood estimate of* $\boldsymbol{\theta}$ *for datapoint* x_n. ($\mathbb{V}_{\boldsymbol{\theta}|\boldsymbol{x}}[\boldsymbol{\theta}]$ *is positive for finite* N.)

For most interesting models, we do not expect such a coincidence. However, in practice, we find WAPDI to be close to zero for datapoints that match the model well. With that, we now turn to computation.

2.4. Computation

Calculating WAPDI is straightforward. The only requirement are samples from the posterior. This is precisely the output of an Markov chain Monte Carlo (MCMC) sampling algorithm. (We used the no-U-turn sampler (Hoffman & Gelman, 2014) for the analyses above.) Other inference procedures, such as variational inference, give an analytic approximation to the posterior (Jordan et al., 1999; Blei et al., 2016). Drawing samples from an approximate posterior also works. (We use this approach for the empirical study in Section 3.)

Equipped with S samples from the posterior, Monte Carlo integration (Robert & Casella, 1999) gives unbiased estimates of the quantities in Equation (1). The variance of these estimates decreases as $\mathcal{O}(1/S)$; we assume S is sufficiently large to cover the posterior (Gelman et al., 2014). We default to $S = 1000$ in our experiments. Algorithm 1 summarizes these steps.

Algorithm 1: Calculating WAPDI.

Input: Data $\boldsymbol{x} = \{x_n\}_1^N$, model $p(\boldsymbol{x}, \boldsymbol{\theta})$.

Output: WAPDI for each datapoint x_n.

Draw S samples $\{\boldsymbol{\theta}\}_1^S$ from posterior (approximation) $p(\boldsymbol{\theta} \mid \boldsymbol{x})$.

for n in *1, 2, ..., N* **do**

Estimate $\log \mu(n), \sigma^2_{\log}(n)$ from samples $\{\boldsymbol{\theta}\}_1^S$.

Store and return

$$\text{WAPDI}(n) = \frac{\sigma^2_{\log}(n)}{\log \mu(n)}.$$

end

3. Experimental Study

We now explore three real data examples using modern machine learning models: voting preferences, supermarket shopping, and population genetics.

3.1. Voting preferences: a hierarchical logistic regression model

In 1988, CBS conducted a U.S. nation-wide survey of voting preferences. Citizens indicated their preference towards the Democratic or Republican presidential candidate. Each individual also declared their gender, age, race, education level, and the state they live in; 11 566 individuals participated.

Gelman & Hill (2006) study this data through a hierarchical logistic regression model. They begin by modeling gender, race, and state; the state variable has a hierarchical prior. This model is easy to fit using automatic differentiation variational inference (ADVI) within Stan (Kucukelbir et al., 2015; Carpenter et al., 2015). (Model and inference details in Appendix C.)

VOTE	R	R	R	R	R	R	R	R	R	R
SEX	F	F	F	F	F	F	F	F	F	F
RACE	B	B	B	B	B	B	B	B	B	B
STATE	WA	WA	NY	WI	NY	NY	NY	NY	MA	MA

Table 1. Lowest predictive accuracy.

VOTE	D	D	D	D	D	D	D	D	D	D
SEX	F	F	F	F	F	M	M	M	M	M
RACE	W	W	W	W	W	W	W	W	W	W
STATE	WY	WY	WY	WY	WY	WY	WY	DC	DC	NV

Table 2. Worst WAPDI values.

Tables 1 and 2 show the individuals with the lowest predictive accuracy and WAPDI. The nation-wide trend predicts that females (F) who identify as black (B) have a strong preference to vote democratic (D); predictive accuracy identifies the few individuals who defy this trend. However, there is not much to do with this information; the model identifies a nation-wide trend that correctly describes most female black voters. In contrast, WAPDI points to parts of the dataset that the model fails to describe; these are datapoints that we might try to explain better with a revised model.

Most of the individuals with poor WAPDI live in Wyoming, the District of Columbia, and Nevada. We focus on Wyoming and Nevada. The average WAPDI for Wyoming and Nevada are -0.057 and -0.041; these are baselines that we seek to improve. (The closer to zero, the better.)

Consider expanding the model by modeling age. Introducing age into the model with a hierarchical prior reveals that older voters tend to vote Republican. This helps explain Wyoming voters; their average WAPDI improves from -0.057 to -0.04; however Nevada's average WAPDI remains unchanged. This means that Nevada's voters may not follow the national age-dependent trend. Now consider removing age and introducing education in a similar way. Education helps explain voters from both states; the average WAPDI for Wyoming and Nevada improve to -0.041 and -0.029.

WAPDI thus captures interesting datapoints beyond what predictive accuracy reports. As expected, predictive accuracy still highlights the same female black voters in both expanded models; WAPDI illustrates a deeper way to evaluate this model.

3.2. Supermarket shopping: a hierarchical Poisson factorization model

Market research firm IRi hosts an anonymized dataset of customer shopping behavior at U.S. supermarkets (Bronnenberg et al., 2008). The dataset tracks 136 584 "checkout" sessions; each session contains a basket of purchased items. An inventory of 7 903 items range across categories such as carbonated beverages, toiletries, and yogurt.

What items do customers tend to purchase together? To study this, consider a hierarchical Poisson factorization (HPF) model (Gopalan et al., 2015). HPF models the quantities of items purchased in each session with a Poisson likelihood; its rate is an inner product between a session's preferences $\boldsymbol{\theta}_s$ and the item attributes $\boldsymbol{\beta}$. Hierarchical priors on $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ simultaneously promote sparsity, while accounting for variation in session size and item popularity. Some sessions contain only a few items; others are large purchases. (Model and inference details in Appendix D.)

A 20-dimensional HPF model discovers intuitive trends. A

few stand out. Snack-craving customers like to buy Doritos tortilla chips along with Lay's potato chips. Morning birds typically pair Cheerios cereal with 2% skim milk. Yoplait fans tend to purchase many different flavors at the same time. Tables 3 and 4 show the top five items in two of these twenty trends.

Item Description	Category
Brand A: 2% skim milk	milk rfg skim/lowfat
Cheerios: cereal	cold cereal
Diet Coke: soda	carbonated beverages
Brand B: 2% skim milk	milk rfg skim/lowfat
Brand C: 2% skim milk	milk rfg skim/lowfat

Table 3. Morning bird trend.

Item Description	Category
Yoplait: raspberry flavor	yogurt rfg
Yoplait: peach flavor	yogurt rfg
Yoplait: strawberry flavor	yogurt rfg
Yoplait: blueberry flavor	yogurt rfg
Yoplait: blackberry flavor	yogurt rfg

Table 4. Yoplait fan trend.

Sessions where a customer purchases many items from different categories have low predictive accuracy. This makes sense as these customers do not exhibit a trend; mathematically, there is no combination of item attributes $\boldsymbol{\beta}$ that explain buying items from disparate categories. For example, the session with the lowest predictive accuracy contains 117 items ranging from coffee to hot dogs.

WAPDI highlights a different aspect of the HPF model. Sessions with poor WAPDI contain similar items but exhibit many purchases of a single item. Table 5 shows an example of a session where a customer purchased 14 blackberry Yoplait yogurts, but only a few of the other flavors.

Item Description	Quantity
Yoplait: blackberry flavor	14
Yoplait: strawberry flavor	2
Yoplait: raspberry flavor	2
Yoplait: peach flavor	1
Yoplait: cherry flavor	1
Yoplait: mango flavor	1

Table 5. A session with poor WAPDI value.

This indicates that the Poisson likelihood assumption may not be flexible enough to model customer purchasing behavior. Perhaps a negative binomial likelihood could model this kind of spiked activity better. Another option might be to keep the Poisson likelihood but increase the hierarchy of the probabilistic model; this approach may identify item attributes that explain such purchases. In either case, WAPDI identifies a valuable aspect of the data that the HPF struggles to capture: sessions with spiked activity. This is a concrete direction for model revision.

3.3. Population genetics: a mixed membership model

Do all people who live nearby have similar genomes? Not necessarily. Population genetics considers how individuals exhibit ancestral patterns of mutations. Begin with N individuals and L locations on the genome. For each location, report whether each individual reveals a mutation. This gives an $(N \times L)$ dataset $\boldsymbol{x}$ where $x_{nl} \in \{0, 1, 2, 3\}$. (We assume two specific forms of mutation; 3 encodes a missing observation.)

Mixed membership models offer a way to study this (Pritchard et al., 2000). Assume K ancestral populations $\boldsymbol{\phi}$; these are the mutation probabilities of each location. Each individual mixes these populations with weights $\boldsymbol{\theta}$; these are the mixing proportions. Place a beta prior on the mutation probabilities and a Dirichlet prior on the mixing proportions.

We study a dataset of $N = 324$ individuals from four geographic locations and focus on $L = 13\,928$ locations on the genome. Figure 5 shows how these individuals mix $K = 3$ ancestral populations. (Data, model, and inference details in Appendix E.)

WAPDI reveals three interesting patterns of mismatch here. First, individuals with poor WAPDI values have many missing observations; the worst 10% of WAPDI have 1 344 missing values, in contrast to 563 for the lowest 10% of predictive scores. We may consider directly modeling these missing observations.

Second, ASW has two individuals with poor WAPDI; their mutation patterns are outliers within the group. While the average individual reveals 272 mutations away from the median genome, these individuals show 410 and 383 mutations. This points to potential mishaps while gathering or pre-processing the data.

Last, MEX exhibits good predictive accuracy, yet poor WAPDI values compared to other groups. Based on predictive accuracy, we may happily accept these patterns. Yet WAPDI highlights a serious issue with the inferred populations. The blue and red populations are almost twice as correlated across genes (0.58) as the other possible combinations (0.24 and 0.2). In other words, the blue and red populations represent similar patterns of mutations at the same locations. These populations, as they stand, are not necessarily interpretable. Revising the model to penalize correlation may be a direction worth pursuing.

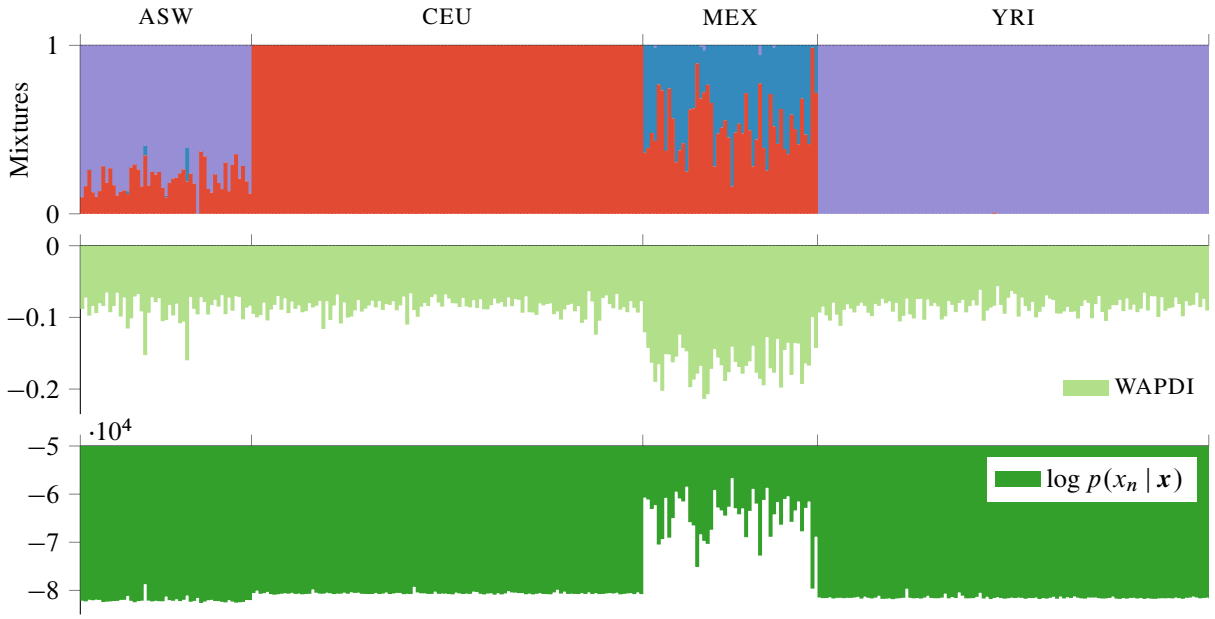

Figure 5. Individuals of African ancestry in southwest U.S. (ASW) and Mexican ancestry in Los Angeles (MEX) exhibit a mixture of two populations. In contrast, Utah residents with European ancestry (CEU) and members of the Yoruba group in Nigeria (YRI) are mostly uniform.

4. Discussion

A posterior dispersion index (PDI) identifies informative forms of model mismatch that compliments predictive accuracy. By highlighting which datapoints exhibit the most uncertainty under the posterior, a PDI offers a new perspective into evaluating probabilistic models. Here, we show how one particular PDI, the widely applicable posterior dispersion index (WAPDI), reveals promising directions for model improvement across a range of models and applications.

The choice of WAPDI is practically motivated; it comes for free as part of the calculation of WAIC. This highlights how PDIs are complimentary to tools such as predictive accuracy, cross validation, and information criteria. While PDIs and predictive accuracy assess model mismatch at the datapoint level, cross validation and information criteria indicate model mismatch at the dataset level.

PDIs provide a relative comparison of datapoints with respect to a model. Can PDIs be thresholded to identify "problematic" datapoints? One approach in this direction draws inspiration from posterior predictive checks (PPCs) (Rubin, 1984; Gelman et al., 1996). A PPC works by hallucinating data from the posterior predictive and comparing properties of the hallucinated data to the observed dataset. Comparing PDI values in this way could lead to a meaningful way of thresholding PDIs.

There are several research directions. One is to extend the notion of a PDI to non-exchangeable data. Another is to leverage the bootstrap to extend this idea beyond probabilistic models. Computationally, ideas from importance sampling could reduce the variance of PDI computations for very high dimensional models.

A promising direction is to study PDIs under the viewpoint of scoring rules (Dawid & Musio, 2014). Understanding the decision theoretic properties of a PDI as a loss function could lead to alternative objectives for inference.

Finally, we end on a reminder that PDIs are simply another tool in the statistician's toolbox. The design and criticism of probabilistic models is still a careful, manual craft. While good tools can help, an overarching obstacle remains to pursue their adoption by practitioners. To this end, making these tools easier to use and more automatic can only help.

Acknowledgments

We thank Dustin Tran, Maja Rudolph, David Mimno, Aki Vehtari, Josh Vogelstein, and Rajesh Ranganath for their insightful comments. This work is supported by NSF IIS-1247664, ONR N00014-11-1-0651, DARPA PPAML FA8750-14-2-0009, DARPA SIMPLEX N66001-15-C-4032, and the Alfred P. Sloan Foundation.

References

Betancourt, Michael. A unified treatment of predictive model comparison. *arXiv preprint arXiv:1506.02273*, 2015.

Bishop, Christopher M. *Pattern Recognition and Machine Learning*. Springer New York, 2006.

Blei, David M. Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application*, 1:203–232, 2014.

Blei, David M, Kucukelbir, Alp, and McAuliffe, Jon D. Variational inference: a review for statisticians. *arXiv preprint arXiv:1601.00670*, 2016.

Bronnenberg, Bart J, Kruger, Michael W, and Mela, Carl F. The IRi marketing data set. *Marketing Science*, 27(4), 2008.

Carpenter, Bob, Gelman, Andrew, Hoffman, Matt, Lee, Daniel, Goodrich, Ben, Betancourt, Michael, Brubaker, Marcus A, Guo, Jiqiang, Li, Peter, and Riddell, Allen. Stan: a probabilistic programming language. *Journal of Statistical Software*, 2015.

Chwialkowski, Kacper, Strathmann, Heiko, and Gretton, Arthur. A kernel test of goodness of fit. *arXiv preprint arXiv:1602.02964*, 2016.

Davison, Anthony Christopher. *Statistical models*. Cambridge University Press, 2003.

Dawid, A. P. *Probability Forecasting*. John Wiley & Sons, Inc., 2006. ISBN 9780471667193.

Dawid, Alexander Philip and Musio, Monica. Theory and applications of proper scoring rules. *Metron*, 72(2):169–183, 2014.

Fano, Ugo. Ionization yield of radiations II. *Physical Review*, 72(1):26, 1947.

Gelman, Andrew and Hill, Jennifer. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2006.

Gelman, Andrew, Meng, Xiao-Li, and Stern, Hal. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6(4):733–760, 1996.

Gelman, Andrew, Carlin, John B, Stern, Hal S, Dunson, David B, Vehtari, Aki, and Rubin, Donald B. *Bayesian Data Analysis*. CRC Press, 2013.

Gelman, Andrew, Hwang, Jessica, and Vehtari, Aki. Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6):997–1016, 2014.

Gopalan, Prem, Hofman, Jake M, and Blei, David M. Scalable recommendation with hierarchical Poisson factorization. *UAI*, 2015.

Gretton, Arthur, Fukumizu, Kenji, Teo, Choon H, Song, Le, Schölkopf, Bernhard, and Smola, Alex J. A kernel statistical test of independence. *NIPS*, 2007.

Hayden, Robert W. A dataset that is 44% outliers. *J Stat Educ*, 13(1), 2005.

Hoel, Paul G. On indices of dispersion. *The Annals of Mathematical Statistics*, 14(2):155–162, 1943.

Hoffman, Matthew D and Gelman, Andrew. The No-U-Turn sampler. *Journal of Machine Learning Research*, 15(1): 1593–1623, 2014.

Jordan, Michael I, Ghahramani, Zoubin, Jaakkola, Tommi S, and Saul, Lawrence K. An introduction to variational methods for graphical models. *Machine Learning*, 37(2): 183–233, 1999.

Koopmans, Lambert H, Owen, Donald B, and Rosenblatt, JI. Confidence intervals for the coefficient of variation for the normal and log normal distributions. *Biometrika*, 51(1/2):25–32, 1964.

Kucukelbir, Alp, Ranganath, Rajesh, Gelman, Andrew, and Blei, David M. Automatic variational inference in Stan. *NIPS*, 2015.

Lloyd, James R and Ghahramani, Zoubin. Statistical model criticism using kernel two sample tests. *NIPS*, 2015.

Murphy, Kevin P. *Machine Learning: a Probabilistic Perspective*. MIT Press, 2012.

Piironen, Juho and Vehtari, Aki. Comparison of Bayesian predictive methods for model selection. *Statistics and Computing*, pp. 1–25, 2015.

Pritchard, Jonathan K, Stephens, Matthew, and Donnelly, Peter. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

Robbins, Herbert. The empirical Bayes approach to statistical decision problems. *Annals of Mathematical Statistics*, 1964.

Robert, Christian P and Casella, George. *Monte Carlo statistical methods*. Springer, 1999.

Rubin, Donald B. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 12(4):1151–1172, 1984.

Vehtari, Aki and Lampinen, Jouko. Bayesian model assessment and comparison using cross-validation predictive densities. *Neural Computation*, 14(10):2439–2468, 2002.

Vehtari, Aki, Ojanen, Janne, et al. A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142–228, 2012.

Vehtari, Aki, Tolvanen, Ville, Mononen, Tommi, and Winther, Ole. Bayesian leave-one-out cross-validation approximations for Gaussian latent variable models. *arXiv preprint arXiv:1412.7461*, 2014.

Vehtari, Aki, Gelman, Andrew, and Gabry, Jonah. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *arXiv preprint arXiv:1507.04544*, 2016.

Watanabe, Sumio. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11:3571–3594, 2010.

Watanabe, Sumio. Bayesian cross validation and WAIC for predictive prior design in regular asymptotic theory. *arXiv preprint arXiv:1503.07970*, 2015.

Winkler, Robert L. Scoring rules and the evaluation of probabilities. *Test*, 5(1):1–60, 1996.