Goodness-of-Fit Tests for High-Dimensional Discrete Distributions with Application to Convergence Diagnostics in Approximate Bayesian Inference

Feras A. Saad[†], **Cameron E. Freer**[†], **Nathanael L. Ackerman**[‡], and **Vikash K. Mansinghka**[†] [†]*Massachusetts Institute of Technology,* [‡]*Harvard University*

Abstract

We present a new family of goodness-of-fit tests that is specialized to high-dimensional discrete distributions. The proposed test is readily implemented using a simple simulation-based procedure and can be computed in linear time, and it can be customized by the practitioner using knowledge of the underlying data domain. Unlike most existing statistics, the proposed test statistic is distribution-free and has an exact (non-asymptotic) sampling distribution. We establish consistency of the test by showing the statistic is distributed as a discrete uniform if and only if the samples are drawn from the candidate distribution. We use the test to assess the convergence behavior of sampling algorithms for approximate Bayesian inference of random partitions in Dirichlet process mixture models.

1. Introduction

We address the problem of testing whether a dataset $y_{1:n} ::= \{y_1, y_2, \dots, y_n\}$ of observed samples was drawn from a candidate probability distribution $\mathbf{p}$. This problem, known as goodness-of-fit testing, is of fundamental interest and has applications in a variety of areas such as checking the quality of posterior predictive simulations from learned Bayesian models [GMS96]; assessing convergence of Markov Chain Monte Carlo simulation algorithms [CGR06]; verifying learned probability density functions in high-energy physics [Wil10] and astronomy [Pea83] data; and testing associations between disease prevalence and gene mutation frequencies in genetic association studies [LK09]. While goodness-of-fit tests for univariate and multivariate continuous distributions have received great attention in the literature, tests for high-dimensional discrete distributions that are theoretically rigorous, customizable using domain knowledge, and practical to implement remain less explored.

We propose to test whether the sample $y_{1:n}$ was drawn from a given discrete distribution $\mathbf{p}$ on the basis of the rank statistic of each y_i with respect to new samples from $\mathbf{p}$. In particular, if $y_{1:n}$ is a sample from the candidate distribution, we expect that the rank statistic of each y_i with respect to a new dataset $X_{1:m}$ simulated from $\mathbf{p}$ is uniformly distributed over the integers $\{0, 1, \dots, m\}$. When the ranks of the y_i show a significant deviation from uniformity, the samples are likely from a different distribution. Since $\mathbf{p}$ is discrete, a key step in the approach is ensuring uniqueness of rank statistics by pairing each y_i with a uniform random variable u_i. We call the resulting statistic the Stochastic Rank Statistic (SRS).

The proposed goodness-of-fit test using the SRS has several desirable statistical properties: it is distribution-free; exactly distributed as a discrete uniform; consistent; simulation-based (avoids computing $\mathbf{p}(x)$ explicitly); and offers the practitioner flexibility in deciding the set of properties on which the observed data $y_{1:n}$ will be checked to agree with samples from $\mathbf{p}$ (manifested in the choice of the total order on $\mathcal{T}$ used to compute the ranks).

2. The Goodness-of-Fit Problem

Problem 1 *Let $\mathbf{p}$ be a candidate discrete distribution over a finite or countably infinite domain $\mathcal{T}$. Given observations $y_{1:n} ::= \{y_1, \dots, y_n\}$ drawn i.i.d. from an unknown distribution $\mathbf{q}$ over $\mathcal{T}$, is there sufficient evidence to reject the hypothesis $\mathbf{p} = \mathbf{q}$?*

In the parlance of statistical testing, we have the following null and alternative hypotheses: $\mathsf{H}_0 ::= [\mathbf{p} = \mathbf{q}]$; and $\mathsf{H}_1 ::= [\mathbf{p} \neq \mathbf{q}]$. A statistical test $\phi_n \colon \mathcal{T}^n \to \{\mathsf{reject}, \mathsf{not\ reject}\}$ says, for each size n dataset, whether to reject or not reject the null hypothesis H_0. For a given significance level $\alpha ::= \Pr\{\phi_n(Y_{1:n}) = \mathsf{reject} \mid \mathsf{H}_0\}$ i.e. to be the probability of incorrectly declaring reject. For a given α, the performance of the test ϕ_n is characterized by its power $\beta ::= \Pr\{\phi_n(Y_{1:n}) = \mathsf{reject} \mid \mathsf{H}_1\}$ to correctly declare reject.

Most existing goodness-of-fit tests for discrete distributions use statistics that apply to domains $\mathcal{T}$ which are either nominal (unordered) or ordinal (ordered): Pearson chi-square [Pea00], likelihood-ratio test [Wil38], multinomial test [Hor77] discrete Kolmogorov-Smirnov [PS77], Cramer–von Mises [CLS94]. These methods require $\mathcal{T}$ to be finite and explicit access to the probabilities $\mathbf{p}(x)$. The most common of these statistics is the Pearson chi-square, although in practice, it is used in non-sparse settings with a small domain. This test suffers from statistical issues when $|\mathcal{T}|$ is large. When the dataset size $n \ll |\mathcal{T}|$, there will be a large number of outcomes with negligible expectation. For discrete random vectors with d entries each having k categories (so that $|\mathcal{T}| = k^d$), chi-squared tests become unreliable even when $d > 4$ and $k > 5$ [MG10]. When $(\mathcal{T}, \prec)$ is a totally ordered finite set, [CLS94] define a family of goodness-of-fit statistics based on discretized versions of the test statistics for continuous data. The sampling distributions of these statistics are only known asymptotically and they have distribution-dependent rejection regions, thus requiring estimating the rejection region for each new distribution $\mathbf{p}$. Moreover, they require access to the CDF, which is intractable in large domains (even when $\mathbf{p}(x)$ is easy to compute).

3. A Family of Exact and Distribution-Free Goodness-of-Fit Tests

Our proposed method for the goodness-of-fit problem combines (i) the intuition from existing methods for ordinal data that the deviation between the expected CDF and empirical CDF of the sample serves as a good signal for goodness-of-fit, with (ii) the flexibility of test statistics in resampling-based Monte Carlo tests [Goo04] to define, using an ordering $\prec$ on $\mathcal{T}$, characteristics of the distribution that are of interest to the experimenter.

Theorem 1 *Let $\mathcal{T}$ be a finite or countably infinite set, let $\prec$ be a strict total order on $\mathcal{T}$, let $\mathbf{p}$ and $\mathbf{q}$ be two probability distributions on $\mathcal{T}$, and let m be a positive integer. Consider:*

$$X_0 \sim \mathbf{q} \tag{1}$$

$$X_1, X_2, \dots, X_m \sim^{\text{iid}} \mathbf{p} \tag{2}$$

$$U_0, U_1, U_2, \dots, U_m \sim^{\text{iid}} \mathsf{Uniform}(0,1) \tag{3}$$

$$R = \textstyle\sum_{j=1}^{m} \mathbb{I}\left[X_j \prec X_0\right] + \mathbb{I}\left[X_j = X_0, U_j < U_0\right]. \tag{4}$$

Then $\mathbf{p} = \mathbf{q}$ if and only if for all $m \geq 1$, the rank R is uniformly distributed on the set of integers $[m+1] ::= \{0, 1, 2, \dots, m\}$.

The random variable R in Eq. (4) determines the rank of $X_0 \sim \mathbf{q}$ within a size m sample $X_{1:m} \sim^{\text{iid}} \mathbf{p}$ with ties broken stochastically (hence the name stochastic rank statistic). Note that in the case where all the X_i are almost surely distinct, the forward direction of Theorem 1, which establishes that if $\mathbf{p} = \mathbf{q}$ then the rank R is uniform for all $m \geq 1$, is easy to show and is known in the statistical literature [ANS13, p. 5]. However, no existing results make the connection between rank statistics and discrete random variables over countable domains with ties broken stochastically. Nor do they establish that $\mathbf{p} = \mathbf{q}$ is a *necessary* condition for uniformity of R (across all m beyond some integer), and can therefore be used as the basis of a consistent goodness-of-fit test. Algorithm 1 formally describes the testing procedure, using multiple observations to test for uniformity of R.

Algorithm 1 Discrete Goodness-of-Fit Testing Procedure

```
Input: { simulator for candidate distribution $\mathbf{p}$ over finite or countable sample space $T$;
       { observed samples $\{y_1, y_2, \dots, y_n\}$ sampled i.i.d. from unknown distribution $\mathbf{q}$;
       { strict total order $\prec$ on $T$, of any order type;
       { number $m \geq 1$ of datasets to resimulate;
       { significance level $\alpha$;
Output: Decision whether to reject $\mathsf{H}_0 : \mathbf{p} = \mathbf{q}$ against $\mathsf{H}_1 : \mathbf{p} \neq \mathbf{q}$ at significance level $\alpha$.
for $i = 1, 2, \dots, n$ do
    $X_1^{(i)}, X_2^{(i)}, \dots, X_m^{(i)} \sim^{\text{iid}} \mathbf{p}$
    $U_0^{(i)}, U_1^{(i)}, \dots, U_m^{(i)} \sim^{\text{iid}} \mathsf{Uniform}(0,1)$
    $r_i \leftarrow \sum_{k=1}^{m} \mathbb{I}[X_k^{(i)} \prec y_i] + \mathbb{I}[X_k^{(i)} = y_i, U_k^{(i)} < U_0^{(i)}]$
Compute $p$-value of observed ranks $\{r_1, \dots, r_n\}$ under uniform distribution on $\{0, 1, 2, \dots, m\}$.
return reject if $p \leq \alpha$, else not reject.
```

4. Application: MCMC Convergence for Dirichlet Process Mixtures

We now apply the proposed goodness-of-fit procedure to assess the sample quality of data structures obtained from approximate sampling algorithms over combinatorially large domains with intractable probabilities. In recent work, [TBS+18] describe a general procedure for validating inference from Bayesian algorithms that can generate posterior samples. More specifically, for a prior $\pi(\theta)$ over the parameters and likelihood function $\pi(x|\theta)$ over the measurements, observe that integrating the posterior over the Bayesian joint distribution returns the prior distribution, i.e.: $\pi(\theta) = \int \left[\pi(\theta|x')\pi(x'|\theta')\mathrm{d}x'\right] \pi(\theta')\mathrm{d}\theta'$.

By simulating datasets $x \sim \pi(x)$ from their marginal distribution and running posterior inference $\pi(\theta|x)$ over the parameters, the "data-averaged" posterior is identically distributed to the prior. The applications in [TBS+18] focus on checking the sample quality of univariate marginals for continuous parameters θ. We extend the procedure to the case of discrete latent variables in a high-dimensional space. In particular, we repeatedly sampled $n = 100$ data points $\{x_1, \dots, x_n\}$, simulated from a Dirichlet process mixture model over five dimensional data. The likelihood model $p(x_{i,1}, \dots, x_{i,5}|z_i)$ for each data point i is a product of independent Gaussians. From simulation-based calibration, the posterior distribution $\pi(z_{1:n}|x_{1:n})$ over mixture assignments $z_{1:n}$ is exactly equivalent to the CRP prior $\pi(z_{1:n})$. In Figure 1 we use the ordering over partitions in Algorithm 2 to characterize the goodness-of-fit (with respect to the true posterior) of approximate partition samples $\hat{z}_{1:n}$ (which can take $\approx 10^{115}$

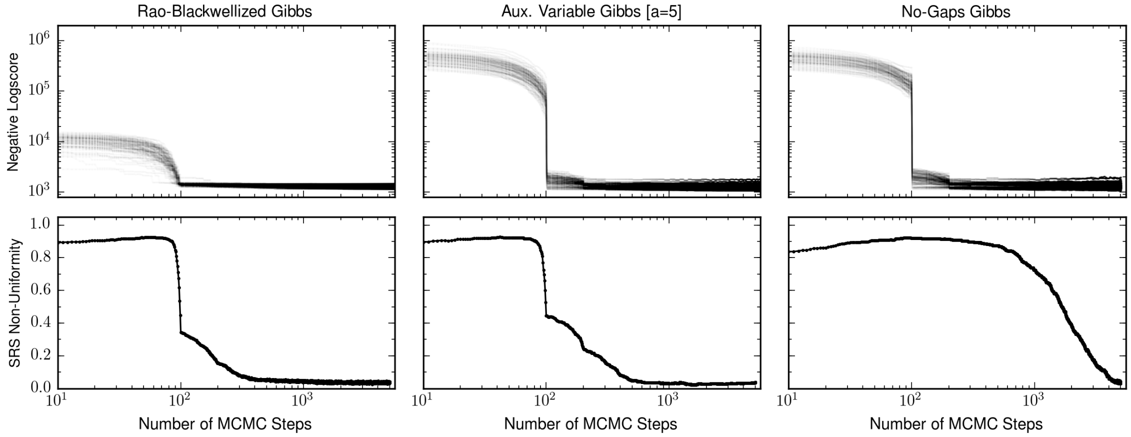

Figure 1: The uniformity of the SRS (bottom row) with simulation-based calibration [TBS+18] is used to compare the behavior of three MCMC sampling algorithms for the discrete partition structure of a Dirichlet process mixture model. Partitions are compared using the linear order in Algorithm 2. The SRS (bottom row) captures convergence behavior that is not found by standard diagnostics such as the logscore (top row).

Algorithm 2 Total order $\prec$ on the set of partitions Π_N

```
Input: { Partition $\pi ::= \{\pi_1, \pi_2, \dots, \pi_k\} \in \Pi_N$ with $k$ blocks.
       { Partition $\nu ::= \{\nu_1, \nu_2, \dots, \nu_l\} \in \Pi_N$ with $l$ blocks.
Output: LT if $\pi \prec \nu$; GT if $\pi \succ \nu$; EQ if $\pi = \nu$.
  if $k < l$ then return LT                                  ▷ $\nu$ has more blocks
  if $k > l$ then return GT                                  ▷ $\pi$ has more blocks
  $\tilde{\pi} \leftarrow$ blocks of $\pi$ sorted by value of least element in the block
  $\tilde{\nu} \leftarrow$ blocks of $\nu$ sorted by value of least element in the block
  for $b = 1, 2, \dots, l$ do
    if $|\tilde{\pi}_b| < |\tilde{\nu}_b|$ then return LT               ▷ $\tilde{\nu}_b$ has more elements
    if $|\tilde{\pi}_b| > |\tilde{\nu}_b|$ then return GT               ▷ $\tilde{\pi}_b$ has more elements
    $\pi'_b \leftarrow$ values in $\tilde{\pi}_b$ sorted in ascending order
    $\nu'_b \leftarrow$ values in $\tilde{\nu}_b$ sorted in ascending order
    for $i = 1, 2, \dots, |\pi'_b|$ do
      if $\pi'_{b,i} < \nu'_{b,i}$ then return LT                  ▷ $\pi'_b$ has smallest element
      if $\pi'_{b,i} > \nu'_{b,i}$ then return GT                  ▷ $\nu'_b$ has smallest element
  return EQ
```

different values) from Rao–Blackwellized Gibbs, Auxiliary Variable Gibbs, and No-Gaps Gibbs (Algorithms 3, 8, and 4 of [Nea00], respectively).

The top row of Figure 1 shows standard trace plots of the logscore (unnormalized posterior) where each line corresponds to a given run of the Markov chain. The bottom row shows the evolution of the uniformity of the SRS. While the logscores typically stabilize after 100 MCMC steps (one epoch through all the observations) and suggest little difference across the three algorithms, the SRS shows that the sample quality of the mixture assignments of the No-Gaps algorithm is inferior to those from the other two algorithms up until roughly 5,000 MCMC steps. The SRS plot verifies the observation from [Nea00] that the No-Gaps sampler has inefficient mixing behavior (it excessively rejects proposals on singleton clusters).

Acknowledgments

This research was supported by the DARPA SD2 program (contract FA8750-17-C-0239); grants from the MIT Media Lab, the Harvard Berkman Klein Center Ethics and Governance of AI Fund, and the MIT CSAIL Systems That Learn Initiative; as well as an anonymous philanthropic gift.

References

[ANS13] Mohammad Ahsanullah, Valery B. Nevzorov, and Mohammad Shakil. *An Introduction to Order Statistics*. Atlantis Studies in Probability and Statistics. Atlantis Press, 2013.

[CGR06] Samantha R. Cook, Andrew Gelman, and Donald B. Rubin. Validation of software for Bayesian models using posterior quantiles. *Journal of Computational and Graphical Statistics*, 15(3):675–692, 2006.

[CLS94] V. Choulakian, R. A. Lockhart, and M. A. Stephens. Cramer–von Mises statistics for discrete distributions. *Canadian Journal of Statistics*, 22(1):125–137, 1994.

[GMS96] Andrew Gelman, Xiao-Li Meng, and Hal Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6:733–807, 1996.

[Goo04] Phillip I. Good. *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. Springer Series in Statistics. Springer-Verlag, 2004.

[Hor77] Susan Dadakis Horn. Goodness-of-fit tests for discrete data: A review and an application to a health impairment scale. *Biometrics*, 33(1):237–247, 1977.

[LK09] Cathryn M. Lewis and Jo Knight. Introduction to genetic association studies. In Ammar Al-Chalalbi and Laura Almasy, editors, *Genetics of Complex Human Diseases: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, 2009.

[MG10] Alberto Maydeu-Olivares and Carlos Garcia-Forero. Goodness-of-fit testing. In Penelope Peterson, Eva Baker, and Barry McGaw, editors, *International Encyclopedia of Education*, volume 7, pages 190–196. 2010.

[Nea00] Radford M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

[Pea00] Karl Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 5:157–175, 1900.

[Pea83] J. A. Peacock. Two-dimensional goodness-of-fit testing in astronomy. *Monthly Notices of the Royal Astronomical Society*, 202(3):615–627, 1983.

[PS77] Anthony N. Pettitt and Michael A. Stephens. The Kolmogorov–Smirnov goodness-of-fit statistic with discrete and grouped data. *Technometrics*, 19(2):205–210, 1977.

[TBS+18] Sean Talts, Michael Betancourt, Daniel Simpson, Aki Vehtari, and Andrew Gelman. Validating Bayesian inference algorithms with simulation-based calibration. *arXiv preprint*, (arXiv:1804.06788), 2018.

[Wil38] Samuel S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, 1938.

[Wil10] M. Williams. How good are your fits? Unbinned multivariate goodness-of-fit tests in high energy physics. *Journal of Instrumentation*, 5(09):P09004, 2010.