Publication Year	2023
Acceptance in OA	2024-05-24T12:01:59Z
Title	Bayesian jackknife tests with a small number of subsets: application to HERA 21 cm power spectrum upper limits
Authors	Wilensky, Michael J., Kennedy, Fraser, BULL, PHILIP, Dillon, Joshua S., Abdurashidova, Zara, Adams, Tyrone, Aguirre, James E., Alexander, Paul, Ali, Zaki S., Baartman, Rushelle, Balfour, Yanga, Beardsley, Adam P., BERNARDI, Gianni, Billings, Tashalee S., Bowman, Judd D., Bradley, Richard F., Burba, Jacob, Carey, Steven, Carilli, Chris L., Cheng, Carina, DeBoer, David R., de Lera Acedo, Eloy, Dexter, Matt, Eksteen, Nico, Ely, John, Ewall-Wice, Aaron, Fagnoni, Nicolas, Fritz, Randall, Furlanetto, Steven R., Gale-Sides, Kingsley, Glendenning, Brian, Gorthi, Deepthi, Greig, Bradley, Grobbelaar, Jasper, Halday, Ziyaad, Hazelton, Bryna J., Hewitt, Jacqueline N., Hickish, Jack, Jacobs, Daniel C., Julius, Austin, Kariseb, MacCalvin, Kern, Nicholas S., Kerrigan, Joshua, Kittiwisit, Piyanat, Kohn, Saul A., Kolopanis, Matthew, Lanman, Adam, La Plante, Paul, Liu, Adrian, Loots, Anita, MacMahon, David Harold Edward, Malan, Lourence, Malgas, Cresshim, Malgas, Keith, Marero, Bradley, Martinot, Zachary E., Mesinger, Andrei, Molewa, Mathakane, Morales, Miguel F., Mosiane, Tshegofalang, Murray, Steven G., Neben, Abraham R., Nikolic, Bojan, Nuwegeld, Hans, Parsons, Aaron R., Patra, Nipanjana, Pieterse, Samantha, Razavi-Ghods, Nima, Robnett, James, Rosie, Kathryn, Sims, Peter, Swarts, Hilton, Thyagarajan, Nithyanandan, van Wyngaarden, Pieter, Williams, Peter K. G., Zheng, Haoxuan
Publisher's version (DOI)	10.1093/mnras/stac3484
Handle	http://hdl.handle.net/20.500.12386/35124
Journal	MONTHLY NOTICES OF THE ROYAL ASTRONOMICAL SOCIETY
Volume	518

Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **518,** 6041–6058 (2023) https://doi.org/10.1093/mnras/stac3484
Advance Access publication 2022 November 29

Bayesian jackknife tests with a small number of subsets: application to HERA 21 cm power spectrum upper limits

Michael J. Wilensky [1,2★] Fraser Kennedy,[2] Philip Bull,[1,2,3] Joshua S. Dillon [4] Zara Abdurashidova,[4] Tyrone Adams,[5] James E. Aguirre [6] Paul Alexander,[7] Zaki S. Ali,[4] Rushelle Baartman,[5] Yanga Balfour,[5] Adam P. Beardsley,[8,9] Gianni Bernardi,[5,10,11] Tashalee S. Billings,[6] Judd D. Bowman,[8] Richard F. Bradley,[12] Jacob Burba [13] Steven Carey,[7] Chris L. Carilli,[14] Carina Cheng,[4] David R. DeBoer,[15] Eloy de Lera Acedo [7] Matt Dexter,[15] Nico Eksteen,[5] John Ely,[7] Aaron Ewall-Wice [4,16] Nicolas Fagnoni [7] Randall Fritz,[5] Steven R. Furlanetto [17] Kingsley Gale-Sides,[7] Brian Glendenning,[18] Deepthi Gorthi,[4] Bradley Greig [19] Jasper Grobbelaar,[5] Ziyaad Halday,[5] Bryna J. Hazelton,[20,21] Jacqueline N. Hewitt,[22,23] Jack Hickish,[15] Daniel C. Jacobs,[8] Austin Julius,[5] MacCalvin Kariseb,[5] Nicholas S. Kern,[4,23] Joshua Kerrigan,[13] Piyanat Kittiwisit [3] Saul A. Kohn [6] Matthew Kolopanis,[8] Adam Lanman [13] Paul La Plante,[4,6] Adrian Liu,[24] Anita Loots,[5] David Harold Edward MacMahon,[15] Lourence Malan,[5] Cresshim Malgas,[5] Keith Malgas,[5] Bradley Marero,[5] Zachary E. Martinot,[6] Andrei Mesinger [25] Mathakane Molewa,[5] Miguel F. Morales,[20] Tshegofalang Mosiane,[5] Steven G. Murray [8] Abraham R. Neben,[23] Bojan Nikolic,[7] Hans Nuwegeld,[5] Aaron R. Parsons,[4] Nipanjana Patra,[4] Samantha Pieterse,[5] Nima Razavi-Ghods,[7] James Robnett,[14] Kathryn Rosie,[5] Peter Sims [24] Hilton Swarts,[5] Nithyanandan Thyagarajan [14,26] Pieter van Wyngaarden,[5] Peter K. G. Williams [27,28] and Haoxuan Zheng [23]

Affiliations are listed at the end of the paper

Accepted 2022 November 23. Received 2022 October 31; in original form 2022 August 18

ABSTRACT

We present a Bayesian jackknife test for assessing the probability that a data set contains biased subsets, and, if so, which of the subsets are likely to be biased. The test can be used to assess the presence and likely source of statistical tension between different measurements of the same quantities in an automated manner. Under certain broadly applicable assumptions, the test is analytically tractable. We also provide an open-source code, CHIBORG, that performs both analytic and numerical computations of the test on general Gaussian-distributed data. After exploring the information theoretical aspects of the test and its performance with an array of simulations, we apply it to data from the Hydrogen Epoch of Reionization Array (HERA) to assess whether different sub-seasons of observing can justifiably be combined to produce a deeper 21 cm power spectrum upper limit. We find that, with a handful of exceptions, the HERA data in question are statistically consistent and this decision is justified. We conclude by pointing out the wide applicability of this test, including to CMB experiments and the H_0 tension.

Key words: methods: data analysis – methods: statistical – software: data analysis – dark ages, reionization, first stars – cosmology: observations.

1 INTRODUCTION

Evaluating the presence and significance of statistical tension is a core part of the process of reconciling independent measurements of the same physical quantity. A lack of tension permits us to combine measurements in order to place improved constraints on quantities of interest, while the presence of tension raises the prospect of contamination of some or all of the data by previously unidentified systematic effects, or new physical effects not anticipated in the theoretical model being used to interpret the data.

A topical example within the cosmology community is the 'Hubble tension' (Verde, Protopapas & Jimenez 2013; Knox & Millea 2020; Di Valentino et al. 2021; Freedman 2021; Riess et al. 2021), which arises from apparent inconsistencies in the measured value of the Hubble parameter, H_0, obtained from different observational probes. In this case, the fact that there is a tension between at least some of

★ E-mail: michael.wilensky@manchester.ac.uk

© The Author(s) 2022.
Published by Oxford University Press on behalf of Royal Astronomical Society. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/mnras/article/518/4/6041/6852951 by guest on 22 April 2024

the measurements is not in dispute; a difference of almost 5σ is found between the reported values of H_0 from the Planck CMB mission and local distance ladder measurements for example (Riess et al. 2021). The question instead is which of the points are discrepant, and why. By finding distinct subsets of the measurements that are internally statistically consistent (and so can be combined to produce a more precise measurement), but in tension with other subsets, we can hope to identify a systematic bias or physical phenomenon that explains the tension (e.g. Di Valentino et al. 2021; Blanchard et al. 2022). Similar tensions between other cosmological parameters have also been tentatively identified and studied (e.g. Lin & Ishak 2017; Raveri & Hu 2019; Lemos et al. 2021).

Assessing the probability that a given measurement is discrepant is difficult (and potentially ill-defined) without some knowledge of the expected distribution of the measurements and possible biases, particularly when only a handful of measurements are available. A Bayesian approach lends itself particularly well to this kind of question, as it allows us to set out the statistical questions we are asking of the data in an explicit manner, e.g. by requiring definite statements of the models/hypotheses and prior assumptions. A wide variety of Bayesian approaches have previously been applied to cosmological tensions (e.g. Press 1997; Liddle 2007; Verde et al. 2013; Harrison et al. 2015; Lucy 2016; Seehars et al. 2016; Lin & Ishak 2017; Handley & Lemos 2019; Raveri & Hu 2019; Blanchard et al. 2022), and tend to share some common features. For instance, many employ model comparison tests between a null hypothesis and a variety of alternative model scenarios, integrating over a chosen prior (with an appropriate penalty/'Occam factor' to penalize more flexible models) to determine which scenario has overall higher odds given the available data (e.g. Liddle 2007; Trotta 2007). Some also employ methods that measure a 'distance' between the empirical distribution of the measured values and the expected/ideal distribution (e.g. Lucy 2016; Nicola, Amara & Refregier 2019), but choose different ways of quantifying the distance and determining what counts as a significant difference. These latter two points, plus different strategies for choosing (or even disregarding) priors, account for much of the variety in these approaches.

Most of the references that we have included above focus on assessing tensions between the values of cosmological parameters inferred from different experiments. Another common situation that requires a similar type of assessment of statistical consistency is the combination of different subsets or 'seasons' of data from a single experiment. Particularly in CMB and radio (21 cm intensity mapping) experiments, long integration times are required to reach sufficient signal-to-noise on the faint signals that are being targeted. This often requires observations that stretch over multiple observing seasons, taken across several years and different times of the year, and possibly even different instrument configurations etc. The potential for discrepancies due to varying systematic effects between subsets of observations is therefore relatively high, and so one is usually interested in performing 'jackknife' tests that attempt to determine whether the subsets are consistent with one another. Assuming that they are, the data can then be combined in order to improve sensitivity. Otherwise, the discrepancy can be taken as evidence of systematic contamination that must be addressed.

We note that Bayesian approaches seem to be less commonly used for this kind of assessment however. Instead, frequentist simulation-based methods seem to be more prevalent, for instance ones in which many simulations of a null hypothesis are passed through the same data analysis pipeline and compared at the level of the measured statistic (e.g. a power spectrum). The measured value of each data point (e.g. each power spectrum bandpower) is then compared with the distribution of simulated null hypothesis measurements to calculate a p-value or 'Probability-to-Exceed' (PTE) (e.g. Planck Collaboration VII 2020; Abbott et al. 2022; Ade et al. 2022), which gives an estimate of whether each data point can be considered an 'outlier' or not, subject to a potentially large list of implicit assumptions. Part of the reason for the popularity of this approach is likely to be the relative complexity of the 'forward model' of the data analysis process, which can involve many non-trivial steps, hence lending itself to a simulation-based method. There is also the question of whether a more careful (and somewhat more cumbersome) Bayesian approach is overly complex when only approximate agreement between subsets of data is typically considered sufficient to proceed with averaging them together.

Certain types of observation, such as CMB B-mode searches and 21 cm intensity mapping, now routinely involve data with very large dynamic ranges, for which a more controlled assessment of statistical consistency may ultimately be required in order to recover the extremely faint target signals from combinations of observing seasons. In this work, we develop a Bayesian jackknife test that has such an application in mind. The test is constructed to determine whether any subsets of the data are biased, and if so, which subsets. The general framework is flexible in that the analyst may sculpt the bias hypotheses by choosing appropriate priors, and is not limited to a particular set of bias configurations. We implement this framework in an open-source PYTHON software package CHIBORG,[1] named as an alternative to 'χ-by-eye' approaches to similar problems. In essence, the test computes the evidence (marginal likelihood) for each bias hypothesis, and combines these into a posterior probability distribution (PPD). The most probable hypothesis can then be identified, or, depending on the application, a loss function may be minimized over the posterior to decide on the outcome of the test. While similar approaches have been developed before, e.g. to assess cosmological parameter tensions, we place an additional focus on the decision-making aspect of the test, pointing out how data analysts can use it to select or exclude subsets of the data for further averaging, as well as to hunt down systematic effects.

Alternative Bayesian approaches to our statistical consistency checks include fitting a hierarchical model where (for example) the explanatory parameter is a covariance parameter between the underlying means of the data subsets,[2] and/or performing posterior predictive checks to see if a model inferred from a subset of the data can produce simulated data sets that are consistent with the remaining measured data (the subset not used for the inference). Both of these procedures are introduced and explored in a broader statistical context in Gelman et al. (2004), which reviews Bayesian data analysis in general. Hierarchical modelling is more specifically explored in Gelman, Hill & Yajima (2012) and Isi, Farr & Chatziioannou (2022), while Rubin (1984), Meng (1994), Gelman, Meng & Stern (1996), Doux et al. (2021), and Moran, Cunningham & Blei (2022) focus on posterior predictive methods. In posterior predictive checks, one simulates fictitious data by sampling its PPD as inferred from (a subset of) the data, and then exposes the original data to a well-defined consistency test within the simulated data such as a χ^2 test (whose sampling distribution can be built up from the ensemble of simulated data). They are appealing since they are well-formulated

[1] https://github.com/mwilensky768/chiborg

[2] In some respects this resembles the analysis suggested in this work, but we have elected to only consider a discrete selection of covariance models, whereas a typical hierarchical fitting procedure would allow this parameter to vary continuously.

Downloaded from https://academic.oup.com/mnras/article/518/4/6041/6852951 by guest on 22 April 2024

and clearly assess the question of statistical consistency. However, implementations that focus on the posterior predictive p-value suffer from the need to arbitrate statistical significance, i.e. it is arbitrary when to accept or reject the hypothesis of statistical consistency. We address this by integrating our inference into a decision-making framework. The relative consistency between the data subsets is represented by a set of hypotheses each with an associated Bayesian posterior. These posteriors can be weighed against a loss function based on inference error in order to determine the optimal conclusion for the needs at hand. In a hierarchical modelling scheme, one could in principle perform the same exercise after obtaining the posterior for the explanatory parameter. Another important strength of our jackknife test is that it assesses the question of statistical consistency by pointing directly to the problematic data subsets in an easily interpreted fashion. In other words, once the data are known to contain inconsistencies, the next most immediate question is often 'which data are inconsistent?'

We have in mind a particular scenario in the analysis of 21 cm array data from the Hydrogen Epoch of Reionization Array (HERA) in The HERA Collaboration (2022), which motivated the design of this test, and its application when only small numbers of subsets of data are available (so that the distribution of the subsets is difficult to determine empirically). Specifically, we use the test in order to develop a (post hoc) statistical justification for coherently combining several disjoint subsets of data into one data set for the purposes of placing an improved upper limit on the cosmological 21 cm power spectrum signal. We explain this in more detail in what follows. For specific reviews on the subject of 21 cm cosmology, we direct the reader to Furlanetto, Oh & Briggs (2006), Morales & Wyithe (2010), and Liu & Shaw (2020), among others.

This paper is laid out as follows. In Section 2, we lay down a theoretical foundation for the test. We also explain from an information theoretical standpoint how to design the test for different applications. We then apply the test to a suite of simple simulations in Section 3 in order to explore the performance and behaviour of the test in different circumstances. In Section 4, we use the test on simulated HERA power spectra using the HERA validation pipeline (Aguirre et al. 2022), as well as the actual power spectrum measurements from The HERA Collaboration (2022). Finally, we provide a summary of the work and our conclusions in Section 5.

2 MATHEMATICAL FORMALISM

In this section, we establish the mathematical formalism for detecting biases in data. We begin by stating the problem and defining our test in terms of a null hypothesis and alternate hypotheses, stating what each of these means mathematically. We then analytically derive the posterior probabilities of these hypotheses given some data.

2.1 Problem statement

We consider the general problem of inferring whether a set of normally distributed data points are drawn about the same mean. We have in mind a particular choice of data in the HERA power spectrum estimation pipeline, and use this example in order to guide the discussion with a concrete scenario. However, the formalism could be applied in other data analysis problems, such as determining whether different measurements of the Hubble parameter are statistically consistent.

Consider a scenario in which the HERA analysts have carefully prepared data for the estimation of the 21 cm power spectrum signal, and have chosen to separate the data into N observing epochs, all disjoint (e.g. 20 nights, separated into $N = 4$ disjoint sets of five nights), and expected to be similar in quality. For the final power spectrum limits, these epochs will be combined into one measurement. However, as an auxiliary data set, they form power spectrum estimates for each epoch independently. We may then ask the question, 'given these power spectrum estimates, do we still believe these epochs are similar in character?'

Clearly, we must be more specific since there are many ways in which the data can be inconsistent with one another. In the case of 21 cm power spectrum estimation, we are often concerned with whether there is a systematic bias. The framework presented in this paper is flexible in that it can answer different statistical questions about the relationships of the potential biases based on which hypotheses are considered. For instance, one question that it can answer is, 'which subsets of the data are likely to possess a significant bias of unknown strength?' Another is, 'given a particular strength of bias that can affect some data but not others (for whatever the reason), how likely is it that this bias is present?' Yet another is, 'given that we believe some subsets of the data are biased, which partitions of these subsets have biases that are similar to one another?'

In the following subsections, we outline a Bayesian hypothesis test to answer these questions. We then provide a formulation of the marginal likelihoods corresponding to a general set of relationships between potential biases of unknown strength. From these likelihoods and corresponding priors, we calculate the posterior odds of various hypothetical bias relationships against one another, allowing us to form a conclusion about the nature of the biases within the data.

2.2 Phrasing the hypothesis test

We are given N data points arranged into a vector, $\mathbf{d}$, and suppose that each datum may be biased by some unknown value, $\boldsymbol{\varepsilon}$. In the absence of a bias, each datum is drawn about the same mean value, μ_0 (potentially unknown), according a Gaussian distribution with known covariance $\mathbf{C}_0$. We define this as our null hypothesis, $\mathcal{H}_0$. In the HERA example, the data would be the bandpower estimates for a single wave mode, and $\mathbf{C}_0$ would be a diagonal matrix where each entry would be the estimated variance of the corresponding bandpower.

In the HERA problem, there are multiple choices for the interpretation of μ_0. For example, one could choose a value from which systematics are expected to be subtracted from theoretical work (such as the validation pipeline accompanying the data analysis software), a chosen reionization model to be investigated, or even the value 0. In this final case, a rejection of the null hypothesis for the real component of the HERA bandpower is ambivalent between whether it is a detection of the reionization signal or some systematic causing the bias, and one would need to program specific hypotheses about the bias (or perform an independent analysis) in order to distinguish between the two.

In general, the scale of the problem is set by the size of the error bars, and we observe that differences in choice of μ_0 that are much smaller than the statistical error associated with the data make imperceptible differences in the test results. This has potential consequences regarding a choice of prior on μ_0, explained in Section 2.4.

To answer a particular statistical question about the biases, we formulate hypotheses about them, and compare the posterior probabilities among one another using the 'odds' of one hypothesis to another, conditional on the data and known covariance. Mathe-

Downloaded from https://academic.oup.com/mnras/article/518/4/6041/6852951 by guest on 22 April 2024

matically, we write

$$P(\mathcal{H}_i|\mathbf{d}, \mathbf{C}_0) = \frac{P(\mathbf{d}|\mathcal{H}_i, \mathbf{C}_0)P(\mathcal{H}_i|\mathbf{C}_0)}{P(\mathbf{d}|\mathbf{C}_0)}, \tag{1}$$

where $\mathcal{H}_i$ is shorthand for a specification of knowledge about the biases, formally detailed in the next subsection. In the formalism in this paper, we do not use our knowledge of the data covariance to inform the prior of the hypothesis, though we imagine that there can be good reasons for doing otherwise. This allows us to write

$$P(\mathcal{H}_i|\mathbf{C}_0) = P(\mathcal{H}_i). \tag{2}$$

If there are Q hypotheses in consideration (including the null hypothesis), then the denominator of equation (1) is simply calculated as

$$P(\mathbf{d}|\mathbf{C}_0) = \sum_{i=0}^{Q-1} P(\mathbf{d}|\mathcal{H}_i, \mathbf{C}_0)P(\mathcal{H}_i). \tag{3}$$

2.3 Evaluating the posterior

In order to evaluate the posterior probability of any hypothesis, we must handle the nuisance parameters by marginalizing over corresponding priors in order to produce the likelihood. For instance, if we expected tbe data to possess a common mean that takes the value μ_0, we would say

$$P(\mathbf{d}|\mathbf{C}_0, \mu_0, \mathcal{H}_0) = \frac{\exp\left(-\frac{1}{2}(\mathbf{d}-\mu_0\mathbb{1})^T\mathbf{C}_0^{-1}(\mathbf{d}-\mu_0\mathbb{1})\right)}{\sqrt{\det(2\pi\mathbf{C}_0)}}, \tag{4}$$

where $\mathbb{1}$ is a vector for which every component is equal to 1. Similarly, if we knew the data to be drawn from a biased Gaussian, with bias given by the vector $\boldsymbol{\varepsilon}$, we would write

$$P(\mathbf{d}|\mathbf{C}_0, \mu_0, \boldsymbol{\varepsilon}, \mathcal{H}_i) = \frac{\exp\left(-\frac{1}{2}(\mathbf{d}-\mu_0\mathbb{1}-\boldsymbol{\varepsilon})^T\mathbf{C}_0^{-1}(\mathbf{d}-\mu_0\mathbb{1}-\boldsymbol{\varepsilon})\right)}{\sqrt{\det(2\pi\mathbf{C}_0)}}. \tag{5}$$

In general, we need not assume particular values for the test. We instead propose that the means and biases may take a range of values, with weights assigned by prior probabilities:

$$P(\mathbf{d}|\mathbf{C}_0, \mathcal{H}_0) = \int_{\mathbb{R}} \mathrm{d}\mu_0 P(\mu_0)P(\mathbf{d}|\mathbf{C}_0, \mu_0, \mathcal{H}_0) \tag{6}$$

$$P(\mathbf{d}|\mathbf{C}_0, \mathcal{H}_i) = \int_{\mathbb{R}} \mathrm{d}\mu_0 P(\mu_0) \int_{\mathbb{R}^N} \mathrm{d}\boldsymbol{\varepsilon}\, P(\boldsymbol{\varepsilon}|\mathcal{H}_i)P(\mathbf{d}|\mathbf{C}_0, \mu_0, \boldsymbol{\varepsilon}, \mathcal{H}_i), \tag{7}$$

where $P(\mu_0)$ and $P(\boldsymbol{\varepsilon}|\mathcal{H}_i)$ are suitable priors that reflect a combination of our knowledge about the system and in some part the nature of the types of biases we would like to inspect. We have implicitly assumed that $P(\mu_0)$ is the same regardless of which hypothesis is true, since the change in that parameter under a violation of the null hypothesis is explicitly parametrized by $\boldsymbol{\varepsilon}$.

If one chooses Gaussian priors for these nuisance parameters, then the calculation has a closed-form solution. We acknowledge that in some instances this choice of prior may seem inappropriate. In the HERA example at hand, if μ_0 is meant to represent the value of the 21 cm EoR power spectrum signal for a particular wave mode, then assigning non-zero probability to a negative value would be incorrect in a very strict treatment. If this value is well known, then the width of the prior might be such that negative values are negligibly probable anyway. However, the current state of the field is that this value is not particularly well-constrained for any wave mode. Fortunately, this is a one-dimensional integral, and so reasonably arbitrary positive-definite priors can be marginalized over numerically. There are also a number of positive-definite priors with shape and scale parameters that can yield a closed-form solution when integrated against the Gaussian likelihood, should this be the chosen interpretation for μ_0. We show the derivation using Gaussian priors.

An alternative interpretation for μ_0 in the HERA context when applying this result is that it is a common bias which we expect each epoch to be limited by e.g. the expected strength of residual cable reflections or radio frequency interference (RFI) based on our understanding of systematic mitigation. An alternative hypothesis then reflects an instance where systematic mitigation was not as effective as expected in one or more epochs (or potentially far stronger than expected in the case of a negative bias, indicating signal loss).

We explicitly write our priors like so:

$$P(\mu_0) = \frac{\exp\left(-(\mu_0-\mu_p)^2/\sigma_p^2\right)}{\sqrt{2\pi\sigma_p^2}} \tag{8}$$

$$P(\boldsymbol{\varepsilon}|\mathcal{H}_i) = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{\varepsilon}-\boldsymbol{\mu}_{\boldsymbol{\varepsilon},i})^T\mathbf{C}_{\boldsymbol{\varepsilon},i}^{-1}(\boldsymbol{\varepsilon}-\boldsymbol{\mu}_{\boldsymbol{\varepsilon},i})\right)}{\sqrt{\det(2\pi\mathbf{C}_{\boldsymbol{\varepsilon},i})}}, \tag{9}$$

where we allow for potential correlations in the bias vector by way of $\mathbf{C}_{\boldsymbol{\varepsilon},i}$. In other words, each alternative hypothesis, $\mathcal{H}_i$, is represented by a mean value, and a relationship between variations in the bias parameters. Having these degrees of freedom allows the analyst to sculpt the statistical question. For instance, taking the limit as $\mathbf{C}_{\boldsymbol{\varepsilon},i}$ goes to the $\mathbf{0}$ matrix produces a delta function in the marginalization, which is equivalent to conditioning on a bias equal to $\boldsymbol{\mu}_{\boldsymbol{\varepsilon},i}$. As another example, if one already believes that all the data are biased, then one may compare a hypothesis with diagonal covariance against one with a highly degenerate covariance, where the degeneracies probe exact (or nearly exact) equality relationships between the biases. We discuss this in more detail in Section 2.4.

Since equation (6) is a limiting case of equation (7), we focus on equation (7). The integral over $\boldsymbol{\varepsilon}$ is a multidimensional Gaussian convolution. The result, which can be derived by completing the square and some matrix manipulation (or the convolution theorem), is essentially identical to the one-dimensional case. We obtain

$$f(\mu_0) \equiv \int_{\mathbb{R}^N} \mathrm{d}\boldsymbol{\varepsilon}\, P(\boldsymbol{\varepsilon})P(\mathbf{d}|\mathbf{C}_0, \mu_0, \boldsymbol{\varepsilon}, \mathcal{H}_i) = \frac{\exp\left(-\frac{1}{2}(\mathbf{d}-\mu_0\mathbb{1}-\boldsymbol{\mu}_{\boldsymbol{\varepsilon},i})^T(\mathbf{C}_{\boldsymbol{\varepsilon},i}+\mathbf{C}_0)^{-1}(\mathbf{d}-\mu_0\mathbb{1}-\boldsymbol{\mu}_{\boldsymbol{\varepsilon},i})\right)}{\sqrt{\det(2\pi(\mathbf{C}_{\boldsymbol{\varepsilon},i}+\mathbf{C}_0))}}. \tag{10}$$

We must then integrate this function, which is Gaussian in μ_0, against the μ_0 prior. However, it is not a probability density function in μ_0, so we must be a little careful. By inspection of the quadratic term, the 'variance' of this Gaussian function, $\tilde{\sigma}_i^2$, is defined by

$$\frac{1}{\tilde{\sigma}_i^2} = \mathbb{1}^T(\mathbf{C}_{\boldsymbol{\varepsilon},i}+\mathbf{C}_0)^{-1}\mathbb{1}. \tag{11}$$

To see the 'mean' of this Gaussian function, $\tilde{\mu}_i$ we do a standard maximization procedure:

$$\left.\frac{\mathrm{d}\ln f}{\mathrm{d}\mu_0}\right|_{\tilde{\mu}_i} = \mathbb{1}^T(\mathbf{C}_{\boldsymbol{\varepsilon},i}+\mathbf{C}_0)^{-1}(\mathbf{d}-\tilde{\mu}_i\mathbb{1}-\boldsymbol{\mu}_{\boldsymbol{\varepsilon},i}) = 0. \tag{12}$$

This is satisfied when

$$\tilde{\mu}_i = \tilde{\sigma}_i^2\mathbb{1}^T(\mathbf{C}_{\boldsymbol{\varepsilon},i}+\mathbf{C}_0)^{-1}(\mathbf{d}-\boldsymbol{\mu}_{\boldsymbol{\varepsilon},i}). \tag{13}$$

Downloaded from https://academic.oup.com/mnras/article/518/4/6041/6852951 by guest on 22 April 2024

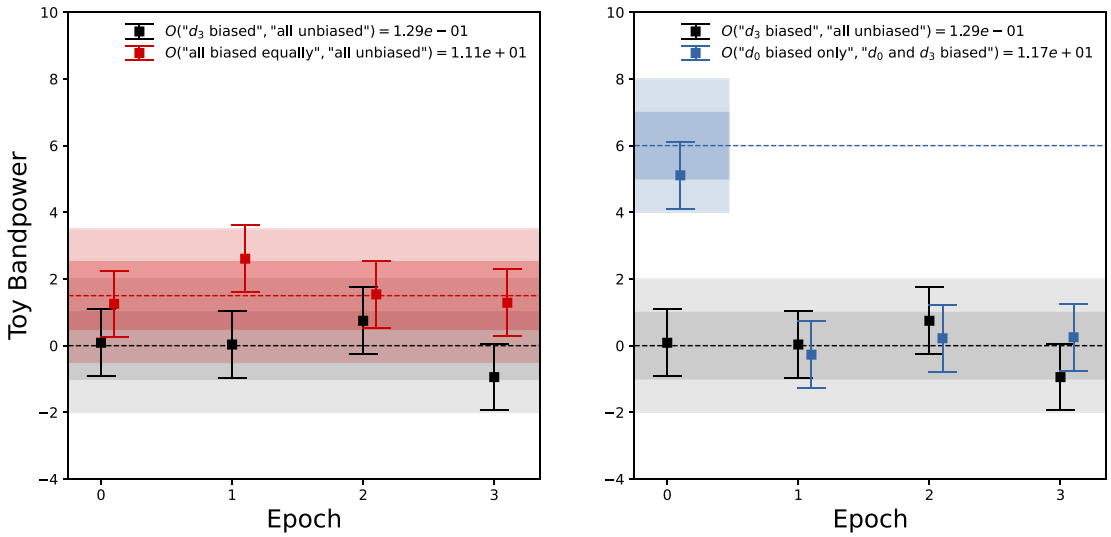

Figure 1. Toy representation of the type of problem that is solvable by the formalism in this paper. Depicted are several pseudorandom draws of normally distributed random variables, colour coded according to their mean (dashed), and their 1σ and 2σ confidence interval (shaded). Definitively concluding whether there is a significant bias in any given data set by eye (doing a χ-by-eye) is difficult in some circumstances. By formalizing different bias scenarios as Bayesian hypotheses, we are able to produce the odds of different scenarios given the data (legend). Even using very broad priors about how strong the biases could be, we are able to resolve each correct hypothesis against its nearest competitor. Using more informative bias priors (which results in more specific versions of each hypothesis), we can bring the odds listed for the red data over 100, and the odds listed for the blue data to an extremely high exponent.

Completing the square and performing the integral over μ_0 yields

$$P(\mathbf{d}|\mathbf{C}_0, \mathcal{H}_i) = \frac{\exp\left(-\frac{1}{2}\left(\tilde{\mu}_i - \mu_p\right)^2/(\sigma_p^2 + \tilde{\sigma}_i^2)\right)}{\sqrt{2\pi(\tilde{\sigma}_i^2 + \sigma_p^2)}} \times \exp\left(-\frac{1}{2}\left[(\mathbf{d} - \boldsymbol{\mu}_{\boldsymbol{\varepsilon},i})^T(\mathbf{C}_0 + \mathbf{C}_{\boldsymbol{\varepsilon},i})^{-1}(\mathbf{d} - \boldsymbol{\mu}_{\boldsymbol{\varepsilon},i}) - \tilde{\mu}_i^2/\tilde{\sigma}_i^2\right]\right)\frac{\sqrt{2\pi\tilde{\sigma}_i^2}}{\sqrt{\det(2\pi(\mathbf{C}_0 + \mathbf{C}_{\boldsymbol{\varepsilon},i}))}}. \quad (14)$$

Since this is a probability distribution in $\mathbf{d}$, and is of Gaussian form, it can be written

$$P(\mathbf{d}|\mathbf{C}_0, \mathcal{H}_i) = \frac{\exp\left(-\frac{1}{2}(\mathbf{d} - \boldsymbol{\mu}_i')^T\mathbf{C}_i'^{-1}(\mathbf{d} - \boldsymbol{\mu}_i')\right)}{\sqrt{\det(2\pi\mathbf{C}_i')}}. \quad (15)$$

In one final exercise in completing the square, one finds that

$$\boldsymbol{\mu}_i' = \boldsymbol{\mu}_{\boldsymbol{\varepsilon}} + \mu_p\mathbb{1} \quad (16)$$

$$\mathbf{C}_i'^{-1} = (\mathbf{C}_0 + \mathbf{C}_{\boldsymbol{\varepsilon},i})^{-1} - \frac{(\mathbf{C}_0 + \mathbf{C}_{\boldsymbol{\varepsilon},i})^{-1}\mathbb{1}\mathbb{1}^T(\mathbf{C}_0 + \mathbf{C}_{\boldsymbol{\varepsilon},i})^{-1}}{\tilde{\sigma}_i^{-2} + \sigma_p^{-2}} \quad (17)$$

Now that the likelihoods are calculated, all that is necessary to produce the posterior probabilities is to multiply through by $P(\mathcal{H})$ and normalize by the total evidence (equation 3). We explore decision rules in more detail in Section 3, but the simplest decision rule from these posteriors is to choose the maximum a posteriori (MAP) hypothesis as the one 'preferred' by the data.

We also make some remarks about the form of the likelihood. A particularly interesting quantity to consider is the logarithm of the odds of two hypotheses:

$$\log O(\mathcal{H}_i, \mathcal{H}_j) = \log\left(\frac{P(\mathcal{H}_i|\mathbf{d}, \mathbf{C}_0)}{P(\mathcal{H}_j|\mathbf{d}, \mathbf{C}_0)}\right) \quad (18)$$

If one conditions on a particular mean and bias and uses a flat prior on $\{\mathcal{H}_i\}$, then this quantity reduces to a difference in χ^2 statistics with two different assumed models – a familiar model comparison procedure from frequentist statistics. We can therefore think of comparing hypotheses this way as a Bayesian generalization to comparing the χ^2 statistic of the data in the face of different models. In other words, this is the quantified version of the 'informed χ-by-eye' encouraged in the example we plot in Fig. 1.

2.4 Choosing parameters for the test

In this section, we discuss parameter choices for each part of the test, which includes choice of prior probabilities, hyperparameters (parameters describing the priors) for continuous priors, and odds thresholds. We begin by discussing the choice of prior for μ_0 in terms of the HERA problem.

2.4.1 Choosing the prior for μ_0

We break the HERA problem into three scenarios and provide suggestions for each one. We note that when the measurement is significantly noise-dominated, the prior on μ_0 does not drive the outcome of the inference provided it is less than the noise level.

Residual systematics dominate the signal: In this case, a Gaussian prior may be appropriate. For non-zero μ_0, the interpretation would be that μ_0 represents a common value to which the epochs are expected to be biased by systematics. Negative values for some modes may represent large amount of signal loss. One could also condition on $\mu_0 = 0$, in which case a failure of the test would indicate the presence of significant systematics.

Signal dominates the systematics, but is poorly constrained: In this case, a Gaussian prior can be inappropriate due to the true signal power being strictly positive. There are at least two options. One option is to implement numerical integration techniques that can handle a non-Gaussian prior. Another option is to instead condition on $\mu_0 = 0$ and use a hypothesis for testing where all epochs are biased identically (or nearly so). In this scenario, ruling the data as all biased

Downloaded from https://academic.oup.com/mnras/article/518/4/6041/6852951 by guest on 22 April 2024

identically indicates that something signal-like has been located that is consistent between the epochs. Only further examination of the data can tell whether this is likely to be the signal of interest or some other, unexpected systematic.

Signal dominates the systematics, and is well constrained: In this case, we can use a Gaussian prior, where the width represents the strength of the constraint. The bias prior mean and/or width would need to be larger than this constraint and the error bar for the test to provide statistically meaningful results. A test failure would indicate an extreme malfunction in systematic mitigation. This is functionally the same as when systematics dominate the signal. All that changes is the meaning of the nuisance parameter.

We provide an open-source PYTHON code called CHIBORG for calculation of the posterior using analytically tractable priors as well as those that require numerical computation. This software is complete with documentation and example notebooks that explore various input and output options of the test to help users learn how to formulate the test for their purposes.

2.4.2 *Choosing the bias priors for alternate hypotheses*

The aim of this section is to define the subspace of bias hyperparameters that produce a well-defined hypothesis test using the information theoretical concept of 'mutual information'. This requires concepts from information theory such as entropy of random variables, which we briefly review in Appendix A. In short, we wish to restrict our attention to scenarios (characterized by Gaussian bias priors with particular means and variances) that can be meaningfully distinguished from the null hypothesis and each other. Some biases will be too weak to distinguish from expected noise fluctuations around the null hypothesis, in which case our test will be inconclusive. By first finding the range of bias scenarios that are in principle distinguishable, we can substantially 'sharpen' our test by only considering situations where a decisive result is possible.

In the following text, we use the *mutual information* to determine the range of effects that are distinguishable using our hypothesis test. The basic idea is that we can obtain at most $S(P(\mathcal{H}))$ bits of information from the hypothesis test, where $S(P(x))$ is the Shannon entropy of a random variate x (we may also write $S(x)$). Taking a two-hypothesis test with a flat prior as an example, this equates to (equation A2) $S_2(1/2) = 1$ bit of information. With 1 bit of information, we would be able to exactly specify the state of the random variable, $\mathcal{H}$.[3] For some data, specifically those where strong biases are common, the mutual information is close to maximal. This means that the average information provided by each application of the hypothesis test is sufficient to distinguish between data that come from the various hypotheses. Conversely, it is difficult to distinguish between a weak bias and what may just be a noise fluctuation; much less information is available in this case. Note that since the actual data observed by the analyst may deviate from the priors, we should be wary of saying that the information gained in practice is necessarily reflected by this relationship. We use this only as a rough guide for determining which priors are actually distinguishable with the volume of data used in the test.

In Fig. 2, we display the mutual information between the various test hypotheses and hypothetical data sets drawn according to the bias priors reflected by the choice of mean and width shown on the axes. We assume that each run of the test only has access to four data points, and condition on the underlying mean being equal to 0. Our hypothesis set can be represented by 16 4-bit strings, where a 1 in the jth entry of the string indicates that the jth datum is biased, and a 0 indicates it is unbiased. The corresponding bias prior is formulated such that the mean vector has non-zero entries only for those indices in which the string is non-zero. The bias prior covariance is diagonal, and constructed in the same way as the mean. We will sometimes refer to this style of hypothesis set as 'considering all diagonal hypotheses'. Parameter choices closer to the bottom-left corner of the plot produce bias priors that are, in a formal sense, close to the null hypothesis.[4] Parameter choices for which the mutual information is low indicate choices of hypotheses that are not reliably distinguishable from one another. Regions in which the mutual information is nearly maximal (4 bits in this example) indicate hypothesis sets for which the test can reliably decide which of the hypotheses is most likely. The corresponding posterior distribution over the hypotheses will usually be strongly concentrated at this hypothesis, i.e. it will be a low-entropy state.

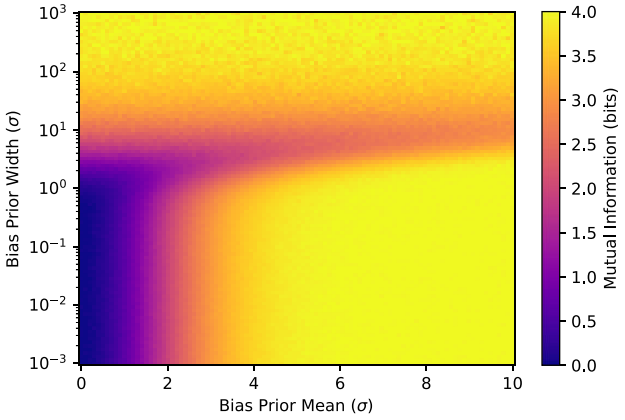

Figure 2. The mutual information provided by a hypothetical data set drawn according to the priors, as a function of bias prior width and mean (in units of the noise standard deviation, σ). The mutual information represents the reduction in uncertainty between two random variables when one is known – in this case the data and the hypotheses (Appendix A). Each draw has four data points (four epochs in the HERA example), and we consider all 16 basic bias configurations (see the main text), leading to a total of 4 bits that are required to specify the state of the system. We assume an exactly zero-mean null hypothesis. Small fluctuations are visible due to Monte Carlo estimation error, whose size is proportional to the bias prior width, hence the increased estimator noise at the top of the plot. In general, we can use this as a landscape for which hypotheses are distinguishable from one another. Bias priors corresponding to the bottom-left sector of the plot produce data that are only weakly biased. When these biases are approximately the size of the statistical error of the data or smaller, the test is, on average, unable to distinguish between any hypotheses. If biases are concentrated away from the origin, some confusion can arise if the bias prior is wide enough to have significant density at the origin.

With only a few data points, Fig. 2 suggests that we may only be able to probe scenarios in which biases are frequently more than a few multiples of the error bar. In other words, it may only decisively point out visibly obvious statistical tensions. We remark that this is a rigorous, quantitative alternative to 'χ-by-eye' that allows an analyst to tune and objectively assess statistical questions with confidence.

[3]Said another way, to communicate the state of the random variable $\mathcal{H}$, we require the full use of 1 bit on average. See Appendix A.

[4]In the binary case (two hypotheses), the plotted quantity is known as the 'Jensen–Shannon Divergence', and its square root satisfies the properties of a mathematical metric.

Downloaded from https://academic.oup.com/mnras/article/518/4/6041/6852951 by guest on 22 April 2024

For the HERA use case, there are hundreds of groups of two to four data points that would need to be assessed in this way, and so this test acts as a form of data reduction. We also point out that an easy avenue for combining more data and therefore increasing the information provided by the test is to combine wave modes in the power spectrum. Formalism for this procedure is contained in Appendix B.

In considering a set of priors for the alternate hypotheses, a useful choice for small data sets is one in which every combination of bias configurations is considered (as used in Fig. 2). If zero-mean priors are used, this requires 2^N hypotheses, while non-zero-mean priors may need as many as 3^N hypotheses if all combinations of positive and negative biases for that mean are probed (assuming a single mean is used). If the analyst possesses genuine prior information that can eliminate some of these possibilities, then clearly some of these hypotheses need not be considered. Once a set of biased data has been identified, it may then be useful to introduce a second stage in the test, where the corresponding covariance matrix of the MAP hypothesis is partitioned to identify common biases among the biased subset (Appendix C). For larger data sets where this exponential growth in hypotheses is impractical, one can imagine repeatedly bisecting the data and running the test with a small hypothesis set until a confident positive result for a bias configuration is located.[5] For example, one could use only two hypotheses: one with no biases, and one where all epochs are biased independently. Since data where only some fraction of its points are biased can still be more consistent with this alternate hypothesis than the null one, this can be a useful way to reduce a data set with a coarser hypothesis test before applying a more targeted one.

The bias priors are meant to reflect hypothetical scenarios that the analyst is interested in knowing. Furthermore, since only a finite number of hypotheses is considered, the MAP decision rule can only choose the hypothesis that is closest to modelling the data within the context of the considered hypotheses. An important implication of this concept is that even if a prior corresponding to some scenario seems to grossly misrepresent the content of the data, it may yet be a better match than, say, the null hypothesis. For example, having an excessively broad prior, e.g. greater than 70 times the size of the error bar, but zero mean, does not prohibit the decision rule from classifying a 3σ outlier as a significant bias. This means that in order to assess fairly general propositions, such as 'the data contain at least one bias', one need not have an exhaustively precise model of how that bias should look, so long as the alternate hypothesis bias priors capture the meaning of 'significance' intended by the analyst (and the analyst is cautioned not to consider hypotheses indistinguishable from the null as being significant). In other words, the marginalization procedure (over the bias priors) endows the test with a degree of flexibility that allows the analyst to sculpt the test to be most sensitive to whichever scenarios they are most concerned about. However, the corresponding answer from the test is limited in scope, in that it can only judge whether some hypothesis is *more likely* than another in consideration, e.g. 'it is more likely that there is at least one bias than that there are none, when these are the types of bias configurations in consideration'.

[5]One notion of 'confidence' is that the maximum a posteriori hypothesis 'dominates' the other hypotheses with a substantial odds value, say 10:1 for all hypotheses. This naturally leads to using entropy as a measure of confidence.

2.4.3 Choosing the prior probability for the hypotheses

The final choice of prior is that regarding each hypothesis, $P(\mathcal{H})$, i.e. the prior probability that each of the (mutually exclusive) hypotheses is true. This is a primarily subjective prior unless one has a model for how often anomalous, unexpected biases should occur in the data. There are many conceivable choices depending on the preferred line of questioning. For instance, one might adopt a flat prior over all bias configurations in order to show no preference to any one configuration. However, this will give preference to conclusions regarding certain overall numbers of biases. Mathematically, there is only one way for all N data points to be simultaneously biased (though perhaps at a different level in each datum), but there are $\binom{N}{k}$ ways for there to exist k biases within N data points. This means that with a flat prior over the 2^N diagonal hypotheses, if one asks the question 'how many data are biased?' rather than 'which combination of biases is most like the data?' then they are a priori more likely to find an answer closer to $N/2$ (where the combinatorial factor is highest) than something closer to N or 0. A similar statement holds for the various correlated partitions of any one of these bias macrostates (Appendix C). In what follows, we choose flat priors at every stage. For the HERA problem, this expresses our assumption that each epoch is just as likely to be biased as each other epoch, which we reason from the fact that data from each epoch were processed identically (exposed to the same selection cuts and systematic mitigation methods).

In subsequent sections, we apply these results to an array of examples of ranging realism.

3 RESULTS WITH TOY DATA

In this section, we apply the formalism to some basic simulations to build intuition. We then apply different decision rules to the posteriors and subsequently examine their performance as classifiers. In these first examples, we examine the case when only a few data are combined, since this is the most analogous to the HERA example.

3.1 Formulating a decision rule

Once the posterior probabilities of each hypothesis are calculated, one may then be interested in formulating a decision about which hypothesis to assume. A decision problem is typically formulated by determining an objective function that is extremized according to the goals of the analyst. We define a loss function, $L(\mathcal{H}_i, \mathcal{H}_j)$, and minimize the expected loss over the posterior. Formally,

$$\langle L(\mathcal{H}_i)\rangle = \sum_{j=0}^{Q-1} L(\mathcal{H}_i, \mathcal{H}_j) P(\mathcal{H}_j|\mathbf{d}, \mathbf{C}_0). \tag{19}$$

The quantity $L(\mathcal{H}_i, \mathcal{H}_j)$ describes the specific loss associated with behaving as if $\mathcal{H}_i$ is true when in fact $\mathcal{H}_j$ is true. $L(\mathcal{H}_i)$ is then the expected loss of behaving as if $\mathcal{H}_i$ is true.

Defining a loss function is sometimes arbitrary, although an objective basis for one can also exist. For instance, in the HERA problem, one might ascribe a loss function based on how much wasted reanalysis occurs when an epoch is misidentified as containing a bias, assuming one goes back to check on the data through some independent analysis when a bias is identified. The usual loss function chosen in a wide variety of analyses throughout the general scientific literature is the sum of squared errors. To apply this to our problem we would need to formulate some sort of 'squared distance' between our various hypotheses. There are various candidates from

Downloaded from https://academic.oup.com/mnras/article/518/4/6041/6852951 by guest on 22 April 2024

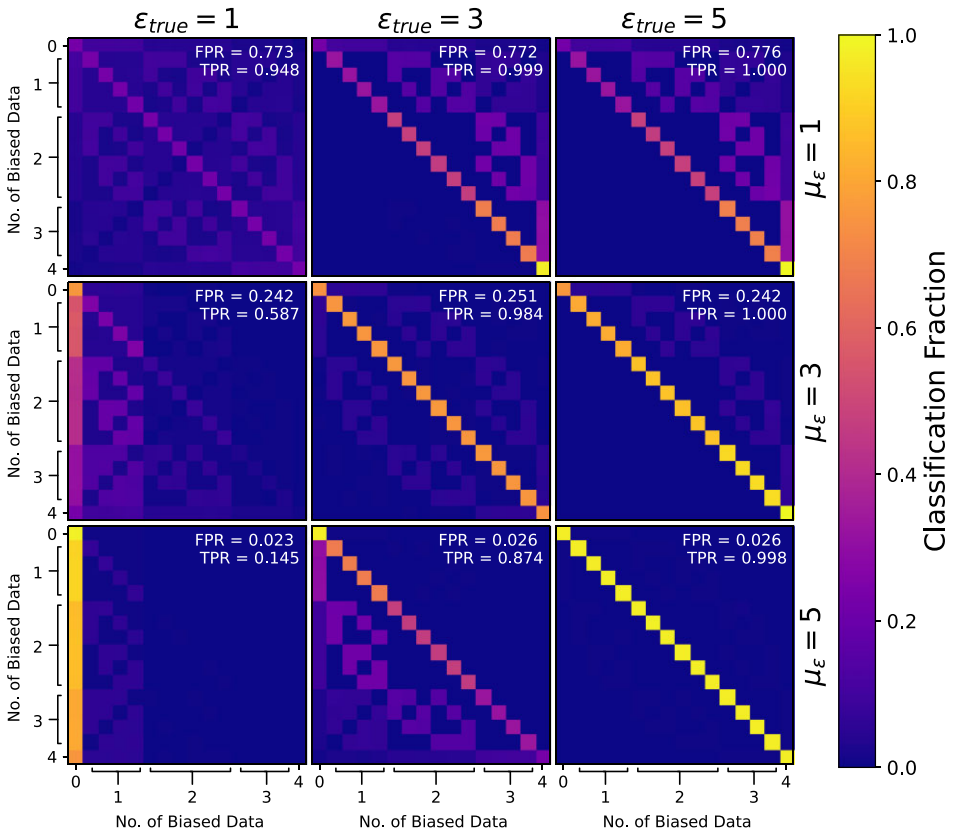

Figure 3. Classification matrices for the $N = 4$ system investigating the 16 basic bias combinations. The bias strength in the data is given by the column of the panel (labels at the top of each column), and the assumed bias strength is given by the row (labels at the right of each row), each taking the values 1, 3, and 5 in units of the noise standard deviation. Within each panel is a 16×16 matrix, where the column index indicates the particular bias configuration that was chosen by the MAP rule, and the row index indicates the actual simulated bias configuration. The colour indicates the fraction of samples simulated according to a particular state (row index) that was classified as the state indicated by the column index. In other words, diagonal entries represent proper identifications, and off-diagonal entries represent classification errors of some form. The annotations beside the axes tell the number of biases in the indicated state. In the upper right corner of each panel is listed the 'false positive rate' (fraction of unbiased data classified as somehow biased) and the 'true positive rate' (fraction of biased data classified as containing at least one bias, even if the biased states are incorrectly identified). When the parameters describing the alternate hypotheses do not lie in the region of high mutual information, there is an extremely large false positive rate. While using alternate hypotheses that are highly distinguishable generically decreases the true positive rate, the ratio of true positives to false positives increases, indicating that one can be more confident in positive classifications with this choice of hypothesis set.

the divergences of information theory, e.g. the Kullback–Liebler divergence or the Jensen–Shannon divergence between any two hypotheses. The latter is equivalent to the mutual information in a binary test between the two hypotheses in question. If we choose a set of hypotheses that are all mutually distinguishable as advised in Section 2.4, then the squared distance (measured in bits in this case) will be approximately 1 bit between any two hypotheses. In other words, there will be equal loss for any misclassification produced by the algorithm, and 0 loss for a proper classification (identifying exactly which epochs are biased in the HERA problem). Using such a loss function, we would obtain

$$\langle L(\mathcal{H}_i)\rangle = 1 - P(\mathcal{H}_i|\mathbf{d}, \mathbf{C}_0), \tag{20}$$

which is minimized for the hypothesis whose posterior probability is maximal. In other words, this just reproduces the MAP decision rule.[6]

[6]Even if the hypotheses are not all highly distinguishable, some hypothesis sets result in roughly equal Jensen–Shannon divergences anyway, which will also produce the MAP rule.

A loss function of particular relevance to the HERA problem might be to penalize false negatives more severely compared to other misclassifications. While identifying a bias where there are none produces concern for the quality of the power spectrum limit, a false negative fails to prevent a lower quality limit. On the other hand, the analysts may also worry about producing selection bias, in which case a loss function that minimizes false positives would be relevant. In general, there is a trade-off between false negatives and false positives, as we show in exploration below.

3.2 Classification performance

We explore the classification performance of the jackknife test by simulating large ensembles of biased data and examining various metrics. First, we use the loss function of equation (19) (MAP decision rule) and count how many instances of each bias configuration are mapped into each bias hypothesis. We also examine this ensemble in terms of entropy. We then inspect the true and false positive rates of a decision rule where false negatives are strongly penalized.

In Fig. 3, we show an array of 'confusion matrices'. We generate an ensemble of Gaussian random vectors of length $N = 4$ under each of the 16 possible bias configurations, with a standard deviation of

Downloaded from https://academic.oup.com/mnras/article/518/4/6041/6852951 by guest on 22 April 2024

1 in each dimension and no correlations. We choose a true bias of (exactly) 1, 3, and 5, i.e. $\varepsilon_{\rm true} \in \{1, 3, 5\}$. We choose a set of alternate hypotheses with μ_ε equal to 1, 3, and 5 as well, and $\mathbf{C}_{\boldsymbol{\varepsilon}}$ nearly 0. Then, for each simulated data set and choice of hypotheses, we count how many data are most likely to come from each of the various hypotheses. From this, we can also calculate the false positive rate, which is the fraction of unbiased sample vectors that are classified as containing at least one biased component, as well as the true positive rate, which we define as the fraction of vectors with at least one component identified as biased (even if the truly biased components are misidentified). All off-diagonal entries in the matrix represent misclassifications, and the diagonal entries show proper classifications.

Unsurprisingly, when biases are strong (rightmost column), the true positive rate is extremely high. However, if the alternate hypotheses are not sufficiently distinct from the null hypothesis (upper row), the MAP rule suffers from frequent misclassification within the biased states. The problem is more prominent for configurations with fewer biases. In addition, the false positive rate is extremely large. The false positive rate is, in general, not a function of the true biases, since the null hypothesis is identical regardless of $\varepsilon_{\rm true}$. The extra aggression exemplified in the top row is also manifest in the pattern of misclassified true positives. When such misclassifications occur, they generally conclude that there are more biases present than in reality. In other words, for the top row, the most common misclassification is that states with one bias are classified as having two biases, and two bias states are misclassified as having three, etc. Usually the correct bias components are identified, but an extra bias is proposed. This is caused by an upward statistical fluctuation in the unbiased state appearing as a bias, which is a significantly less likely scenario when the assumed bias level is large compared to the noise. For the bottom row of matrices, the misclassification pattern is the opposite, except for when biases are extremely weak compared to the alternate hypotheses. When biased data are misclassified, they are often classified as containing one less bias then they truly possess. However, once the biases are sufficiently weak, the most common misclassification is a false negative (bottom-left matrix). Such confusion matrices offer a way of choosing hyperparameters for the alternative hypotheses in terms of one's appetite for various types of failures, rather than from the theoretical considerations presented in Section 2.4. However in general we see that these two methods correspond with one another in a logical manner.

In Fig. 4, we use entropy as a measure of the 'confidence' of the posterior distribution. The basic idea is that if the posterior is strongly concentrated on a small number of states, the sample entropy for that data draw

$$\hat{S}(\mathcal{H}|\mathbf{d}^{(i)}) \equiv -\sum_{j=0}^{Q-1} P(\mathcal{H}_j|\mathbf{d}^{(i)}) \log_2 \left(P(\mathcal{H}_j|\mathbf{d}^{(i)})\right) \tag{21}$$

will be low. If we average over all $\mathbf{d}^{(i)}$ in the ensemble, then we have the sample entropy of the data set under that particular data generating distribution,

$$\hat{S}(\mathcal{H}|\mathbf{d}) \equiv -\frac{1}{D}\sum_{i=1}^{D} \hat{S}(\mathcal{H}|\mathbf{d}^{(i)}), \tag{22}$$

D being the number of samples drawn. This acts as a measure of how certain the algorithm is of its classifications on average. We repeat a similar simulation process as for Fig. 3, but on a more finely spaced grid of μ_ε and $\varepsilon_{\rm true}$. Additionally, we mix all bias configurations at a given $\varepsilon_{\rm true}$ in equal proportions. We generally observe lower entropies for higher μ_ε and $\varepsilon_{\rm true}$. For low $\varepsilon_{\rm true}$, and high μ_ε, these low-entropy states are 'confident misclassifications'.

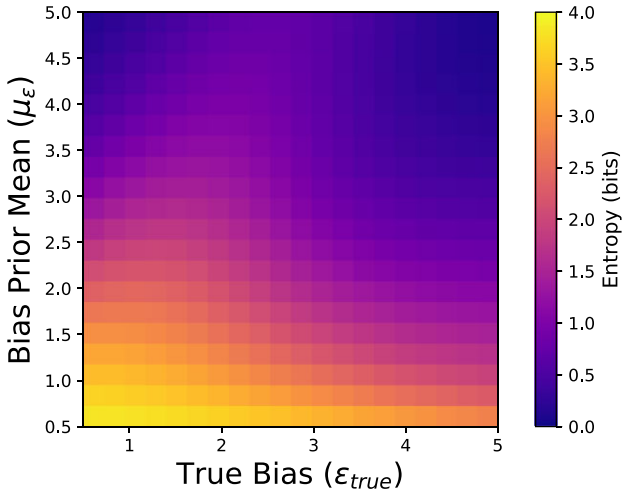

Figure 4. Entropy of the posterior distribution under different choices of alternative hypotheses and true biases. Regions of high entropy towards the lower left reflect uncertain conclusions, while regions of lower entropy express more certainty (generally fewer competing hypotheses).

This is reasonable behaviour, since weak biases should be difficult to distinguish from noise when one is expecting strong biases. For example, if one is specifically expecting 5σ biases and observes a datum at 1σ, it would be unreasonable to be in doubt about which hypothesis is most likely. In a complementary vein, if one expects weak biases and weak biases are often present, one should expect to be highly uncertain about the state of the system. This corresponds to the region of high entropy in the lower left of the plot.

So far, we have not used much the loss function formalism. A loss function is highly contextual. A particularly important type of misclassification in the HERA problem is a false negative (i.e. concluding no biases when at least one is present). If a bias can be identified before coherently averaging the data to form a final power spectrum, then in principle the data that led to the subspectra on which the test is being run can be reanalysed and potentially cleared of systematics. Conversely, failing to identify a bias can lead to a systematically dominated power spectrum upper limit that could have otherwise been avoided. One may use the loss function in order to more severely penalize such misclassifications and therefore avoid them more often in exchange for more false positives. Alternatively, one might be concerned with selection bias effects, such as in the HERA problem where we deploy the jackknife test on a measurement that is extremely close to the final scientific product. In that case, one could construct a loss function where false positives are punished more severely, so that one does not inadvertently label unbiased data as biased.

To illustrate this with math, we can think of equation (18) as a loss matrix $\mathbf{L}$ acting on a posterior probability vector $\mathbf{p}$, where the ijth entry of $\mathbf{L}$ is equal to $L(\mathcal{H}_i, \mathcal{H}_j)$ and the jth component of $\mathbf{p}$ is the posterior probability of the jth hypothesis. To achieve the MAP rule, one would use a loss function given by

$$\mathbf{L}_{\rm MAP} = \mathbb{1}\mathbb{1}^T - \mathbf{I}_Q = \begin{pmatrix} 0 & 1 & 1 & \dots & 1 \\ 1 & 0 & 1 & \dots & 1 \\ 1 & 1 & 0 & \dots & 1 \\ \vdots & & & \ddots & \end{pmatrix}, \tag{23}$$

where $\mathbf{I}_Q$ is the $Q \times Q$ identity matrix. This loss matrix is a symmetric matrix with 0s down the diagonals and 1 for all other entries, i.e. 0

Downloaded from https://academic.oup.com/mnras/article/518/4/6041/6852951 by guest on 22 April 2024

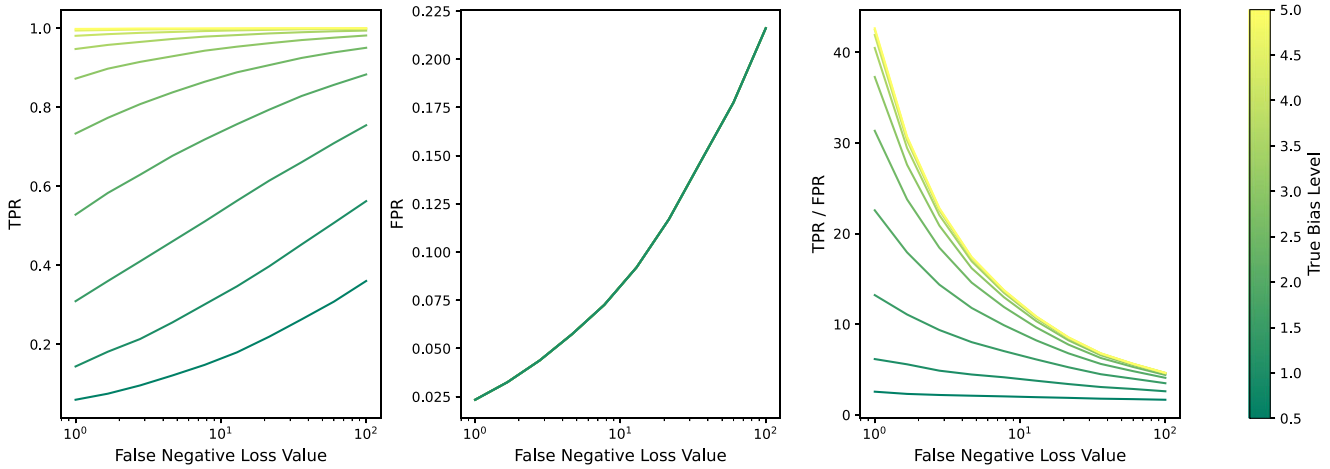

Figure 5. True positive rates, false positive rates, and their ratio when using a bias prior mean of five times the error bar, as a function of the strength of loss for false negatives (equation 23). For larger loss values, we can obtain strong gains in true positive rates for smaller biases in exchange for a larger false positive rate. The odds of true to false positives decrease monotonically as the loss strength goes up, but reasonable odds can be maintained for relatively weak biases at an appreciable increase in true positive rate.

loss for proper classifications and a unit loss for any misclassification regardless of its type. This reproduces equation (19) exactly. To produce a lower false negative rate at the expense of more false positives, one changes the top row of this matrix to some false negative loss value, $L > 1$, except for the diagonal entry. In math,

$$\mathbf{L}_{\text{low FNR}} = \begin{pmatrix} 0 & L & L & \dots & L \\ 1 & 0 & 1 & \dots & 1 \\ 1 & 1 & 0 & \dots & 1 \\ \vdots & & & \ddots & \end{pmatrix}. \tag{24}$$

Similarly, to achieve a lower false positive rate, at the expense for more false negatives, one replaces the first column with some loss value, $L > 1$. Explicitly,

$$\mathbf{L}_{\text{low FPR}} = \begin{pmatrix} 0 & 1 & 1 & \dots & 1 \\ L & 0 & 1 & \dots & 1 \\ L & 1 & 0 & \dots & 1 \\ \vdots & & & \ddots & \end{pmatrix}. \tag{25}$$

In Fig. 5, we explore the effect of modulating a false negative penalty while holding all other misclassifications at a constant loss value. In other words, we vary L in equation (23). When $L = 1$, this is the same as the MAP rule. Using the same simulation suite as Fig. 4, we analyse the true and false positive rate as we adjust the strength of the loss for false negatives relative to that of other misclassifications. We only use alternate hypotheses with $\mu_\varepsilon = 5\sigma$. The left-hand panel shows the true positive rate as a function of loss value for the chosen biases. We see that for weak biases we can appreciably increase the true positive rate. The middle panel shows the false positive rate. Combining this with the left-hand panel, one may tune the false positive rate and then estimate the sensitivity to various biases. For instance, using a loss strength of 10 for false negatives increases the true positive rate for 1σ biases by about 50 per cent of its MAP value (loss strength equal to 1) while keeping the false positive rate at less than 10 per cent. The right-hand panel shows the ratio of true to false positive rate. If for some reason the prevalence of each bias configuration is known (or if one is willing to use prior values), this quickly translates to an odds of a positive result being true versus false. This can also be used to tune the loss function by setting an odds threshold for a given strength of bias. In general, we see that this ratio is monotonically decreasing as a function of loss strength, indicating that the false positive rate generally increases quicker than the true positive rate.

Using a large array of simulated data has allowed us to thoroughly explore the statistical properties underlying this test. Specifically, we have illustrated a number of tools one can use in order to tune the test to the question at hand. We have primarily focused on the form of the jackknife test in which one is interested in all 2^N bias configurations without asking about interbias similarities ($\mathbf{C}_\varepsilon$ diagonal). In principle, the manner of investigation presented in this section may be used for any hypothesis set, e.g. those formalized in Appendix C, or a more specific hypothesis set defined by the analyst's problem. We first turn to simulations from the HERA validation pipeline in order to investigate the performance of the test in a controlled yet realistic setting, and then examine the results of the test on the measured HERA bandpowers.

4 RESULTS FOR HERA SIMULATIONS AND DATA

In this section, we apply the jackknife test to the simulated and measured HERA bandpowers from their first observing season, as constructed and analysed in The HERA Collaboration (2022). The data for the observing season, spanning 94 nights after initial data quality cuts, were split into four epochs of roughly equal size. The data were then analysed identically, and power spectra were formed on a per-epoch basis. Accompanying the HERA data analysis pipeline is a comprehensive validation pipeline designed to verify the efficacy of the data analysis techniques using thorough and realistic simulations of the experiment (Aguirre et al. 2022). This includes simulation of foregrounds, fiducial EoR signals, thermal noise, calibration, known systematic effects, systematic mitigation techniques, and so on. For The HERA Collaboration (2022), all 94 nights passing initial data selection were simulated and separated into epochs, mimicking the exact data analysis procedure. This allows us to apply the jackknife test to both the simulated bandpowers as well as the measured bandpowers.

Downloaded from https://academic.oup.com/mnras/article/518/4/6041/6852951 by guest on 22 April 2024

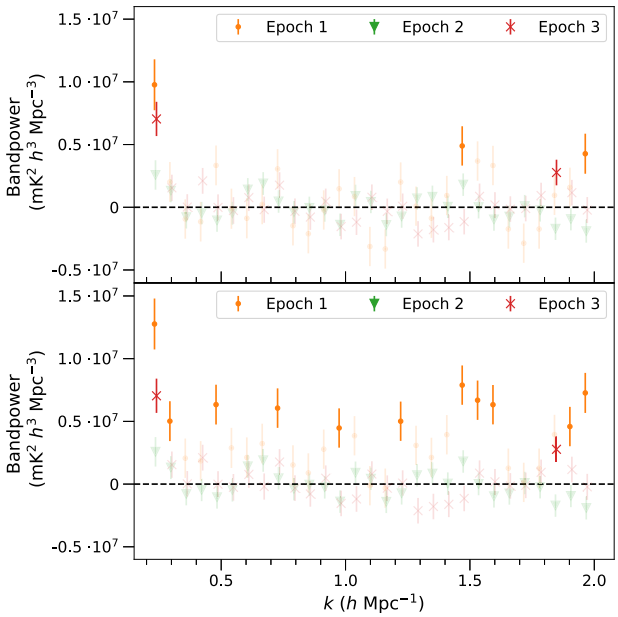

Figure 6. Top: Bandpowers directly from the HERA validation pipeline, simulating foregrounds, noise, systematics, and systematic subtraction (no EoR signal), for band 1, field E (top). Bottom: Same bandpowers, except we artificially add a bias to the epoch 1 measurements of 3×10^6 mK^2 h^3 Mpc^{-3} (approximately two times the size of the error bar for most modes). Opaque data have been identified as biased according to the MAP decision rule (see Section 4.1 for test details).

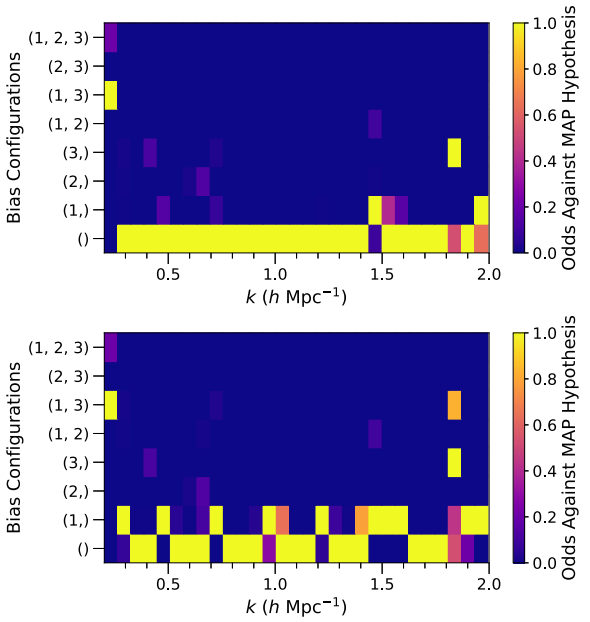

Figure 7. Top: The odds against the MAP hypothesis based on the posterior distribution for each mode when considering the raw validation bandpowers (top panel of Fig. 6). As expected, the raw validation results are largely consistent with zero-mean noise (the plot is concentrated at the empty bias configuration with relatively few competing hypotheses). Bottom: When a bias is artificially added to epoch 1, we see that the hypothesis that 'epoch 1 is biased' is highly favoured compared to other hypotheses for about half the modes.

We first show the results of the jackknife test on validation simulations where foregrounds are simulated, but no EoR signal is present. This effectively produces a data set where the null hypothesis is satisfied almost everywhere. We then modify the outputs of these simulations by adding a constant bias to one of the epochs on all modes, which is an effect that could arise, for example, due to the presence of ultra-faint RFI (Wilensky et al. 2020). Note we do not actually use the validation pipeline to simulate the effect of RFI (or any other systematic that might add a constant bias); we are just artificially introducing the bias directly to the end-stage bandpowers. Finally, we show the results of the jackknife test on the actual measured bandpowers, demonstrating a majority null result, thereby justifying (post hoc) the decision to coherently average the epochs together to form one final power spectrum.

4.1 Results with HERA validation simulations

In Figs 6 and 7, we show the results of the jackknife test for a single band and field of power spectra made using the HERA validation pipeline. These simulations included foregrounds, noise, instrumental/analysis systematics, and systematic mitigation, but no EoR signal. The jackknife settings are contained in the bulleted list below.

(i) **Hypothesis Set**: all diagonal
(ii) **Bias Mean/Width**: $6\sigma_i \pm \sigma_i$
(iii) **Null Mean/Width**: 0 ± 0
(iv) **Decision Rule**: MAP

We test all diagonal hypotheses (i.e. all epoch combinations, but without asking whether some are degenerate as in Appendix C), using a prior mean and width that are sized based on the per-epoch error bar. We use the MAP decision rule for determining biased epochs, which is equivalent in the preceding section's formalism to using a constant loss function over the range of possible misclassifications. Fig. 6 shows the bandpowers with (bottom) and without (top) the addition of an artificial bias to epoch 1 that is roughly two times the size of the error bar for most modes. Fig. 7 shows the posterior probability of each bias configuration divided by the posterior probability of the MAP hypothesis, i.e. the odds against the MAP hypothesis. Such a plot allows us to see how competitive various hypotheses are with the MAP hypothesis. This shows us that realistic simulations intended to be somewhat consistent with the null hypothesis at sufficiently high k do appear to demonstrate this consistency. Since small residual systematics still exist despite mitigation, we do not expect perfect consistency with the null hypothesis. When we artificially add a bias to epoch 1, the MAP solution shifts towards identifying epoch 1 as biased for about half the modes, as expected.

With these settings, we expect a false positive rate of about 1.5 per cent, and a true positive rate for biases of 2σ of 49 per cent. When we artificially add a 2σ bias, the data are roughly consistent with this true positive rate. Examining validation outputs from all bands and fields, and accounting for differences in number of hypotheses between tests, we find that there are more data identified as biased than expected from the false positive rate, which is between 1 per cent and 2 per cent depending on the field (higher rates for more epochs). About 10 per cent of the validation bandpowers over all bands and fields are identified as biased. Slightly less than half of the identified bandpowers are in the lowest k-mode. While efforts have been made to exclude contributions from the foreground wedge to these modes, it is possible that the cuts that provide these exclusions (The HERA Collaboration 2022) may be imperfect, and that some residual foreground signal exists in these bins. The remaining identified data points appear somewhat uniformly scattered throughout the rest of the wave numbers. This excess above the false positive rate suggests a potential low-level

Downloaded from https://academic.oup.com/mnras/article/518/4/6041/6852951 by guest on 22 April 2024

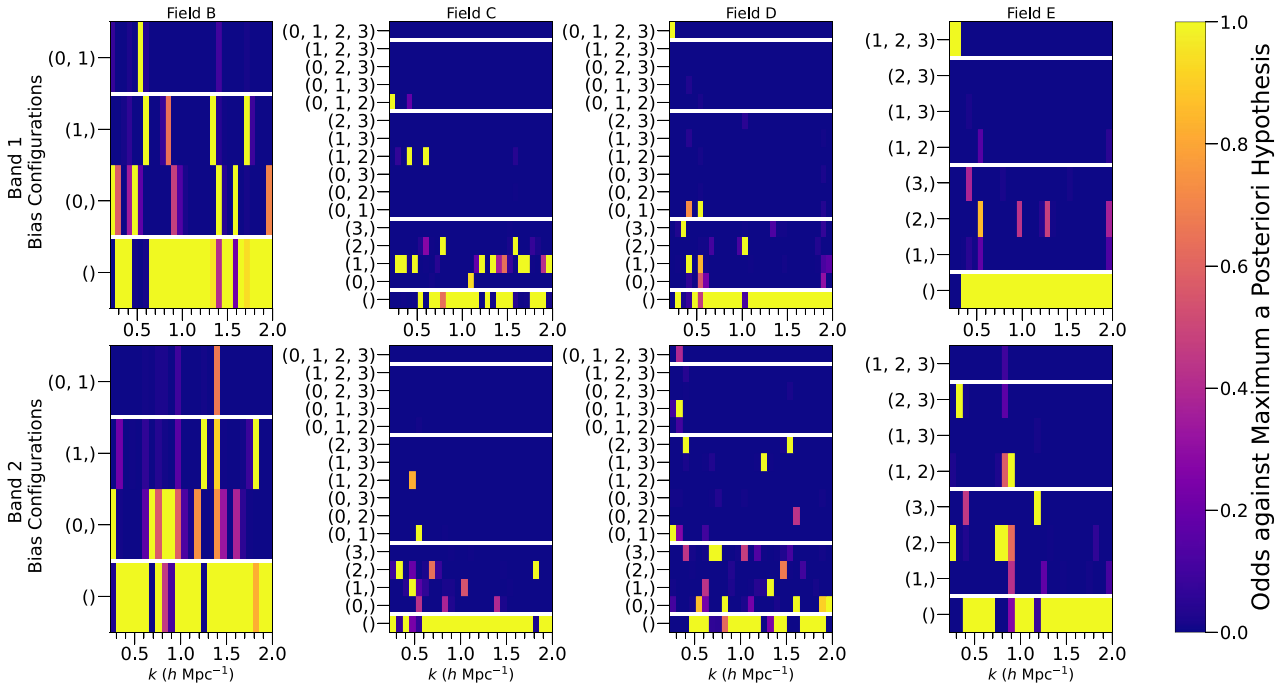

Figure 8. Odds against the MAP hypothesis for (positive) bias configurations (vertical axes – contents of the tuple indicate which epochs are biased under the hypothesis) in the measured bandpowers for each observed frequency band (panel rows), celestial field (panel columns), and spherical wave number (horizontal axes). The horizontal white lines delineate the borders between sections of the plot where certain numbers of epochs are biased. We observe that the null hypothesis is most likely in the majority of instances, suggesting that the data are unlikely to contain significant biases in the bandpowers for most bands, fields, and modes. Interestingly, epoch 1 is often identified as a biased bandpower in band 1, field C with few or no competing hypotheses on most modes. This can be visually confirmed in Fig. 10, where epoch 1 demonstrates a consistent positive bias across most k-modes.

violation of the null hypothesis, such as residual systematics that are incompletely subtracted. For example, biases at half the error bar have a true positive rate of about 13 per cent. With the exception of these potential nuances, we observe that the jackknife test performs more or less within expectation and proceed to applying it on the actual measured bandpowers.

4.2 Results with HERA data

Fig. 8 shows the posterior odds against the MAP hypothesis for each wave mode in the power spectra of the observed bands and fields in The HERA Collaboration (2022). There are two bands: one extending over 117.19–133.11 MHz and another extending over 152.25–167.97 MHz. The fields span roughly 17 h of right ascension, with two small gaps between fields A and B as well as B and C. Fields C, D, and E are contiguous. They are not all equally sized, both in terms of angular extent and total data volume contributing. Additionally, due to the large angular extent of the field collection, some epochs do not have measurements for some fields. This is reflected in the thermal uncertainties of the measured bandpowers. For this test, we use the same bias hypotheses as in the validation test, whose parameters are listed in Section 4.1. The null hypothesis, that the data are consistent with zero-mean noise, is the MAP solution in the majority of instances, suggesting that the data are unlikely to possess substantial positive biases[7] for most k-modes in the bands and fields that were tested. Furthermore, there is usually only a small number of competing hypotheses compared to the size of the configuration space.

This is summarized neatly by the entropy of the posteriors, which we show in Fig. 9. In most instances, the entropy is less than 1 bit. Interpreting 2^H as the effective number of competing hypotheses, having less than 1 bit of entropy suggests that there is usually less than 1 hypothesis with comparable probability to the MAP solution, though this may be manifest as multiple other hypotheses each with low-to-moderate probability. The relatively low entropy of the posteriors is largely a consequence of choosing alternative hypotheses that are highly distinguishable by the data, cf. Figs 2 and 4.

Fig. 8 also allows one to easily see whether a particular epoch is often problematic, such as epoch 1 in band 1, field C. Since the full posterior is available (normalized against the MAP), we can also read off trends in the competitive hypotheses. For instance in band 1, field B, the odds of a scenario where only epoch 1 is biased and the null hypothesis is close to 1 in a pair of high k-modes, and epoch 1 is highlighted for two other modes in the same panel. This suggests that if we were to combine information over k-modes using the formalism in Appendix B, then we may observe fewer null results.

In Fig. 10, we show the measured bandpowers for each band and field, along with 1σ error bars. Opaque data were those identified as biased according to the MAP solution, while transparent data are classified as unbiased. Comparing to Fig. 8, we can visually confirm that epoch 1 consistently measures bandpowers that are several error widths above 0. Given that the false positive rate with these settings is roughly 6 per cent, we are confident that most of these identified biases are true positives. Most EoR models produce power spectrum signals that are orders of magnitude beneath the thermal noise on these modes, meaning that these biases are probably residual

[7]We also searched for negative biases, for which the algorithm reports a small handful of negative outliers among all the data.

Downloaded from https://academic.oup.com/mnras/article/518/4/6041/6852951 by guest on 22 April 2024

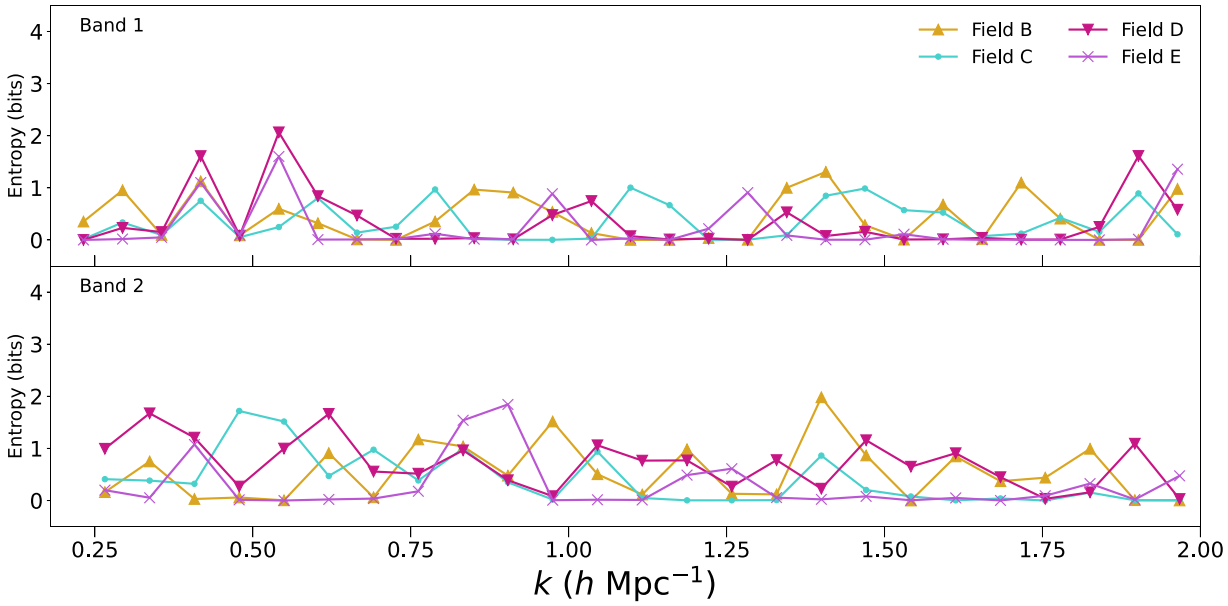

Figure 9. Entropy of the posterior for each band and field as a function of wave number. The points of low entropy correspond to posteriors with very few competing hypotheses, i.e. very certain inferences. Higher entropy states, such as band 2, field B, $k = 1.40\,h\,\mathrm{Mpc}^{-1}$, indicate uncertain inferences. As observable in Fig. 10, this particular mode has two positive-valued bandpowers slightly more than $2\sigma_i$ away from 0 – dubious evidence of a potential bias. Ultimately the test reasons it is unbiased under the chosen test parameters.

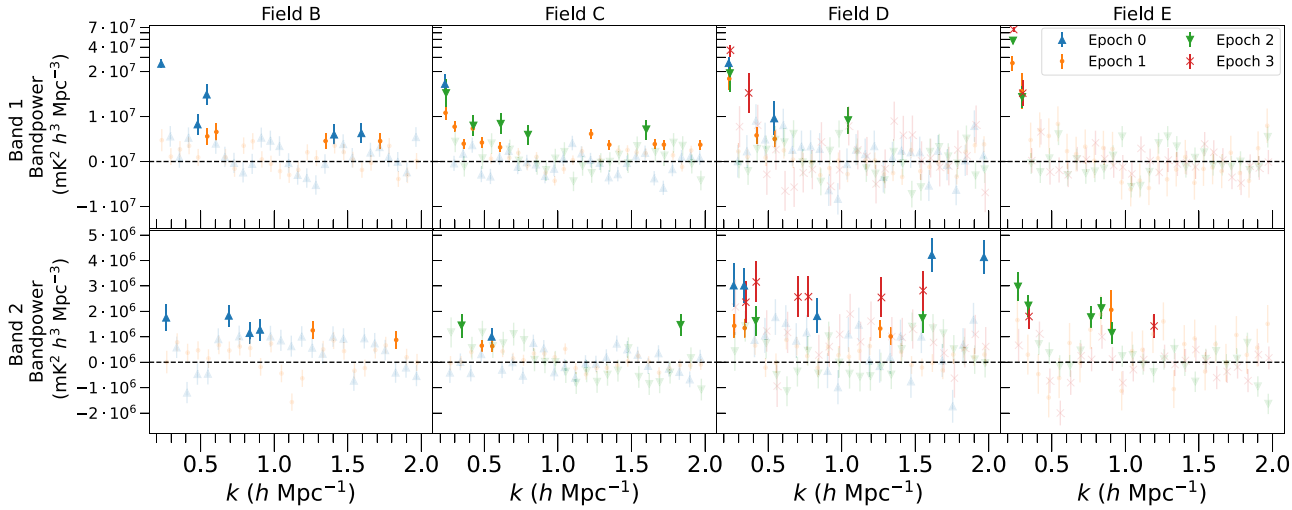

Figure 10. Measured bandpowers and 1σ error bars from The HERA Collaboration (2022). In the top row, we have compressed the upper range of the plot using a logarithmic scale above 2×10^7 in order to show some extreme outliers without obscuring the clarity at lower power. Data identified as positively biased according to the MAP decision rule are opaque, while data classified as unbiased are transparent. Points from different epochs are slightly staggered for visual clarity. Note that the two lowest limits belong to band 1, field D and band 2, field C, which are quite well-behaved under the jackknife test in general. The particular modes from which the lowest limits come ($k = 0.36\,h\,Mpc^{-1}$ and $k = 0.34\,h\,Mpc^{-1}$, respectively) do show strong evidence of biases in one epoch each. Note that in field C, we have removed epoch 3 from the plot for visual clarity, since it contained substantially less data in this particular field than other epochs, leading to large error bars that obscure the finer details in the plot.

systematic effects. This plot highlights the ability of the jackknife test to quantitatively identify likely outliers in circumstances where visual identification would be difficult. We also remark that since the data are considered jointly and under several different models, the test is more surgical and gives different results than a simple z-score cut. Expanding on this point, the Bayesian formalism makes the conclusion more readily interpreted than a decision based solely on a z-score, which directly translates to a p-value under the null hypothesis. A p-value under one hypothesis is non-specific to other models. This test provides more specific information about how the null hypothesis could be violated: it points the analyst in a particular direction when the null hypothesis is rejected. Moreover, it provides a specific criterion for when the null hypothesis should be rejected.

The decision to average the epochs together in each band and field in The HERA Collaboration (2022) was made prior to the application of this hypothesis test. Therefore, any justification that this test would

Downloaded from https://academic.oup.com/mnras/article/518/4/6041/6852951 by guest on 22 April 2024

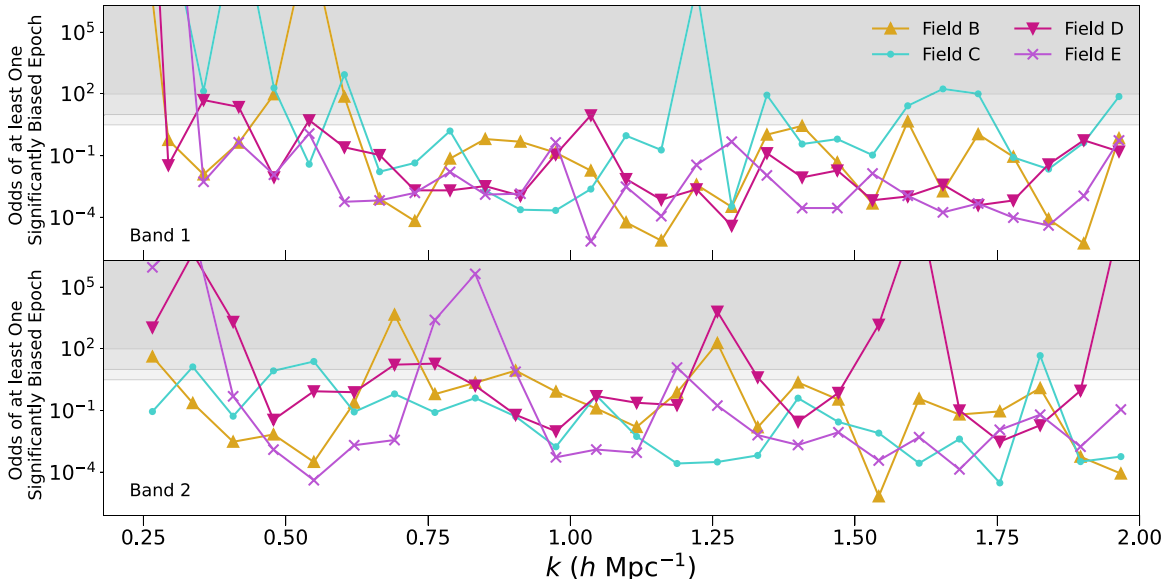

Figure 11. Odds of at least one significantly biased epoch in each band and field as a function of wave number (reproduced from The HERA Collaboration 2022). This is calculated by summing the posterior over all bias configurations and dividing by the posterior probability of the null hypothesis. Regions on the plot are shaded according to the odds interpretation presented in Kass & Raftery (1995). The majority of points show no decisive evidence of significant per-epoch biases (darkest shade, odds $>10^2$), except at lower k-modes and with a few exceptions at higher k. A modest number of points lie in the region of strong evidence for bias (middle shade, odds between 10^1 and 10^2), and a handful of points lie in the small lightly shaded region between $10^{1/2}$ and 10^1, which is a region of substantial but not strong evidence for bias. This majority null result in combination with the typically low entropies of the inferences essentially qualifies as a pass, except potentially for band 1, field C and band 2, field D.

lend to such a decision is post hoc, and serves strictly as validation of the decision. As with any decision, one may approach this in a variety of ways. For example, one might consider the power spectrum as a whole and make the decision based on a summary statistic that consistently responds for a large number of k-modes. Following this line of thinking, we construct the 'odds of at least one significantly biased epoch' per band, field, and mode by summing the posterior over all bias configurations and dividing by the posterior probability of the null hypothesis, displayed in Fig. 11. We might then decide a coherent average is unjustified if there is significant odds of a biased epoch on a large number of modes.[8] Inspecting Fig. 11, we see that for most bands and fields, and modes, there is counterevidence (odds less than 1) or weak evidence of a significantly biased epoch. Exceptions include the lower order k-modes, where we suspect some residual foreground contamination may be present, as well as band 1, field C and band 2, field D, which appear to have consistent evidence of bias on a large number of modes. In summary, we consider the result of this jackknife test as largely justifying the decision to coherently average the epochs together in The HERA Collaboration (2022).

5 SUMMARY AND CONCLUSIONS

We presented the formalism for a jackknife test for identifying biases from a small collection of subsets of data. The Bayesian formalism is flexible in that different questions may be asked by posing different scenarios. These scenarios, or hypotheses, are described by prior probability distributions on the possible biases. The test essentially proceeds by calculating the evidence of each hypothesis. These evidences are marginal likelihoods of the hypotheses, and thus using Bayes' theorem, we can combine them into a PPD over the hypotheses. A decision about whether a significant bias has been identified in a given scenario is made by specifying a loss function and minimizing its expected value over the posterior.

We show the explicit calculation of the marginal likelihoods when Gaussian priors on the biases and underlying mean under the null hypothesis are used. This is not a necessary feature of the framework, but is appropriate in many cases and permits extremely fast computation the posteriors. This allows for rapid exploration of the performance qualities of the jackknife test, such as true and false positive rates. A key concept in our understanding of the test is Shannon entropy, which acts as a metric of certainty for the test. Related concepts such as mutual information allowed us to make a good theoretical guess for what types of bias hypotheses would provide informative jackknife tests. We then validated this theory by simulating the classification problem and evaluating confusion matrices for the simulated data. In principle, these exploratory techniques are available for whichever bias scenarios one may wish to probe, however hypotheses for which analytic marginalization fails will take longer in order to perform the relevant exploration.

A particular hypothesis set that we spend a great deal of time exploring is one in which any of N data points may be biased by some unknown amount whose mean and (diagonal) covariance we specify. This produces 2^N hypotheses over which the posterior must be calculated. This is a fairly general hypothesis set that allows one to agnostically probe which data are most likely to be biased without specifying potential equivalencies between the biases. One may of course choose a subset of these hypotheses in order to sculpt a more specific question, or more generally, adjust the prior weights in order to favour some hypotheses over others. Furthermore, one may establish equivalence relations between the biases using the formalism that we present in Appendix C.

[8]One could eliminate this mode counting step and generate posterior odds for the whole spectrum by applying the formalism in Appendix B.

Downloaded from https://academic.oup.com/mnras/article/518/4/6041/6852951 by guest on 22 April 2024

As a proof of concept, we applied the jackknife method to a HERA data set and accompanying validation simulations used for a power spectrum upper limit in The HERA Collaboration (2022). The purpose of the test in this instance was to determine (post hoc) whether coherently combining subsets of the limit data was statistically justified. The basic idea is that if a particular subset of data consistently showed evidence of bias, then it would have been prudent to exclude those data from the coherent average since this would indicate that it had high levels of systematic contamination. We found that in most instances there was little evidence of consistent bias for any given epoch except in a minority of band/observing field combinations. We therefore conclude that the choice to coherently average the data together was indeed justified, except potentially at lower k-modes. In future applications to the HERA experiment, we aim to implement the multi-k-mode formalism in Appendix B, as well as apply the test in earlier stages of the pipeline such as the LST-averaged visibilities. We suspect that this jackknife test will be a valuable tool for assessing statistical tension in other cosmological data sets as well, for instance between different seasons or detector sets in CMB experiments, different dishes/baselines of 21 cm intensity mapping experiments, and so on. Analytic and numerical versions of the test have been made available to the community in the CHIBORG software.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under grants #1636646 and #1836019 and institutional support from the HERA collaboration partners. This research is funded in part by the Gordon and Betty Moore Foundation through Grant GBMF5212 to the Massachusetts Institute of Technology. HERA is hosted by the South African Radio Astronomy Observatory, which is a facility of the National Research Foundation, an agency of the Department of Science and Innovation. This result is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 948764; PB and MJW).

PB and FK acknowledge support from STFC Grant ST/T000341/1. JSD gratefully acknowledges the support of the NSF AAPF award #1701536. NK gratefully acknowledges support from the MIT Pappalardo fellowship. PK and MGS acknowledge support from the South African Radio Astronomy Observatory (SARAO; www.sarao.ac.za) and the National Research Foundation (Grant No. 84156). YQ would like to acknowledge the support from the High Performance Computing centers of the Scuola Normale Superiore (Italy), the Council for Scientific and Industrial Research (South Africa), the OzSTAR national facility (Australia), and XSEDE (U.S.) for computational resources. JM acknowledges computational resources and support on the supercomputer Cedar at Simon Fraser University, which is managed by Compute Canada and funded by the Canada Foundation for Innovation (CFI). Parts of this research were supported by the Australian Research Council Centre of Excellence for All Sky Astrophysics in 3 Dimensions (ASTRO 3D), through project number CE170100013. SH acknowledges support from STFC studentship 2277533 in project ST/T505985/1. JBM is supported by a Clay Fellowship at the Smithsonian Astrophysical Observatory. GB acknowledges funding from the INAF PRIN-SKA 2017 project 1.05.01.88.04 (FORECaST), support from the Ministero degli Affari Esteri e della Cooperazione Internazionale – Direzione Generale per la Promozione del Sistema Paese Progetto di Grande Rilevanza ZA18GR02, and the National Research Foundation of South Africa (Grant Number 113121) as part of the ISARP RADIOSKY2020 Joint Research Scheme, from the Royal Society and the Newton Fund under grant NA150184, and from the National Research Foundation of South Africa (grant No. 103424). EdLA acknowledges the funding support of the UKRI Science and Technology Facilities Council SKA grant. AL acknowledges support from the New Frontiers in Research Fund Exploration grant program, the Canadian Institute for Advanced Research (CIFAR) Azrieli Global Scholars program, a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant and a Discovery Launch Supplement, the Sloan Research Fellowship, and the William Dawson Scholarship at McGill.

DATA AVAILABILITY

The simulations used in Sections 2–3 can be reconstructed using example notebooks as part of the publicly available HERA_STATS repository.[9] The validation outputs are available in the H1C_IDR3_VALIDATION_NOTEBOOKS repository.[10] See The HERA Collaboration (2022) for data availability of the per-epoch bandpowers and associated statistical uncertainties.

REFERENCES

Abbott T. M. C. et al., 2022, Phys. Rev. D, 105, 023520
Ade P. A. R. et al., 2022, ApJ, 927, 77
Aguirre J. E. et al., 2022, ApJ, 924, 85
Blanchard A., Héloret J.-Y., Ilić S., Lamine B., Tutusaus I., 2022, preprint (arXiv:2205.05017)
Cover T. A., Thomas J. A., 2005, Elements of Information Theory. John Wiley & Sons, Ltd., Hoboken, NJ
Di Valentino E. et al., 2021, Class. Quantum Gravity, 38, 153001
Doux C. et al., 2021, MNRAS, 503, 2688
Freedman W. L., 2021, ApJ, 919, 16
Furlanetto S. R., Oh S. P., Briggs F. H., 2006, Phys. Rep., 433, 181
Gelman A., Meng X.-L., Stern H., 1996, Stat. Sin., 6, 733
Gelman A., Carlin J. B., Stern H. S., Rubin D. B., 2004, Bayesian Data Analysis, 2nd edn. Chapman and Hall/CRC, London
Gelman A., Hill J., Yajima M., 2012, J. Res. Educ. Eff., 5, 189
Handley W., Lemos P., 2019, Phys. Rev. D, 100, 043504
Harrison D., Sutton D., Carvalho P., Hobson M., 2015, MNRAS, 451, 2610
Isi M., Farr W. M., Chatziioannou K., 2022, Phys. Rev. D, 106, 024048
Kass R. E., Raftery A. E., 1995, J. Am. Stat. Assoc., 90, 773
Kern N. S. et al., 2020, ApJ, 888, 70
Knox L., Millea M., 2020, Phys. Rev. D, 101, 043533
Lemos P. et al., 2021, MNRAS, 505, 6179
Liddle A. R., 2007, MNRAS, 377, L74
Lin W., Ishak M., 2017, Phys. Rev. D, 96, 083532
Liu A., Shaw J. R., 2020, PASP, 132, 062001
Lucy L. B., 2016, A&A, 588, A19
MacKay D. J. C., 2002, Information Theory, Inference & Learning Algorithms. Cambridge Univ. Press, Cambridge
Meng X.-L., 1994, Ann. Stat., 22, 1142
Morales M. F., Wyithe J. S. B., 2010, ARA&A, 48, 127
Moran G. E., Cunningham J. P., Blei D. M., 2022, Bayesian Anal. Advance Publication, 1
Nicola A., Amara A., Refregier A., 2019, J. Cosmol. Astropart. Phys., 2019, 011
Planck Collaboration VII, 2020, A&A, 641, A7
Press W. H., 1997, in Unsolved Problems in Astrophysics. Princeton Univ. Press, Princeton, p. 49

[9] https://github.com/HERA-Team/hera_stats
[10] https://github.com/HERA-Team/h1c_idr3_validation_notebooks

Downloaded from https://academic.oup.com/mnras/article/518/4/6041/6852951 by guest on 22 April 2024

Raveri M., Hu W., 2019, Phys. Rev. D, 99, 043506
Riess A. G., Casertano S., Yuan W., Bowers J. B., Macri L., Zinn J. C., Scolnic D., 2021, ApJ, 908, L6
Rubin D. B., 1984, Ann. Stat., 12, 1151
Seehars S., Grandis S., Amara A., Refregier A., 2016, Phys. Rev. D, 93, 103507
Sloane N. J. A., OEIS Foundation Inc., 2022, The On-line Encyclopedia of Integer Sequences, available at https://oeis.org/A000110
The HERA Collaboration, 2022, preprint (arXiv:2210.04912)
Trotta R., 2007, MNRAS, 378, 72
Verde L., Protopapas P., Jimenez R., 2013, Phys. Dark Universe, 2, 166
Wilensky M. J., Barry N., Morales M. F., Hazelton B. J., Byrne R., 2020, MNRAS, 498, 265

APPENDIX A: MUTUAL INFORMATION AND ENTROPY

In this appendix, we provide a brief review of mutual information and entropy. For a thorough and accessible introduction to information theory, and particularly these concepts, see Cover & Thomas (2005) and MacKay (2002).

We define the Shannon entropy, S, of a discrete random variable, X, as

$$S(X) = -\sum_x P(x)\log_2(P(x)), \tag{A1}$$

where $P(x)$ is its probability mass function and the sum is over all states obtainable by the random variable. A similar definition exists for continuous random variables, and is referred to as 'differential entropy'. This is interpreted as a measure of uncertainty for the random variable, and is measured in bits if $\log_2$ is used. To understand this interpretation, consider a two-sided coin with probability p of returning a particular side (say 'heads') after a toss. The entropy of the coin is

$$S_2(p) \equiv -p\log_2(p) - (1-p)\log_2(1-p), \tag{A2}$$

which is sometimes referred to as the binary entropy function. This function is plotted in Fig. A1. It is maximal for $p = 1/2$, expressing the fact that the outcome of a fair coin is minimally predictable (or maximally uncertain) compared to any unfair one. In the limit as p goes to 0 or 1, the binary entropy function goes to 0, expressing that such a coin will always occupy a certain state after a toss.

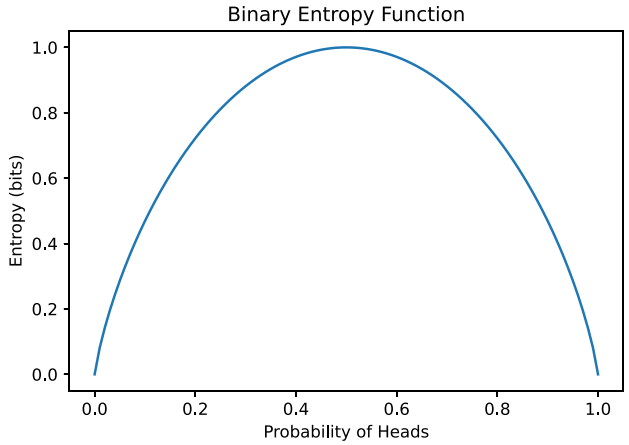

Figure A1. The binary entropy function for a coin toss of varying fairness. The entropy is maximal for a fair coin, and equal to 0 for a coin that always returns one of heads or tails when tossed. Coin flips with highly uncertain outcomes have a high entropy, and inversely, coin flips with highly certain outcomes have low entropy. In general, we can use the entropy as a measure of uncertainty of a random variable.

We define the conditional entropy of the random variable, X, conditioned on another random variable, Y, as

$$S(X|Y) \equiv -\sum_{x,y} P(x,y)\log_2(P(x|y)), \tag{A3}$$

where $P(x, y)$ is the joint probability mass function of X and Y, and $P(x|y)$ is the conditional distribution of X given Y. This should be interpreted as the average remaining uncertainty about X after considering a known value for Y. When X and Y are independent, this quantity is equal to $S(X)$, and when the state of Y completely specifies the state of X, this quantity is equal to 0. For our purposes, we can define the mutual information between X and Y as

$$\begin{aligned} I(X;Y) &\equiv S(X) - S(X|Y) \\ &= S(Y) - S(Y|X), \end{aligned} \tag{A4}$$

where the second equality indicates symmetry in the argument. In concordance with our previously stated interpretations, the mutual information is the average reduction in uncertainty about one random variable after gaining knowledge of the state of the other. If (and only if) X and Y are independent, they have 0 mutual information, and therefore knowledge about one cannot reduce uncertainty about the state of the other. The mutual information is always non-negative (Cover & Thomas 2005).

In an inference problem, we are concerned with how informative the data are for a given set of hypotheses. In other words, we want to know $I(\mathcal{H};\mathbf{d})$, where $\mathcal{H}$ is the set of hypotheses. Due to $P(\mathbf{d})$ being a mixture distribution, this problem prohibits analytic calculation of the mutual information. However, we can quickly estimate it using Monte Carlo methods using the second form of equation (A4). In this case, we have[11]

$$I(\mathcal{H};\mathbf{d}) = S(\mathbf{d}) - \sum_{i=0}^{Q-1} P(\mathcal{H}_i)S(\mathbf{d}|\mathcal{H}_i). \tag{A5}$$

Since the conditional distributions are all Gaussian with known mean and covariance, their individual entropies can be calculated analytically. The left-hand term involves integrating the logarithm of a Gaussian mixture against that mixture. To estimate this term, we generate D data points from the mixture distribution $P(\mathbf{d}) = \sum_i P(\mathbf{d}|\mathcal{H}_i)P(\mathcal{H}_i)$, and then compute the sample entropy:

$$\hat{S}(\mathbf{d}) \equiv -\frac{1}{D}\sum_{j=1}^{D}\log_2\left(\sum_{i=0}^{Q-1} P(\mathbf{d}^{(j)}|\mathcal{H}_i)P(\mathcal{H}_i)\right), \tag{A6}$$

where $\mathbf{d}^{(j)}$ is the jth sample of the mixture. The mutual information between a hypothetical data set and the hypothesis may be estimated as

$$\hat{I}(\mathcal{H};\mathbf{d}) = \hat{S}(\mathbf{d}) - \sum_{i=0}^{Q-1}\frac{1}{2}\log_2(\det(2\pi e\mathbf{C}_i'))P(\mathcal{H}_i). \tag{A7}$$

APPENDIX B: EVALUATING THE POSTERIOR FOR VECTOR VALUED MEASUREMENTS

In Section 2.3, we showed how to evaluate the posterior probability for a situation in which multiple measurements were made of a single quantity, such as several measurements of a single wave mode in the 21 cm power spectrum. However, it is sometimes of interest

[11] For brevity, we are dropping $\mathbf{C}_0$ in this expression.

Downloaded from https://academic.oup.com/mnras/article/518/4/6041/6852951 by guest on 22 April 2024

to combine information about several measured quantities, such as examining several modes in the power spectrum simultaneously. This will be useful in HERA since systematics can be mildly delocalized (cable reflections; Kern et al. 2020) or totally delocalized (RFI; Wilensky et al. (2020) in the power spectrum. The general result of this consideration is that one takes the result from Section 2.3 and evaluates all elements of vector and matrix quantities blockwise. In more detail, let $\mathbf{d}_j$ be a column vector of length M representing the jth measurement of the M quantities (wave modes) in consideration. Then let[12]

$$\mathbf{d} = \begin{pmatrix} \mathbf{d_0} \\ \mathbf{d}_1 \\ \vdots \\ \mathbf{d}_{N-1} \end{pmatrix} \tag{B1}$$

and define

$$\boldsymbol{\mu} \equiv \begin{pmatrix} \mu_0 \\ \mu_1 \\ \vdots \\ \mu_{M-1} \end{pmatrix} \tag{B2}$$

to hold the expected values of the separate quantities (power in each wave mode) under the null hypothesis. We then pack $\boldsymbol{\varepsilon}$ in the same way as $\mathbf{d}$, and each element of $\mathbf{C}_0$ and $\mathbf{C}_{\boldsymbol{\varepsilon},i}$ is replaced by a matrix of shape $M \times M$. For the HERA example, each diagonal block represents the covariance between the different wave-modes within a given epoch, and the off-diagonal blocks describe correlations between both wave modes and different epochs.

Since the common mean is now vector valued, the final marginalization is over the vector, $\boldsymbol{\mu}$. This means that in the case of $\mathcal{H}_i$, the marginalization over the biases must also report something over that vector. The correct answer is to do the following replacement:

$$\mu_0 \mathbb{1} \rightarrow \mathbb{1} \otimes \boldsymbol{\mu}. \tag{B3}$$

The symbol, $\otimes$, represents the Kroneckcer product. This replacement takes $\boldsymbol{\mu}$ and concatenates N copies of it columnwise. We then call the result of the bias marginalization $f(\boldsymbol{\mu})$, which is now an improperly normalized multivariate Gaussian over $\boldsymbol{\mu}$. Proceeding as in the previous section, one may see that this Gaussian has covariance matrix, $\tilde{\mathbf{C}}_i$, equal to

$$\tilde{\mathbf{C}}_i^{-1} = \sum_{m=0}^{N-1} \sum_{n=0}^{N-1} (\mathbf{C}_0 + \mathbf{C}_{\boldsymbol{\varepsilon},i})^{-1}_{mn}, \tag{B4}$$

where each element of the sum is the mn'th block of shape $M \times M$ in the matrix $(\mathbf{C}_0 + \mathbf{C}_{\boldsymbol{\varepsilon},i})^{-1}$. This is really just the blockwise generalization of equation (10), where in that case the operation sums over scalar elements of the inverse covariance matrix, rather than its blocks. In other words, if one takes equation (10) and writes it in terms of elementwise addition rather than in matrix notation, one achieves the identical expression as above but with the meaning of the symbols slightly changed. We can package this using the Kronecker product to obtain

$$\tilde{\mathbf{C}}_i^{-1} = (\mathbb{1} \otimes \mathbf{I}_M)^T (\mathbf{C}_0 + \mathbf{C}_{\boldsymbol{\varepsilon},i})^{-1} (\mathbb{1} \otimes \mathbf{I}_M), \tag{B5}$$

[12]The exact packing order is inconsequential as long as it is consistent between quantites, but this is a natural packing order for the HERA problem since epochs are uncorrelated, meaning this packing produces a block-diagonal $\mathbf{C}_0$.

where $\mathbf{I}_M$ is the $M \times M$ identity matrix. Similarly, equation (12) generalizes to

$$\tilde{\boldsymbol{\mu}} = \tilde{\mathbf{C}}_i (\mathbb{1} \otimes \mathbf{I}_M)^T (\mathbf{C}_0 + \mathbf{C}_{\boldsymbol{\varepsilon},i})^{-1} (\mathbf{d} - \boldsymbol{\mu}_{\boldsymbol{\varepsilon},i}). \tag{B6}$$

Following all this though, equations (15) and (16) generalize to

$$\boldsymbol{\mu}_i' = \boldsymbol{\mu}_{\boldsymbol{\varepsilon},i} + \mathbb{1} \otimes \boldsymbol{\mu}. \tag{B7}$$

$$\begin{aligned} (\mathbf{C}_i')^{-1} = {} & (\mathbf{C}_0 + \mathbf{C}_{\boldsymbol{\varepsilon},i})^{-1} - (\mathbf{C}_0 + \mathbf{C}_{\boldsymbol{\varepsilon},i})^{-1} (\mathbb{1} \otimes \mathbf{I}_M) \\ & \times \left(\mathbf{C}_p^{-1} + \tilde{\mathbf{C}}_i^{-1} \right)^{-1} (\mathbb{1} \otimes \mathbf{I}_M)^T (\mathbf{C}_0 + \mathbf{C}_{\boldsymbol{\varepsilon},i})^{-1}, \end{aligned} \tag{B8}$$

where we have implicitly generalized the prior over the common mean to a multivariate Gaussian.

For interpretations of $\boldsymbol{\mu}$ where Gaussian priors are inappropriate, the final multivariate marginalization can present significantly more numerical difficulty if many quantities (modes) are considered simultaneously.

APPENDIX C: PARTITIONING A BIAS CONFIGURATION INTO CORRELATED AND UNCORRELATED SUBSETS

In most of this paper, we consider the hypothesis set in which any combination of subsets of data may be biased; each datum has one mean value it may take when biased, the bias priors have diagonal covariances, and only data biased under the hypothesis have non-zero covariance entries. Due to these last two points, we refer to this as the 'diagonal hypothesis set'. In certain applications, it may be beneficial to 'partition' each diagonal hypothesis into correlated and uncorrelated subsets. As we show, this generates a huge number of hypotheses, however it may be useful for investigating particular systematic effects that are expected to affect some data and not others.

As a simple example, we consider the $N = 2$ case with equal bias prior widths in all hypotheses. As pointed out in the main text, the null hypothesis corresponds to the bias covariance

$$\mathbf{C}_{\boldsymbol{\varepsilon},0} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}. \tag{C1}$$

Assuming equal width in the bias priors for the alternative hypotheses, we can write the rest of the diagonal entries down as

$$\{\mathbf{C}_{\boldsymbol{\varepsilon},i} : 1 \leq i \leq 3\} = \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right\}. \tag{C2}$$

This set of hypotheses provides no constraint on the relative equivalence between the biases. We may consider a fifth hypothesis with covariance, $\mathbf{C}_{\boldsymbol{\varepsilon},4}$, defined by

$$\mathbf{C}_{\boldsymbol{\varepsilon},4} \equiv \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}. \tag{C3}$$

Since this matrix is degenerate, this hypothesis establishes an exact equivalence relation between the two bias parameters of the form

$$\varepsilon_1 = \varepsilon_0 + b, \tag{C4}$$

where b is a constant in terms of the bias prior means for those data. A slope can be added using different prior widths for the data, and the equivalence can be softened to a correlation by choosing a correlation coefficient less than 1, thus removing the degeneracy. These are all the hypotheses for the $N = 2$ case, since no other hypotheses in this set may be further partitioned.

Let us now take a brief foray into the $N = 3$ case before generalizing to arbitrary N. When all three diagonal elements are 'on', we have a

Downloaded from https://academic.oup.com/mnras/article/518/4/6041/6852951 by guest on 22 April 2024

set of five matrices to consider (let us denote this set, R):

$$R = \left\{ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}, \times \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \right\}. \tag{C5}$$

These represent the five possible equivalence relations between three numbers, where the diagonal matrix reflects that they are all independent, the middle three matrices represent scenarios where two have an exact equivalence, while the third is unrelated, and the final matrix represents a scenario in which all are exactly related. Notice however that we must also partition all $\binom{3}{2}$ matrices in which only two diagonal elements are 'on', which is described by the previous example (two matrices for each pair – one with off-diagonals 'on' and one with them 'off' i.e. the diagonal case). Of course, we must also include the remaining $\binom{3}{1}$ hypotheses where only one element is 'on', and the null hypothesis in which none are on.

In other words, the total number of hypotheses for the Nth case, denoted B_{N+1}, is such that

$$B_{N+1} = \sum_{k=0}^{N} \binom{N}{k} B_k \tag{C6}$$

$$B_0 = 1, \tag{C7}$$

where this last relation reflects that there is only one null hypothesis.

This turns out to define a set of well-studied numbers known as Bell numbers (Sloane & OEIS Foundation Inc. 2022). The Nth bell number describes how many ways there are to partition a set of N elements, or equally well, the total number of equivalence relations on a set of such a size. They also arise in the moments of the Poisson distribution. For our Nth case, we must sum over all sub-partitions as well as the final case where all diagonal elements are on, meaning we have B_{N+1} hypotheses in consideration for a jackknife with N data points. The Bell numbers increase extremely rapidly. The first few are 1, 1, 2, 5, 15, 52, 203, 877, 4140, ...

Due to this extremely rapid increase, we do not recommend an exhaustive search over all of the hypotheses. Rather, we recommend surgically choosing a subset of them where equivalence relations are expected. For instance, if two data are expected to be affected by a similar systematic that is not present in other data within the jackknife, then one might include the partition in which this equivalence relation between only these two is posed (e.g. different experimental methods of constraining the Hubble parameter).

[1]*Jodrell Bank Centre for Astrophysics, University of Manchester, Manchester M13 9PL, UK*
[2]*Astronomy Unit, Queen Mary University of London, London E1 4NS, UK*
[3]*Department of Physics and Astronomy, University of Western Cape, Cape Town 7535, South Africa*
[4]*Department of Astronomy, University of California, Berkeley, CA 94720-3411, USA*
[5]*South African Radio Astronomy Observatory, Black River Park, 2 Fir Street, Observatory, Cape Town 7925, South Africa*
[6]*Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, PA 19104-6396, USA*
[7]*Cavendish Astrophysics, University of Cambridge, Cambridge Cavendish Astrophysics: CB3 OHE, UK*
[8]*School of Earth and Space Exploration, Arizona State University, Tempe, AZ 85287-1404, USA*
[9]*Department of Physics, Winona State University, Winona, MN 55987, USA*
[10]*INAF-Istituto di Radioastronomia, via Gobetti 101, I-40129 Bologna, Italy*
[11]*Department of Physics and Electronics, Rhodes University, PO Box 94, Grahamstown 6140, South Africa*
[12]*National Radio Astronomy Observatory, Charlottesville, VA 22903, USA*
[13]*Department of Physics, Brown University, Providence, RI 02096, USA*
[14]*National Radio Astronomy Observatory, Socorro, NM 87801, USA*
[15]*Radio Astronomy Lab, University of California, Berkeley, CA 94720-3411, USA*
[16]*Department of Physics, University of California, Berkeley, CA 94720-3411, USA*
[17]*Department of Physics and Astronomy, University of California, Los Angeles, CA 90095-1547, USA*
[18]*National Radio Astronomy Observatory, Socorro, NM 87801-0387, USA*
[19]*School of Physics, University of Melbourne, Parkville, VIC 3010, Australia*
[20]*Department of Physics, University of Washington, Seattle, WA 98195-1560, USA*
[21]*eScience Institute, University of Washington, Seattle, WA 98195-1560, USA*
[22]*MIT Kavli Institute, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*
[23]*Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02142, USA*
[24]*Department of Physics and McGill Space Institute, McGill University, 3600 University Street, Montreal, QC H3A 2T8, Canada*
[25]*Scuola Normale Superiore, I-56126 Pisa, Italy*
[26]*Commonwealth Scientific and Industrial Research Organisation (CSIRO), Space & Astronomy, P. O. Box 1130, Bentley, WA 6102, Australia*
[27]*Center for Astrophysics, Harvard & Smithsonian, Cambridge, MA 02138, USA*
[28]*American Astronomical Society, Washington, DC20006, USA*

This paper has been typeset from a TEX/LATEX file prepared by the author.

Downloaded from https://academic.oup.com/mnras/article/518/4/6041/6852951 by guest on 22 April 2024