Evaluating Summary Statistics with Mutual Information for Cosmological Inference

Ce Sui [1] **Xiaosheng Zhao** [1] **Tao Jing** [1] **Yi Mao** [1]

Abstract

The ability to compress observational data and accurately estimate physical parameters relies heavily on informative summary statistics. In this paper, we introduce the use of mutual information (MI) as a means of evaluating the quality of summary statistics in inference tasks. MI can assess the sufficiency of summaries, and provide a quantitative basis for comparison. We propose to estimate MI using the Barber-Agakov lower bound and normalizing flow based variational distributions. To demonstrate the effectiveness of our method, we compare three different summary statistics (namely the power spectrum, bispectrum, and scattering transform) in the context of inferring reionization parameters from mock images of 21 cm observations with Square Kilometre Array. We find that this approach is able to correctly assess the informativeness of different summary statistics and allows us to select the optimal set of statistics for inference tasks.

1. Introduction

Statistical inferences in cosmology consist of two parts. Firstly, summary statistics are selected to extract relevant information from raw observed data. These statistics are then used to infer parameters for a given physical model. The choice of summary statistics is crucial for obtaining better constraints on these physical parameters. Many summary statistics have been proposed to extract information from various types of astronomical datasets. For Cosmic Microwave Background (CMB) studies, analyses have largely focused on the power spectrum (Planck Collaboration et al., 2016). However, other cosmological fields, such as the 21 cm signal, are expected to be highly non-Gaussian, necessitating more informative summary statistics. As a result, new summary statistics have been proposed in 21 cm cosmology, such as the bispectrum (Yoshiura et al., 2015; Shimabukuro et al., 2016) and Minkowski Functionals (Gleser et al., 2006; Kapahtia et al., 2021). Furthermore, neural networks are being explored for learning summaries from input images in the context of cosmological inference (Zhao et al., 2022; Prelogović et al., 2022; Charnock et al., 2018).

An important task is to predict and compare the effectiveness of different summary statistics in constraining target physical parameters. There is no universal way to do that in a given inference task, as summary statistics are often derived from different frameworks. Traditional approaches for comparing new statistics often involve posterior comparisons and Fisher analysis (Watkinson et al., 2022; Shimabukuro et al., 2017; Zhao et al., 2022). These methods typically consider one set of fiducial parameters. The former approach relies on inference results from mock observations and assesses the effectiveness of different statistics by examining their posteriors. The latter approach often uses Gaussian assumptions and calculates the Fisher information at the fiducial parameter. However, predictions obtained from these methods only consider a single point in the parameter space, which might not fully capture the overall performance of the statistics. One way to evaluate statistics across a large parameter space is through regression performance. This involves assessing the optimal performance that the statistics can achieve in predicting parameters across the entire parameter range (Zhao et al., 2022; Prelogović et al., 2022). However, this approach only considers point estimates and does not provide a comprehensive analysis of statistical performance.

A similar problem is extensively studied in the machine learning community in the context of representation learning. The goal of representation learning is to find a low-dimensional representation that preserves most of the information from the original data. It has been shown that such task can be framed as a problem of maximizing the mutual information (MI) between the learned summaries and the original data (Devon Hjelm et al., 2018; van den Oord et al., 2018). This suggests that we may use a similar metric to quantify the effectiveness of different summary statistics.

[1]Department of Astronomy, Tsinghua University, Beijing 100084, China. Correspondence to: Ce Sui <suic20@mails.tsinghua.edu.cn>.

ICML 2023 Workshop on Machine Learning for Astrophysics, Honolulu, Hawaii, USA. PMLR 202, 2023.

In this study, we introduce a novel approach to compare different summary statistics by estimating the mutual information between the statistics and target parameters in a given inference task. This method enables a quantitative comparison of the effectiveness of various statistics by considering their statistical dependence on parameters. Unlike many previous methods that focus on a single point in the parameter space, our approach takes into account the entire parameter range, providing a comprehensive evaluation of how well the summary statistics capture the necessary information for inference.

2. Method

In statistical inference we seek to estimate the parameters θ of a physical model given some input observation x. In Bayesian inference, this requires estimating the posterior distribution $p(\theta|x)$. However, the original observation is typically high-dimensional and contains a substantial amount of irrelevant information. To address this, we often utilize summary statistics to compress the data into more compact representations s that contain the most critical information about the parameter.

To achieve optimal performance, summary statistics should capture all the relevant information contained in the original observation x regarding the parameter of interest. This leads to the definition of sufficient statistics as those that satisfy the condition

$$p(\theta|x, s) = p(\theta|s), \tag{1}$$

which implies that once the summary statistics are given, the original data does not provide additional information. If a summary statistic closely resembles the sufficient statistic, it is likely to provide more accurate and reliable information about the underlying parameter.

2.1. Mutual information as a probe of Sufficient Statistics

Mutual information is a fundamental concept in statistical inference and information theory that measures the amount of information one random variable contains about another. Mutual information is defined as

$$\begin{aligned} \mathrm{I}(x; y) &= KL(p(x, y)||p(x)p(y)) \\ &= \mathbb{E}_{p(x,y)}\left[\log \frac{p(x, y)}{p(x)p(y)}\right], \end{aligned} \tag{2}$$

where x and y are random variables, and $p(x, y)$, $p(x)$, and $p(y)$ are their joint and marginal probability distributions, respectively. Mutual information quantifies the difference between the joint distribution and the product of the marginal distributions using the Kullback-Leibler divergence (KLD). If x and y are independent, their mutual information is zero, indicating that knowing x does not provide any information about y. Conversely, if x contains information about y, mutual information is non-zero, and a higher value indicates a stronger dependence between the two variables.

Mutual information can also be used to define sufficient statistics. For Bayesian inference, we can show that Equation 1 is equivalent to

$$I(\theta; x) = I(\theta; s(x)), \tag{3}$$

which implies that sufficient statistics s contain all the information about θ in the original observation x. Thus, by measuring the mutual information between a summary statistic and the target parameters, we can evaluate the sufficiency of the statistic. This interpretation provides a powerful tool for selecting summary statistics and assessing their suitability for use in statistical inference.

2.2. Mutual information estimation

Estimating mutual information is a challenging task in practice as it requires the computation of the KLD between complex distributions, which are often unknown. For many cosmological inference tasks we do not have access to a tractable joint distribution of the physical parameters and the data summaries, $p(\theta, x)$. Hence, it is infeasible to directly evaluate the mutual information between them. However, we can use variational distributions to approximate the real distributions, and obtain variational bounds of mutual information.

Assuming that we have a variational distribution $q(\theta|s)$, we can utilize it to replace the actual conditional distribution $p(\theta|s)$ and prove that it generates a lower bound on MI. Specifically, we have

$$\begin{aligned} I(\theta; s) =& \mathbb{E}_{p(\theta,s)}\left[\log \frac{p(\theta|s)}{p(\theta)}\right] \\ =& \mathbb{E}_{p(\theta,s)}\left[\log \frac{q(\theta|s)p(\theta|s)}{p(\theta)q(\theta|s)}\right] \\ =& \mathbb{E}_{p(\theta,s)}\left[\log \frac{q(\theta|s)}{p(\theta)}\right] \\ &+ \mathbb{E}_{p(s)}[KL(p(\theta|s)\|q(\theta|s))] \\ \geq& \mathbb{E}_{p(\theta,s)}[\log q(\theta|s)] + h(\theta), \end{aligned} \tag{4}$$

where $h(\theta) = -\mathbb{E}_{p(\theta)}[\log p(\theta)]$ is the differential entropy of θ and the last inequality is due to the non-negativity of the KLD. This is often referred as the Barber-Agakov lower bound (Barber & Agakov, 2004; Poole et al., 2019). It is evident from this equation that the replacement of the real conditional distribution with a variational model results in a lower bound on MI, which is only tight when $p(\theta|s) = q(\theta|s)$. In cosmological inference problems, calculating the differential entropy of the prior distribution of θ is relatively straightforward as it is typically tractable.

For the first component, we can fit a highly flexible generative model to a large number of parameter-statistic pairs, yielding a variational distribution that approximates the actual distribution closely. In this work, we use a masked autoregressive flow (MAF; Papamakarios et al., 2017) as the variational distribution. We optimize MAFs by minimizing the negative log-probability, which is equivalent to maximizing the lower bound. In principle, alternative loss functions (e.g. Zeghal et al., 2022) can also be employed.

2.3. Relation with other methods

We can also consider other methods commonly used for comparing data summaries as processes of MI estimation. For instance, we can write MI as

$$I(\theta; s) = \mathbb{E}_{p(s)}[KL(p(\theta|s)\|p(\theta))]. \tag{5}$$

Using this expression, we can estimate MI by assuming $p(\theta|s) = \hat{p}(\theta|s_0)$, where we replace the true conditional distribution with a posterior distribution that we estimate at a particular observation s_0. In this case, the estimated MI becomes $\hat{I}(\theta; s) = KL(\hat{p}(\theta|s_0)\|p(\theta))$, which measures the difference between the estimated posterior and the prior. This is similar to comparing different statistics by examining their posteriors on one mock observation. However, this estimation has high variance since it uses only one estimated posterior distribution to represent the conditional distribution.

Regression performance can also be utilized to formulate a Barber-Agakov lower bound. In one dimensional case if we let $q(\theta|s) = \mathcal{N}\left(f(s), \mathbb{E}_{p(\theta,s)}\left[(\theta - f(s))^2\right]\right)$ in equation 4, where $f(s)$ is a estimator trained to predict θ from given summaries statistics s, then:

$$\begin{aligned} I(\theta; s) \geq & \mathbb{E}_{p(\theta,s)}[\log q(\theta|s)] + h(\theta) \\ = & h(\theta) - 1/2 \log \mathbb{E}_{p(\theta,s)}\left[(\theta - f(s))^2\right] \\ & - 1/2 \log(2\pi e), \end{aligned} \tag{6}$$

where a smaller mean square error produces a larger MI lower bound. This implies that summary statistics that can more accurately predict the parameters in a regression task may contain more information about those parameters. Note that if we further assume the estimator is unbiased and efficient, we can also derive a similar relation between MI and Fisher information (Brunel & Nadal, 1998). In our experiments, we train Light Gradient Boosting Machines (LGBM; Ke et al., 2017) to predict parameters directly from data summaries. We use the regression performance to validate the MI estimation from MAFs.

3. Data

As a demonstration of the this method, we consider an inference problem in 21 cm cosmology, where we need to constrain two Reionization parameters based on mock SKA (Square Kilometre Array) images. The 21 cm lightcones are simulated using the publicly available code `21cmFAST`[1] (Mesinger & Furlanetto, 2007; Mesinger et al., 2011). The simulations were performed on a cubic box of 100 comoving Mpc on each side, with 66^3 grid cells. In this work, we use coeval boxes at redshift 11.76. We consider the following reionization parameters in simulations and inferences:(1) ζ, the ionizing efficiency. We vary ζ as $10 \leq \zeta \leq 250$. (2) $T_{\rm vir}$, the minimum virial temperature of halos that host ionizing sources. We vary this parameter as $4 \leq \log_{10}(T_{\rm vir}/\mathrm{K}) \leq 6$.

To include observational effects, signals with three different levels of signal contamination are considered here: (i) signal with $\mathbf{k}_\perp = 0$ mode removal (removal of the mean for each frequency slice); (ii) signal with the SKA thermal noise; (iii) signal with the SKA thermal noise + residual foreground after foreground removal with the singular value decomposition. The thermal noise is produced using the `Tools21cm`[2] (Giri et al., 2020) package by considering the SKA1-Low configuration. The foreground simulation is based on the GSM-building model in Zheng et al. (2017). We generate 10000 data cubes for each case with different reionization parameters and randomized observational effects.

In this work, we try to evaluate the effectiveness of three different summary statistics in the inference reionization parameters: power spectrum (PS), bispectrum (BS) and scattering transform (ST). For BS, we find that it may contain many uninformative features, which can cause overfitting issues. To address this problem, we conduct feature selection using LGBM by comparing feature importance for parameter estimation in regression tasks. Note that the MI estimates after feature selection are still lower bounds due to data processing inequality. The details of these summary statistics and the process of feature selection are given in Appendix A.

4. Results

Comparisons between different summary statistics: We present our MI estimation results for different data summaries and contamination levels in Figure 1. Our results indicate that the presence of observational effects can significantly reduce the amount of information available about reionization parameters in our mock images. As we increase the level of contamination, the estimated MI decreases accordingly. Furthermore, we find that in all three datasets used in our experiments, the ST is better at extracting physical information than correlation functions. This is an expected result since ST is designed to capture more spatial

[1] https://github.com/andreimesinger/21cmFAST

[2] https://github.com/sambit-giri/tools21cm

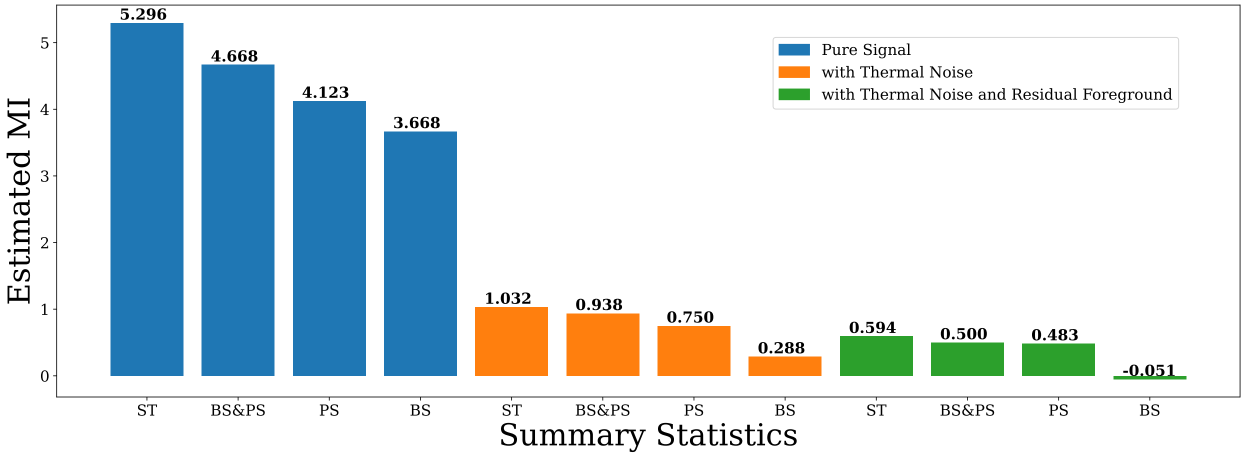

Figure 1. The estimated mutual information (MI) between target parameters and various summary statistics under three levels of contamination: (1) pure signal (blue bars) obtained by removing only the $\mathbf{k}_\perp = 0$ mode, (2) signal contaminated by thermal noise from the SKA telescope (orange bars), and (3) signal contaminated by both thermal noise and residual foreground (green bars). We estimated MI for four different statistics for each dataset: power spectrum (PS), bispectrum (BS), scattering transform (ST), and a combination of BS and PS (BS&PS).

information (Cheng et al., 2020). The performance of the correlation functions is consistent with previous studies (Watkinson et al., 2022; Shimabukuro et al., 2017), where PS and BS are evaluated in a similar inference task. We also notice that the MI estimation of BS in the highest contamination case is negative, while MI should be non-negative by definition. This is because we are estimating a lower bound and the variational distribution is not exact. This result indicates that BS in this case provides nearly no information about the reionization parameters.

Validation with regression tasks: Our results are largely consistent with theoretical interpretations and previous works. To further validate their correctness, we trained LGBM to predict reionization parameters and evaluated the regression performance for each dataset. As mentioned in Section 2.3, regression performance can also be used as an estimator for mutual information by selecting a specific variational distribution. We used the R^2 score to evaluate the regression performance of LGBM and present the results in Figure 2, where we also plotted the previous MI estimates. The two MI estimators are observed to be consistent with each other. However, we note that the R^2 score was unable to show the difference between summary statistics when mutual information was relatively high.

5. Summary

Our study presents a practical framework for evaluating the effectiveness of summary statistics in a specific inference problem. We accomplish this by estimating the mutual information between these statistics and the target parameters.

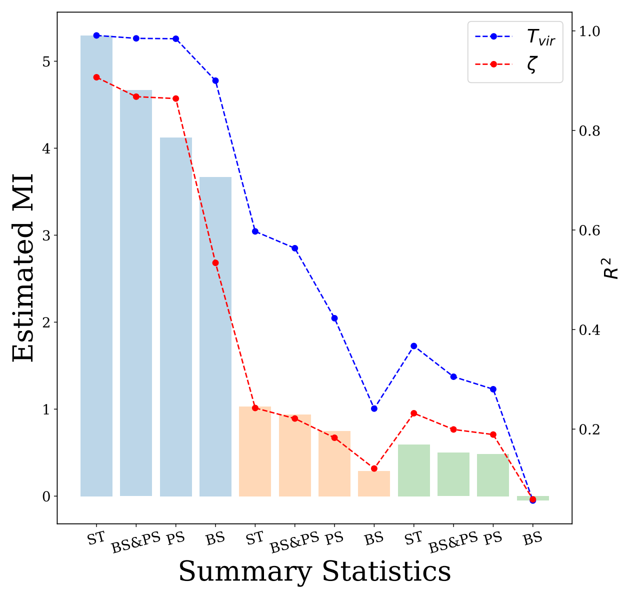

Figure 2. Comparison between the optimal R^2 achieved by different statistics and their MI estimates. The R^2 values(right y-axis) are presented as points overlaid on the MI results. The regression performance for two target parameters, T_{vir} and ζ, is indicated by the blue and red color, respectively.

Unlike existing approaches that rely on inference performance assuming fiducial parameters, our method provides a more robust assessment. We validate this methodology by applying it to the task of inferring reionization parameters from simulated SKA images. Our results demonstrate that the MI estimates agree with previous works and regression-based verification. This novel framework introduces a valuable tool for evaluating the informativeness of summary statistics in the field of cosmology.

Acknowledgements

This work is supported by the National SKA Program of China (grant No. 2020SKA0110401), NSFC (grant No. 11821303), and the National Key R&D Program of China (grant No. 2018YFA0404502). CS thanks Richard Grumitt for useful comments. We acknowledge the Tsinghua Astrophysics High-Performance Computing platform at Tsinghua University for providing computational and data storage resources that have contributed to the research results reported within this paper.

References

Andreux, M., Angles, T., Exarchakis, G., Leonarduzzi, R., Rochette, G., Thiry, L., Zarka, J., Mallat, S., Andén, J., Belilovsky, E., Bruna, J., Lostanlen, V., Hirn, M. J., Oyallon, E., Zhang, S., Cella, C., and Eickenberg, M. Kymatio: Scattering Transforms in Python. *arXiv e-prints*, art. arXiv:1812.11214, December 2018.

Barber, D. and Agakov, F. The im algorithm: a variational approach to information maximization. *Advances in neural information processing systems*, 16(320):201, 2004.

Brunel, N. and Nadal, J.-P. Mutual information, fisher information, and population coding. *Neural computation*, 10(7):1731–1757, 1998.

Charnock, T., Lavaux, G., and Wandelt, B. D. IMNN: Information Maximizing Neural Networks. Astrophysics Source Code Library, record ascl:1804.014, April 2018.

Cheng, S., Ting, Y.-S., Ménard, B., and Bruna, J. A new approach to observational cosmology using the scattering transform. *Monthly Notices of the Royal Astronomical Society*, 499(4):5902–5914, December 2020. doi: 10.1093/mnras/staa3165.

Devon Hjelm, R., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. Learning deep representations by mutual information estimation and maximization. *arXiv e-prints*, art. arXiv:1808.06670, August 2018. doi: 10.48550/arXiv.1808.06670.

Eickenberg, M., Exarchakis, G., Hirn, M., and Mallat, S. Solid harmonic wavelet scattering: Predicting quantum molecular energy from invariant descriptors of 3d electronic densities. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6543–6552, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Eickenberg, M., Exarchakis, G., Hirn, M., Mallat, S., and Thiry, L. Solid harmonic wavelet scattering for predictions of molecule properties. *The Journal of chemical physics*, 148(24):241732, 2018.

Giri, S. K., Mellema, G., and Jensen, H. Tools21cm: A python package to analyse the large-scale 21-cm signal from the epoch of reionization and cosmic dawn. *Journal of Open Source Software*, 5(52):2363, 2020. doi: 10.21105/joss.02363. URL https://doi.org/10.21105/joss.02363.

Gleser, L., Nusser, A., Ciardi, B., and Desjacques, V. The morphology of cosmological reionization by means of Minkowski functionals. *Monthly Notices of the Royal Astronomical Society*, 370(3):1329–1338, August 2006. doi: 10.1111/j.1365-2966.2006.10556.x.

Kapahtia, A., Chingangbam, P., Ghara, R., Appleby, S., and Choudhury, T. R. Prospects of constraining reionization model parameters using Minkowski tensors and Betti numbers. *Journal of Cosmology and Astroparticle Physics*, 2021(5):026, May 2021. doi: 10.1088/1475-7516/2021/05/026.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

Mesinger, A. and Furlanetto, S. Efficient Simulations of Early Structure Formation and Reionization. *The Astrophysical Journal*, 669(2):663–675, November 2007. doi: 10.1086/521806.

Mesinger, A., Furlanetto, S., and Cen, R. 21cmfast: a fast, seminumerical simulation of the high-redshift 21-cm signal. *Monthly Notices of the Royal Astronomical Society*, 411(2):955–972, 2011. doi: 10.1111/j.1365-2966.2010.17731.x. URL http://dx.doi.org/10.1111/j.1365-2966.2010.17731.x.

Papamakarios, G., Pavlakou, T., and Murray, I. Masked Autoregressive Flow for Density Estimation. *arXiv e-prints*, art. arXiv:1705.07057, May 2017. doi: 10.48550/arXiv.1705.07057.

Planck Collaboration, Ade, P. A. R., Aghanim, N., Arnaud, M., Ashdown, M., Aumont, J., Baccigalupi, C., Banday,

A. J., Barreiro, R. B., Bartlett, J. G., Bartolo, N., Battaner, E., Battye, R., Benabed, K., Benoît, A., Benoit-Lévy, A., Bernard, J. P., Bersanelli, M., Bielewicz, P., Bock, J. J., Bonaldi, A., Bonavera, L., Bond, J. R., Borrill, J., Bouchet, F. R., Boulanger, F., Bucher, M., Burigana, C., Butler, R. C., Calabrese, E., Cardoso, J. F., Catalano, A., Challinor, A., Chamballu, A., Chary, R. R., Chiang, H. C., Chluba, J., Christensen, P. R., Church, S., Clements, D. L., Colombi, S., Colombo, L. P. L., Combet, C., Coulais, A., Crill, B. P., Curto, A., Cuttaia, F., Danese, L., Davies, R. D., Davis, R. J., de Bernardis, P., de Rosa, A., de Zotti, G., Delabrouille, J., Désert, F. X., Di Valentino, E., Dickinson, C., Diego, J. M., Dolag, K., Dole, H., Donzelli, S., Doré, O., Douspis, M., Ducout, A., Dunkley, J., Dupac, X., Efstathiou, G., Elsner, F., Enßlin, T. A., Eriksen, H. K., Farhang, M., Fergusson, J., Finelli, F., Forni, O., Frailis, M., Fraisse, A. A., Franceschi, E., Frejsel, A., Galeotta, S., Galli, S., Ganga, K., Gauthier, C., Gerbino, M., Ghosh, T., Giard, M., Giraud-Héraud, Y., Giusarma, E., Gjerløw, E., González-Nuevo, J., Górski, K. M., Gratton, S., Gregorio, A., Gruppuso, A., Gudmundsson, J. E., Hamann, J., Hansen, F. K., Hanson, D., Harrison, D. L., Helou, G., Henrot-Versillé, S., Hernández-Monteagudo, C., Herranz, D., Hildebrandt, S. R., Hivon, E., Hobson, M., Holmes, W. A., Hornstrup, A., Hovest, W., Huang, Z., Huffenberger, K. M., Hurier, G., Jaffe, A. H., Jaffe, T. R., Jones, W. C., Juvela, M., Keihänen, E., Keskitalo, R., Kisner, T. S., Kneissl, R., Knoche, J., Knox, L., Kunz, M., Kurki-Suonio, H., Lagache, G., Lähteenmäki, A., Lamarre, J. M., Lasenby, A., Lattanzi, M., Lawrence, C. R., Leahy, J. P., Leonardi, R., Lesgourgues, J., Levrier, F., Lewis, A., Liguori, M., Lilje, P. B., Linden-Vørnle, M., López-Caniego, M., Lubin, P. M., Macías-Pérez, J. F., Maggio, G., Maino, D., Mandolesi, N., Mangilli, A., Marchini, A., Maris, M., Martin, P. G., Martinelli, M., Martínez-González, E., Masi, S., Matarrese, S., McGehee, P., Meinhold, P. R., Melchiorri, A., Melin, J. B., Mendes, L., Mennella, A., Migliaccio, M., Millea, M., Mitra, S., Miville-Deschênes, M. A., Moneti, A., Montier, L., Morgante, G., Mortlock, D., Moss, A., Munshi, D., Murphy, J. A., Naselsky, P., Nati, F., Natoli, P., Netterfield, C. B., Nørgaard-Nielsen, H. U., Noviello, F., Novikov, D., Novikov, I., Oxborrow, C. A., Paci, F., Pagano, L., Pajot, F., Paladini, R., Paoletti, D., Partridge, B., Pasian, F., Patanchon, G., Pearson, T. J., Perdereau, O., Perotto, L., Perrotta, F., Pettorino, V., Piacentini, F., Piat, M., Pierpaoli, E., Pietrobon, D., Plaszczynski, S., Pointecouteau, E., Polenta, G., Popa, L., Pratt, G. W., Prézeau, G., Prunet, S., Puget, J. L., Rachen, J. P., Reach, W. T., Rebolo, R., Reinecke, M., Remazeilles, M., Renault, C., Renzi, A., Ristorcelli, I., Rocha, G., Rosset, C., Rossetti, M., Roudier, G., Rouillé d'Orfeuil, B., Rowan-Robinson, M., Rubiño-Martín, J. A., Rusholme, B., Said, N., Salvatelli, V., Salvati, L., Sandri, M., Santos, D., Savelainen, M., Savini, G., Scott, D., Seiffert, M. D., Serra, P., Shellard, E. P. S., Spencer, L. D., Spinelli, M., Stolyarov, V., Stompor, R., Sudiwala, R., Sunyaev, R., Sutton, D., Suur-Uski, A. S., Sygnet, J. F., Tauber, J. A., Terenzi, L., Toffolatti, L., Tomasi, M., Tristram, M., Trombetti, T., Tucci, M., Tuovinen, J., Türler, M., Umana, G., Valenziano, L., Valiviita, J., Van Tent, F., Vielva, P., Villa, F., Wade, L. A., Wandelt, B. D., Wehus, I. K., White, M., White, S. D. M., Wilkinson, A., Yvon, D., Zacchei, A., and Zonca, A. Planck 2015 results. XIII. Cosmological parameters. *Astronomy and Astrophysics*, 594:A13, September 2016. doi: 10.1051/0004-6361/201525830.

Poole, B., Ozair, S., van den Oord, A., Alemi, A. A., and Tucker, G. On Variational Bounds of Mutual Information. *arXiv e-prints*, art. arXiv:1905.06922, May 2019. doi: 10.48550/arXiv.1905.06922.

Prelogović, D., Mesinger, A., Murray, S., Fiameni, G., and Gillet, N. Machine learning astrophysics from 21 cm lightcones: impact of network architectures and signal contamination. *Monthly Notices of the Royal Astronomical Society*, 509(3):3852–3867, January 2022. doi: 10.1093/mnras/stab3215.

Shimabukuro, H., Yoshiura, S., Takahashi, K., Yokoyama, S., and Ichiki, K. 21 cm line bispectrum as a method to probe cosmic dawn and epoch of reionization. *Monthly Notices of the Royal Astronomical Society*, 458(3):3003–3011, May 2016. doi: 10.1093/mnras/stw482.

Shimabukuro, H., Yoshiura, S., Takahashi, K., Yokoyama, S., and Ichiki, K. Constraining the epoch-of-reionization model parameters with the 21-cm bispectrum. *Monthly Notices of the Royal Astronomical Society*, 468(2):1542–1550, June 2017. doi: 10.1093/mnras/stx530.

van den Oord, A., Li, Y., and Vinyals, O. Representation Learning with Contrastive Predictive Coding. *arXiv e-prints*, art. arXiv:1807.03748, July 2018. doi: 10.48550/arXiv.1807.03748.

Villaescusa-Navarro, F. Pylians: Python libraries for the analysis of numerical simulations. Astrophysics Source Code Library, record ascl:1811.008, November 2018.

Watkinson, C. A., Greig, B., and Mesinger, A. Epoch of reionization parameter estimation with the 21-cm bispectrum. *Monthly Notices of the Royal Astronomical Society*, 510(3):3838–3848, March 2022. doi: 10.1093/mnras/stab3706.

Yoshiura, S., Shimabukuro, H., Takahashi, K., Momose, R., Nakanishi, H., and Imai, H. Sensitivity for 21 cm bispectrum from Epoch of Reionization. *Monthly Notices of the Royal Astronomical Society*, 451(1):266–274, July 2015. doi: 10.1093/mnras/stv855.

Zeghal, J., Lanusse, F., Boucaud, A., Remy, B., and Aubourg, E. Neural Posterior Estimation with Differentiable Simulator. In *Machine Learning for Astrophysics*, pp. 52, July 2022. doi: 10.48550/arXiv.2207.05636.

Zhao, X., Mao, Y., Cheng, C., and Wandelt, B. D. Simulation-based Inference of Reionization Parameters from 3D Tomographic 21 cm Light-cone Images. *The Astrophysical Journal*, 926(2):151, February 2022. doi: 10.3847/1538-4357/ac457d.

Zheng, H., Tegmark, M., Dillon, J. S., Kim, D. A., Liu, A., Neben, A. R., Jonas, J., Reich, P., and Reich, W. An improved model of diffuse galactic radio emission from 10 MHz to 5 THz. *Monthly Notices of the Royal Astronomical Society*, 464(3):3486–3497, January 2017. doi: 10.1093/mnras/stw2525.

A. Summary Statistics

In this work, we try to evaluate the effectiveness of three different summary statistics in the inference of reionization parameters: power spectrum, bispectrum and scattering transform.

Power spectrum (PS) $P_{21}(k)$ is the most commonly used statistic, defined as:

$$\left\langle\delta_{21}(\boldsymbol{k})\delta_{21}\left(\boldsymbol{k}'\right)\right\rangle = (2\pi)^3\delta^D\left(\boldsymbol{k}+\boldsymbol{k}'\right)P_{21}(\boldsymbol{k}), \tag{7}$$

where $\delta_{21}(k)$ is the Fourier transform of the 21 cm brightness temperature field, δ^D is the Dirac delta function and $\langle ... \rangle$ represents an ensemble average. We use the `Tools21cm` (Giri et al., 2020) package to calculate the spherical averaged power spectrum of the mock observation.

Bispectrum (BS) is the Fourier dual of three-point correlation function, defined as:

$$\left\langle\delta_{21}\left(\boldsymbol{k}_1\right)\delta_{21}\left(\boldsymbol{k}_2\right)\delta_{21}\left(\boldsymbol{k}_3\right)\right\rangle = (2\pi)^3\delta\left(\boldsymbol{k}_1+\boldsymbol{k}_2+\boldsymbol{k}_3\right)B\left(\boldsymbol{k}_1,\boldsymbol{k}_2,\boldsymbol{k}_3\right). \tag{8}$$

The BS is a function of three k vectors that form a closed triangle ($\boldsymbol{k}_1+\boldsymbol{k}_2+\boldsymbol{k}_3$=0). In this work, we consider only isosceles triangles and use the python package `Pylians`[3](Villaescusa-Navarro, 2018) to calculate the bispectrum of mock observation. We normalize BS following Watkinson et al. (2022) to separate non-Gaussian information. In our experiments we also consider a combination PS and BS by directly concatenating them and refer it as BS&PS. For BS, we observe that it may contain some uninformative features (here each feature represents the bispectrum value calculated for a specific triangle configuration), which can lead to overfitting issues. To tackle this problem, we employ feature selection using the LGBM. We first add random noise features to the bispectrum and conduct regression with LGBM. The regression process is repeated multiple times to calculate the mean and variance of the importance of the random noise features. Subsequently, we identify all features that fall within the 3-sigma range of the importance of the random features and exclude them from the fitting of MAFs.

The solid harmonic wavelet scattering transform (ST), first introduced by Eickenberg et al. (2017; 2018), is a method for compressing data for inference by convolving the original fields with a cascade of solid harmonic wavelets, performing non-linear modulus on the convolved fields, and integrating over all coordinates. It is an effective way to capture information at different scales and orientations, producing coefficients that are invariant to both translation and rotation. The ST is implemented with the `Kymatio` [4] (Andreux et al., 2018) package.

[3]https://github.com/franciscovillaescusa/Pylians3
[4]https://www.kymat.io/