Boletim da Sociedade Astronômica Brasileira, **34**, no. 1, 193-197


# Obtaining cosmological covariance matrices with machine learning

Natalí S. M. de Santi[1,2] & L. Raul Abramo[1]

[1] Instituto de Física, Universidade de São Paulo, R. do Matão 1371, 05508-900, São Paulo, Brazil
[2] Center for Computational Astrophysics, Flatiron Institute, 162 5th Avenue, New York, NY, 10010, USA
e-mail: natalidesanti@gmail.com.br and e-mail: abramo@fma.if.usp.br

**Abstract.** Parameter inference is one of the central topics of Cosmology, and the cosmological covariance matrices are fundamental for this task. Estimating accurate and precise covariance matrices requires lots of data – either observations or simulations, both of which may be very costly. In this work we propose a machine learning approach that starts from matrices which are built with only a small amount of data (as little as 50 power spectra), but is able to provide significantly improved covariance matrices, which are almost indistinguishable from the ones built from much larger samples (thousands of spectra). The methodology consists in training convolutional neural networks to denoise the covariance matrices using in the training process a data set made up entirely of spectra extracted from simple, inexpensive halo simulations (mocks). In order to validate the process we have compared the denoised matrices with covariance matrices built with thousands of spectra. We have used several different metrics to compare the matrices, and in particular we measured the ability of the denoised covariance matrices to recover the cosmological parameters. The denoised matrices scored significantly better in all the analyses, being comparable to the validation matrices (built from thousands of spectra).

**Resumo.** Um dos principais tópicos de estudos da Cosmologia reside na inferência de parâmetros comológicos e as matrizes de covariância são uma peça fundamental para sua obtenção. Matrizes de covariância precisas e acuradas são obtidas com o uso de uma enorme quantidade de dados (inúmeras observações ou custosas simulações), o que nem sempre é uma opção viável. Neste trabalho nós propomos o uso de um método de aprendizado de máquina capaz de transformar matrizes construídas com apenas uma pequena quantidade de dados (centenas de espectros) em matrizes significativamente melhoradas, quase indistinguíveis de matrizes feitas com amostras muito maiores (milhares de espectros). A metodologia consiste em treinar redes neurais convolucionais para remover o ruído de matrizes de covariância usando no treinamento um conjunto de dados de matrizes construídas totalmente com espectros de potência obtidos por meio de um gerador de catálogos de halos simples e barato. Para validar esse processo nós comparamos as matrizes de covariância sem ruído obtidas com matrizes feitas com um grande número de espectros. Foram utilizadas várias e diferentes métricas para medir a habilidade do método de recuperar os parêmetros cosmológicos. Por fim, as matrizes sem ruído obtidas apresentaram resultados significativamente melhorados, em todas as análises, comparadas as matrizes de validação.

**Keywords.** cosmological parameters – Methods: statistical – large-scale structure of Universe

## 1. Introduction

Parameter inference reside in one of the most important end products of cosmological data analysis. The cosmological covariance matrices appears in this context, connecting theory and data by quantify the amount for which we expect the measurements of some summary statistic to fluctuate, given the underlying physical phenomena and observational conditions (e.g., survey masks).

Analytical approaches can provide a good first approximation for the covariance matrix, however, both the physical models and the observational conditions are often best represented in terms of simulations. Then, the matrices are computed as the sample covariance, given a set of independent simulations and their summary statistic. In order to get these matrices both precise and accurate (Hartlap, Simon & Schneider 2007) to not bias the parameter estimation (Taylor, Joachimi & Kitching 2013) we need large samples, which require significant computational times, a task which is not always possible (Heavens et al. 2017).

In the case of the power spectrum the sample size that is typically required to fulfill those needs is around $N_s \sim n_k^2$, where $n_k$ is the number of $k$ bins (bandpowers). This number can grow even more with different tracers and all the resulting auto- and cross-spectra. Thus, it is of paramount importance to optimize methods that can estimate efficiently these matrices.

Several efforts have already been made with the goal of obtaining precise matrices using smaller samples (Taylor, Joachimi & Kitching 2013; Schneider et al. 2011). Other option is the use of approximate numerical methods (instead of using the expensive N-body simulations) as PINOCCHIO (Monaco et al. 2013), Lognormal (Agrawal et al. 2017), and ExSHalos (Voivodic, Lima & Abramo 2019). From a different perspective, machine learning (ML) techniques offer alternative solutions to some of these challenges. There are works trying to speed-up the process of producing high-resolution N-body simulations, by starting from lower-resolution ones and others where they try to em-

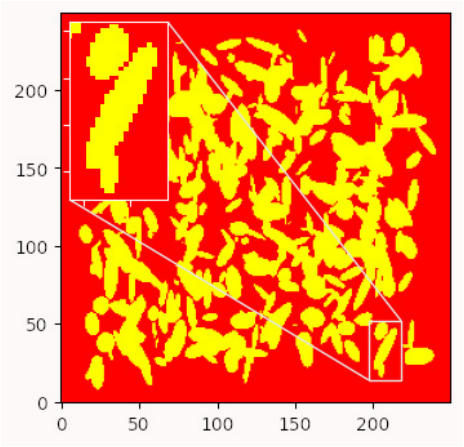


**Figure 1.** Slice of the mask of random ellipsoids. The inset shows a zoom in on three ellipsoids, for a better visualization.

ulate the full non-linear evolution of N-body simulations by inputting only approximate simulations of these (Li et al. 2021; Alves de Oliveira et al. 2020).

In the present paper we propose a new approach, that employs convolutional neural networks (CNN), more specifically image denoising techniques, as a tool to enhance sample cosmological covariance matrices. We show that the final covariance matrices (after denoising) become as precise and accurate as the ones obtained with a much higher number of simulations. This work is organized as follows: in section 2 comprehends the methodology; in section 3 we present the main results of this work; and, in section 4, we discuss the implications of our results and explain future applications.

## 2. Dataset and methodology

### 2.1. The halo catalogs

Excursion Set Halos (ExSHalos) is a new, fast, and parameter-free method to generate dark matter halo catalogs (Voivodic, Lima & Abramo 2019). The code implements the notion of excursion sets (Bond et al. 1991), and then it corrects the positions of the halo using Lagrangian Perturbation Theory (LPT) (Vlah, Seljak & Baldauf 2015). The required inputs are the cosmology, the linear matter power spectrum, and the threshold density for halo formation in linear theory (constant or ellipsoidal collapse barriers). For the ExSHalos mocks we used a linear power spectrum from the Code for Anisotropies in the Microwave Background (`CAMB`) (**?**). We have chosen the constant barrier and used LPT to second order. The cosmology was chosen according to Planck (Planck Collaboration 2020): $\Omega_m = 0.3175$, $\Omega_b = 0.049$, $h = 0.6711$, $n_s = 0.9624$, $\sigma_8 = 0.834$, $M_\nu = 0.0eV$ and $\omega = -1$. The size of the box was $L = 1000$ Mpc$/h$ and we used cubic cells of 1 Mpc/h at a fixed redshift $z = 0$. In total, we have produced 30,000 mock catalogs (hereafter, $N_{max} = 30,000$), but it should be noticed that not all the ML models used in this work needed this whole amount.

### 2.2. The power spectrum

We chose as tracers the halos with masses between $M \in [10^{13.12}, 10^{13.37}]M_\odot/h$, i.e., with a mean mass $\langle M \rangle = 10^{13.25} M_\odot/h$ and the power spectrum as our summary statistic. In this way, for each simulation, the data vector corresponds to a set of $P(k)$, and that is what we use to build the covariance matrices. To mimic real-life effects we have masked the halo maps in a way that attempts at emulating a mask covering some regions of the sky: to reflect a survey's footprint as well as bright stars, cloudy nights, regions with poor seeing, etc (Coupon et al. 2018). Apart from that, here, the mask induce non-trivial correlations between different spectral modes.

We built the mask using randomly placed ellipsoids with random sizes and orientations, in such a way that regions outside those ellipsoids were masked out. The mask occupies approximately $\sim 50\%$ of the total box volume, and we estimated the spectrum on a grid of the entire box, with cells of $4\,h^{-1}$ Mpc on a side. In Figure 1 we show a slice of the mask.

We computed the spectra of both the masked and unmasked catalogs, which are shown (in terms of the mean for 100 maps) in figure 2. The effect of the mask can be seen from this plot, in terms of a suppression of the clustering amplitude on all but the smallest scales. Notice that we have considered the halo bias $b$ for this halo mass bin (according to Tinker's definition (Tinker et al. 2010)), which was computed for 100 ExSHalos catalogs, and averaged over scales in the range $k \in [0.015, 0.2625]$ $h$ Mpc$^{-1}$.

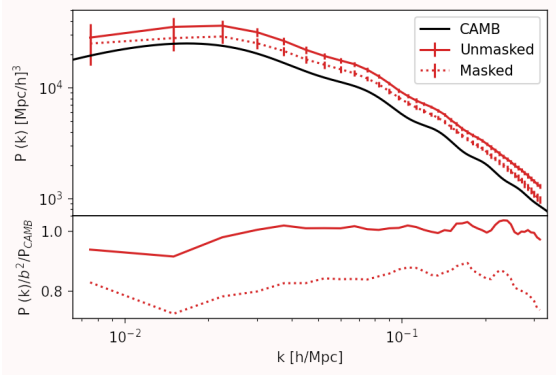


**Figure 2.** Power spectra of the masked (dotted) and unmasked (solid) halos. The linear matter power spectrum is also shown as the solid black line for comparison.

Then, the computed value is: $1.25 \pm 0.04$. We should stress that we treat the bias as a nuisance parameter in the cosmological parameter inference.

### 2.3. The data set of the covariance matrices

The cosmological covariance matrices were computed as:

$$Cov^{(N)}[P(k_i), P(k_j)] = \frac{1}{(N-1)} \sum_{l=1}^{N} \Big[P(k_i)_l - \bar{P}(k_i)\Big]\Big[P(k_j)_l - \bar{P}(k_j)\Big], \quad (1)$$

where $N$ is the number of spectra in the data vector (the sample size), $P(k_i)_l$ is the value of the $l$th spectra for the $i$th bin, and $\bar{P}(k_i)$ is the mean power spectrum.

We have used sets of samples of different sizes to train our ML denoiser. First, we constructed the *input* covariance matrices, which correspond to small sample sizes ($n$ spectra), the ones we would like to enhance with our denoiser - $n \in [50, 250]$, in increments of $\Delta n = 25$. Second, we computed covariance matrices with larger sample sizes ($N$ spectra, the *target* matrices, using $N = 1000$). Third, we compare the cleaned version of the input matrices with the *best* possible covariance matrix, which is computed using the maximum sample size ($N_{max}$ spectra).

We stress that we do not use this total number of spectra to train the ML suites: in each training we used 120 input matrices and 120 target matrices. E.g., in the cases of $n = \{50, 100, 200\}$ input spectra we used $\{6000, 12,000, 24,000\}$ spectra for the entire training process. The main limitation of our model was the computational cost associated with running the ExSHalos simulations.

### 2.4. The ML suite

The ML method that we have used in this work is an image denoising. In this problem, a noisy image $\mathbf{y}$ can be decomposed in an unknown image signal $\mathbf{x}$, and the noise $\nu$. The goal of the denoiser is to produce a cleaned version of the image, $\hat{\mathbf{x}}$, by reducing the noise while keeping its original properties and features while not creating new artifacts in the process (Tian et al. 2019).

The state-of-art of these methods are given by CNNs, or auto-encoders, mainly because they are purely data driven, with no assumption about the nature of the noise (Tian et al. 2019; Chollet 2017). They took a pair of images: a noisy image $\mathbf{y}$ (input) and a clear image $\mathbf{x}$ (target). Then, they learn to recognize what is signal, what is noise, and how to remove that noise, predicting images which are closer (less noisy) to the ones used as

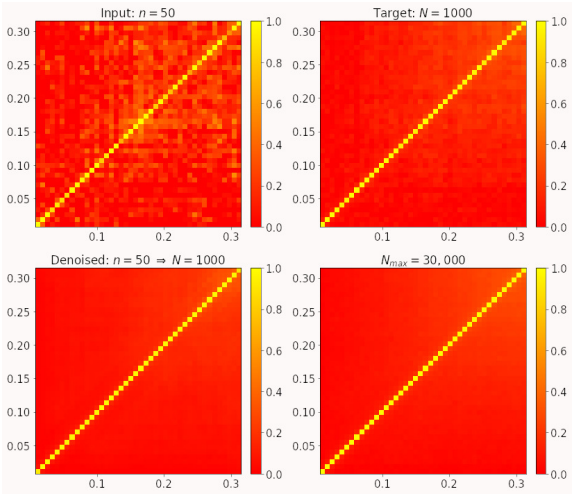


**Figure 3.** Comparison of the (normalized) cosmological covariance matrices. On the top left we have the input matrices with $n = 50$ spectra; on the top right, the target matrices, with $N = 1000$ spectra; on the bottom left we have the denoised matrix, corresponding to the model of: $(n = 50, N = 1000)$; and on the bottom right, the *best* matrix, with $N_{max}$ spectra. In each one of these figures, the axes corresponds to values of $k$, representing the 42 Fourier bins of the power spectrum.

target – i.e., as close as possible to the ground truth images **x**. In the context of covariance matrices the noisy images are the input matrices (built with samples of $n$ spectra) and the target images are the matrices built with samples of $N$ spectra.

A basic auto-encoder has two parts: an encoder, followed by a decoder, i.e., a sequence of convolutional layers that are responsible for extracting the features from the images, capturing the abstraction of their content, and then recovering the features at the end of the process. This is performed minimizing a loss function (Vincent et al. 2008).

In this work, we used different ML models to deal with each combination of input and target matrices with $(n, N)$ spectra. The size of the data set was composed with 120 matrices (240 in total, because each input matrix had its respective target). The entire method was implemented using the keras library (Chollet 2015).

In order to homogenize the entries of the covariance matrices in the training stage, we normalized the matrices using the diagonal of the matrix built with $N_{max}$ according to:

$$Cov_{ij}^{(N)} \rightarrow \frac{Cov_{ij}^{(N)}}{\sqrt{Cov_{ii}^{(N_{max})}, Cov_{jj}^{(N_{max})}}}. \qquad (2)$$

We plug back the normalization to recover the denoised covariance matrices at the end of the process and we impose the symmetry of the covariance matrices, $Cov_{ij} \rightarrow (Cov_{ij} + Cov_{ji})/2$.

## 3. Results

### 3.1. Visualizing the matrices

In figure 3 we have the visual inspection of the matrices, comparing the normalized matrices (according to equation 2): denoised $(n = 50 \Rightarrow N = 1000)$ against the target $(N = 1000)$ and the best-case scenario $(N_{max} = 30,000)$ spectra. It can be seen that the denoised matrices appear almost identical as the best ones $(N_{max})$ and are visually smooth and noiseless.

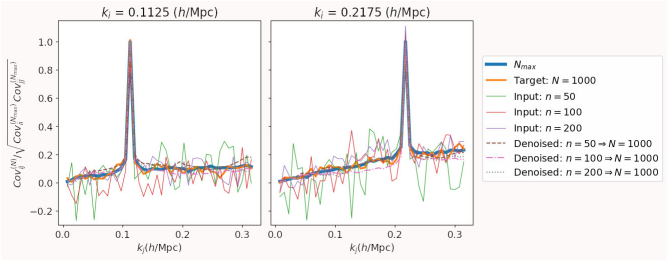


**Figure 4.** Slices of the normalized covariances (according to equation 2) for different fixed values of $k_i$. The peaks correspond to points along the diagonal. The plots show that the denoiser is able to remove the noise (seen in the input matrices), without introducing new features, so the denoised matrices match closely the targets and the *best* matrix (with a sample of $N_{max}$ spectra).

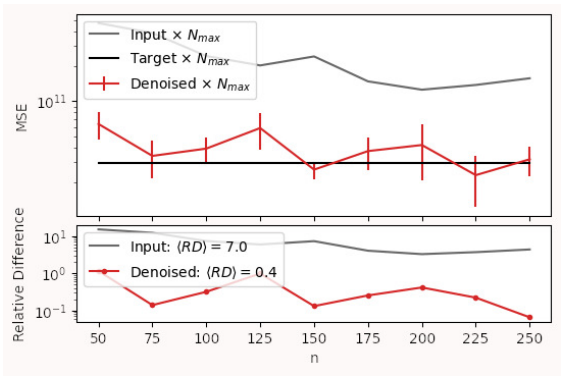


**Figure 5.** MSE for the cosmological covariance matrices, computed by comparing the best sample $(N_{max} = 30,000)$ and the original (input) matrices, in gray, and the denoised ones, in red. As the sample size of the input matrices $(n)$ grow, the agreement between the matrices improve and the MSE becomes smaller. The MSE between the best matrix and the target matrix (with $N = 1000$ spectra) is shown as the black line. The error bars account for the epistemic error (i.e., values obtained for different seeds of the same model). The lower panel shows the relative difference according to equation 3.

A more accurate comparison of the matrices can be glimpsed from comparing slices (rows/columns) of the normalized covariance matrices, in order to show both the diagonal and off-diagonal elements (Blot et al. 2019). In figure 4 we show a few fixed $k_i$ slices of these matrices as a function of $k_j$, with the corresponding values for the input, denoised, target and *best* normalized covariances. All the matrices follow the behavior of the best-case scenario $(N_{max})$, but it is clear that the input matrices are severely affected with noise, especially in the off-diagonal elements. However, after applying the ML denoiser those fluctuations basically disappear, and the off-diagonal elements match the behavior of the target and *best* matrices in all cases.

### 3.2. The MSE between different matrices

We have monitored the improvements in the covariance matrices using the MSE metric:

$$MSE = \frac{1}{N} \sum_{l=1}^{N} \frac{1}{n_k^2} \sum_{i,j=1}^{n_k} (Cov_{ij}^{l} - Cov_{ij}^{(N_{max})})^2 \,, \qquad (3)$$

where $Cov_{ij}^{l}$ are the input, target or denoised covariance matrices, $Cov_{ij}^{(N_{max})}$ is the best matrix, $n_k$ the number of bins of $k$ in

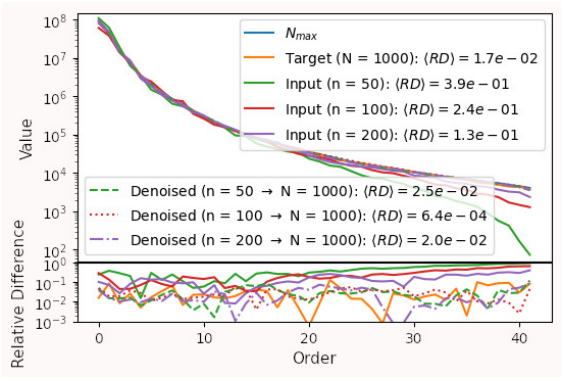


**Figure 6.** Ranked eigenvalues (main plot) and their relative difference (subplot, with the respective values for the matrix with $N_{max}$) for the covariance matrices.

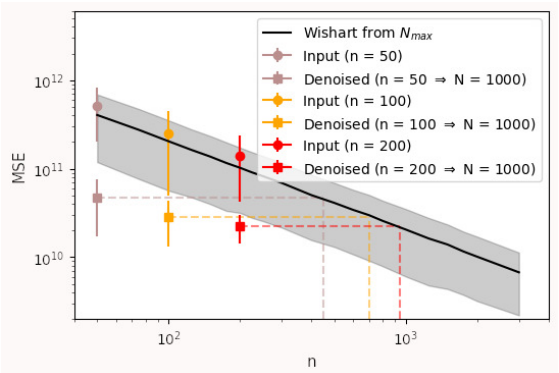


**Figure 7.** MSE comparison between the best matrix with: matrices estimated from the Wishart distribution, in black; the input, in circles (for $n \in [50, 100, 200]$); and denoised matrices, in squares (from $n \in [50, 100, 200] \Rightarrow N = 1000$). The colors corresponds to the number of matrices in the input and resulted denoised matrices. The gray region corresponds to $1\sigma$ deviation for the mean values, in the case of the Wishart matrices. The dashed lines has the intention to guide the reader to see to which number of spectra $n$ the input matrices were taken to their Wishart comparison.

the data vector, and $\mathcal{N}$ is the total number of matrices used in the evaluation. The results are shown in figure 5. The gray line corresponds to the MSE for the original (input) matrices, the red line corresponds to the denoised matrices, and the error bars account for the standard deviation for the results for different random seeds of the ML suite. As a comparison, the black line represents the MSE of the target matrices ($N = 1000$), which is a lower bound for the MSE of the original matrices and provides a sanity check. The lower panel in figure 5 shows the residue between the [input, denoised] and best matrices. Naturally, the MSE decreases when the sample size ($n$) grows. The decrease in MSE that results from applying the denoiser is much larger: the mean values of the residue ($\langle RD \rangle$) show that the denoised matrices deviate by only $\sim 0.4$ from the best matrices, compared with approximately 7.0 for the original input matrices.

### 3.3. The eigenvalues of the matrices

Typically for cosmological covariance matrices the eigenvalues that effectively carry information obey some power law, and as we reach the lower eigenvalues the noise appears as an abrupt change in the scaling of the eigenvalues (Vogeley & Szalay 1996). In figure 6 we show the ranked eigenvalues for covariance matrices. We can see that the eigenvalues of the denoised matrices are much closer to the target and best matrices, while the original input matrices show clear signs of noise in the lower end of the spectrum of eigenvalues, which become more important as we decrease the sample size. The improvements considering RD are a factor of more than $\sim 10$.

### 3.4. An analytical comparison for the ML

We can compare the resulting (denoised) matrices with a model for the probability distribution function for covariance matrices, and which describes how their fluctuations depend on the sizes of the data vector and the sample size. This was done according to the Wishart distribution Wishart (1928); Taylor, Joachimi & Kitching (2013):

$$p(\hat{M}|M,\nu,\eta) = \left(\frac{\nu^{\nu\eta/2}|M|^{-\nu/2}|\hat{M}|^{\gamma/2}}{2^{\nu\eta/2}\Gamma_\eta[\nu/2]}\right)\exp\left[-\frac{\nu\mathrm{Tr}\left(\hat{M}M^{-1}\right)}{2}\right], \quad (4)$$

where $M$ represents the statistical mean of the matrices, $|M|$ is its determinant, $\hat{M}$ is the random variable, the sample covariance matrix, $\gamma = \nu - \eta - 1$, and $\Gamma_\eta[\nu/2]$ the multivariate Gamma function. The parameters of this distribution are the size of the data vector $\eta$, the number of degrees of freedom $\nu$ (the sample size). Therefore, given an "ideal" covariance matrix $M$, this distribution allows us to generate random covariance matrices corresponding to different sample sizes.

We have used the Wishart as an estimator to obtain Wishart matrices with different values for $\nu/n$, analyzing the MSEs among them and the best matrix and the MSE among the denoised matrices and this same best matrix to obtain the effective sample size of the denoised matrices. In figure 7 we plot this MSE (mean and variance). Since the mean MSE values for the denoised matrices are significantly lower compared with the input matrices, according to the Wishart distribution the denoiser is effectively taking matrices with samples of $n = [50, 100, 200]$ spectra, and transforming them into covariance matrices with much larger samples, for $n \in [450, 700, 940]$ spectra, respectively.

### 3.5. Recovering the cosmological parameters

Ultimately, we analyzed the ability of the denoised matrices to recover the fiducial simulated parameters and we have compared these estimation with the parameters coming from the original matrices, to validate our results. The analysis is presented in figure 8. We have explored the parameter space using the Markov Chain Monte Carlo (MCMC) approach, using the emcee library (Foreman-Mackey et al. 2013), data points as a random power spectrum vector, and the different cosmological covariance matrices. We analyzed the the parameters: $H_0$, $\Omega_b$, $\Omega_c$ and the nuisance parameter bias $b$, for the matrices built with the maximum number of spectra, target, denoised and input matrices. In all the analyses we have used 20 walkers and chains of 5000 length (except for the input matrices, for which we have used 6000).

Overall, the parameters were well constrained using the denoised matrices. It is interesting to see that the inference considering the input matrices have a "false" precision (the volumes in the parameter space are smaller, when compared to all the other matrices) and is very inaccurate, because the mean values estimated is highly shifted. Moreover the input matrix with

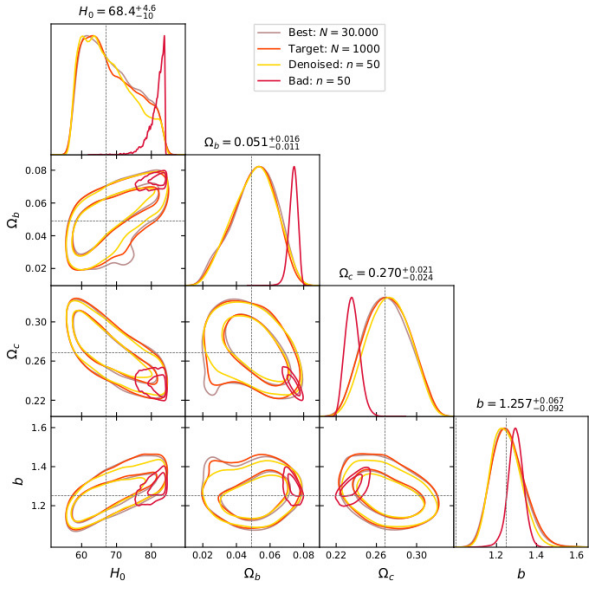


**Figure 8.** Cosmological parameter estimation comparison for different covariance matrices: matrices built using all the spectra available (with $N_{max} = 30,000$ spectra); targets ($N = 1000$ spectra); input (using $n = 50$ spectra); and the denoised ones ($n \Rightarrow N$).

$n = 50$ spectra presents fluctuations on their contours, which are improved/removed in the denoised matrix estimation.

## 4. Discussion and Conclusions

This work presents an efficient approach for the estimation of cosmological covariance matrices. The main idea behind our method is that, starting from matrices built with only hundreds of spectra, we are able to provide covariance matrices that are as good as if they were built with thousands of spectra. We have implemented this method using CNNs as a denoising algorithm, that cleans the noise in the input matrices. Visual inspection (see figures 3 and 4) already shows that the noise was removed, without the introduction of any visible artifacts when compared with the best matrices.

In order to check whether the resulting (denoised) covariance matrices were more similar to the best ones, we performed a series of tests. We started by computing the mean square error (MSE) of equation (3), and verified that the denoiser is able to reduce that indicator by a factor of $\sim 10$ – see figure 5. We also compared the ranked eigenvalues of the matrices, and showed that after denoising the input matrices we recover the main features of the *target* and even of the *best* matrices.

We have shown that the proposed method can be matched in terms of an extrapolation according to the Wishart distribution (Taylor, Joachimi & Kitching 2013; Wishart 1928), by effectively augmenting the size of the sample that underlies the covariance matrix. Finally, the strongest evidence for the power of the denoising technique is provided by the parameter estimation. All the parameters were well constrained in the case of the denoised matrices. We saw improvements (when comparing the input and denoised matrices) for all of them. In particular, the improvements achieved for $H_0$ using the sample with $n = 50$ spectra were of a factor of $\sim 17.9$.

The next steps regarding this project are: (i) to test our machinery in matrices from different cosmologies than the fiducial one, that was used in the training stage (to check if the method can generalize in those situations); (ii) to apply the ML suite in more complex and realistic covariance matrices (redshift space, multiple tracers, and in higher-order statistics).

*Acknowledgements.* We would like to thank the São Paulo Research Foundation (FAPESP), the Brazilian National Council for Scientific and Technological Development (CNPq), and the Coordination for the Improvement of Higher Education Personnel (CAPES) for financial support. NSMS acknowledges financial support from FAPESP, grant 2019/13108-0.

## References

Agrawal, A. et al. 2017, JCAP, 2017, 003
Alves de Oliveira, R. et al. 2020, https://arxiv.org/abs/2012.00240
Blot, L. et al. 2019, MNRAS, 485, 2806
Bond, J. R. et al. 1991, ApJ, 379, 440
Chollet, F. 2017, *Deep Learning with Python*, Manning Publications Company
Chollet, F. R. et al. 2015, https://github.com/fchollet/keras
Coupon, J., et al. 2018, PASJ, 70, S7
Efron, B. 1980, *The Jackknife, the Bootstrap and Other Resampling Plans*, SIAM, Philadelphia
Foreman-Mackey, D. et al, 2-13, PASP, 125, 306
Hartlap, J. , Simon P., & Schneider, P. 2007, A&A, 464, 399
Heavens, A. F. et al. 2017, MNRAS, 472, 4244
Li, Y. et al. 2021, PNAS, 118, 2022038118
Monaco, P. et al. 2013, MNRAS, 433, 2389
Planck Collaboration, Aghanim, N., Akrami, Y. et al, A&A, 641, A6
Schneider, M. D. et al. 2011, ApJ, 737, 11
Taylor, A., Joachimi, B., & Kitching, T. 2013, MNRAS, 432, 1928
Tian, C. et al. 2019, https://arxiv.org/abs/1912.13171
Tinker, J. L. et al. 2010, ApJ, 724, 878
Tukey, J. 1958, Ann. Math. Statist, 29, 614
Vincent, P. , Larochelle, H., Bengio, Y, & Manzagol, P.-A. 2008, Proceedings of the 25th International Conference on Machine Learning, https://dl.acm.org/doi/10.1145/1390156.1390294
Vlah, Z., Seljak, U., & Baldauf, T. 2015, Phys. Rev. D, 91, 023508
Vogeley, M. S. & Szalay, A. S. 1996, ApJ, 465, 34
Voivodic, R. et al. 2019, https://arxiv.org/abs/1906.06630
Wishart, J. 1928, Biometrika, 20A, 1