PHYS788 Final Report

Julian Meunier[1]

[1] *University of Waterloo*
Department of Physics

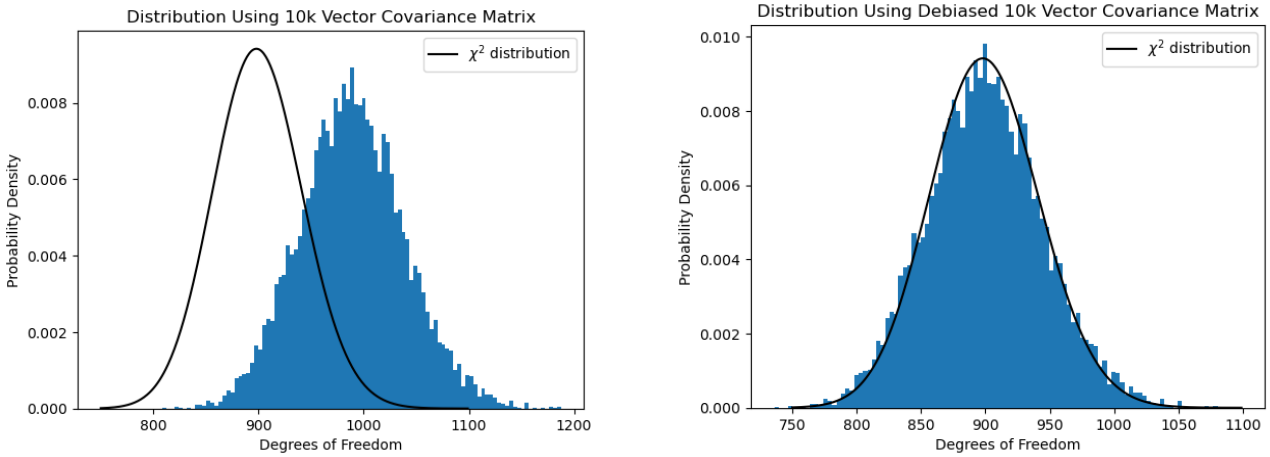
1. ASSIGNMENT 1 - COVARIANCE, χ^2, AND THE HARTLAP FACTOR

In this assignment, we studied covariance matrices through χ^2 distributions and the Hartlap factor. We primarily studied how covariance matrices can be biased and how to properly apply corrective factors and the matrices themselves for other data.

Starting with a cosmological reference model, we generated two sets of 10000 noisy data vectors. With one set we created covariance matrices using 500 to 10000 data vectors. We observed that the resulting distributions of the other data set using these numerical covariance matrices did not match a χ^2 distribution, a result of the covariance matrices being biased since they were generated using correlated data vectors. Also, the results from the 500 vector covariance matrix were nonsensical as there were more degrees of freedom in the data than data vectors used to create the matrix.

We learned about the Hartlap factor, a numerical factor that is applied to a covariance matrix to de-bias it. We found that this worked very well, especially for the covariance matrices using larger amounts of data vectors (see Figure 1). We applied the Hartlap factor instead of others as the computation and implementation are very simple, and the effect of this simple factor is powerful.

Finally, we found that performing the same analysis on the data set used to make the covariance matrices produced strange results. The distributions were shifted significantly and/or exhibited



(a) χ^2 distribution without the Hartlap factor.

(b) χ^2 distribution with the Hartlap factor.

Figure 1: The resulting χ^2 distributions using the 10000 realization numerical covariance matrix, with & without the Hartlap factor.

strange behaviors (like a bimodal distribution). This is due to the fact that the covariance matrix and the data set are correlated, since the covariance matrix was created using that data set.

2. ASSIGNMENT 2 - EMULATOR AND PRINCIPAL COMPONENT ANALYSIS

In this assignment, we studied how to use an emulator to model the cosmological data set, as well as how to use principal component analysis (PCA) to compress data and its effects on the results.

We created an emulator using `CosmoPower` (Spurio Mancini et al. 2021), which uses a neural network to calculate the parameters. The important hyper-parameters of the neural network are the number of nodes & layers, learning rates, batch sizes, patience values, and max epochs for each learning step. We chose four layers of 512 nodes, as well as the values for the other hyper-parameters and number of learning steps through computational trial and error. We chose the learning rates to start relatively large, and to decrease for each learning step, in order to allow for the emulator to appropriately navigate the phase space for the global minimum. We chose a 70%/30% training/validation split.

We chose to use an emulator over different types of machine learning due to its applicability. We chose to use supervised learning as we have some knowledge and expectations of the data and results. Since the data is continuous, a regression model is the best choice, which is why we chose the emulator over a classification model (like decision trees).

The emulator was able to find a good fit on the data (see Figure 2). However, due to the quantity of data used as well as the number of trials it took to figure out appropriate hyper-parameters, the

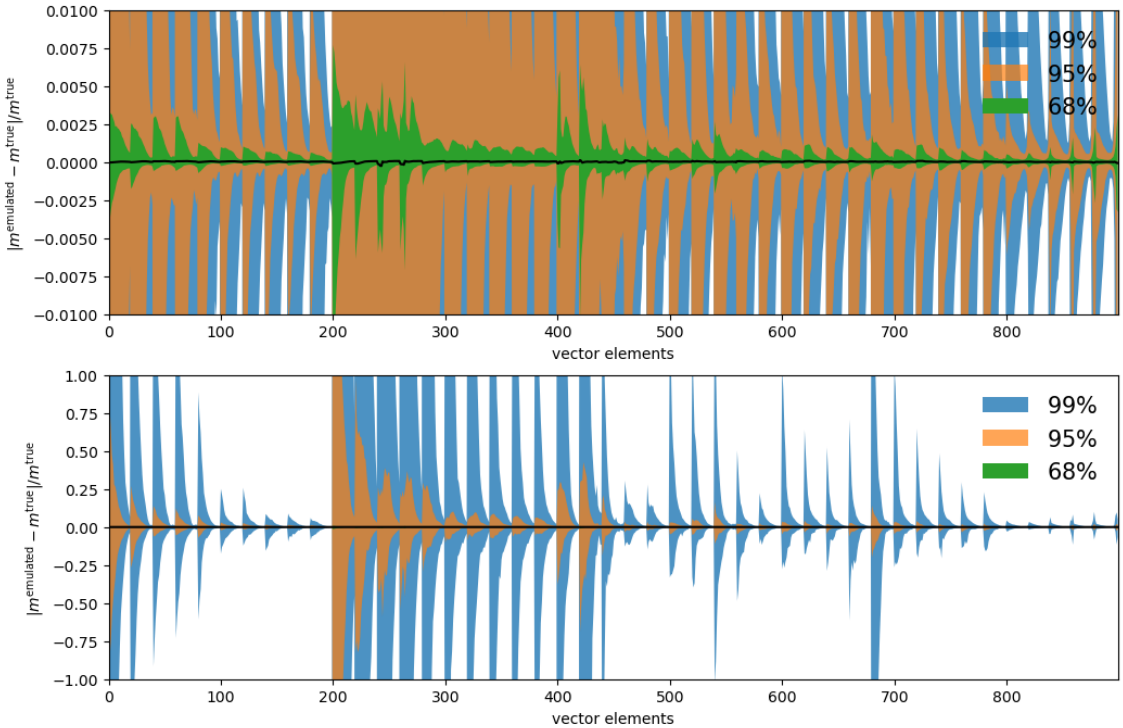
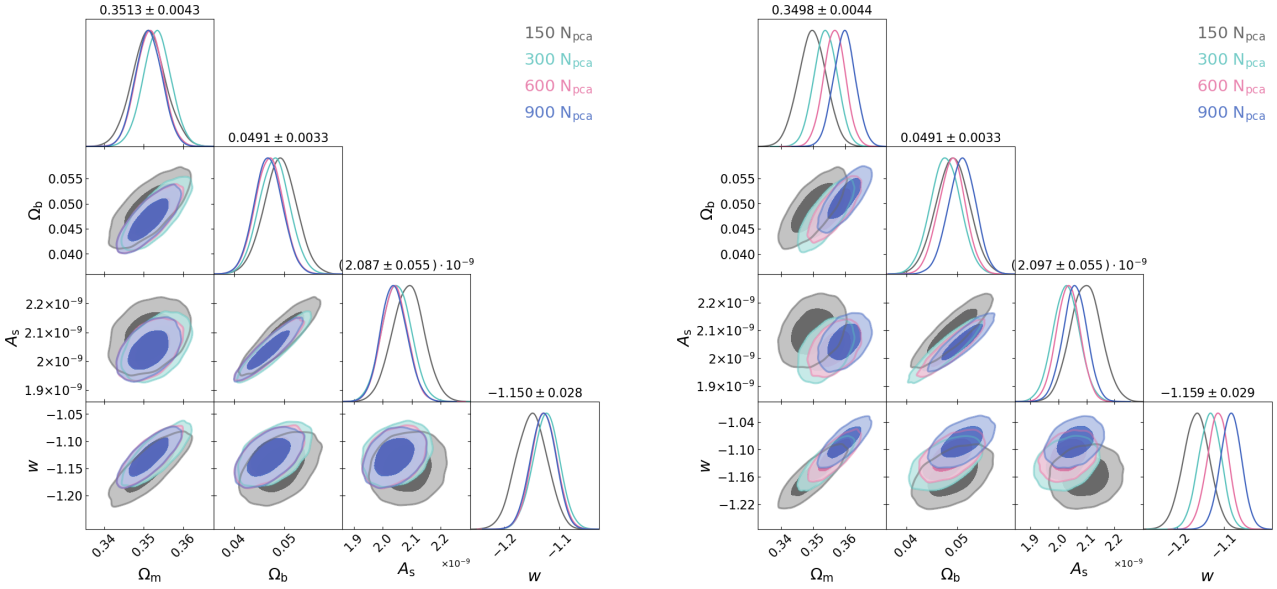


Figure 2: The resulting accuracy of the emulator after training.

computational cost of this endeavour was significant.

We used Fisher Analysis to calculate the covariance of Ω_m and ω. Using PCA compression on the Fisher matrix, and found that roughly 125/525 PCA elements (N_{pca}) were necessary for 10%/1% constraining power. We chose PCA over other types of compression because it maximizes variance in the compressed basis, and the matrix math computation is simple to implement and efficient. With the constraining power results, this makes PCA compression a great performer.

3. ASSIGNMENT 3 - MCMC



(a) Posterior distributions without Hartlap factor applied.

(b) Posterior distributions with Hartlap factor applied.

Figure 3: The resulting posterior distributions using the 1500 realization covariance matrix with PCA compression, with and without the Hartlap factor applied.

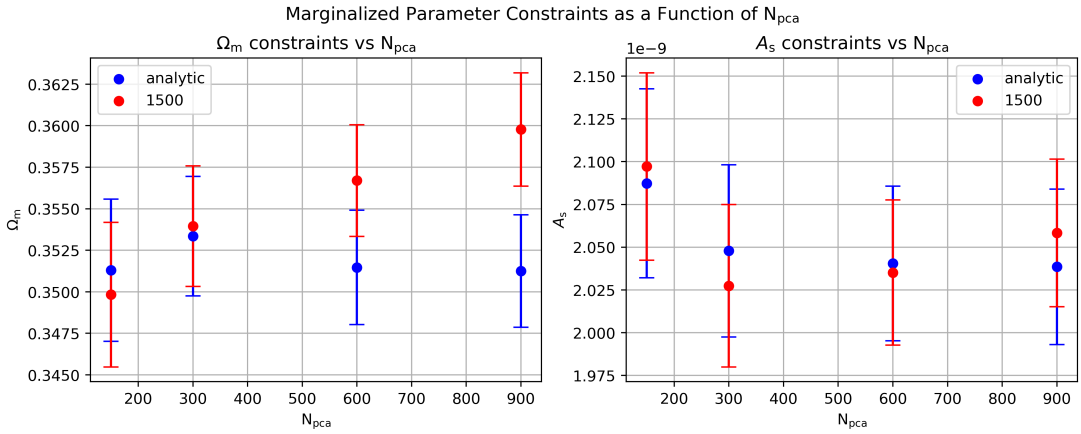
In this assignment, we learned how to use MCMC to calculate posterior distributions for the parameters, and applied the tools we learned in assignments 1 & 2.

We used `emcee` (Foreman-Mackey et al. 2013), with a noisy reference model and the emulator we developed in assignment 2, along with analytical and numerical covariance matrices. We set the priors and likelihood such that the walkers had constant likelihood to spawn only within the range of parameters the emulator was trained on. The random seed was kept fixed to remove any uncertainties from random effects between runs. The hyper-parameters of the MCMC are the total number of steps, walkers, and burning steps. The total number of steps was chosen through trial and error by monitoring the rate at which the MCMC walkers converged. The number of walkers was chosen to balance computational efficiency and precision in the posterior distributions. Finally, the number of burning steps was chosen to be sufficiently small to ignore the initial probing of the

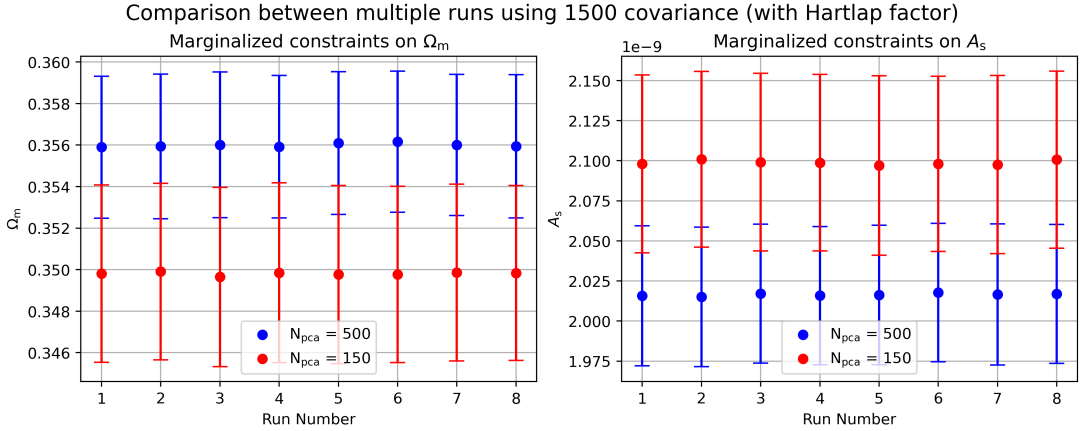
parameter space, but to avoid removing a significant amount of data. We chose to use `emcee` over other MCMC samplers because this sampler is affine, meaning changes in coordinate bases is invariant to covariance between parameters, which is something we need for later work with PCA.

Using the numerical covariance matrices (1500, 3000, 10000 realizations), the results only matched the analytical results with a large realization count and Hartlap factor applied, matching the trend in assignment 1.

We applied PCA compression and the Hartlap factor to the numerical covariance matrices in the MCMC. We found that N_{pca} significantly affected the posteriors, both in mean value and standard deviation. Without the Hartlap factor, the posteriors were all significantly large and overlapping, but far from the analytical results. With the Hartlap factor, only the posteriors using larger numbers of PCA elements (roughly 300+) were sufficiently de-biased. The size of the standard deviation decreased with increasing PCA elements, but the shift in the means caused the overlap between



(a) Marginalized constraints as a function of N_{pca}.



(b) Marginalized constraints for fixed N_{pca} over multiple runs.

Figure 4: The marginalized constraints on Ω_m and ω.

distributions to decrease significantly (see Figure 3). The results of the posteriors from the PCA compression and application of the Hartlap factor to the covariance matrices match our expectations from assignments 1 & 2.

Finally, we analyzed the marginalized constraints on Ω_m and ω as a function of N_{pca}, now using the noise-free reference model. It is to note that we made an error, only changing the reference model to the noise-free model in the PCA compression and omitting the same change to the likelihood function. We found that the mean varied as a function of N_{pca}, however the standard deviation decreased with increasing N_{pca}. In order to verify that the constraints on the parameters were reliable between runs, we chose to run multiple MCMCs with different random seeds at two fixed values for N_{pca}: 150 & 500. We found that the variance in the mean and standard deviations from these runs was multiple orders of magnitude less than when N_{pca} was varied (see Figure 4). This means we can trust our results to be consistent at all scales of N_{pca}.

4. APPLICATIONS TO MY RESEARCH

My research involves modelling the velocity fields of gas clouds surrounding galaxies & clusters, as well as modelling and studying X-Ray spectra for the new telescope XRISM. The most important insight I gained is regarding covariance matrices, and their many intricacies for data analysis. As someone who started with zero experience with statistics before the course, this knowledge will be powerful in the data analysis stages of my projects. Additionally, I could apply MCMC for the modelling of velocity fields to better probe the complicated correlated parameter space, if I find that the current tools I am using are insufficient.

REFERENCES

Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, Publications of the Astronomical Society of the Pacific, 125, 306–312, doi: 10.1086/670067

Spurio Mancini, A., Piras, D., Alsing, J., Joachimi, B., & Hobson, M. P. 2021. https://arxiv.org/abs/2106.03846