School of Physics and Astronomy

Shrinkage Estimation of Covariance Matrices in Cosmological Analysis

Marnix Looijmans
September 2022

Abstract

The covariance matrix is an indispensable quantity for characterizing the uncertainties in correlated multi-dimensional datasets. Since the covariance matrix is usually not known a priori in cosmology, and since only a single data realization is available, it must be determined from mock data realizations. Shrinkage estimation is a class of techniques which are used to estimate the covariance matrix from fewer mock data realizations while still retaining the same accuracy. Two shrinkage estimation algorithms, one linear and one non-linear shrinkage algorithm, were applied to power spectrum measurements. The number of mock data realizations included in the estimation of the covariance estimation was varied and the covariance matrix estimates were compared to a benchmark covariance matrix. Parameter fitting was performed using the various estimates of the covariance matrix and the parameter distributions were compared to a benchmark as well. Both linear and non-linear shrinkage estimation yield promising results. Linear shrinkage is shown to provide accurate estimates of the covariance matrix when up to only twenty mock data realizations are included compared to a data vector of length thirty. Non-linear shrinkage is shown to provide accurate estimates when up to forty mock data realizations are included and is shown to provide biased estimates when the number of mock data realizations is lower than forty.

Supervisor: Dr. Mike Wang

Contents

1 Introduction

In the current decade, many new surveys such as DESI and EUCLID will be launched or are already active and with every newly developed survey, there comes an increase in the precision of its instruments. Especially for measurements of the large-scale structure (LSS) of the universe, which yield datasets that are correlated and high-dimensional, the covariance matrix becomes an indispensable quantity for characterizing the uncertainties in these datasets. Since the covariance matrix is usually not known beforehand in cosmological measurements and only a single data realization is available, it must be determined from mock data realizations. This process turns entries of the covariance matrix themselves into random variables with associated uncertainties [Joachimi, 2017].

This project will explore various methods of estimating the covariance matrix from large mock datasets. These methods can be classified under the category of shrinkage estimators. Traditionally, covariance matrices are directly estimated from the datasets, which yield estimates that are unbiased and work well in the regime where the number of data realizations is much larger than the number of parameters. They perform poorly, however, in the regime where the number of data realizations is roughly as large as or is much smaller than the number of parameters in the datasets [Pope and Szapudi, 2008].

Shrinkage estimation is a promising set of methods to obtain reliable estimates of the covariance matrix in the latter two regimes. The first key aim of this project is to implement two different shrinkage estimation algorithms and verify their implementation with example data, and the second key aim is to apply these techniques to existing galaxy clustering datasets such as the BOSS data releases.

2 Shrinkage Estimators

Two shrinkage estimation algorithms were studied and compared as part of this project. The first one, linear shrinkage estimation, is described and used in Pope and Szapudi [2008] which in turn builds upon the explanation in Schäfer and Strimmer [2005]. The second one, non-linear shrinkage estimation, also referred to as NERCOME[1], is described in Joachimi [2017] building upon work done by Lam [2016].

2.1 Linear shrinkage estimation

Suppose $\mathbf{X}$ is a $p \times n$ matrix where p is the number of random variables and n is the number of data realizations, and let $x_i^{(k)}$ be the entry of the matrix which represents the k-th realization of the i-th random variable. The entries of the empirically estimated sample covariance matrix $\mathbf{S}$ are then given by

$$S_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_i^{(k)} - \bar{x_i})(x_j^{(k)} - \bar{x_j}), \tag{1}$$

where $\bar{x_i}$ and $\bar{x_j}$ are the estimated means of the i-th and j-th random variables from the data realizations, respectively. Secondly, it is necessary to supply a target covariance matrix $\mathbf{T}$, which comes from a model of the data that does not have many free parameters and thus may be biased but has lower variance. A new estimate of the true covariance matrix $\mathbf{C}$ is then constructed by combining $\mathbf{S}$ and $\mathbf{T}$ and is given by

$$\hat{\mathbf{C}} = \lambda\mathbf{T} + (1-\lambda)\mathbf{S}. \tag{2}$$

Here, the hat on $\hat{\mathbf{C}}$ indicates that we are referring to an estimate and λ is called the shrinkage intensity and it takes values from the interval $[0, 1]$. The analytic solution to obtain the best estimate of the shrinkage intensity $\hat{\lambda}^*$ is derived in Ledoit and Wolf [2003] and is given by the equation

$$\hat{\lambda}^* = \frac{\sum_{i,j}\widehat{\mathrm{Var}}(S_{ij}) - \widehat{\mathrm{Cov}}(S_{ij}, T_{ij})}{\sum_{i,j}(T_{ij} - S_{ij})^2}. \tag{3}$$

If $\hat{\lambda}^*$ is less than 0, then $\hat{\lambda}^* = 0$ is enforced and if $\hat{\lambda}^*$ is greater than 1, then $\hat{\lambda}^* = 1$ is enforced.

2.2 NERCOME

Again, let $\mathbf{X}$ be a $p \times n$ matrix where p is the number of random variables and n is the number of data realizations. In this procedure it is assumed that each of the n data realization vectors is mean-subtracted. The sample covariance matrix is then estimated by

$$\mathbf{S} = \frac{1}{n-1}\mathbf{X}\mathbf{X}^T. \tag{4}$$

Observe that this equation is equivalent to equation 1 above, provided that the data vectors are indeed mean-subtracted. The steps of the NERCOME estimator are as follows as described in Joachimi [2017].

[1] Non-parametric eigenvalue-regularized covariance matrix estimator

1. For a given split parameter s which takes integer values $2, 3, ..., n-2$, divide the data matrix $\mathbf{X}$ into an $p \times s$ matrix $\mathbf{X}_1$ and an $p \times (n-s)$ matrix $\mathbf{X}_2$. There are $\binom{n}{s}$ possible ways of choosing s out of n columns.

2. Estimate the sample covariance $\mathbf{S}_1$ and $\mathbf{S}_2$ of $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively.

3. Diagonalize the sample covariance $\mathbf{S}_1$, i.e. $\mathbf{S}_1 = \mathbf{U}_1\mathbf{D}_1\mathbf{U}_1^T$, where $\mathbf{D}_1$ is the matrix which contains the eigenvalues of $\mathbf{S}_1$ on the diagonal and $\mathbf{U}_1$ is the corresponding matrix of eigenvectors.

4. Calculate the estimator
$$\hat{\mathbf{Z}} = \mathbf{U}_1 \operatorname{diag}(\mathbf{U}_1^T\mathbf{S}_2\mathbf{U}_1)\mathbf{U}_1^T \tag{5}$$
and average over different decompositions $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$. The implementation in this project takes the average for each s over 1000 different decompositions unless $\binom{n}{s} < 1000$, in which case the average is taken over all possible different decompositions. Let the average value of $\hat{\mathbf{Z}}$ and $\mathbf{S}_2$ be denoted by $\bar{\mathbf{Z}}$ and $\bar{\mathbf{S}}_2$, respectively.

5. Finally, calculate the quantity
$$Q(s) = \left|\left|\bar{\mathbf{Z}} - \bar{\mathbf{S}}_2\right|\right|_F^2 . \tag{6}$$
This is the Frobenius matrix norm given by $||\mathbf{M}||_F^2 = \operatorname{Tr}(\mathbf{M}\mathbf{M}^T)$.

The optimal value for the split parameter s is determined by minimizing equation 6. The paper by Lam [2016] suggests that we perform the NERCOME procedure for the split locations
$$s \in \{2n^{1/2}, 0.2n, 0.4n, 0.6n, 0.8n, n - 2.5n^{1/2}, n - 1.5n^{1/2}\}$$
whereas the paper by Joachimi [2017] suggests that we perform the procedure for 20 equidistant steps of s in the interval $[0.1n, 0.9n]$ and then determine which of these values minimizes $Q(s)$. The estimated covariance matrix is then given by $\bar{\mathbf{Z}}$. The implementation of NERCOME in this project combines these two suggestions and evaluates $Q(s)$ for the split locations
$$s \in \{2n^{1/2}, 0.1n, 0.15n, 0.2n, ..., 0.85n, 0.9n, n - 2.5n^{1/2}, n - 1.5n^{1/2}\}. \tag{7}$$

One obvious difference between linear shrinkage and NERCOME is that NERCOME does not require a target matrix to be provided by the user. Contrastingly, NERCOME is computationally a more expensive algorithm than linear shrinkage due to the various successive averaging processes that form part of the algorithm.

3 Toy Examples

As part of this project, the shrinkage estimators from the previous section were implemented in a Python module[2] and in order to verify our implementation, toy examples from Pope and Szapudi [2008] and Joachimi [2017] were recreated and compared.

3.1 Uncorrelated toy examples

The first toy example is discussed in Pope and Szapudi [2008] and starts by generating n realizations of p random variables, each following a distribution $X_i^{(k)} = N(0, \sigma^2)$. This project extends the toy example from Pope and Szapudi [2008] by applying NERCOME in addition to linear shrinkage. Setting $p = 18$ and $\sigma = 1.1$ and choosing the target matrix to be the identity matrix, 100 simulations were run for each value of n. The plots below in figure 1 show the average results from the simulations for $n = 40, 400, 4000$ and allow for a comparison of the eigenvalue spectra of the linear shrinkage estimator and the NERCOME estimator. The empirically estimated covariance, the target and the true covariance are also shown.

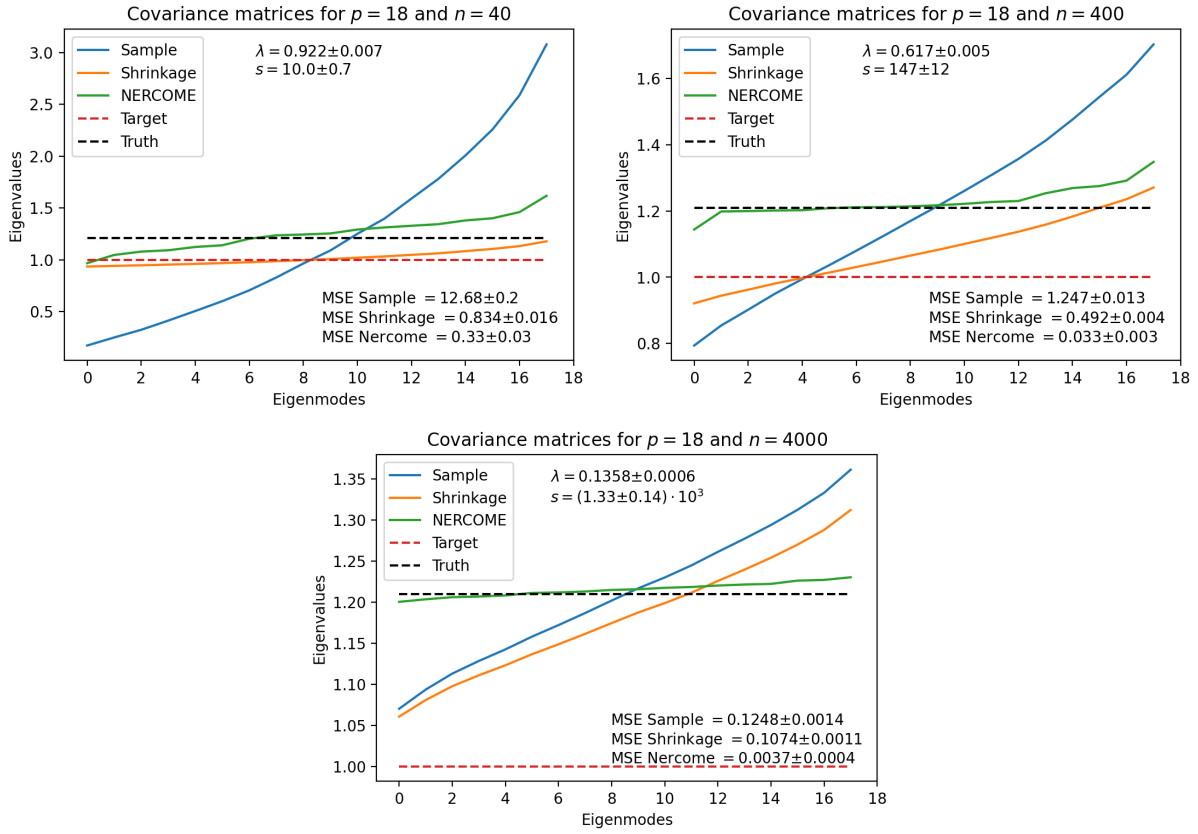

Figure 1: Eigenvalue spectra of linear shrinkage and NERCOME estimators for $p = 18$ and $n = 40, 400, 4000$. The true covariance matrix is a multiple of the identity. The eigenvalue spectra of the sample covariance, target, and true covariance are also shown. Each plot shows the average of 100 simulations.

[2] `https://github.com/marnixlooijmans/shrinkage-estimators-modules`

Looking at the eigenvalue spectra and the mean squared error (MSE) values with respect to the true covariance matrix, it is clear that NERCOME provides a better estimate of the true covariance than the sample or linear shrinkage estimates. The goodness of the linear shrinkage estimate is of course dependent upon the choice of target covariance matrix and a different target will provide a different estimate. The difference in MSE values becomes greater between NERCOME on the one hand, and linear shrinkage and the sample on the other when the number of data realizations, n, is increased. The sample estimate seems to move closer to the linear shrinkage estimate for larger n. The results for linear shrinkage, including the optimal shrinkage intensity values as quoted in the plots, are in agreement with the toy example results in Pope and Szapudi [2008].

The second toy example is discussed in Joachimi [2017] and starts by generating n realizations of p random variables with each random variable following a distribution $X_i^{(k)} = N(0, \sigma^2)$. This project extends the toy example by applying linear shrinkage in addition to NERCOME. In contrast to the first toy example, here the variance was set to $\sigma^2 = 1$ in the first half of the random vectors X_i and $\sigma^2 = 5$ in the second half. The target matrix was chosen to be the matrix with the number 3 on the diagonal. The plots in figure 2 show the eigenvalue spectra for $p = 30$ and $n = 15, 30, 100, 1000$.

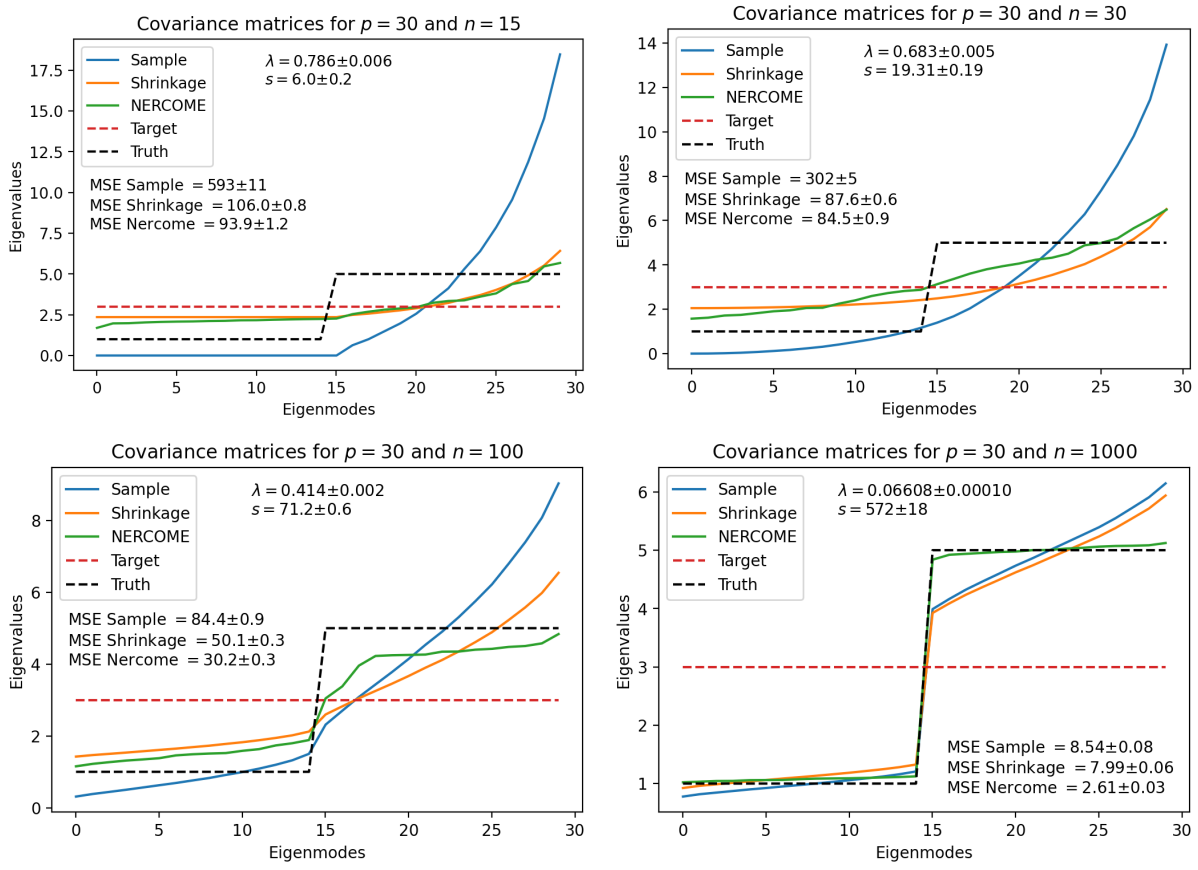

Figure 2: Eigenvalue spectra of linear shrinkage and NERCOME estimators for $p = 30$ and $n = 15, 30, 100, 1000$. The variance is set to $\sigma^2 = 1$ in the first half of the random vectors and $\sigma^2 = 5$ in the second half. The eigenvalue spectra of the sample covariance, target, and true covariance are also shown. Each plot shows the average of 100 simulations.

Again, NERCOME outperforms linear shrinkage and the sample estimate both in terms

of mean squared errors and by looking at the eigenvalue spectra. Both the linear shrinkage and sample estimates are prone to underestimating some of the eigenvalues of the covariance matrix. This could prove problematic in data analysis since the estimated errors associated to the data would be too small and incorrect conclusions could then be drawn. NERCOME provides a close-to-ideal estimate of the eigenvalue spectrum of the true covariance matrix for $n = 1000$ and similarly to the first toy example, the sample estimate seems to move closer to the linear shrinkage estimate for increasing n.

It is important to note here that the performance of linear shrinkage depends on the chosen target matrix. In this case we knew beforehand what the true covariance matrix looked like and this influenced our choice of target. In real data analysis one will not know the true covariance matrix a priori and hence one must resort to existing literature to find a possible target.

3.2 Correlated toy examples

In addition to the two uncorrelated toy examples above, two correlated toy examples were also used to test both algorithms, because cosmological data are often correlated and thus the covariance matrix is generally non-diagonal.

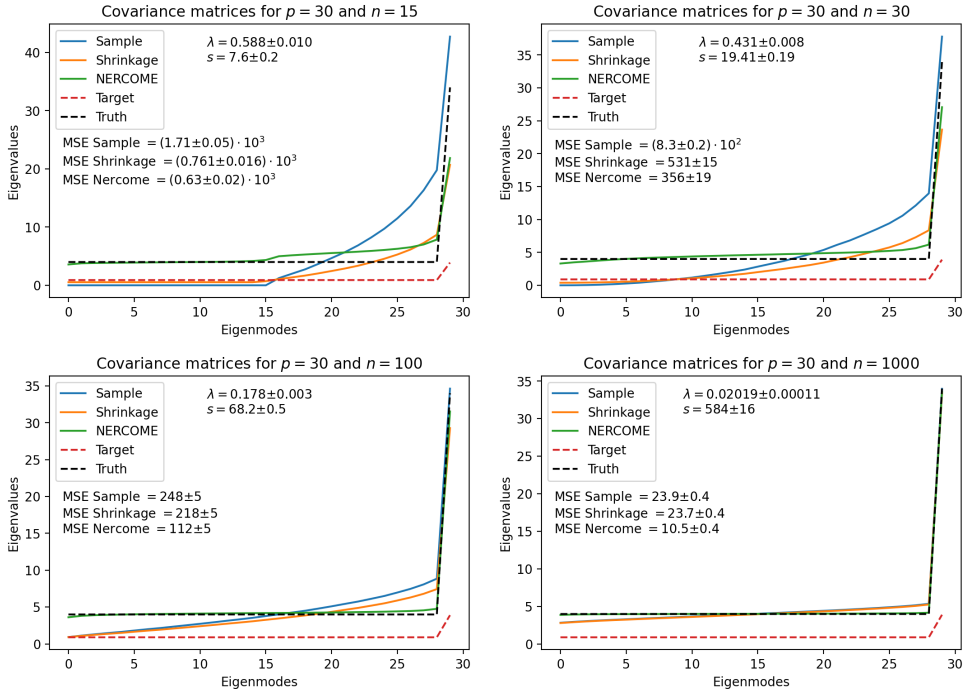

Figure 3: Eigenvalue spectra of linear shrinkage and NERCOME estimators for $p = 30$ and $n = 15, 30, 100, 1000$. The true covariance matrix has 5 on the diagonal and 1 on the off-diagonal entries. The eigenvalue spectra of the sample covariance, target, and true covariance are also shown. Each plot shows the average of 100 simulations.

The third toy example starts by generating a $p \times p$ covariance matrix where the diagonal entries are set to a value of 5 and all off-diagonal entries to a value of 1. We set $p = 30$ and generate n data realizations to which we apply both shrinkage algorithms. The target covariance matrix was chosen to have diagonal entries equal to 1 and all off-diagonal entries equal to 0.1. The eigenvalue spectra are shown in figure 3 for $n = 15, 30, 100, 1000$.

Similar to the first two toy examples, the NERCOME estimate is closer to the true covariance matrix than the linear shrinkage and sample estimates, considering both the MSE values and by looking at the eigenvalue spectra. The linear shrinkage and sample estimate appear to underestimate part of the eigenvalues for low n, which could prove problematic as outlined earlier.

The fourth and last toy example starts by generating a covariance matrix where the off-diagonal entries are equal to 1 and the upper right and lower left entries in the matrix are equal to 0. All entries in between are equal to a value between 1 and 0, decreasing linearly in equal decrement as we mover further away from the diagonal. An example of such a covariance matrix for $p = 5$ is

$$\mathbf{C} = \begin{pmatrix} 1 & 3/4 & 1/2 & 1/4 & 0 \\ 3/4 & 1 & 3/4 & 1/2 & 1/4 \\ 1/2 & 3/4 & 1 & 3/4 & 1/2 \\ 1/4 & 1/2 & 3/4 & 1 & 3/4 \\ 0 & 1/4 & 1/2 & 3/4 & 1 \end{pmatrix}.$$

The same target as in the third toy example was chosen. The eigenvalue spectra for $p = 30$ and $n = 15, 30, 100, 1000$ are plotted in figure 4.

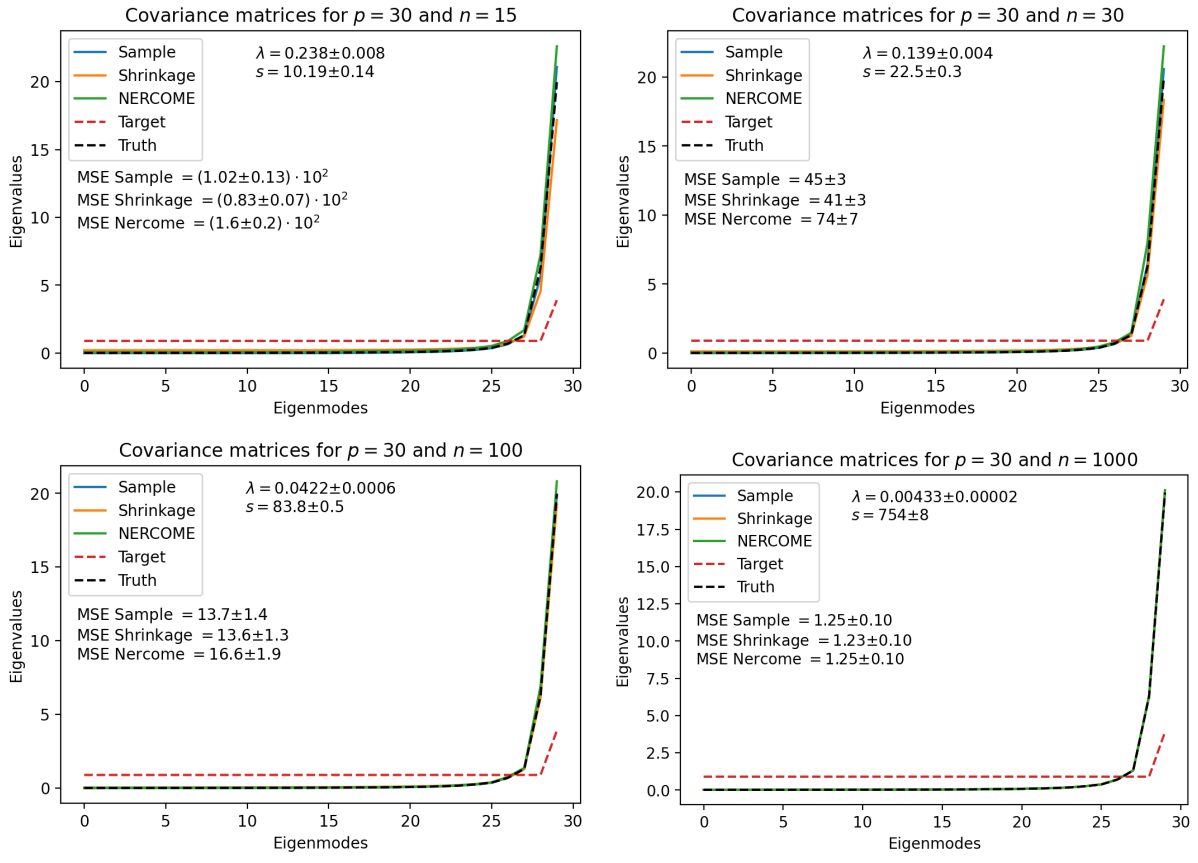

Figure 4: Eigenvalue spectra of linear shrinkage and NERCOME estimators for $p = 30$ and $n = 15, 30, 100, 1000$. The true covariance matrix has entries which decreasing linearly in equal decrement as we mover further away from the diagonal. The eigenvalue spectra of the sample covariance, target, and true covariance are also shown. Each plot shows the average of 100 simulations.

The different eigenvalue spectra in figure 4 are hard to distinguish from each other which limits any analysis to the MSE values with respect to the true covariance matrix. In stark contrast to the other toy examples, the NERCOME estimate is outperformed in terms of MSE values by both the sample and linear shrinkage estimates, although the differences become smaller when n is increased.

There is a major caveat to using the MSE values for assessing the goodness of the covariance matrix estimates. Joachimi [2017] explains that one is usually more interested in minimizing the uncertainty and bias of the entries of the inverse covariance matrix, which is used in likelihood analyses, instead of minimizing the mean squared error. Therefore, in the next section we perform a power spectrum likelihood analysis where we use different estimates of the covariance matrix for comparison purposes.

4 Power Spectra

The BOSS DR12 NGC z1 power spectrum dataset [Alam et al., 2015], [Reid et al., 2016], and corresponding mock catalogs [Kitaura et al., 2016], were used to find estimates of the covariance matrix, which were subsequently used in likelihood analyses. The ultimate goal was to vary the number of mock data realizations and observe both the effects on the covariance matrix itself and also on the fitting of cosmological parameters in likelihood analyses. The power spectrum data used in this section was downloaded from here[3].

4.1 Covariance matrix

In order to generate the covariance matrix for the power spectrum, we made use of a selection of the 2048 mock catalogs provided by Beutler and McDonald [2021]. Both the NERCOME and linear shrinkage algorithm were applied to n mock catalogs and the estimated covariance matrices obtained from this procedure were compared to the sample covariance matrix (Eq. 4) computed from the same n mock catalogs and also to the covariance matrix provided by Beutler and McDonald [2021]. For linear shrinkage we chose the following target from Schäfer and Strimmer [2005],

$$T_{ij} = \begin{cases} S_{ij}, & \text{if } i = j; \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

This target only shrinks the off-diagonal entries of the sample covariance matrix and leaves the diagonal entries, i.e. the variances, intact. It is important to keep in mind that this target comes from the same model as the sample covariance matrix instead of a different model of high bias and low variance. The key question we want to answer is if we can obtain an estimate of the covariance matrix which is close to the estimate of Beutler and McDonald [2021] but using fewer mocks. In other words, Beutler and McDonald's estimate is used as the benchmark.

We made use of the `pk_tools` module[4] provided by Beutler to read the mock power spectrum data. A random selection of n mock data files were selected and imported where every 10 bins were combined. Only the $\mathbf{P_0}, \mathbf{P_2}, \mathbf{P_4}$ multipoles were included in the analysis and hence the final mock data vector was

$$\mathbf{P}^{(i)} = \begin{pmatrix} \mathbf{P_0}^{(i)} \\ \mathbf{P_2}^{(i)} \\ \mathbf{P_4}^{(i)} \end{pmatrix}, \tag{9}$$

which contained 120 elements. Here, i refers the i-th mock data vector out of n. All the n mocks were then combined into a single $120 \times n$ matrix

$$\mathbf{P}^{\text{mock}} = \begin{pmatrix} \mathbf{P}^{(1)} & \mathbf{P}^{(2)} & \dots & \mathbf{P}^{(n)} \end{pmatrix}. \tag{10}$$

This data matrix was then used as argument in the NERCOME and linear shrinkage method and additionally the sample covariance matrix was calculated. The eigenvalue spectra of these two estimates are shown in figure 5 for $n = 100, 120, 200$. The covariance

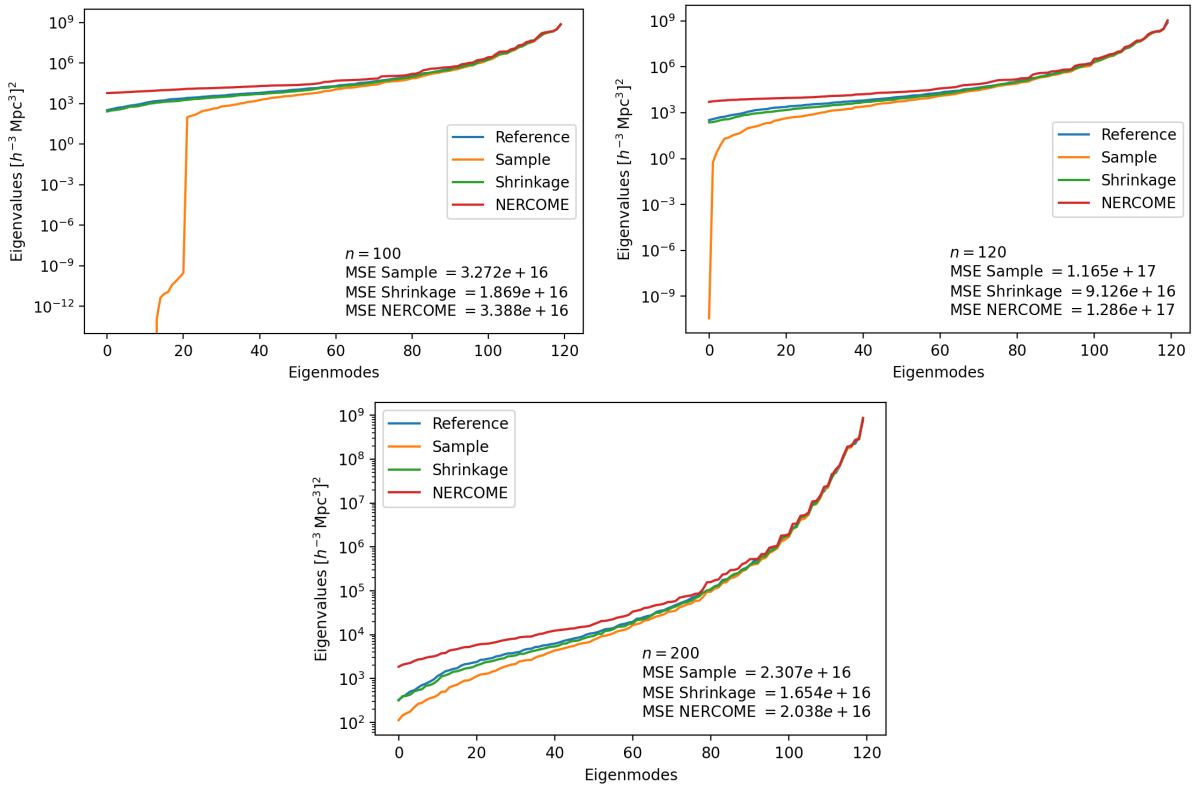

Figure 5: Eigenvalue spectra of covariance matrices generated from $n = 100, 120, 200$ mocks.

matrix provided by Beutler and McDonald [2021], indicated by the label 'reference', is also shown where only the entries corresponding to the even multipoles were included.

The cases in figure 5 cover the cases where the number of mocks n is less than, equal to, and greater than the number of bins. In all three cases, the eigenvalue spectra for the reference covariance matrix and the linear shrinkage estimate are close to each other whereas eigenvalue spectra for the NERCOME and sample estimates deviate more from the reference. The NERCOME estimate overestimates the reference eigenvalue spectrum and the sample estimate underestimates the reference eigenvalue in all three cases. In particular, underestimation could be problematic since this would imply underestimating the uncertainties in the data. Looking at the MSE values, it is clear that the linear shrinkage estimate has the smallest mean squared error compared to the reference. Also notice that in the $n = 200$ case, the MSE for NERCOME is less than the MSE for the sample whereas in the $n = 100$ and $= 120$ cases, the MSE for the sample is less than the MSE for NERCOME.

4.2 Likelihood analysis

One is usually not interested in the covariance matrix as a result in itself, but rather as a quantity required for likelihood analyses in order to sample cosmological parameters. In this project, we sampled the Hubble parameter h, the bias parameter b and the growth

[3] `https://fbeutler.github.io/hub/deconv_paper.html`
[4] `https://github.com/fbeutler/pk_tools`

rate f, which all change the power spectrum. The likelihood function for power spectrum measurements is

$$\mathcal{L} \propto \exp\left[-\frac{1}{2}(\mathbf{P}^{\rm conv} - \mathbf{WMP}^{\rm true,flat-sky})^T \mathbf{C}_{\rm conv}^{-1}(\mathbf{P}^{\rm conv} - \mathbf{WMP}^{\rm true,flat-sky})\right], \quad (11)$$

where $\mathbf{C}_{\rm conv}$ is the covariance matrix, $\mathbf{W}$ is the window function matrix, $\mathbf{M}$ is the wide-angle transformation matrix, $\mathbf{P}^{\rm true,flat-sky}$ is a vector of monopole, quadrupole and hexadecapole models (no wide-angle effects), and $\mathbf{P}^{\rm conv}$ is the vector of measured power spectrum monopole, dipole, quadrupole, octopole and hexadecapole, [Beutler et al., 2019] and [Beutler and McDonald, 2021].

The measured power spectrum vector $\mathbf{P}^{\rm conv}$ and the $\mathbf{W}$ and $\mathbf{M}$ matrices for BOSS DR12 NGC z1 are derived in Beutler and McDonald [2021] with the wide-angle formalism suggested by Beutler et al. [2019].

The vector $\mathbf{P}^{\rm true,flat-sky}$ is the model data consisting of monopole, quadropole, and hexadecapole models and does not include wide-angle effects. This vector is a function of our sampling parameters h, b, and f. The standard kaiser RSD Model [Saito, 2016] was used to find values for the vector components and the monopole, quadrupole and hexadecapole models are given by

$$P_0(k) = \left(b^2 + \frac{2}{3}bf + \frac{1}{5}f^2\right) P_m(k), \quad (12)$$

$$P_2(k) = \left(\frac{4}{3}bf + \frac{4}{7}f^2\right) P_m(k), \quad (13)$$

$$P_4(k) = \frac{8}{35}f^2 P_m(k), \quad (14)$$

where $P_m(k)$ was taken to be the `cosmology.LinearPower` function from the `nbodykit` package[Hand et al., 2018]. The `cosmology` class was initialized with `cosmology.Plank15` values, i.e. the Planck15 cosmological model, except we let the Hubble parameter h vary. Observe that the other sampling parameters are included in the Kaiser model equations above. The redshift parameter of `cosmology.LinearPower` was set to $z_1 = 0.38$ which is the effective redshift from Beutler and McDonald [2021] for the BOSS DR12 NGC z1 dataset. The functions above were then evaluated at values for k which come from the observed data file. The analysis was limited to $k < 0.1\,h\,\mathrm{Mpc}^{-1}$ to ensure that `nbodykit` would produce accurate model predictions. This resulted in a data vector of length 30. To match the length of the data vector, new covariance matrix estimates of 30×30 needed to be generated by only including mock data for $k < 0.1\,h\,\mathrm{Mpc}^{-1}$ here as well.

The next step involved MCMC sampling of the parameters h, b, and f. To this end, the `zeus` package was used. The MCMC sampling process was done with various covariance matrices while keeping all other parameters the same during each run. The covariance matrices used were the 30×30 estimates from the NERCOME and linear shrinkage estimators as well as the sample covariance matrix estimates. The estimates were generated from $n = 20, 30, 40, 100$ mocks. We also used the covariance matrix provided by Beutler and McDonald [2021], only including the entries corresponding to $k < 0.1\,h\,\mathrm{Mpc}^{-1}$ and to the even multipoles, which we again used as the benchmark. A total of 6 walkers were used in the MCMC analyses, each taking 1000 steps.

5 Results and Discussion

After the MCMC analysis processes finished, the first half of the evaluations in each chain were removed to cut off the burn-in phase. The second half of the evaluations in each of the six chains were combined to form three arrays of length 3000, one for each parameter. The distributions of these sampled parameter values are plotted as corner plots in the figures below for which we used the `corner` package [Foreman-Mackey, 2016]. The distributions sampled using the covariance matrix provided by Beutler and McDonald [2021], referred to as 'reference', is used as benchmark to compare the sampling results from other covariance matrix estimates to.

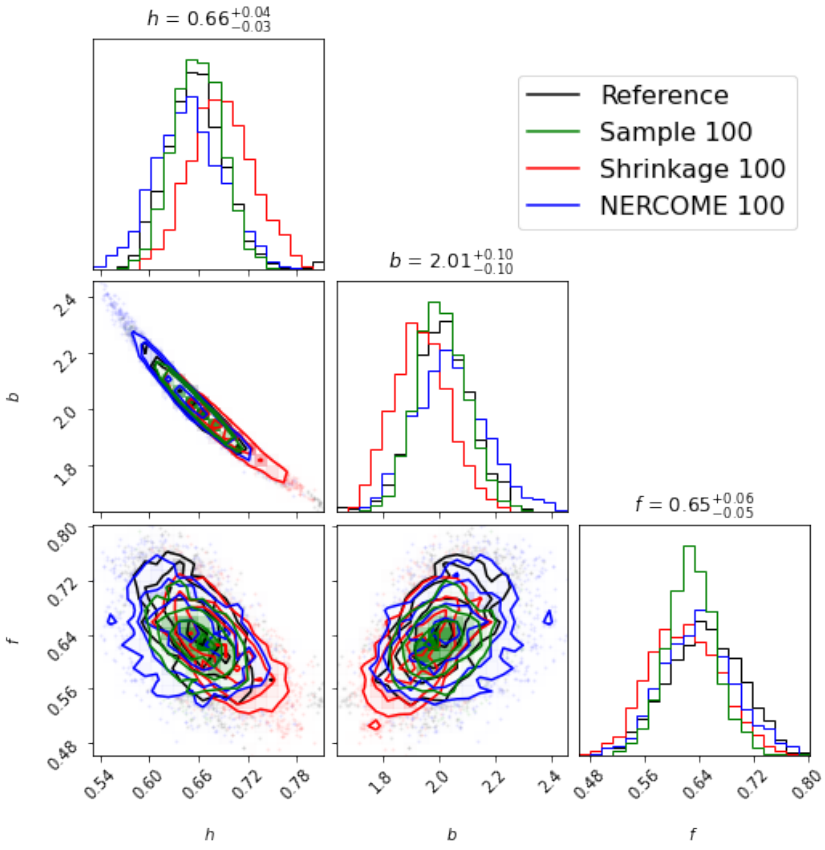

Figure 6: Posterior constraints of the Hubble parameter h, bias parameter b, and growth parameter f. The parameters were fitted using different covariance matrix estimates for $n = 100$ mocks. The subplot titles are the means of the distributions from the reference covariance matrix.

In the $n = 100$ case in figure 6 above, the parameter distributions for h and b from all three covariance matrix estimates are reasonably close to the reference distribution. The distribution for f from the sample covariance matrix underestimates the uncertainty since the distribution is too narrow compared to the reference distribution. The distributions from linear shrinkage appear to be shifted by a small amount with respect to the reference. Looking at the similarity of the sampling distributions from NERCOME and the reference distributions, the NERCOME estimate of the covariance matrix provides slightly more accurate parameter distributions than the linear shrinkage estimate and the sample estimate.

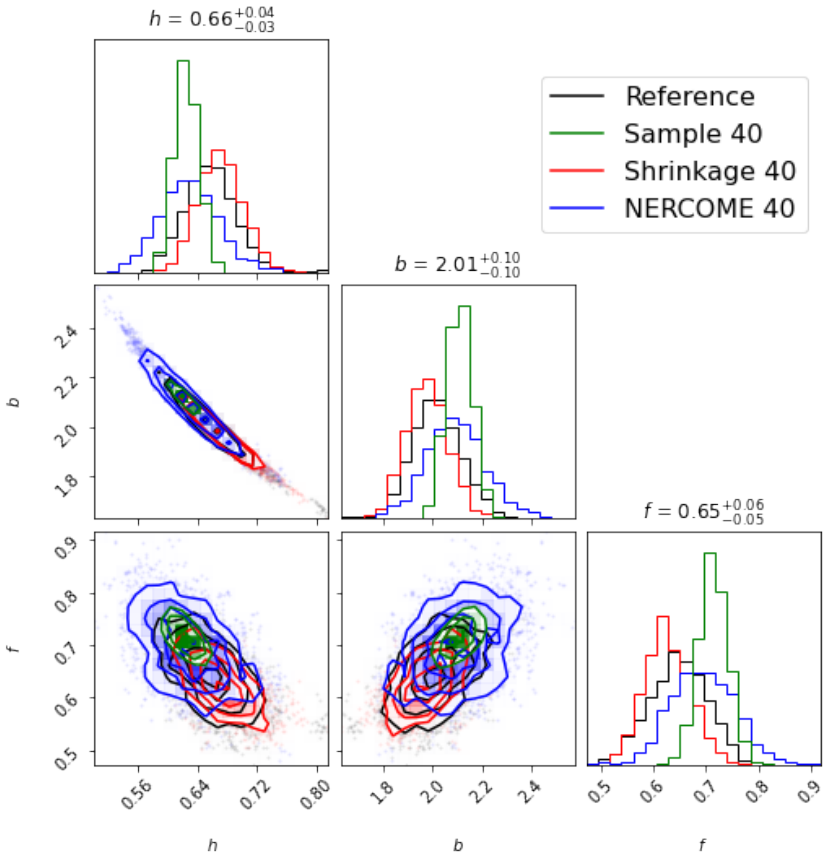

Figure 7: Posterior constraints of the Hubble parameter h, bias parameter b, and growth parameter f. The parameters were fitted using different covariance matrix estimates for $n = 40$ mocks. The subplot titles are the means of the distributions from the reference covariance matrix.

It is clear from figure 7 above that when $n = 40$ mocks, the sample covariance matrix estimate significantly underestimates the uncertainties in the data. This is because the distributions for h, b, and f are too narrow in comparison to the reference distributions. An explanation for this is that we did not apply the Hartlap factor (Eq. 16 in Hartlap et al. [2007]). An unbiased covariance matrix estimate does not yield an unbiased precision matrix estimate upon inversion. In order to find an unbiased precision matrix estimate, a corrections factor must be applied which is often referred to as the Hartlap factor.

Moreover, for $n = 40$, both the distributions from the linear shrinkage and NERCOME estimates for all three parameters are slightly shifted with respect to the reference distribution. Also observe that for all three parameters the distribution from linear shrinkage is slightly narrower than the reference and the distribution for NERCOME is slightly wider than the reference. This could mean that linear shrinkage is prone to underestimation of the uncertainties of the parameters when only 40 mocks are included. In summary, both NERCOME and linear shrinkage provide reasonable approximations of the parameter distributions in the $n = 40$ case.

For the following two cases, no parameter fitting was performed using the sample covariance matrix generated from $n = 20$ and $n = 30$ mocks. Notice that in these two cases, the number of mocks n is equal to or less than the number of bins p. The MCMC sampling process using the sample covariance matrix from $n = 20$ mocks failed to converge

and instead unexpectedly quit after roughly 6% of steps were taken. The walkers of the MCMC analysis using the $n = 30$ sample covariance matrix traced out chains of constant values instead of distributions of values.

The inverse of the sample covariance matrix for $n = 20$ and $n = 30$ contained negative eigenvalues. This was thought to be the main cause of the issues described above and we attempted to fix that by flipping the signs of the negative eigenvalues, turning them into positive eigenvalues and the inverse covariance matrix was then reconstructed from the eigenvectors and new eigenvalues. This did not have any effect on the MCMC sampling and the sampling still failed for the $n = 20$ and $n = 30$ sample covariance matrix.

Another solution that we attempted was setting the negative eigenvalues of the inverse covariance matrix to zero. This did not make a difference either. We conclude that the sample covariance matrix estimate for $n = 20$ and $n = 30$ becomes singular and hence the estimate fails to provide any useful approximation to the true covariance matrix. Therefore no parameter distributions for the sample estimate from $n = 20$ and $n = 30$ mocks are provided.

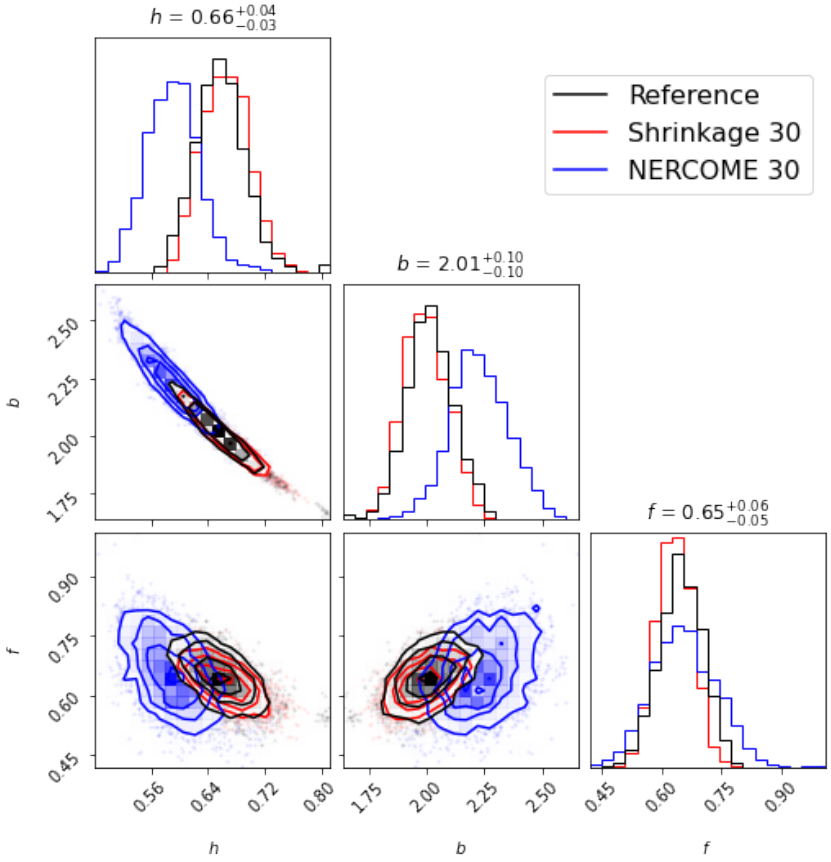

Figure 8: Posterior constraints of the Hubble parameter h, bias parameter b, and growth parameter f. The parameters were fitted using different covariance matrix estimates for $n = 30$ mocks. The subplot titles are the means of the distributions from the reference covariance matrix.

Looking at the distributions in figure 8 where $n = 30$ mocks, it is clear that when we use the the linear shrinkage covariance matrix estimate, we obtain parameter distributions for h, b, and f that are significantly closer to the reference distributions than when

we used the NERCOME covariance estimate. The linear shrinkage estimate does not significantly underestimate the uncertainties of the parameters whereas the NERCOME estimate overestimates the uncertainties. Additionally, the distributions from the NERCOME estimate are shifted by a large amount with respect to the reference distributions.

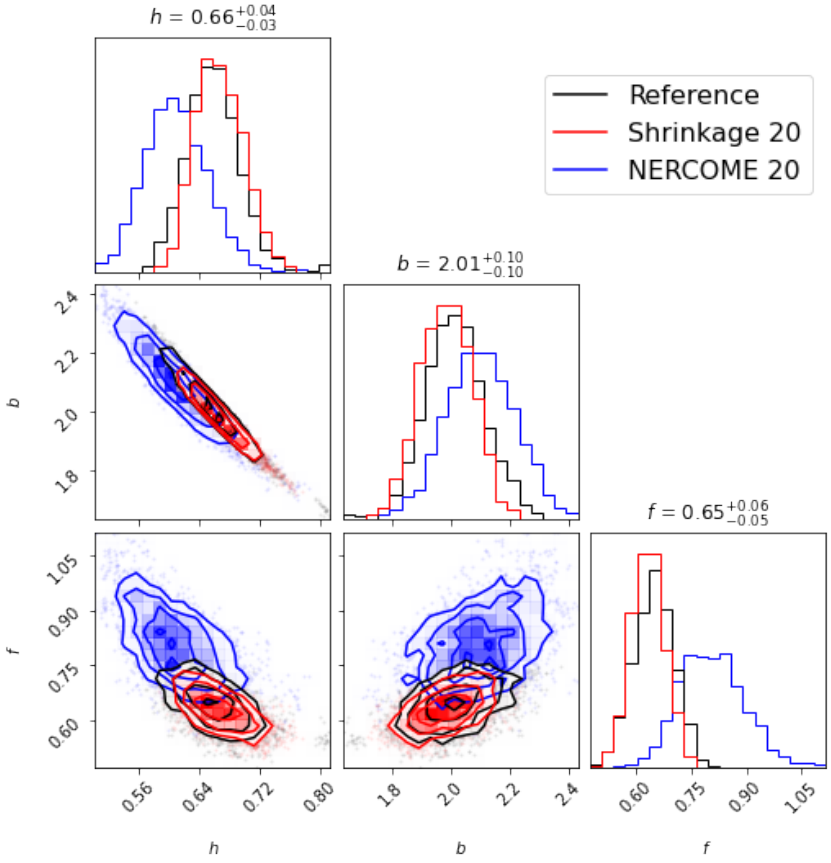

Figure 9: Posterior constraints of the Hubble parameter h, bias parameter b, and growth parameter f. The parameters were fitted using different covariance matrix estimates for $n = 20$ mocks. The subplot titles are the means of the distributions from the reference covariance matrix.

When only 20 mock data realizations were included, the results are similar to the $n = 30$ case. In figure 9, the distributions from linear shrinkage are much more similar to the reference distributions than the distributions from NERCOME. Also, the NERCOME estimate overestimates the uncertainties of the parameters and the parameter distributions from NERCOME are shifted with respect to the reference distributions. There are a few caveats to the results in this project however.

Looking at the NERCOME procedure, in this project we applied NERCOME on power spectrum measurements whereas Joachimi [2017] applied NERCOME on cosmic weak lensing. In figure 9 and 8 we see that the parameter distributions from NERCOME for $n = 20$ and $n = 30$, respectively, are biased. The bias is minimal for $n = 40$ and $n = 100$. There is agreement between these findings and the results in Joachimi [2017] where the bias becomes high for very small numbers of mock data realizations. We therefore recommend the NERCOME estimator only for situations where the number of data realizations is higher than the number of elements of the data vector. There appears to be consistency in terms of the direction of the bias for the different parameters across

all the corner plots. The means of the fitting distributions for h from NERCOME are lower than those from the reference. This is true for all values of n. Similarly, the means of the distributions for b and f from NERCOME are both higher than the means of b and f from the reference for all n.

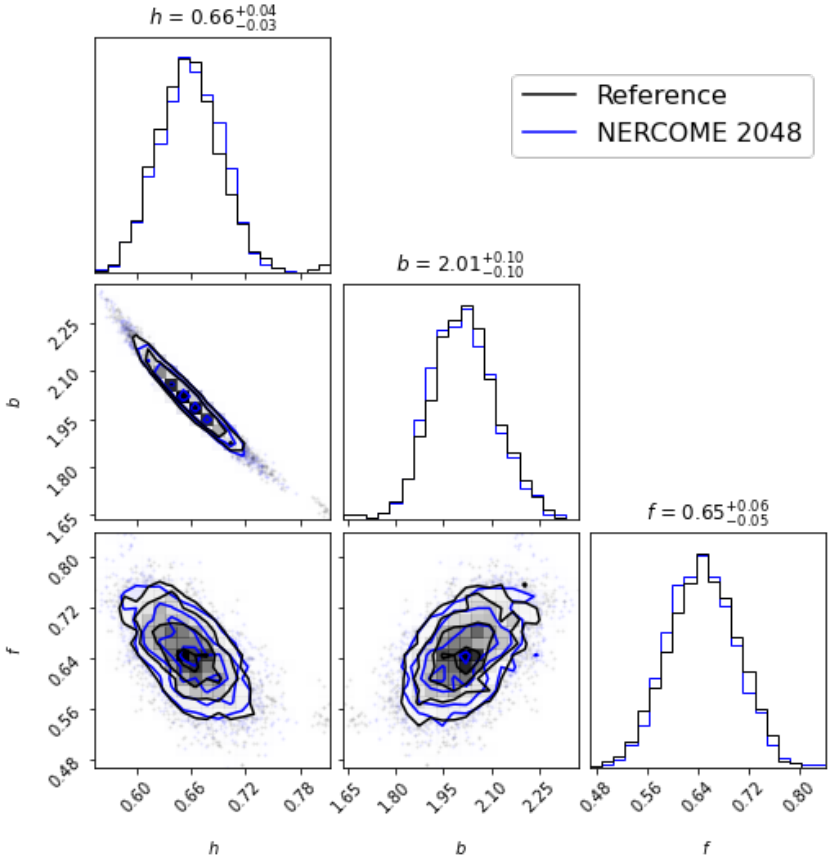

Figure 10: Comparison of parameter fitting distributions of h, b, and f for the reference covariance matrix and the NERCOME estimate from all 2048 mocks. The plot titles are the means of the distributions from the reference covariance matrix.

In order to see if this was induced by the design of the NERCOME algorithm or by noise in the mock data realizations, parameter fitting was also performed where we used a covariance matrix generated by the NERCOME algorithm from all 2048 mocks available. These results are displayed in figure 10. We can clearly see that the fitting distributions from the NERCOME covariance matrix estimate matches the reference distributions almost exactly. This indicates that the bias is induced by the limited number of mock catalogs available when $n = 20$ or $n = 30$ and is not an inherent issue in the NERCOME algorithm. It is inconclusive why the fitting distributions from the NERCOME covariance matrix estimate from $n = 20$ and $n = 30$ mocks are biased and more work needs to be done to explore this.

Both in this project and in Pope and Szapudi [2008], linear shrinkage outperforms traditional covariance matrix estimation techniques. This project used the sample estimator and Pope and Szapudi [2008] used Monte Carlo and Jackknife estimators. We used a target which differs from the one used in Pope and Szapudi [2008] in that we only included the diagonal entries of the sample estimate, i.e. variances, whereas Pope and Szapudi [2008] used a mix of variances and actual model power spectrum predictions as entries

(Eq. 29 in Pope and Szapudi [2008]). It is not known under what conditions our target consisting of variances on the diagonal 8 performs well and more research needs to be done to explore this.

Linear shrinkage also outperforms NERCOME for cases where there are fewer mock data realizations available than data vector elements. We therefore recommend linear shrinkage in these cases over NERCOME as it provides better accuracy in parameter fitting. Additionally, the linear shrinkage procedure is computationally more efficient than NERCOME due to the various averaging processes which form part of NERCOME. However, due to remaining uncertainty surrounding the choice of target in linear shrinkage, we stand by our recommendation of NERCOME for situations where we the number of mocks is greater than the number of data vector elements.

Finally, we draw attention to the fact that the power spectrum likelihood function (Eq. 11) depends on the inverse covariance matrix instead of the covariance matrix. Perhaps better accuracy can be achieved by focusing on estimating the inverse of the covariance matrix directly instead of estimating the covariance matrix itself and subsequently inverting the estimate.

6 Conclusion

The purpose of this research project was to explore if shrinkage estimation techniques can be used to find estimates of covariance matrices which outperform sample estimates. We have shown that we can indeed find such estimates using fewer mock data realizations and still obtain parameter distributions that are similar to distributions obtained when all 2048 mock data realizations were used to find the sample estimate. Both the linear shrinkage and NERCOME algorithms are promising. In cases where we have a number of mock data realizations that is higher than the number of elements in our data vector, we recommend the NERCOME algorithm. When we have fewer mock data realizations than elements in our data vector, we recommend linear shrinkage with the diagonal entries of the sample covariance matrix on the diagonal of the target covariance matrix.

Reducing the number of mock data realizations necessary to do data analysis in cosmology is in the interest of the cosmological community as generating such mocks is computationally very expensive. Future galaxy surveys such as DESI and EUCLID will have a higher precision than past surveys and hence we can expect that the number of mock data realizations necessary needs to increase accordingly if we want to draw conclusions with similar precision, provided the same tools will be used such as the sample covariance matrix estimate. Employing shrinkage estimation instead of the sample estimate reduces the number of mocks required and therefore reduces computational costs involved.

7 Acknowledgements

I acknowledge financial support from the School of Physics & Astronomy at the University of Edinburgh through the award of a Career Development Summer Scholarship. I would also like to express my sincere gratitude to my supervisor, Mike, for offering me the opportunity to do this project and for his guidance and advice during the course of it. Finally, I want to say thanks to the cosmology research group, and to Florian in particular, for helpful discussions and for including me in the weekly meetings and coffee chats.

8 Personal Statement

Over the course of this project, I have had the opportunity to develop skills, both technical and non-technical, which will help me in my future career. First of all, I have improved my coding skills in Python and I learned about statistical tools such as likelihood analysis and MCMC methods. Moreover, I have practiced my presentation skills by giving a total of three presentations about my work and I have demonstrated my ability to work in a team by sharing intermediate progress and results during group meetings and by consulting with other research group members whenever I had questions. Most importantly, I have increased my critical thinking and problem-solving capabilities by analyzing and interpreting my results and coming up with creative solutions to solve problems. Finally, by writing this project report, I have had the opportunity to practice my writing skills.

9 Lay Summary

Usually, physicists and astronomers make multiple measurements of the same event to determine the accuracy of their measurements. If all the values are close to each other, then the measurement was accurate. The uncertainties in the measurements are described by looking at the deviation of all the measured values from the expected value. This no longer works when one can only make one measurement, which is usually the case in cosmology. Cosmologists therefore simulate events and make simulated measurements to make up for the lack of real measurements. Performing these simulations on a computer takes a lot of time so ideally one wants to limit the number of simulations required while still having enough simulated values to make accurate predictions. In this project, I looked at methods of estimating uncertainties which are as good as usual, but which require fewer simulations than are normally required in cosmology.

References

Shadab Alam et al. THE ELEVENTH AND TWELFTH DATA RELEASES OF THE SLOAN DIGITAL SKY SURVEY: FINAL DATA FROM SDSS-III. *The Astrophysical Journal Supplement Series*, 219(1):12, July 2015. ISSN 0067-0049. doi: 10.1088/0067-0049/219/1/12. URL https://doi.org/10.1088/0067-0049/219/1/12. Publisher: American Astronomical Society.

Florian Beutler and Patrick McDonald. Unified galaxy power spectrum measurements from 6dFGS, BOSS, and eBOSS. *Journal of Cosmology and Astroparticle Physics*, 2021(11):031, November 2021. ISSN 1475-7516. doi: 10.1088/1475-7516/2021/11/031. URL http://arxiv.org/abs/2106.06324. arXiv:2106.06324 [astro-ph].

Florian Beutler, Emanuele Castorina, and Pierre Zhang. Interpreting measurements of the anisotropic galaxy power spectrum. *Journal of Cosmology and Astroparticle Physics*, 2019(03):040–040, March 2019. ISSN 1475-7516. doi: 10.1088/1475-7516/2019/03/040. URL http://arxiv.org/abs/1810.05051. arXiv:1810.05051 [astro-ph].

Daniel Foreman-Mackey. corner.py: Scatterplot matrices in python. *The Journal of Open Source Software*, 1(2):24, jun 2016. doi: 10.21105/joss.00024. URL https://doi.org/10.21105/joss.00024.

Nick Hand, Yu Feng, Florian Beutler, Yin Li, Chirag Modi, Uros Seljak, and Zachary Slepian. nbodykit: an open-source, massively parallel toolkit for large-scale structure. *The Astronomical Journal*, 156(4):160, September 2018. ISSN 1538-3881. doi: 10.3847/1538-3881/aadae0. URL http://arxiv.org/abs/1712.05834. arXiv:1712.05834 [astro-ph].

J. Hartlap, P. Simon, and P. Schneider. Why your model parameter confidences might be too optimistic – unbiased estimation of the inverse covariance matrix. *Astronomy & Astrophysics*, 464(1):399–404, March 2007. ISSN 0004-6361, 1432-0746. doi: 10.1051/0004-6361:20066170. URL http://arxiv.org/abs/astro-ph/0608064. arXiv:astro-ph/0608064.

Benjamin Joachimi. Non-linear shrinkage estimation of large-scale structure covariance. *Monthly Notices of the Royal Astronomical Society: Letters*, 466(1):L83–L87, March 2017. ISSN 1745-3925, 1745-3933. doi: 10.1093/mnrasl/slw240. URL http://arxiv.org/abs/1612.00752. arXiv:1612.00752 [astro-ph].

Francisco-Shu Kitaura et al. The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: mock galaxy catalogues for the BOSS Final Data Release. *Monthly Notices of the Royal Astronomical Society*, 456(4):4156–4173, March 2016. ISSN 0035-8711. doi: 10.1093/mnras/stv2826. URL https://doi.org/10.1093/mnras/stv2826.

Clifford Lam. Nonparametric eigenvalue-regularized precision or covariance matrix estimator. *The Annals of Statistics*, 44(3), June 2016. ISSN 0090-5364. doi: 10.1214/15-AOS1393. URL https://projecteuclid.org/journals/annals-of-statistics/volume-44/issue-3/Nonparametric-eigenvalue-regularized-precision-or-covariance-matrix-estimator/10.1214/15-AOS1393.full.

Olivier Ledoit and Michael Wolf. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, page 19, 2003.

Adrian C. Pope and István Szapudi. Shrinkage Estimation of the Power Spectrum Covariance Matrix. *Monthly Notices of the Royal Astronomical Society*, 389(2):766–774, September 2008. ISSN 00358711, 13652966. doi: 10.1111/j.1365-2966.2008.13561.x. URL http://arxiv.org/abs/0711.2509. arXiv:0711.2509 [astro-ph].

Beth Reid et al. SDSS-III Baryon Oscillation Spectroscopic Survey Data Release 12: galaxy target selection and large-scale structure catalogues. *Monthly Notices of the Royal Astronomical Society*, 455(2):1553–1573, January 2016. ISSN 1365-2966, 0035-8711. doi: 10.1093/mnras/stv2382. URL http://academic.oup.com/mnras/article/455/2/1553/1112409/SDSSIII-Baryon-Oscillation-Spectroscopic-Survey.

Shun Saito. Galaxy clustering in redshift space, June 2016. URL https://wwwmpa.mpa-garching.mpg.de/~komatsu/lecturenotes/Shun_Saito_on_RSD.pdf.

Juliane Schäfer and Korbinian Strimmer. A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1), January 2005. ISSN 1544-6115, 2194-6302. doi: 10.2202/1544-6115.1175. URL https://www.degruyter.com/document/doi/10.2202/1544-6115.1175/html.