Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **472,** 4244–4250 (2017)
Advance Access publication 2017 September 12

doi:10.1093/mnras/stx2326

Massive data compression for parameter-dependent covariance matrices

Alan F. Heavens,[1]* Elena Sellentin,[1,2] Damien de Mijolla[1] and Alvise Vianello[1]

[1] *Imperial Centre for Inference and Cosmology (ICIC), Astrophysics, Imperial College, Blackett Laboratory, Prince Consort Road, London SW7 2AZ, UK*
[2] *Département de Physique Théorique, Université de Genève, Quai Ernest-Ansermet 24, CH-1211 Genève, Switzerland*

Accepted 2017 September 5. Received 2017 September 5; in original form 2017 July 20

ABSTRACT

We show how the massive data compression algorithm MOPED can be used to reduce, by orders of magnitude, the number of simulated data sets which are required to estimate the covariance matrix required for the analysis of Gaussian-distributed data. This is relevant when the covariance matrix cannot be calculated directly. The compression is especially valuable when the covariance matrix varies with the model parameters. In this case, it may be prohibitively expensive to run enough simulations to estimate the full covariance matrix throughout the parameter space. This compression may be particularly valuable for the next generation of weak lensing surveys, such as proposed for *Euclid* and Large Synoptic Survey Telescope, for which the number of summary data (such as band power or shear correlation estimates) is very large, $\sim 10^4$, due to the large number of tomographic redshift bins which the data will be divided into. In the pessimistic case where the covariance matrix is estimated separately for all points in an Monte Carlo Markov Chain analysis, this may require an unfeasible 10^9 simulations. We show here that MOPED can reduce this number by a factor of 1000, or a factor of $\sim 10^6$ if some regularity in the covariance matrix is assumed, reducing the number of simulations required to a manageable 10^3, making an otherwise intractable analysis feasible.

Key words: methods: data analysis – methods: statistical.

1 INTRODUCTION

Many problems concern data which are Gaussian-distributed, either as a result of some underlying physical process, or by virtue of the central limit theorem. The sampling distribution then depends only on the mean and the covariance matrix of the data, and inference of model parameters then follows with the use of a likelihood which is a multivariate Gaussian function of the data. One challenge that can be considerable is if the covariance matrix cannot be calculated readily, and the experiment has to be simulated and the covariance matrix estimated from the simulated data. In principle this is not difficult, but it can be expensive to do, since at least $p + 3$ simulations are required, where p is the number of data. If the number of simulations is less than this, the expectation of the precision matrix (the inverse of the covariance matrix) diverges. Ideally one would like many more than $p + 3$, in order for the estimated covariance matrix to be precise. Furthermore, if the covariance matrix depends on the model parameters, then it may be a severe challenge: for Bayesian inference using, for example Monte Carlo Markov Chains (MCMCs), the covariance matrix might in the worst case be estimated at each point in parameter space that is sampled.

If it is impractical to perform so many simulations, then some savings may be made by regularizing the behaviour of the covariance matrix, but in addition we can markedly improve the situation by reducing the number of data points p, in some cases by orders of magnitude. In the same spirit, Asgari & Schneider (2015) proposed a more modest level of linear compression of COSEBI statistics. In general, this will lose information, but we previously published an algorithm MOPED[1] (Heavens, Jimenez & Lahav 2000) which can massively reduce the number of data points, without losing information, in the sense that the Fisher matrix is unchanged by the data compression, subject to certain conditions. The MOPED algorithm reduces the size of the data set from p to m, where m is the number of parameters in the model, and this can be a dramatic reduction in the data set size with little or no loss of information. It has been successfully applied to determine the star formation history of galaxies (Reichardt, Jimenez & Heavens 2001; Heavens et al. 2004; Panter et al. 2007), and investigated for data compression in the cosmic microwave background (Gupta & Heavens 2002; Zablocki & Dodelson 2016) and in gravitational waves (Graff, Hobson & Lasenby 2011).

MOPED is therefore an interesting candidate to tackle the issue of experiments with large data sets, relatively few model parameters

* E-mail: a.heavens@imperial.ac.uk

[1] Massively Optimised Parameter Estimation and Data compression.

and covariance matrices which need to be simulated. In this paper, we explore how effective MOPED can be in such situations, finding that it can reduce enormously the computational requirements to analyse such experiments, at the expense of a small increase in the parameter errors compared with the ideal, but unattainable, analysis.

The second element in this paper is that when the covariance matrix is estimated, then the true covariance matrix needs to be marginalized over, as shown by Sellentin & Heavens (2016), leading to a modified t-distribution. This leads to a modification of the credible regions, increasing them at low credibility levels but maintaining a compact core. The situation may be more complicated if one has some prior knowledge of the covariance matrix, or there is one part of it which is known. This subject has become very topical in the light of the expected data set size of future cosmology surveys such as *Euclid* and the Large Synoptic Survey Telescope (LSST), and as a result, much attention is being devoted to this issue. As one application, it is expected that next-generation photometric surveys will be split into ∼10 tomographic bins of redshift, so with ∼25 band-powers in frequency (or separations, if configuration-space statistics such as correlation functions are used), then the total number of summary data, including auto- and cross-correlations, and E and B (or ξ_+ and ξ_-), is $\sim 6 \times 10^3$, or higher if one also investigates $E-B$ correlations to test isotropy.

See Sellentin & Heavens (2017) and Blot et al. (2016), Dodelson & Schneider (2013), and Percival et al. (2014), Taylor & Joachimi (2014) for assessments of the increase in errors due to uncertainties in the covariance matrix, and for further discussion, see Joachimi (2017), Friedrich & Eifler (2016) and Padmanabhan et al. (2016), Petri, Haiman & May (2016), Pope & Szapudi (2008).

2 THE MOPED ALGORITHM

Here, we review and extend the MOPED algorithm as originally presented in Heavens et al. (2000). MOPED forms linear combinations of the data $\boldsymbol{x}$ (which has length p), using a set of MOPED vectors $\boldsymbol{b}_\alpha$, where $\alpha = 1 \ldots m$, and m is the number of parameters, each of which is contained in an ordered list represented by a vector $\boldsymbol{\theta}$. They compress the data to a set of MOPED coefficients

$$y_\alpha = \boldsymbol{b}_\alpha^T \boldsymbol{x}. \qquad (1)$$

The MOPED vectors are chosen in sequence, according to the following algorithm: the first is the linear combination which minimizes the expected conditional error on parameter $\boldsymbol{\theta}_1$. i.e. it maximizes the matrix element $\mathbf{F}^y_{11}$, where $\mathbf{F}^y_{\alpha\beta} = -\langle \partial^2 \ln L^y / \partial\theta_\alpha \partial\theta_\beta \rangle$ is the Fisher matrix for the $\boldsymbol{y}$ data set, and $L^y = p(\boldsymbol{y}|\boldsymbol{\theta})$ is the likelihood of the compressed data. Subsequent $\boldsymbol{b}_\alpha$ vectors are chosen to maximize $\mathbf{F}^y_{\alpha\alpha}$ $(\alpha > 1)$, subject to the $\boldsymbol{b}$ vectors being orthogonal to the previous vectors, in the specific sense that the y_α are uncorrelated. This requires $\boldsymbol{b}_\alpha^T \mathbf{C} \boldsymbol{b}_\beta = \delta_{\alpha\beta}$, if we also normalize the MOPED coefficients to unit variance.

For Gaussian data, the Fisher matrices ($\mathbf{F}^y$ and the analogue $\mathbf{F}^x$ for the original data set) are computed from (Tegmark, Taylor & Heavens 1997)

$$\mathbf{F}_{\alpha\beta} = \frac{1}{2}\mathrm{Tr}\left[\mathbf{C}^{-1}\mathbf{C}_{,\alpha}\mathbf{C}^{-1}\mathbf{C}_{,\beta} + \mathbf{C}^{-1}\left(\boldsymbol{\mu}_{,\alpha}\boldsymbol{\mu}_{,\beta}^T + \boldsymbol{\mu}_{,\beta}\boldsymbol{\mu}_{,\alpha}^T\right)\right]. \qquad (2)$$

where $\boldsymbol{\mu} \equiv \langle \boldsymbol{x}(\boldsymbol{\theta}) \rangle$ or $\langle \boldsymbol{y}(\boldsymbol{\theta}) \rangle$ are length p or length m expected data vectors for the full or compressed data sets, respectively. $\mathbf{C}$ is the $p \times p$ or $m \times m$ covariance matrix of the data, and a comma indicates a partial derivative with respect to the labelled parameter. Generally if there is no superscript on $\mathbf{C}$, it will refer to the original data vector $\boldsymbol{x}$, but we will identify $\mathbf{C}$ with a superscript x or y if extra clarity is required. We assume that $\boldsymbol{\mu}$ and its derivatives can be computed via theoretical or computational methods.

In its previous applications, MOPED has made the assumption that the covariance matrix is independent of the parameters. This assumption is relaxed in this paper, and we can properly account for the parameter dependence. MOPED also requires a fiducial set of parameters to be chosen, since the Fisher matrix depends on derivatives of $\boldsymbol{\mu}$ as well as the covariance matrix. The solutions for the optimized weighting vectors in equation (1) are

$$\boldsymbol{b}_1 = \frac{\mathbf{C}^{-1}\boldsymbol{\mu}_{,1}}{\sqrt{\boldsymbol{\mu}_{,1}^T \mathbf{C}^{-1}\boldsymbol{\mu}_{,1}}} \qquad (3)$$

and

$$\boldsymbol{b}_\alpha = \frac{\mathbf{C}^{-1}\boldsymbol{\mu}_{,\alpha} - \sum_{\beta=1}^{\alpha-1}\left(\boldsymbol{\mu}_{,\alpha}^T \boldsymbol{b}_\beta\right)\boldsymbol{b}_\beta}{\sqrt{\boldsymbol{\mu}_{,\alpha}^T \mathbf{C}^{-1}\boldsymbol{\mu}_{,\alpha} - \sum_{\beta=1}^{\alpha-1}\left(\boldsymbol{\mu}_{,\alpha}^T \boldsymbol{b}_\beta\right)^2}} \qquad 1 < \alpha \le m, \qquad (4)$$

where $\mathbf{C}$ and $\boldsymbol{\mu}_{,\alpha}$ are evaluated at the fiducial parameter set.

It can be shown (Heavens et al. 2000) that if the fiducial parameters coincide with the true parameters, and the Fisher matrix is dominated by the second term of equation (2), then the compression is locally lossless, defined by $\mathbf{F}^x = \mathbf{F}^y$. In the case that the covariance matrix does not depend on the parameters, the covariance matrix of the compressed data is by construction very simple everywhere. If we define $\mathbf{B}$ to be a $p \times m$ matrix of which the columns are the $\boldsymbol{b}$ vectors, then the compressed data vector is $\boldsymbol{y} = \mathbf{B}^T\boldsymbol{x}$, and from the orthogonality condition of the $\boldsymbol{b}$ vectors,

$$\mathbf{C}^y = \mathbf{B}^T\mathbf{C}\mathbf{B} = \mathbf{I}_m \qquad (5)$$

i.e. the $m \times m$ identity matrix. This makes parameter inference with MOPED extremely fast, as the likelihood involves only $O(m)$ operations, rather than the $O(p^3)$ operations for the full data set, provided that the covariance matrix is independent of the model parameters. Note that the method is completely general: the data and the model can be anything. In this paper, we use as an illustrative example the pixellized intensities of a galaxy image as the data vector, and an exponential light profile as the model. Another example, which we do not explore, is to take as the data vector the estimates of the shear correlation functions in a weak lensing analysis. A specific example of this is the CFHTLenS analysis of Heymans et al. (2013), which had $p = 210$ shear correlation measurements and a model with $m = 6$ parameters, so the gains would be considerable in this case. However, since various assumptions in deriving equation (5) are violated in practice, in this paper we do not assume that $\mathbf{C}^y$ is the identity matrix, but we estimate it with simulations (see Section 4).

3 PARAMETER INFERENCE WITH ESTIMATED COVARIANCE MATRICES

The fact that the covariance matrix is not known but is estimated changes the likelihood function. As pointed out by Kaufman (1967), even if the estimated covariance matrix is unbiased, its inverse is not, and needs to be multiplied by a factor $\alpha = (N - p - 2)/(N - 1)$ to make it so, where p is the number of data and N the number of simulations. This was introduced into astronomy by Hartlap, Simon & Schneider (2007). In fact it is strictly incorrect to retain the Gaussian form and use an unbiased inverse covariance matrix; rather one should marginalize over the true covariance matrix, given its estimate. Since the estimate, which we denote by $\mathbf{S}$, follows a Wishart distribution, this can be done analytically, and the solution is given by Sellentin & Heavens (2016), yielding a likelihood which is a modified t-distribution:

$$\ln p(\boldsymbol{x}|\boldsymbol{\mu}, \mathbf{S}, N) = \text{const.} - \frac{N}{2} \ln\left[1 + \frac{(\boldsymbol{x}-\boldsymbol{\mu})^T \mathbf{S}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{N-1}\right]. \quad (6)$$

In the limit $N \gg p$, this approaches the original Gaussian distribution, but in general it has a narrower core and wider tails.

If we form linear combinations of the original data, as here with the MOPED compression, $\boldsymbol{y} = \mathbf{B}^T\boldsymbol{x}$, the estimated covariance matrix of the compressed data $\mathbf{S}^y$ is also Wishart distributed (with scale matrix $\mathbf{C}^y/n_c$ and degrees of freedom $n_c = N_c - 1$, where N_c is the number of simulations of the compressed data). The same marginalization then applies, and the likelihood of $\boldsymbol{y}$ is given by the t-distribution of equation (6) but with $\boldsymbol{x} \to \boldsymbol{y}$, $\mathbf{S} \to \mathbf{S}^y$ and $N \to N_c$, and $\boldsymbol{\mu}$ is the expectation value of $\boldsymbol{y}$. However, the big advantage of the compression is that the enlargement of the credible regions due to the uncertainty in the covariance matrix is small provided only that $N_c \gg m$, which requires far fewer simulations than when using the full data set, if $p \gg m$.

4 METHOD FOR A COVARIANCE MATRIX THAT DEPENDS ON MODEL PARAMETERS

It is important to realize that we can choose to make any linear compression of the data, whether it is locally lossless or not. Hence, we can apply a MOPED compression to the data even if the assumptions in its derivation are violated so that the compression is not optimal. The parameter inference would still be entirely valid; the credible regions would just be larger than they could be. Past investigations (Heavens et al. 2000; Gupta & Heavens 2002) have shown that in practical cases, the increase in parameter credible reasons is usually negligibly small. There is a subtlety in that the inference will be correct provided that the compressed covariance matrix is correct. In typical MOPED applications, the compressed covariance matrix has been assumed to be fixed at the identity, and this gives very rapid inference. However, it is an approximation if $\mathbf{C}$ depends on parameters. If more accuracy and precision are required, an iterative solution is to find the most probable parameters and then repeat the MOPED data compression with the solution as the fiducial model. In practical applications, we have not found this to be necessary, but strictly we should use the correct covariance matrix appropriate for the position in parameter space. This is what we do in this paper.

An alternative to the assumption that the compressed covariance matrix is the identity is to compute it directly from $\mathbf{B}$ and $\mathbf{C}$, or equivalently to simulate $\boldsymbol{x}$ and then form $\boldsymbol{y}$ by matrix multiplication, and estimate $\mathbf{S}^y$ from the simulated $\boldsymbol{y}$ vectors. This then dispenses with an approximation, but the cost of the matrix operations in normal applications then negates the massive computational speed advantage of MOPED. However, in this paper we are considering the situation where the time is dominated by the time to estimate the covariance matrix, not to evaluate the likelihood, so the matrix operations used to generate $\boldsymbol{y}$ come at negligible cost.

The method we advocate is this: create N simulated data sets, $\boldsymbol{x}_{(i)}$; $i = 1 \ldots N$, for a fiducial set of model parameters, in order to obtain an unbiased estimate $\mathbf{S}$ for the full (fiducial) covariance matrix:

$$\mathbf{S} = \frac{1}{N-1}\sum_{i=1}^{N}(\boldsymbol{x}_{(i)} - \bar{\boldsymbol{x}})(\boldsymbol{x}_{(i)} - \bar{\boldsymbol{x}})^T. \quad (7)$$

We then use this to pre-compute a set of MOPED compression vectors, using equations (3) and (4) but with $\mathbf{C}$ replaced by its estimate $\mathbf{S}$. This set will be close to optimal, provided that the chosen fiducial model is correct, and the covariance matrix $\mathbf{S}$ has been estimated sufficiently from many simulations to be a good approximation to $\mathbf{C}$. After this point, we keep the MOPED $\boldsymbol{b}$ vectors fixed, and do not vary them during the parameter inference phase. If this preliminary step is already too expensive in terms of computer time, then an alternative approach is to use an approximate covariance matrix, perhaps theoretically generated on the basis of assumptions that do not precisely hold. The MOPED vectors would not be optimal in this case, but may be close enough that the information loss is small.

When inferring parameters (via say MCMC chains), we again make an estimate of a covariance matrix, but this time we form $\mathbf{S}^y$ as an estimate for the compressed data covariance matrix $\mathbf{C}^y$, using

$$\mathbf{S}^y = \frac{1}{N_c-1}\sum_{i=1}^{N_c}(\boldsymbol{y}_{(i)} - \bar{\boldsymbol{y}})(\boldsymbol{y}_{(i)} - \bar{\boldsymbol{y}})^T. \quad (8)$$

The advantage that we have is that we require only $m + 3$ or more simulations, rather than the typically much larger $p + 3$. If the simulations are expensive, this could still be a considerable cost, depending on how much the covariance matrix depends on the parameters, but it may make the analysis feasible when otherwise it might be essentially impossible (if $p \gg m$, as is typical).

5 EXAMPLE PROBLEM

Let us illustrate with a simple $m = 2$ parameter model, representing a circularly symmetric image of a galaxy with an exponential surface brightness profile. Ignoring complications of finite pixel size, the model is that the pixel brightness values are

$$\boldsymbol{\mu}(\boldsymbol{r}) = A\exp(-a|\boldsymbol{r}|) \quad (9)$$

where $\boldsymbol{r}$ is the pixel position vector, of length n_{pix}. A and a are the model parameters. Purely for simplicity in this illustrative example, we assume that the true covariance matrix is proportional to the identity, $\mathbf{C} = \sigma^2\,\mathbf{I}_p$, where σ^2 is the pixel variance. In this initial example, we will not vary σ^2 with the parameters.

In order to estimate $\mathbf{C}$, we generate N simulated data sets and evaluate $\mathbf{S}$ via equation (7). For a given N, we compute the two MOPED vectors using equation (3) and (4), with $\mathbf{C}$ replaced by $\mathbf{S}$. We then generate a test image and compute the likelihood of A and a using the compressed data. In this example, we estimate the compressed covariance only once, from N_c simulations of the full data set, which are then compressed, and use equation (8). We then compute the posterior of A and a given the estimated compressed covariance matrix, by analytically marginalizing over the unknown covariance matrix, and using the likelihood of Sellentin & Heavens (2016).

In Fig. 1, we show contours of the likelihood, for a case when the covariance matrix is independent of parameters, so the compressed covariance matrix is estimated only once. We see that MOPED is very effective when the covariance matrix is known and the fiducial model is the correct one. In more realistic cases, when the covariance matrices for the full data and for the compressed data have to be estimated, then the compression is not locally lossless except in the limits $N, N_c \to \infty$. The large size of the outer contour in the bottom right panel comes from the broad wings of the Sellentin & Heavens (2016) likelihood, whereas the contours containing ~68 per cent and ~95 per cent of the posterior are not much larger than in the ideal case.

If the covariance matrix depends on the parameters of the model, then the analysis is much more challenging. The covariance matrix may need to be estimated separately each time a new point in

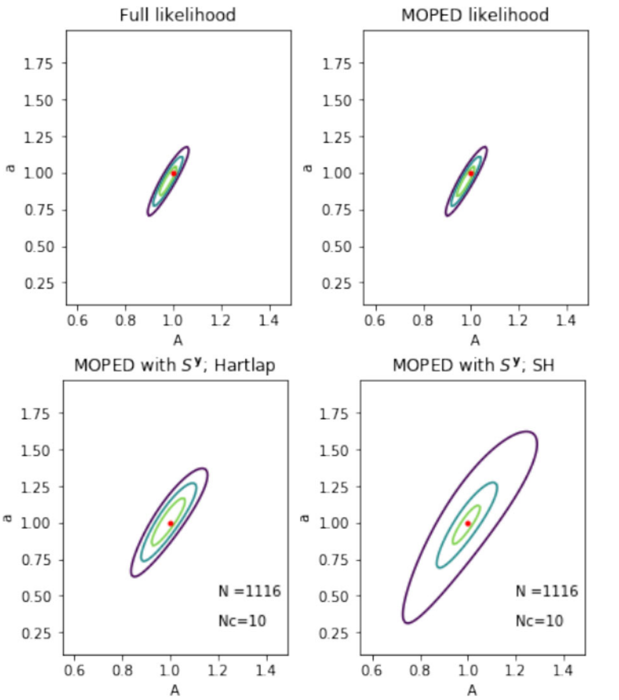

Figure 1. Posteriors for the parameters in the model $\mu = A\,\exp(-ar)$ for 400 pixels and noise per pixel of 0.1. True values ($A = 1, a = 1$) are marked by the red dot, and contours are at levels $\delta\,\ln L = -2.3, -6.2, -11.8$, corresponding to 1σ, 2σ, 3σ, two-parameter credible regions of Gaussian likelihoods. From top left, clockwise: full likelihood with known covariance matrix (best possible case); MOPED compression, using correct full and compressed covariance matrix, and correct fiducial model; compressed analysis using the likelihood of Sellentin & Heavens (2016), with 1116 simulations used to determine the MOPED vectors, and only 10 to estimate the compressed covariance matrix; same, with a Gaussian likelihood, using the Hartlap et al. (2007) scaling, where the inner contour is too large and the outer one is too small. Here, we assume that the covariance matrix is parameter-independent, so we estimate it only once.

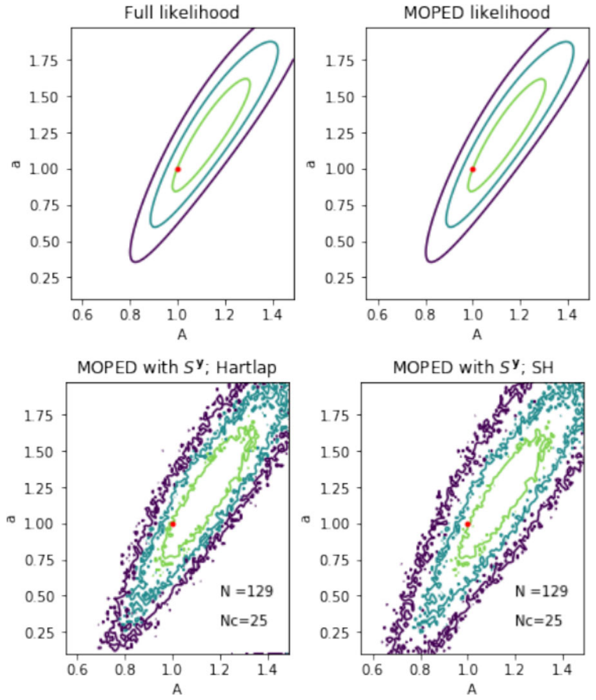

Figure 2. Similar to Fig. 1, except that we estimate the compressed covariance matrix separately at each point in the parameter space grid. This would be required if the covariance matrix varied with the parameters of the model, when brute force estimation of the covariance matrix everywhere might be impractical without the data compression proposed here. There are 25 pixels in this example, the MOPED vectors are determined from 129 simulations, and 25 simulations are used to estimate the compressed covariance matrix at each point.

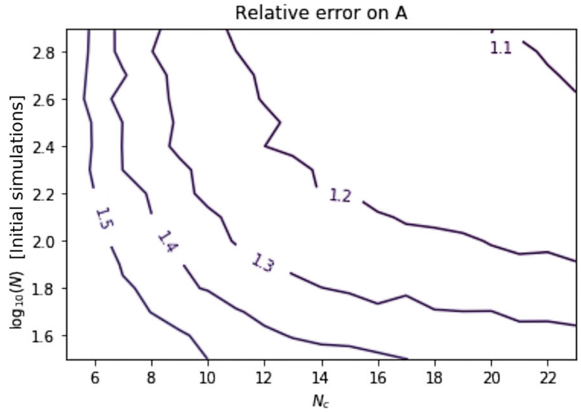

Figure 3. Posterior standard deviation for the amplitude parameter A, relative to the ideal error when the covariance matrix is known and the full data set is used. The contour labels refer to the relative increase of the standard deviation. The vertical axis is the number of initial simulations used to estimate the MOPED vectors. The horizontal axis is the number of simulations used to compute the covariance matrix at different points in parameter space. In this case, the image is a square of $p = n_{\rm side}^2 = 25$ pixels, $\sigma = 0.3$ and the plot is averaged over 500 realizations.

parameter space is considered. Fig. 2 is an illustration of this, where we estimate the compressed covariance matrix afresh at every point in the parameter grid. Since the estimated covariance matrix is a random object, this adds noise to the posterior, which might benefit from some smoothing. In practice some sort of regularization procedure would almost certainly be employed for the covariance matrix, which would smooth the contours.

In Figs 3 and 4, we show the relative increase in error compared with the ideal case (where we use the true covariance matrix and the full data set, or indeed the MOPED compressed data assuming the correct fiducial model and covariance matrix; they are essentially identical), as a function of the number of simulations N and N_c, or the Hartlap parameters $\alpha = (N - p - 2)/(N - 1)$ and $\alpha_c = (N_c - m - 2)/(N_c - 1)$. To produce these figures, we simulate images and compute the marginal credible regions by integration of the 2D posterior, and average over 500 realizations. In Fig. 3, we see that there is little to be gained on increasing N beyond 200 ($\log_{10}N = 2.3$), for this relatively small image of $p = 25$ pixels.

In these examples, we have chosen data sets of different sizes, $p = 400$ and $p = 25$ pixels. The time for inversion of a $p \times p$ matrix scales as p^3 (although iterative techniques may be faster), whereas the size of the compressed data set is $m = 2$ in both cases, so the time taken is less dependent on the original data set size. The timings would scale as p^2, arising both from the time required to generate each sample image used in the estimation of the compressed likelihood, and also from the scalar products which compress the image data. The generation of the MOPED vectors scales as p^3, but this is done only once, and not at each point where the posterior is computed.

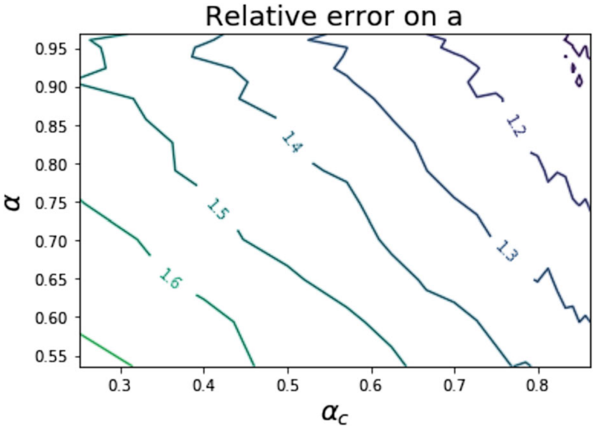

Figure 4. As in Fig. 3, but here for the scalelength parameter a, and plotted against $\alpha = (N - p - 2)/(N - 1)$ (y axis) and $\alpha_c = (N_c - m - 2)/(N_c - 1)$ (x axis), where $m = 2$ is the number of compressed data, and $p = 25$ is the number of pixels.

5.1 A more complex model explored with MCMC

In Fig. 5, we show the effect of MOPED compression on a more complex four-parameter model, which we explore with more typical MCMC techniques. In this model, the model represents a circular exponential profile disc, seen at an angle, and resulting in a surface brightness distribution (see Heavens, Alsing & Jaffe 2013 for more details):

$$\mu(r, \psi | a, \epsilon, \phi, A) = A \exp\left[-a\, r \sqrt{1 + \epsilon^2 - 2\epsilon \cos 2(\psi - \phi)}\right] \quad (10)$$

where A is the central surface brightness, a is the inverse semimajor axis, ϕ its position angle and ϵ is the (magnitude of the) ellipticity of the galaxy. r and ψ are polar coordinates about the centre of the galaxy, whose position is assumed to be known. Posteriors for the parameters are obtained using Stan (Carpenter et al. 2017). An image is generated with $a = A = \phi = 1.0$ and $\epsilon = 0.25$, on a 10×10 grid. In this case, we make the covariance matrix parameter dependent, assuming white noise, but with a pixel variance that depends on the central amplitude parameter: $\sigma^2 = 0.01A$.

In Fig. 5, we plot a comparison of the different possibilities for analysing this data set. In the top left, we plot the posterior gained from the full data set of size $p = 100$, using the known covariance matrix. Flat priors on the parameters were assumed. This panel depicts the maximal information content on the parameters to be measured. In the bottom left of Fig. 5, we still assume the correct covariance matrix is known, but we now apply MOPED compression. The MOPED compression vectors are hence determined from the correct covariance matrix, with the fiducial model coinciding with the true model. The covariance for the compressed data set is then the 4×4 identity matrix. In the top right, we see the effect of determining the covariance matrix of the full data set from 1000 simulations, for the purpose of determining the MOPED vectors. The MOPED compression vectors are therefore not quite optimal, but we still assume the compressed covariance matrix is the identity. Finally, in the bottom right we show the actual target of MOPED compression in cosmology: parameter dependence in the covariance matrix is now included, and each time a compressed covariance matrix is estimated from 10 simulations only but at each point in the chain. The compressed covariance matrix is then marginalized over using the Sellentin & Heavens (2016) likelihood. The likelihood is computed with Stan in a simple hierarchical model, where the covariance matrix is a random object. Note that 10 simulations are not in the asymptotic regime where the compressed covariance matrix is very well determined, so we expect to see a degradation of the errors. We also see that the effect of sampling is to obscure the variability that is apparent in Fig. 2. Note that in this last case, the outer contours are again broadened because of the marginalization over the true covariance matrix. The inner contours are only moderately larger than in the other figures, reflecting the small core and broad wings of the Sellentin–Heavens likelihood. The result of the non-optimal MOPED vectors, and the marginalization over the compressed covariance matrix are to increase the errors, by approximately 50–100 per cent in this case. However, the compression has now successfully accomplished the otherwise unfeasible task of computing the parameter-dependence of the (now compressed) covariance matrix at each point of the MCMC chain.

6 CONCLUSIONS

In this paper, we have considered the relatively common situation of parameter inference from Gaussian-distributed data (of length p), where the covariance matrix is not directly calculable, but has to be simulated. In the case where the covariance matrix varies with parameters, this can lead to a requirement for an unfeasibly large number of simulations, especially if the covariance matrix were to be evaluated separately at each sample point in the m-dimensional parameter space. We have shown that this can be speeded up by a very large factor, with little loss of information, by compressing the data using the MOPED algorithm first. The algorithm proposed is to run a very large number $N \gg p + 2$ of simulations with the model parameters kept fixed at some fiducial values, if this is feasible, and to use the resulting estimated covariance matrix for the full data set to define a set of near-optimal MOPED data compression vectors, which are then kept fixed. When sampling the parameter space, using MCMC for example, the much smaller compressed covariance matrix may be estimated accurately from far fewer simulations, requiring only $N_c > m + 2$, which is typically much less than p.

It is clear that there is some trade-off between running many simulations to define the MOPED vectors, and running more simulations during the MCMC phase, but Figs 3 and 4 indicate that it is likely that the best strategy will be to run $N \gg p$ simulations for a fiducial parameter choice, since an accurate full covariance matrix delivers MOPED vectors which are closer to optimal, and which thus require fewer compressed simulations when the parameter space is sampled. However, it may be that this reduction in the number of simulations is still inadequate, and there are various possibilities to overcome this.

For the MOPED vectors, it may be adequate to have an approximate full covariance matrix, determined without simulations. It may not yield optimal compression vectors, but the compression is likely still to be useful. Secondly, to reduce the number of simulations for the compressed stage, one could use some interpolation in the parameter space, estimating the compressed covariance matrix only at a relatively small number of locations.

Emulator-based methods, for example based on a Latin hypercube, may be effective (Heitmann et al. 2009, 2010), even with only ∼100 simulations. Such a scheme was proposed by Morrison & Schneider (2013), using Gaussian processes to interpolate between the covariance matrices. An alternative approximate approach would be to estimate the covariance matrix at a fiducial point, and estimate the generator of the linear part of the variation

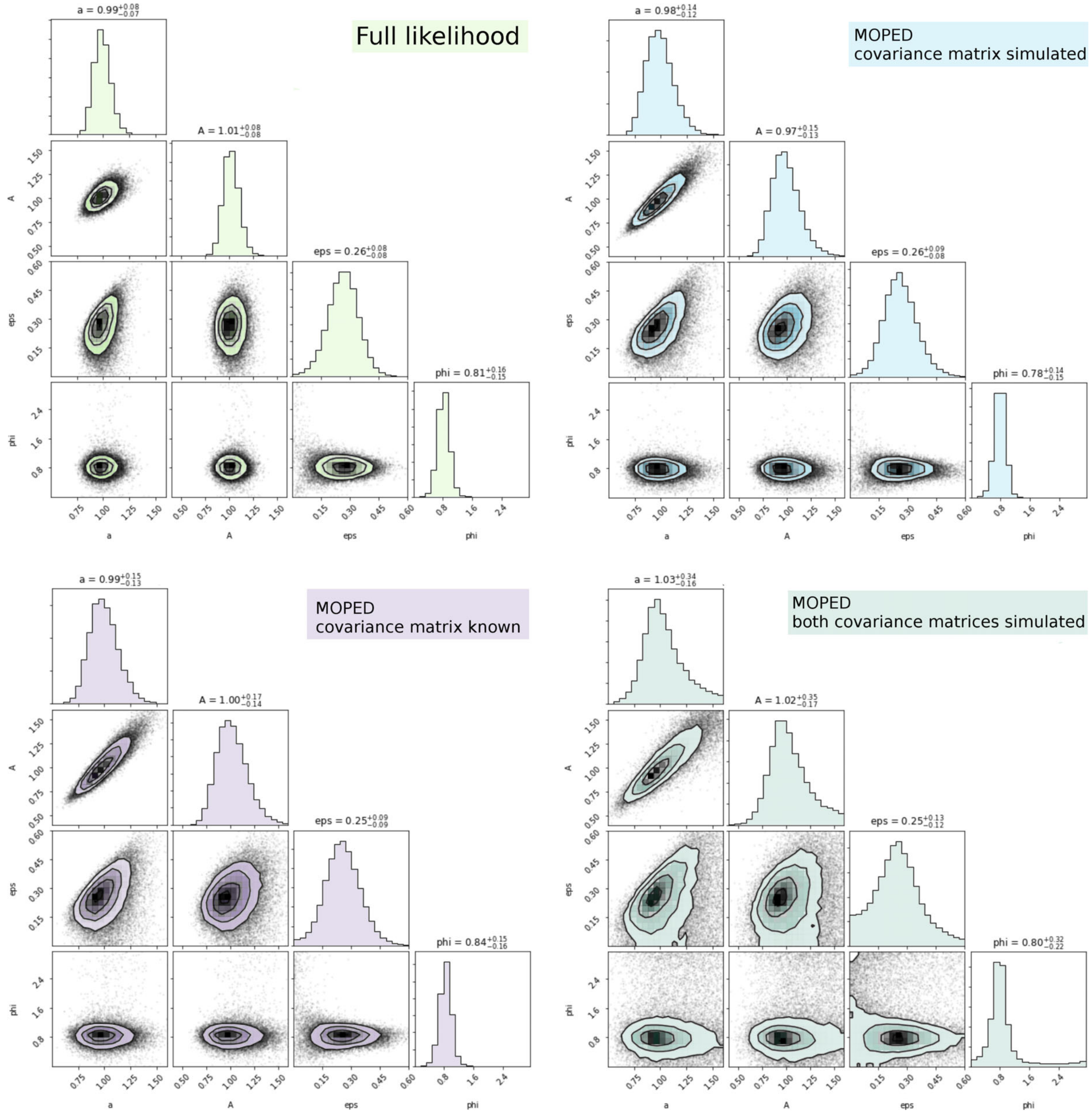

Figure 5. Top left: full likelihood of the four parameters a, A, ϵ, ϕ of the model of equation (10), given a 10 × 10 galaxy image as the $p = 100$ data vector. 10 000 points are generated after burn-in, using Hamiltonian Monte Carlo NUTS with Stan. Bottom left: likelihood of parameters using $m = 4$ MOPED compressed data. MOPED compression vectors have been computed on the basis of the true full covariance matrix. The compressed covariance matrix is assumed to be the identity matrix. Uncertainties are increased since the distribution has been marginalized, and MOPED only guarantees that the distribution is unchanged near the peak. Top right: MOPED compression vectors computed on the basis of an estimated covariance matrix from 1000 simulated data sets. The compressed covariance matrix is assumed to be the identity matrix. In this case, the constraints are essentially as good as if the true covariance matrix was known, since many simulations were used. Bottom right: now also including parameter-dependence in the covariance matrix with the 4 × 4 compressed covariance matrix estimated from 10 simulations at each MCMC point. The credible regions are moderately larger, but the general aim is to make such calculations feasible at all, in cases where it would essentially be impossible to compute the full likelihood.

of the covariance matrix with parameters, and using it to extrapolate to other locations in parameter space (Reischke, Kiessling & Schaefer 2017).

An additional advantage of this radical data compression is that the central limit theorem may assist in giving the compressed data a near-Gaussian sampling distribution, although there is no guarantee that the summary statistics can be grouped into large iid subsets. Furthermore, it will be far easier to explore numerically the sampling distribution in a small number of dimensions rather than in the original very high dimensional space, to test the Gaussian assumption.

We see clear applications, including, but not limited to, the analysis of weak lensing data from the *Euclid* and LSST photometric surveys, where the number of summary statistics is expected to be $\sim 10^4$, and the number of cosmological parameters only ~ 10, so reductions in the number of simulations by a factor of 1000 is feasible, or by a factor of 10^6 with emulation techniques as well (see Table 1).

Table 1. Number of simulations required, for numbers typical of future *Euclid* or LSST weak lensing surveys, with $p = 5000$ summary statistics, $m = 6$ cosmological parameters, and an MCMC chain of length 10^5. 100 emulator points are assumed.

Estimating C^y at	Emulator locations	Each MCMC point	Comments
No compression	10^6	10^9	Estimating $\mathbf{C}^x$ for each MCMC sample is overkill
MOPED compression, using simulated $\mathbf{C}^x$	10^4	10^6	Preferred option
MOPED compression, using analytic/theoretical $\mathbf{C}^x$	10^3	10^6	Sub-optimal, but reduces simulation requirements

ACKNOWLEDGEMENTS

ES is funded by a Deutscher Akademischer Austauschdienst (DAAD) research fellowship of the German Academic Exchange Service. AFH is partially funded by the Science and Technology Facilities Council, by grant ST/N000838/1.

REFERENCES

Asgari M., Schneider P., 2015, A&A, 578, A50

Blot L., Corasaniti P. S., Amendola L., Kitching T. D., 2016, MNRAS, 458, 4462

Carpenter R. et al., 2017, J. Stat. Softw., 76, 1

Dodelson S., Schneider M. D., 2013, Phys. Rev. D, 88, 063537

Friedrich O., Eifler T. F., 2017, MNRAS, preprint (arXiv:1703.07786)

Graff P., Hobson M., Lasenby A., 2011, MNRAS, 413, L66

Gupta S., Heavens A. F., 2002, MNRAS, 334, 167

Hartlap J., Simon P., Schneider P., 2007, A&A, 464, 399

Heavens A. F., Jimenez R., Lahav O., 2000, MNRAS, 317, 965

Heavens A. F., Panter B. D., Jimenez R., Dunlop J. S., 2004, Nature, 428, 625

Heavens A. F., Alsing J., Jaffe A., 2013, MNRAS, 433, L6

Heitmann K., Higdon D., White M., Habib S., Williams B. J., Lawrence E., Wagner C., 2009, ApJ, 705, 104

Heitmann K., White M., Wagner C., Habib S., Higdon D., 2010, ApJ, 715, 186

Heymans C. et al., 2013, MNRAS, 432, 2433

Joachimi B., 2017, MNRAS, 466, L83

Kaufman G. M., 1967, Some Bayesian Moment Formulae Report No. 6710. Center for Operations Research and Econometrics, Catholic University of Louvain, Heverlee, Belgium

Morrison C. B., Schneider M. D., 2013, J. Cosmol. Astropart. Phys., 11, 9

Padmanabhan N., White M., Zhou H. H., O'Connell R., 2016, MNRAS, 460, 1567

Panter B. D., Jimenez R., Heavens A. F., Charlot S., 2007, MNRAS, 378, 1550

Percival W. J. et al., 2014, MNRAS, 439, 2531

Petri A., Haiman Z., May M., 2016, Phys. Rev. D, 93, 063524

Pope A. C., Szapudi I., 2008, MNRAS, 389, 766

Reichardt C., Jimenez R., Heavens A. F., 2001, MNRAS, 327, 849

Reischke R., Kiessling A., Schaefer B., 2017, MNRAS, 465, 4016

Sellentin E., Heavens A. F., 2016, MNRAS, 456, 132

Sellentin E., Heavens A. F., 2017, MNRAS, 464, 4658

Taylor A., Joachimi B., 2014, MNRAS, 442, 2728

Tegmark M., Taylor A. N., Heavens A. F., 1997, ApJ, 480, 22

Zablocki A., Dodelson S., 2016, Phys. Rev. D, 93, 083525

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