

## Notes from Pub Board

Some of the writing in the abstract/introduction should be polished up, there are grammatical issues throughout and it could do with another pass. For example: How best to check these crucial ingredients in the abstract reads poorly, the data vector increasing should be the length of the data vector increasing, and discussion of the length of the tomographic bins should be referring to a data vector rather than a bin. Tidying the language would improve the overall impact of the paper. This is in general true throughout the paper but particularly so in the introduction.

I have added more specific comments below, please do let me know if you have any questions or concerns about these. We would like to ask to see an updated draft for a final quick check before submission/posting.

### Specific comments:

- the data vector has 227 elements, varying with angular separation, and different pairs of tomographic redshift bins  
I would maybe remove the second comma to make it clear that this is 227 total (e.g. varying both with angular separation and with pairs of tomographic redshift bins). **[Done]**
- Apart from MOPED, we will also be analysing the covariance matrix with three other compression methods: the first involves discarding the modes associated with the lowest eigenvalues.  
presumably first involves performing an eigenmode decomposition and then discarding... **[Done]**
- The chosen covariance matrix has a dimension of  $190 \times 190$ , and so, for one eigenmode, we can compress it to 10% of its original size, yielding 190 independent elements.  
just to be clear, youre compressing at the map level and then retaining only the eigenmode with the largest eigenvalue? **[Changed to KL mode.]**
- We thus begin with 200 data points for each  $\xi_+(\theta)$  and  $\xi_-(\theta)$   
this confuses me because it seems like you have 200 data points for each theta point, when you really mean you have 200 theta points. Perhaps you could say 200 data points for each statistic. **[Done]**
- Line 140: The 68% CL constraints are as follows:  
you have 16 free parameters but have stated the results of three. Perhaps say why these values are chosen as your metric for comparison and if youre marginalizing over the other parameters or say that these are a selection of the parameters. A full set of constraints in a table might be preferable if you are constraining everything. **[We added a paragraph explaining why we're only showing these three parameters.]**
- We first diagonalise the covariance matrix in order to calculate its eigenvalues and sort them in decreasing order. Since the highest eigenvalues are said to be most informative [24], we want to remove the lowest ones. To do that, we replace the lowest eigenvalues with lower values (nine orders of magnitude lower), thus removing their effective contribution; we then transform back to the original basis and perform a

cosmological analysis with the new covariance matrix, to constrain the parameters of our model.

why are you reducing it by a factor of  $10^9$ , rather than setting it to zero? Once you've transformed back into the original basis are you hitting issues with inversion of the matrix? **[Setting the eigenvalues to zero results in a semi-positive definite matrix, which is what we want to avoid.]**

More generally, what does it even mean to have a low eigenvalue in this context? Surely the covariance depends on the amplitude of the signal which means a high signal could give a high covariance which doesn't tell you where the constraining power is. It would be good to have a statement about this. **[We added more information on this. The lowest eigenvalues are usually characterising numerical noise.]**

- Paragraph starting 162: constraints are significantly broader for the two parameters shown.

you haven't introduced the figure yet and need to specify where they are shown. In fact figure 3 shows 3 parameters. **[Done]**

I also don't get the discussion about consistency with removing a specific percentage of the data (given that you're really showing the importance of specific eigenmodes). Are you trying to say that this is consistent with the idea that removing data weakens constraints? While removing data from the covariance matrix weakens the information on the covariance matrix itself, it's not clear to me that this should necessarily make the constraints broader in all cases. Am I missing something here? **[Precisely, we meant that removing those modes meant removing information from the covariance matrix, which would imply larger constraints.]**

Inconsistent with the notion that they are irrelevant. Please state what they are. Do you mean the modes with low eigenvalues? **[We meant the modes with low eigenvalues.]**

- this shows another method is called for. Does this show that? It could simply show that you are not retaining enough of the modes and compressing too far. It is only true that this shows lack of validity in the method in the case that you need to retain exactly 10% of the data, which seems somewhat arbitrary. **[We updated the text stating this.]**

- If  $C$  were diagonal, then the eigenvectors would simply be the  $T_i$ s themselves, do you mean a function of the  $T_i$ s? **[No, we mean the  $T_i$ s themselves, or the  $D_i$ s for the data vector.]**
- This analysis was also applied to the 200 modes with the highest SNR, and yielded similar results  
which analysis? Removing the modes or characterising their dependence on parameters? Why would you expect this to produce similar results, when surely removing the modes with the highest signal to noise should give significantly worse constraints and their dependence on parameters should be different. **[We rewrote the paragraph to describe the analysis mentioned. We find that the constraining power is weakened for all three parameters.]**
- The left plot in Figure 4 shows the diagonal elements of the signal part and the noise part of the spectrum. The right-hand panel shows the spectrum of the modes with the largest signal to noise

is the right hand plot showing the spectrum or the S/N ratio? the figure axis says S/N as does the caption . **[Changed the text to S/N for the right hand plot.]**

- Line 255:  
what does the  $i$  refer to in the superscript of  $\kappa$ ? If the bins, then perhaps say 4 tomographic bins  $I$  **[It refers to the bins, we made it clearer in the text.]**
- Figure 4: according to the text the left hand plot shows the tomographic bins and the right hand plot shows the KL modes, but theyre labelled in the same way. Is this correct? This should also be detailed in the figure captions. **[We labelled it differently.]**
- Line 273: what does  $j$  refer to? If this is the cross-spectrum between bins  $I$  and  $j$  this should be specified here. **[Yes, it refers to the cross-spectrum, we made it clearer in the text.]**
- Line 284: You have not introduced the concept of  $E$  before this point. I was entirely unclear that this symbol refers to the KL eigenmode until it was described on the next page. **[Done]**
- Figure 5:  
what exactly are the row vectors of  $R$ ? Can you specify that this is  $E/N$  if this is the case? **[It is now specified it in the caption]**
- Line 313 vs e.g. 305: Inconsistent use of eta vs eta(theta) throughout the paper. **[We believe the reviewer meant  $\xi$ , and have corrected for that.]**
- Line 294:  
so the power spectrum of the new observable  $D_\ell$ . This makes it sound like the new observable is  $D_\ell$  not  $b_{lm}$ . **[Done]**
- We should point out that these KL-modes are uncorrelated, so the power spectrum of the new observable  $D_\ell$  is a diagonal matrix, with  $1+\text{SNR}$  of the corresponding eigenmodes on the diagonal elements. Since the KL decomposed modes of shear power spectrum are uncorrelated, we can make a compression here by taking only the first one or two modes with the highest SNR. By doing so, we compress ten tomographic bin-pairs to one or two  
Can I assume here that you are actually computing  $D_\ell^{ab}$  for the correlation between mode  $a$  and  $b$  of this decomposition and it is in this space that you mean this is a diagonal matrix? This treatment of different modes of the decomposition hasnt been made clear in the notation or text to me if so. I think this refers back to the  $p$  values so perhaps the notation  $D_\ell^{pp}$  would make this clearer exactly what the superscripts in Figure 4 and the discussion here is referring to. **[This is true, we changed to  $D_\ell^{pp'}$ ]**
- Figure 5: You discussed previously that its important to use the 1st mode, the 2nd mode and the cross-correlation of the 1st and 2nd mode. Why is that cross-correlation not shown here in favour of the 3rd mode? **[We believed it would be more informative to show that, for each mode, the curves follow roughly the same tendency, independent of  $\ell$ .]**

- Line 322: is  $E$  normalised or not, and are you plotting  $E$  or  $E/N$ ?  
Also it may be better to say increasing opacity rather than increasing darkness of the colour. **[Done]**
- Line 338: likelihood analysis  
suggest as detailed in section A as its been a while and the reader might need referring back. **[Done]**
- Line 358:  
- rather than saying the compression here ... I would strongly advise saying the MOPED compression scheme ... **[Done]**
- Line 389: We have apparently captured from the initial set of  $(227 \times 228)/2 = 25,878$  independent elements of the covariance matrix a small subset (only 136) of linear combinations of these 26k elements that really matter. If two covariance matrices give the same set of  $C_{\alpha\beta}$ , it should not matter whether any of the other thousands of elements differ from one another.  
This is a very strong statement. Isn't it more true that it should not matter in estimating the Fisher matrix, but it can still make a difference to the likelihood contours? **[This statement is true. We are able to reproduce different covariance matrices from the same C (by tweaking the compression scheme to make it invertible) and they all give similar constraints.]**
- Perhaps even more importantly, there are much fewer points on this plot, meaning that MOPED reduces the number of elements that need to be compared  
this is true by design however? Is this not a statement of methods rather than a result? **[It's a statement, we changed it slight to reflect that.]**
- Line 489:  
may be worthwhile to add here if  $\delta$  is a single number for all elements, or if it is computed elementwise. **[Done]**
- Eq. 14: where is the  $\alpha, \beta$  in the RHS of the equation? **[We removed the subscript.]**
- Line 604: since the Fisher matrix is identical for both the original and compressed ones, the compression scheme is lossless.”  
it came up in the previous referee reports that this is not technically true. (see Alan Heavens second referee report) **[We removed this sentence.]**

# Review by internal referees

## Second Reports

Alan Heavens, August 4

With apologies for the delay in getting comments to you, here goes. This draft is much improved over the first one, and the focus is clearer. I think in particular the comparison between the field-level (KL-like) compression and summary-stats level (MOPED) is interesting. The field-level is not the most general one, but if I understood correctly, is Davides elegant weighting of tomographic fields. I am still not convinced of the motivation for the large eigenvalue/small S/N compressions, and I guess the results rather bear that out, but I don't see much harm in testing some straw person cases. It is possible that a referee will take a different view.

### Abstract

The covariance matrix is a four-point function only if the statistics are two-point functions. **[Done]**

It wasn't immediately clear to me how compression helps identify the significant parts of the covariance matrix. Does this use the magnitude of the elements in the matrix which translates the original covariance matrix to the compressed one? This would make sense. **[Exactly.]**

In the last line, would inaccuracies be more suitable than changes? **[Done]**

### Introduction

In the first paragraph, the analysis only requires a covariance matrix if the likelihood is assumed to be gaussian. This is a big assumption (and not necessarily always good). A bit more explanation is needed here. **[We reworded this to reflect the more general case.]**

MOPED is not truly lossless, but only in the restricted sense that the Fisher matrix of the compressed data is the same as for the full dataset, if certain conditions are met. For a given data set, this usually makes the likelihood surface near the peak almost identical, and in practice the similarity often extends right into the wings of the distribution, as you find here in Fig. 8. **[Fixed this.]**

In the list of applications, can you add Mootoovaloo et al. (arXiv:2005.06551) for an application to weak lensing? It is the one I mentioned was coming, in my earlier report. **[Done]**

### Section II B

Louca and Sellentin (arXiv:2007.07253) actually argue for removing the highest S/N modes (to reduce biases). This is worth referring to. **[We did a similar analysis removing the 200 modes with highest SNR and found similar results. We chose not to display those here so as not to produce plots with confusing information. We point out, however, that this might have been different if we had not removed such a large number of modes. We added a paragraph referring to this.]**

almost 200%  $\rightarrow$  the errors of almost 200%. **[Done]**

### Section II C

The sentence saying that the eigenvalues would be the data points themselves doesn't make sense to me. Do you mean they would be the diagonal elements themselves? **[We mean that the eigenvectors would simply be the  $T_i$ s (or on the data the  $D_i$ s); i.e., they would not be linear combinations of either.]**

After (3), Modes...can be discarded rather pre-empts the conclusion (that in fact they can't, without losing information). I'd reword this so it doesn't suggest that they can. **[Done]**

### Section II D

It is worth pointing out that the linear compression is not the most general one, since the weights don't depend on  $l$  or  $m$ . As Davide found, it can be very effective nonetheless. **[Done]**

Typo: existence  $\rightarrow$  existence. **[Done]**

### Section II E

(After equation 12). In TTH we only considered compression for one parameter for the known covariance matrix case. The extension from one to more than one is not as trivial as it appears (for example, we suggested a recipe for this for the KL known mean case, and in fact it is often a very poor solution). The MOPED paper showed that all elements of the Fisher matrix could be reproduced simultaneously, and not just a single element, or just the diagonal elements. I'd suggest rewording to This compression was first suggested by Tegmark et al. for a single parameter only. The non-trivial extension to multiple parameters, where the full Fisher matrix is reproduced with the compressed data, is the MOPED algorithm. **[Done]**

### Section III B

In the top panel of Fig. 10, it would be helpful to plot GCM/FCM, so the top plot resembles more the bottom, rather than being flipped. **[We made the changes also to Fig. 9, which has a similar plot.]**

### Section IV

I think I mentioned this before, but one way to ensure that  $C$  is positive definite is to perturb the matrix log of  $C$ . It's worth perhaps offering it as an alternative, and giving the argument (I think you had one) for why you don't do it this way. **[Done]**

### Section V

In The next step was to shrink... I don't think you decompose the shear power spectrum, but rather the data vector. i.e. you do this at the field level (albeit its spherical harmonic transform). 1-point, rather than 2-point. **[Done]**

You might also want to mention in connection with Harry Mootoaloos paper, that to speed up the analysis one needs a fast way to make theoretical predictions for the compressed data (you don't generally want to generate full data sets and do repeated linear compression, since that is slow, and sometimes as slow as analysing the full data). Harry used Gaussian processes to generate the compressed theory. **[Done]**

David Alonso, April 4

I thank the authors for the revised manuscript. The paper reads much better, and its main results and conclusions are much clearer and useful. Just a couple of minor points:

- I may have missed this in the text (in which case I apologize): is Fig. 12 showing the results for one of the 50 realizations of the covariance matrix? Or the mean of the constraints for all of them? **[We generate 50 different covariance matrices with the desired error budget, then take the mean covariance out of those. The parameter constraints are performed only for that covariance matrix.]**
- Do you have any intuition why  $A_{IA}$  seems to be more sensitive to covariance errors than any other parameters? I understand why this is the case in the context of KL transforms, but not in general. **[We know that the  $A_{IA}$  is not well constrained by the data (even with the full covariance matrix), this just worsens when we introduce noise to the compressed covariance matrix. We corrected Figure 12, since it was actually showing the  $4\sigma$  interval. We find that, while the loss of constraining power for this parameter is more exacerbated than the others, it is not as bad as we originally thought. The increase is only of about 30% (although most are lower than that), an improvement when compared to what was previously reported.]**

**Jonathan Blazek, August 16**

My apologies to the authors for this delayed review. I thank the authors for their work in addressing my first round of comments, and I agree with the other reviewers that the paper has been significantly improved. I don't have much to add to the other comments, but here are a few things:

- Related to David's question, it would be great if a bit more insight could be provided into why the  $A_{IA}$  parameter is so sensitive. I suspect this is related to the fact that the covariance has no IA contribution, but I'm not sure how to best test this. In any event, it must be the case that the sensitivity to noise in the elements/eigenvectors is related to the sensitivity to low S/N modes - perhaps there is a simple mathematical way to make this connection. **[We discuss this in our reply to David's question. We politely ask to refer to that.]**
- line 12: ways of compression  $\rightarrow$  compression methods. **[Done]**
- line 539: loss of 200% of constraining power. This wording is confusing. **[We meant that the  $2\sigma$  intervals increased by about 200%.]**
- Sec IV: (Related to one of Alan's points above) It's not entirely clear (to me) why this method of perturbing the compressed cov matrix is a realistic test of how errors would arise in actual applications. Could this be motivated a bit more? **[The perturbations were introduced as a way to imitate the noise from simulations. In that sense, the tolerance tells us how the amount of noise affects the parameter constraints.]**

## First Reports

Alan Heavens, March 22

As I see it, the paper has two broad aims: to explore a number of ways to compress weak lensing data; and to look at the sensitivity of results to covariance matrix errors, including seeing which elements of the covariance matrix are most important for parameter inference, and which should therefore be computed especially accurately if possible.

The paper makes a number of good points, but it has some problems, and this report is rather critical, but I hope constructively so. It might be worth stepping back and thinking what the main messages of the paper should be, what results can be presented that are genuinely new, and if some of the studies can be more systematic than now.

### Summary

The executive summary is that the compression side tries a few things, and concludes that the best way is to use MOPED. Since MOPED is known to be optimal and cant be beaten in terms of the Fisher matrix (subject to some conditions), this shouldnt be a surprise. For suggestions of how to present something new here, see details below.

For the covariance matrix studies, this is an interesting question to tackle. The concern I have is that its not clear how to do this in a systematic way, and I think that a referee will consider it to be too ad hoc to draw general conclusions from. This needs to be strengthened. I suggest a small technical improvement to avoid non-positive definite matrices, but I dont have any good ideas for a more systematic general approach.

### Detailed comments

The first compression method involves modifying the highest or lowest eigenvalues of the covariance matrix, effectively discarding the eigenmodes with the highest or lowest eigenvalues. Its perhaps not surprising that this doesnt work very well, since the procedure knows nothing about the sensitivity of the covariance matrix to the parameters, and there is no formal way of ordering the modes by contribution to the information on each parameter. The second compression method is better, based around S/N eigenmodes (Vogeley and Szalay 1996 should be referenced). Here linear combinations of the field values are considered (field here corresponds to shear measurements on a number of tomographic slices). Compression to the highest S/N modes is optimal in one sense, if (and only if) it is the amplitude of the (various) power spectra that is the (single) target. This is a more principled approach, and it is not surprising that it does better. It is still not very good, as Fig. 4 shows, because in general, we want to infer multiple parameters, whose influence on the power spectra is more complicated. This problem was tackled by Tegmark, Taylor, Heavens (1997; TTH), and we found the modes that would be optimal for any single parameter, by including the sensitivity of the power spectra to the parameters. However, the challenge is to find a compression that works for all parameters simultaneously. This is not at all trivial. In TTH we proposed combining the optimal modes determined for each parameter, and it turns out to offer something, but in the end it does not work very well, for reasons that I can explain if needed. The S/N eigenmodes do seem to be quite effective, in a restricted setting when applied in the radial direction only, as Davide Alonso found. Here we see that they are limited in this context, perhaps because once again they have no knowledge of the sensitivity of the signal to the parameters, so are therefore almost inevitably sub-optimal.

The final approach is to work with linear combinations of the two-point functions (rather



than of the field). This uses a principled approach, and is the MOPED algorithm from Heavens, Jimenez, Lahav (2000), just without the (optional) Gram-Schmidt orthogonalization. A number of authors have rediscovered it since then, such as Gualdi et al., but the radical compression to the number of parameters goes back to HJL. TTH found the solution for a single parameter, but (as seen above in the compression of field values) the extension to multiple parameters is not trivial at all; guaranteeing that the Fisher matrix is unchanged on compression is not even obviously possible. What HJL showed mathematically is that in fact it is, and the credible regions can be unchanged (at least at the level of the Fisher approximation), even with such a radical compression, and this is what Fig.11 basically shows. This of course is wonderful, but I wonder if it tells us anything new we already know that MOPED works (and that there is nothing that can do better for gaussian data in terms of Fisher matrix).

To present something new here, it might be best to angle it as the first analysis that applies MOPED to a weak lensing situation with  $> 15$  parameters. In the interest of full disclosure, we have a paper in late draft stage that applies MOPED to weak lensing data. Given the LSST internal review process, our paper will likely appear on arxiv before this one. However, it has fewer parameters, so you might want to stress the  $> 15$  parameters aspect. **[We thank you for this information. We have included this point in the paper; we also make it clear that we're using MOPED (minus the normalising factor).]**

A suggestion for the covariance matrix study, to ensure that the perturbed matrix is positive definite, is to take the matrix log of the fiducial covariance matrix, and then to modify the elements of the log matrix in some way (how to do this systematically, I'm not sure). Taking the matrix exponential will then give a positive-definite matrix for sure. Its still a bit ad hoc, but avoids that particular problem. **[One of the issues here is how to introduce an error to the log matrix that would be similar to what we expect to see in the original matrix. A naive assumption would be that**

$$C_{\alpha\beta}^{new} = (1 + \delta)C_{\alpha\beta}^{old}$$

$$\rightarrow \log(C_{\alpha\beta}^{new}) = \log(1 + \delta) + \log(C_{\alpha\beta}^{old}) .$$

**We know this to be wrong, since obtaining the matrix logarithm is not as simple as that, so this approach would be incorrect. Also, applying error to the elements of the log matrix (like a 10% error) produces a covariance matrix with elements several orders of magnitude larger than the original one. This would work if we diagonalised the matrix and applied the transformation to the  $\log$  of its eigenvalues, but since we already have a section for modifying the eigenvalues, we fear this would be repetitive.]**

Smaller points are:

The comparison of the elements in Fig. 2 is a bit misleading (or rather, the conclusions are). Assuming that the small elements (which show the biggest fractional difference) are off-diagonals, then it probably doesn't much matter if they are  $10^{-4}$  or  $10^{-8}$  of the diagonal elements they are effectively zero. If they are on the diagonal, then it's a different matter, so it is worth checking. **[We added a more thorough investigation into these differences, pointing out that the biggest difference are in the smallest elements and correspond to off-diagonal terms.]**

The compression in Fig. 8 shows that for the compressed data, the two covariance matrices lead to compressed covariance matrices which agree much better. This is encouraging, and expected since the linear combinations involved will be dominated by the larger elements, which agree closely, and differences in the small elements are largely irrelevant. **[These are the results we were hoping for.]**

For the MOPED compression, it is suggested that the full covariance matrix is used to determine the fiducial model where the derivatives are computed, but this is slow. It is much faster to choose a rough fiducial model and find the maximum likelihood from a preliminary MOPED analysis and then repeat with this as the fiducial model. The process could be repeated as many times as necessary, but its never been needed in my experience. **[The fiducial values used were the bestfit values obtained in the previous analysis with the full covariance matrices. We agree with your comments, but we already had those results at hand.]**

Im sorry not to be able to suggest a route that makes the covariance matrix study likely to be strong enough, but I hope that these comments are constructive. **[These points have helped immensely, and we are very thankful. We hope that we have now been able to provide a draft with clearer and more conclusive results.]**

**David Alonso, April 3**

My overall comments would go along the same lines as Alan's, so I won't repeat them here. My main worry is the fact that there doesn't seem to be a clear lesson learnt or take-home message with regards to the problem of identifying the most relevant components of the covariance matrix. This is a hard problem! One possible angle that I didn't see explored here is whether MOPED would be able to circumvent errors in the covariance matrix. I.e. how much do the contours change if you use the two different covariances? How about their associated minimum  $\chi^2$ ? **[We explored this in more detail with the addition of the constraints on the intrinsic alignment parameter  $A_{IA}$ . We generated 9 new covariance matrices by introducing error to the compressed Full Covariance Matrix (FCM, which we previously referred to as CL), and saw that, while the constraints on  $\Omega_m$  and  $S_8$  did not alter by much, there was an alarming loss of constraining power for  $A_{IA}$ ]**

Beyond this, I have a number of specific comments. Here they are in order:

- **L32** Individual  $\rightarrow$  independent? **[Done]**
- **L43** I'd remove this paragraph from the introduction. **[Done]**
- **L47** You seem to use the present continuous a lot we are trying. I think I'd go for the present simple. **[Done]**
- **L113** Could you explain in a bit more detail what the differences are between the two covariances? I.e. is it that one is missing the non-Gaussian parts? **[We explain this better in the text now. The Gaussian Covariance Matrix (GCM) only consists of the Gaussian parts.]**
- **Eq. 1** Does  $T$  here contain the noise bias contribution? I imagine not, but I wanted to check. **[This is just the formal definition.]**

- In both sections IV and V you seem to arbitrarily choose 200 as the number of modes to throw away. Can you justify this? **[We modified the text to make this clearer. This is done to achieve a reduction of about 90% of the covariance matrix, so we have a fairer comparison between these methods and the other, more complicated compression schemes.]**
- **Fig. 5** This would benefit from larger labels and legend. You could even consider making it a double-column figure. **[Done]**
- **L249** compression at the ell space - I think I'd say map-level compression. **[Done]**
- **Eq. 6** Use  $\langle$  and  $\rangle$  instead of  $<$  and  $>$ . **[Done]**
- **L314** It is worth justifying that ell-independent  $U^{ij}$  is a good approximation (which it usually is for cosmic shear). **[Low ell spread is because of the cosmic variance, and is also lower weighted by the weighted average, high ell shows great consistency in Figure 5.]**
- **Eq. 10** It'd be good to justify the 2ell+1 weighting. **[Justified by the number of modes for each ell.]**
- **Fig. 7** I think  $W_{ij}$  should be  $U^{ij}$  in the caption. **[Done]**
- **L354** is  $\rightarrow$  are **[Done]**
- **L354** The result also makes sense in terms of the cumulative nature of lensing (i.e. the further away the source, the more lensed it is), although the thinning out of the galaxy density at higher redshifts makes the 4th bin have a lower weight than the 3rd one. **[Done]**
- **Fig. 8** This plot is tiny! Do you need the right panel? **[Done]**
- **Fig. 10** Seems to appear in the text before Fig. 9. Actually, I'm not sure you discuss figure 9. **[This was removed.]**
- **Eq. 12** The transformation is parameter-dependent. I assume you perform it at a fiducial cosmological model? **[We do indeed. We used the best-fit results from the full cosmological analysis and take those to be our fiducial cosmology everywhere.]**
- **Eq. 13** Probably worth emphasizing that this is the Fisher matrix of your likelihood. **[Done]**
- **L415** Where does eighty thousand come from? **[We were referring to the elements of the covariance matrix. We fixed this in the text.]**
- The shape and variance of the posterior is not dependent on the derivative, but the best-fit value shifts according to the point where the derivative is taken. I'm a bit confused by this. Can you justify this statement? **[We've removed this paragraph from the paper. The point of it was to explain that we needed a cosmological analysis had to be done first, in order to get the bestfit values.]**
- You should try to reorganize the figures so they appear a bit closer to where they are discussed. **[Done]**

- **L449** We find that the mean values of the parameter constraints for the two methods agree to  $1\sigma$ , which shows that they are equivalent to each other. I don't understand this statement in the context in which it appears (the comparison between covariance matrices). **[Since we are comparing the two compressed covariance matrices, there are two ways of doing this: we can either compress them using the same weight (remember that the inverse of the covariance matrix is used for this result), or with a different weight for each (the weight would be obtained with the corresponding covariance matrix). We wanted to check how different weights impact the parameter constraints and found the differences to be small (as compared to the difference between the constraints for the two matrices).]**
- **L496** The qualitative jump between 25% and 30% errors is hard to digest. What is the cause of this? Also, did you do this for a single realization of the noisy covariance or for several of them? **[We redid the analysis with 50 realizations of the covariance matrix. The jump was most likely due to a large fluctuation, which was corrected with more realizations. The new analysis brought new aspects to light, which are pointed out in the new draft.]**
- **L510** Do you mean  $\delta > -1$ ? **[We meant  $|\delta| < 1$ .]**

**Jonathan Blazek, April 6**

I have followed with interest the progress of this important work, and I congratulate the authors on completing the paper. Please let me know if you have questions about the below comments/suggestions. Note that I deliberately wrote my comments before reading the other reviews, so there is some overlap.

One general comment before specifics. The paper has two main goals: 1) comparing different covariance codes/methods and 2) exploring data compression. I find that these two threads are not always clearly distinguished. For instance, much of the data compression is done only on the CL covariance matrix, and the conclusions about the comparison part are not given much attention. It could help to slightly restructure where the results are presented, or to add a few words making clear when CL-only is being used and when an important comparison is being done.

- For the reason mentioned above, the abstract should be more clear. What are the main goals and conclusions? Also start along the path in the first sentence sounds tentative for the start of an abstract. **[Updated the abstract to include more information about our goals and conclusions.]**
- Reference numbers are out of order. **[The references are in alphabetical order, that's why they're not in increasing order in the text.]**
- **L34:** As data sets increase  $\rightarrow$  you mean the number of elements in the data vector, not necessarily the overall amount of observational data. **[We replaced *data sets* with *data vectors*.]**
- **L68** Missing section symbol. **[Done]**

- **L83** Emphasize here (or elsewhere) that this analysis is done in configuration rather than Fourier space. Is there any reason we would expect the qualitative results to be different in a Fourier space analysis? **[Map level compression actually works better in Fourier space, but we don't think MOPED is going to differ much in two setups. The reason we use real space is because it is in line with the DES Y1 analysis.]**
- **Table 1** Here, and in the relevant parts of the text, it may improve clarity to add the  $_{IA}$  subscript to the A and eta variables. Also, you can cite the Troxel DES cosmic shear paper when mentioning the IA params. **[Done]**
- **L84** DES Y1 is the state-of-the-art for cosmic shear data sets, at least right now. You could mention this.
- **L97** Are all of these citations for CosmoSIS needed? **[These are the references that CosmoSIS spits out after each run. We thought it would be polite to include all of them.]**
- **L140** The difference of 4 orders of magnitude is casually discussed without further explanation. Although your later results show that this is mostly due to non-important elements, more explanation at this stage would be helpful. Why is this happening? How can we plausibly think that these two cov matrices are roughly the same when there are such large differences. **[We try to address this by taking into consideration Alan's comments as well. The elements that show greatest discrepancies are the smallest ones (of order  $10^{-18}$ , this is probably why the constraints are not all that different. As can be seen, the diagonal elements (and the larger ones), have smaller differences.]**
- **Paragraph beginning at L163** I was a bit lost by the logic of this paragraph. If possible, provide more explanation of what you are doing and why. **[We removed the figure and the associated text.]**
- **L179 (and elsewhere)** Why 200 eigenmodes? This number is not explained, and the results depend strongly on how many eigenmodes are kept. Similar question about the S/N analysis. **[We try to explain this better in the text. This is done to achieve a reduction of about 90% of the covariance matrix, so we have a fairer comparison between these methods and the other, more complicated compression schemes.]**
- **L183** Throwing away 90% of the information... By what metric? Where does this number come from? **[Change to reducing the number of elements in the covariance matrix by 90%.]**
- **Fig. 4 (and elsewhere)** The blue/purple/magenta color scheme is tough to parse. They look similar, especially with overlapping contours. **[Changed one of the blueish color to green. The goal is trying to be colorblind-friendly as well as monochrome copy friendly.]**
- **Fig. 4 caption** You probably don't need all of the detail on how you remove these modes (which is included in the text anyway). You can just say that you remove those modes. **[Done]**

- **L210** very small. How is that quantified? **[We meant the lowest 200.]**
- **L226** signal OF these modes... **[Done]**
- **L234** vanilla What does this mean? **[Changed to 'simple'.]**
- **L237** Explain that the shaded region shows the modes excluded in the previous analysis. **[This is in the text.]**
- **L249** Compression at the ell-space... This sentence is confusing. **[Changed to map-level compression.]**
- **L267-269** Eliminates even more modes the numbers are the same. These ideas seem to be in contradiction. Which point do you want to emphasize? **[This paragraph was removed.]**
- **Eq. 5** I dont understand the indices. What is the summation convention? **[They are really just matrix multiplication, simplified to matrix representation.]**
- **L298** double summation What is meant? **[It really just meant  $\sum_i \sum_j$ .]**
- **Eq. 8** This is the standard xi+- expression, which is not how it is described in the text above. **[Changed so that it is only meant to be used to derive the  $\xi_{+/-}$  for the compressed data.]**
- **L323-324** Wrong symbol. Should be arcmin, not deg. **[Done]**
- **L342** increasing shades is confusing. I also found the plot itself hard to parse, since the contrast between different ell values isnt very much. **[Made it so it is a little easier to see the change in color.]**
- **L345** What is  $W_\ell$ ? **[It should be  $U_\ell$ .]**
- **L347** You say earlier that the transformation has to be ell independent, but then discuss here that this is some ell dependence. Is this a problem? If not, why not? **[For low ells, there is the cosmic variance that is inevitable, but they are lower weighted so it is fine. For higher ells, they converge to an acceptable level.]**
- **L353-354** I found this sentence confusing. **[Added more specifications of what the results reveal.]**
- **L369** First two KL modes and their cross correlation? **[Yes, and changed.]**
- **L415** Do you mean elements or linear combinations of the elements? **[We meant the elements.]**
- **L428** This discussion of parameter space volume is tough to understand without discussing the number of parameters. Also, do you want to claim that the other parameters agree well when they are only consistent at the 2sigma level? **[We wanted to be very conservative here. The cosmological and IA parameters show an agreement of > 99% (with the exception of  $\alpha_{IA}$ ). The real problem are the systematic parameters.]**

- **Fig. 7** As I mentioned before, the different colors are pretty subtle and hard to parse. Also, you should mention the ell range spanned by the different color shades. [Changed the color to blue-green-orange, should be better. Also specify the meaning of different shade level in the caption.]
- **Fig. 8** Larger text, like in Fig 12. Also, why are some of the diag elements below 1, when all of the others are above? I didnt see a discussion of this anywhere. [Specified in Section III A in v2.]
- **L443** Extra space in i.e. [Done]
- **L451** Agreeing to  $1\sigma$  does not imply equivalence. [Changed this.]
- **A general comment about figures** There are several similar figures, and they dont always appear right next to where they are described in the text. So it would help if each caption included info on which method of compression is being used. [Done]
- **Fig. 9** Where in the text is this figure mentioned? [We removed this figure.]
- **Section VII** Is this always with the CL covariance? Is there any tolerance testing for the other one? [The tolerance testing is done only on CL/FCM. We believe the results will be similar, if done with GCM.]
- **L475** Clarify that the eigenvalues here are not the original eigenvalues used in the compression method. [Done]
- **L479** What is this extra step which is mentioned a couple of places? Is it simply a check on the final cov matrix? If so, does this introduce a selection bias on the otherwise random noise in the cov? [We made this clearer in the text. For the elements, a constant is added to the diagonal and in the case of the eigenvalues,  $|\delta| > 1$ .]
- **L517** What is the takeaway of this result and this final sentence? The question of noise in the eigenvalues is just dropped. What does this result mean? [We have given more emphasis in this new draft.]
- **Conclusion** Say more about the results of the comparison between the two methods. What have you learned? [We included a discussion about this in the new version.]
- **L538** Is it actually untrue that these modes carry less information, or is it simply that they do carry some information? [We see that they do carry relevant information, since we lost constraining power for all the parameters by applying this cut (this wasn't the case for SNR, for instance).]
- **L549** The phrasing of IA being sensitive to low S/N may be misleading. You mean low S/N in cosmic shear the cov doesnt include IA at all. [Done]
- **L559** consecutive  $\rightarrow$  resulting? [Done]
- **Fig. 14** Its hard to see whats going on since we dont know how the x axis is ordered. [We added labels to the x-axis, showing the parts corresponding to  $\xi_+$  and to  $\xi_-$ .]

- **Fig. 15** Hard to see black vs blue lines. [We replace the figure, and changed the colours.]
- **Fig 15** englobes  $\rightarrow$  covers. [Done]
- **L577** This is very interesting - is it truly lossless? [We included more information regarding this: it is lossless only if the noise in the data does not depend on the model parameters.]
- **L592** The sudden degradation in constraining power when going from 25% to 30% errors is surprising (to me, at least) and an interesting result. Are you able to explain what is going on? [The results shown in the paper had been for only one realization of the perturbed covariance matrix. We redid the analysis for 50 of those, and the results are now different – the previous results were likely due to statistical fluctuations.]
- **L595:** What is happening with the eigenvalue noise? Are there interesting lessons from this result? [Introducing noise to the eigenvalues was done in an attempt to ensure positive definiteness of the covariance matrix. The constraints obtained are similar to those obtained when the perturbation is applied to the elements themselves. The interesting lesson is what happens to the  $A_{IA}$ : even though the error in the final elements is small, the constraining power is largely lost for this parameter.]