Reviewer comments:

Reviewer #1: I am not very happy with the authors' response to my criticism of their use of 12 four-week dummies to capture seasonality. First, they note that this parameterization assumes the seasonal effect to be constant \*within\* each four-week bucket, which is good - but they do not seem to note the other effect I noted: that in this parameterization, \*different\* four-week buckets are treated as \*completely unrelated\*. This parameterization assumes that there is no reason why June and July should be any more similar than June and December, which is prima facie ecologically very doubtful. Also, they claim in their response that "the models are estimated with a comparably large sample (e.g., 160 weeks), where the loss of 12 degrees of freedom is not an issue". First, 12 df are indeed a problem for models with just 160 observations. A good rule of thumb is having 20 observations per df, and here we have about 15. But we actually have far fewer than 15 observations per df!

The truly monumental four-line equation (7) alone seems to imply 52 dfs if we consider only a single interacting product (M=N=P=1). The entire system is hopelessly overparameterized, and no, using the Lasso does not mean that an ecologically invalid parameterization (see above) suddenly becomes ecologically valid.

I would definitely suggest that the entire analysis be re-run with a far more parsimonious and ecologically valid seasonal model. However, this would be a lot of work, and I believe that the paper is useful even with the current seasonal model. Nevertheless, I do not want readers to come away with the impression that this approach is defensible. (It isn't. There is no argument in favor, except for "we already did it this way", and that it is easy to interpret - but so is the humour theory in medicine.) Consequently, the least I urge is a \*much\* stronger statement about the limitations of the seasonal model.

We thank the reviewer to point out the limitation of the four-week dummy variables and the potential benefit of the trigonometric variables in capturing the seasonal effect. We now re-conduct the entire analysis using trigonometric variables to account for the seasonal effect. The final results are all consistent with the results which we obtained previously (e.g., including the comparative forecasting performance between different models for all the scenarios, and also for the results of the DM test). We find that the models with trigonometric variables do have higher forecasting accuracy compared to the previous models where we capture the seasonal effects using deterministic four-week dummy variables. We thank the reviewer for the suggestion which is indeed helpful.

Section 2, the literature review, should be revisited. Section 2.1 is one long paragraph. Please break it up and structure paragraphs logically. Section 2.2 is entitled "The effect of marketing activities including price and promotions", but all of section 2.1 is already about promotional modelling. Please reorganize. In addition, please critically review the literature cited and consider cutting literature that does not consider \*forecasting\*.

We now reorganize section 2.1 into two paragraphs: the first paragraph introduces the studies which forecast retailer product sales for the promoted period and the non-promoted period separately; the second paragraph introduces the studies which do not split the forecast period.

We also revise section 2.2 to highlight the evidence which suggests that the effects of marketing activities change over time. We now cut the reivew of the studies which focus on previously well-known effect of the marketing activities.

We now critically review the studies we cited by highlighting their limitations.

Eqs. 12 & 13: since all the MASEs have the same denominator, the formulas can equivalently be written in terms of MAEs, which may be a bit more intuitive.

We thank the reviewer to point this out. We now add a note “ In Equation (12) and (13), all the MASE’s have the same denominator, thus the percentage reductions of the MASE is equivalent to the percentage reductions of the MAE”.

Table 6: please indicate in the table caption that positive numbers refer to forecast improvements with respect to the benchmark (same for Figure 3). By which logic are entries bolded?

Revised: we have added notes to indicate that positive numbers refer to forecast improvement.

We now remove the bolded format in Table 6. Also, we explain the logic why we select the product categories and show their results in Figure 3(a) and 3(b). e.g., “Figures 3(a) and 3(b) show the boxplots for the percentage reduction in the MASE for selective product categories where the two methods respectively produce the biggest improvement in forecasting performance compared to the ADL-intra model..”

Here is another possible explanation for structural changes in response to markting activities: the IRI dataset spans four years. It is quite possible that promotions have changed during this time, even if they are all labeled "feature" and "display". For instance, display racks may have been redesigned, or features moved from weekly flyers to mobile apps. (Condensing the enormous amount of different promotions at a typical retailer into just two categories "feature" and "display" is a bit of a laugh, too, but that is a feature of the IRI dataset.)

We thank the reviewer to point this out. We describe this in section 2.2 as one of the reasons marketing activities may change. We add the following sentence:

“The effect of the marketing activities can also change depending on how retailers communicate their marketing events. For example, retailers may promote products through mobile applications and adopt new prominent promotional shelf tags, which can make the promotions more effective (M. Dinner, van Heerde, & Neslin, 2015). The effect of the marketing activities can also change due to an update of their content and format. For example, retailers tend to launch promotional events of a wide range of types such as multi-buy promotions, store flyers, mobile apps, billboard advertising, and temporary price reduction (TPR), or TPR for shopper-card holders only. Retailers may initially promote a product with ‘Buy One Get One Free’ but then update the content to ‘Buy One Get the Second for Half Price” months later. They may change the format of the feature advertising from weekly store flyers to mobile apps and also redesign the racks of their display. These many changes in the content and format of marketing activities can be expected to lead to changes in consumer response.

The authors keep referring to the base-lift method as the industry standard. I beg to differ. There are lots of econometric methods (see section 5.3 in Fildes et al., 2018), which should definitely outperform base-lift. In particular since the authors' description of base-lift implies that the lift factor is taken from the last promotion, regardless of whether it had a feature or a display, or whether the price reduction was in any way comparable to the price reduction in the promotion we are forecasting for. The base-lift method, in this particular setup, is little more than a straw man that may improve on the simplest naive forecast, but certainly not by much, and it is certainly not state of the art, and Appendix B of Fildes et al. (2018) indicates that it is not all that common: none of the five cases tabulated use it.

We now revise it by introducing the Base-lift model as “a model which has been used as the benchmark model in previous studies”. We also cite Fildes, et al. (2018) in the last section of the study and highlight that nowadays industrial practitioners tend to take the advantages of econometric models (and this is why we compared the forecasting performance of our proposed methods with those conventional econometric models which have similar specifications but overlooks the problem of structural change). We add the following sentence to the last section:

“In this study, we also compare the forecasting performance of our proposed methods with conventional econometric models which have similar specifications but overlook the structural change problem. The ADL-intra-EWC method and the ADL-intra-IC method outperform the ADL-intra model, and the ADL-own-EWC method and the ADL-own-IC method outperform the ADL-own model. We conduct the comparison to highlight the benefit of taking into account the problem of structural change as some industrial practitioners have tried to take advantage of conventional econometric models (Fildes, et al., 2018).”

I am still not convinced by section 8. By the time I had arrived here, I was well and truly confused, and this would have been a good point to stop the paper. Instead, we get a PCA (is this a PCA?) that adds yet another layer of abstraction, and all of it post hoc. I find this section weak and unconvincing. Please consider cutting it.

We now remove this section. We summarize some of the findings in the last section, and we make it clear that the results themselves are exploratory and worth further analysis.

Please proofread the manuscript carefully. There are many mentions of "Mariana" which should be "Mariano". Often, citations indicate that an author has different first names and/or initials in the literature database, which leads EndNote to believe that these refer to different people, and to erroneously disambuguate them. In the list of references, Wildt et al. seems to have been edited by someone named "E. proceedings".

Revised. We have proofread the manuscript and corrected the reference list.

Reviewer #2: I am happy to see that the authors managed to overcome many deficiencies of the original draft. Although, originally I thought that the manuscript was to be rejected, to my surprise this version of the paper is much improved in all aspects hopefully through my comments and those of the other reviewer.

The literature review is now on spot, contributions are clearly explained, methodology is vastly easier to read through and it is easy to understand the process that the authors follow. As a result, the whole positioning of the manuscript in the sales forecasting respective literature along with the OR one overall is successful. Therefore, I would now recommend that this manuscript is accepted with minor revisions, as long as the editor agrees.

Those revisions are merely the polishing of the paper in terms of the writing. I would suggest one more check for typos and reduction of 'we...' across the manuscript.

We thank the reviewer’s advice and we have revised, proofread, and polished the manuscript.