|  |
| --- |
| **A list of responses to the reviewers' comments**  We want to thank the two anonymous reviewers for their helpful comments and valuable suggestions. We have carefully read through the reviewers’ reports and revised the manuscript based on their suggestions. We have now completely revised the manuscript, and we have positively taken into account all the comments. We believe that the paper has improved substantially with their contributions.  In addition to the modifications based on the reviewers’ suggestions, we highlight the following major changes in the revised manuscript:   1. We have completely revised the introduction section to emphasize the purpose of our research and highlight the contributions. 2. We have completely revised the literature review section. 3. We have completely revised the sections which explain the structural change problem and the methods. We have now merged the two sections. We have replaced some of the analytical equations for the EWC method with intuitive explanations. We have also described the limitations of the EWC method and the IC method, and accordingly, we highlight that their performance for retailer product sales is an empirical question. 4. We have added a more explicit description for the dataset. 5. We have more explicitly described the rationale of the three stages of the modeling process including the details of how we detect the presence of structural change. 6. We have added the description for the Base-lift method. 7. We have also updated more appropriate presentations for the error measures across all the products. 8. We have now reconducted the whole evaluation using a different setting for the sequential Chow test. For example, we now conduct the test for up to 95% of the observations, compared to the previous version where we conduct the sequential Chow test for up to 70% of the observations. The results suggest little difference in the models’ forecasting performance. Our proposed methods consistently have the best forecasting performance. 9. We have replaced he Wilcoxon SR test with the Diebold-Mariana Test. 10. We have described more explicitly, for each category, the improved forecasting performance by our proposed methods compared to the model with similar specifications but overlook the structural change problem. We show the distribution of the improved forecasting performance at SKU level for selected product categories. 11. We have now completely revised the section which explores the determinants of the improve forecasting performance by the proposed methods compared to the ADL-intra model. We remove some of the statistical measures (e.g., Skewness and Kurtosis, etc.) and construct five factors based on the remaining nine measures. The findings are consistent compared to those in the previous version. However, we tune down our claim and emphasize that the findings are only exploratory.   Please see our detailed responses to the reviewers’ comments as follows:  Reviewer #1: This is a nice paper that investigates the influence of structural change (see one point I make below) on retail data in the IRI data set. It proposes to augment established Autoregressive Distributive Lag (ADL) models, by either fitting multiple models to shorter and shorter terminal segments of the series, then averaging the forecast (so data points after the conjectured structural break are weighted more heavily), or by attempting to estimate a bias from the terminal in-sample fits and correcting for this estimated bias. The forecast accuracy is improved in either way. I have few important points to make and mostly recommend toning down some overly enthusiastic claims.    Medium points:  - Highlights: "Retailer product sales from a wide range of product categories" is not very useful.  We have now removed this point from the highlights.  - "Structural breaks" suggests a sudden and abrupt change in a parameter, like a step change in the overall level of the time series. I find this choice of words somewhat unhappy. On the one hand, in a retail environment, I would typically not expect a sudden abrupt change (except in exceptional cases, like a new store opening next door), but rather a gradual one, which would also be more consistent with drivers of such change that the authors discuss, e.g., shifts in lifestyle. On the other hand, there is a vast literature on detecting structural changes in time series (e.g., look at the documentation of the strucchange R package on CRAN), but the authors nowhere apply such a test (which would not be overly useful in my opinion, per above). Thus, maybe "structural change" instead of "structural break" would be a more useful word for the concept.  We thank the reviewer for this helpful advice. We change the term from structural break to structural change and add the following footnote:  “The term of ‘structural change’ is used interchangeably with the term of ‘structural break’ in the literature. In this study, we use the term “structural change” as in the retailer context we expect the effect of the marketing activities to change sometimes gradually rather than but sometimes in a sudden and abrupt way. We thank one of the anonymous reviewers for to pointing this out.”  Previously we described the sequential Chow test we used to detect the presence of structural change in the Appendix of the supplementary material. We have now described how we conduct the test explicitly in section 6 and we also address the possibility of implementing alternative structural change tests.  - The EWC is very similar to estimating a single model with weighted observations (or residuals), with the weights decreasing as we go further into the past. I wonder whether a direct approach like this would be feasible. It would have the advantage of returning a single parameter vector that could be examined and analyzed, whereas in EWC, one has to look at multiple such vectors, one from each of the models we average.  What the reviewer has mentioned is a possibility. We have evaluated the forecasting performance of ADL models with time-varying parameters (TVP) which has the same rationale described here. However, the preliminary results are mixed. One possible explanation is that the model may have some sophisticated structures which are not robust to the retailer data at SKU level. We add the following to the last section:  “Also, an alternative to the ADL-intra-EWC method and the ADL-intra-IC method is to directly model the change in the effect of the marketing activities, such as the time-varying parameter model. However, a disadvantage of this method is that we need to make strong assumptions of how the effects of the marketing activities change. For example, Foekens, Leeflang, and Wittink (1999) modeled the effect of marketing activities as a linear function of previous promotional activities. Their models were not developed for forecasting purposes.”  - None of the accuracy measures used is minimized in expectation by an unbiased forecast (see the introduction of Kolassa, 2016, IJF). I am somewhat concerned that the increase in accuracy may have come at the cost of bias. I would recommend using either a measure that is minimized in expectation by an unbiased forecast, e.g., a scaled RMSE, or assessing bias in some way.  We have now added the scaled MSE as one of the error measures, and our proposed methods consistently have the best forecasting performance in general. The increase in the accuracy comes from the reduction of the bias but at the cost of increased forecasting error variance. The IC method adds the estimated bias back to the forecasts, which directly reduces the bias but increases the forecast error variance. The EWC method combines the forecasts from different estimation windows. Compared to the full estimation window, other estimation windows are all smaller but with more recent information (e.g., closer to the forecast origin). Forecasts based on these ‘smaller’ estimation windows are less biased (e.g., these estimation windows contain fewer pre-structural change data) but potentially with higher forecast error variance (e.g., fewer estimation information). Thus, the final forecasts will be less biased and potentially with higher forecast error variance. Whether or not the final forecasts by the EWC method can be more accurate depends on the trade-off between the reduced forecast bias and the potentially increased forecast error variance.    - I see no discussion of how the estimation was actually carried out. Was some third-party software used, e.g., an R package? Please cite software and packages used (with version numbers). Software authors are entitled to recognition.  We estimate the model parameters using the OLS estimator, and we implement the estimation using the MODEL procedure with macros in SAS. 9.4. We have added this in section 6 and section 7.  - p. 11: what is "four-week seasonality"? Does this refer to cutting the year into 13 four-week periods (13\*4=52), then expending 12 dummies? If so, this corresponds to fitting a periodic \*step\* function, which stays constant for four weeks and then may change sharply when we enter the next four-week bin. This is discretizing a continuous variable (namely, time). Don't do this. The disadvantages of discretization have been well documented over the years - apart from the almost certainly ecologically invalid step fit I discuss, it also expends far too many degrees of freedom. Instead, use a number of periodic spline transforms of time. Three or four parameters expended here would be much better invested than twelve in discretization. Further reading here:  https://stats.stackexchange.com/questions/230750/when-should-we-discretize-bin-continuous-independent-variables-features-and-when  https://stats.stackexchange.com/questions/41227/justification-for-low-high-or-tertiary-splits-in-anova (see the links in Glen\_b's answer)  In this study, we use the deterministic four-week dummy variables (e.g., 12 four-week dummy variables for the 52 weeks) to capture the seasonality which cannot be captured by the holiday event dummy variables (e.g., Christmas, New Year’s Day, etc.). This approach has the limitation that the effect is assumed to stay constant within the four-week bins and has a cost of degrees of freedom. However, in our context, 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.  We think the performance of alternative methods for the seasonality depends on the characteristics of the data. In our study, we focus on weekly retailer data at SKU level which have unique characteristics. e.g., product sales have high variations, the effect of marketing activities change overtime, and less ‘seasonal’ compared to data from other industries (e.g., the electricity demand data), and product sales may be more driven by promotional events. Also, we propose sophisticated multi-stage model specification strategies. The methods initially include a large number of independent variables and then attempt to recursively simplify the model’s specification. Thus, the performance of any other methods to capture the seasonality for retailer product sales becomes an empirical question, especially when integrated with sophisticated model simplification strategies.  We thank the reviewer for the suggestion on how to improve the model’s performance using the alternative methods for seasonality. For example, there are other methods such as periodic spline functions and the seasonal exponential smoothing etc., which have been found useful in modeling seasonality in other fields such as electricity demand. We also thank the reviewer for the reference where the natural splines method is compared with the traditional deterministic method based on the simulation data. We address the limitation of our current approach in the last section as an avenue for further research.  - p. 19: To be honest, I do not find the introduction and discussion of the ADL-EWC-IC model convincing, since it was created after analyzing the performance of the separate models on subsamples. It is always easy to build a new model ex post that appears to perform well, but this is little better than data snooping. Can the ADL-EWC-IC model be compared to the other models on previously unseen data? If not, please label this discussion explicitly as exploratory. Similarly, please revisit the discussion in the first paragraph on p. 25.  We thank the reviewer for this useful comment and suggestion. We have now evaluated the forecasting performance of the ADL-EWC-IC model based on previously unseen data. That is, based on 1605 SKU’s from the same product categories but from a different set of 28 stores. The results are consistent (the results are now shown in Table 5). We take the reviewer’s suggestion and describe the ADL-EWC-IC model as “exploratory.”  - p. 23: In discussing a statistically significant positive coefficient for "Randomness and growth" in ADL-intra-EWC, the authors write that "This suggests that our proposed models tend to be more advantageous for the SKUs which are difficult to forecast and exhibit a trend in sales". I do not understand this. The dependent value modeled is MASE, so a positive coefficient of 0.4 should mean that a unit increase in "Randomness and growth" should be associated with a 0.4 unit (up to multiplication by 100, per the footnote to Table 7) \*increase\* in MASE. That is, ADL-intra-EWC should perform \*worse\* than the benchmark for high-"Randomness and growth" series, not \*better\*. Please clarify. Same for the rest of the discussions on the same page, and the third paragraph on p. 25.  We have now completely revised this section. The dependent variable is not the MASE but the percentage reduction of the MASE by the ADL-intra-EWC model and the ADL-intra-IC model compared to the ADL-intra model. e.g.,    where and are the percentage reduction of the MASE by the ADL-intra-EWC model and the ADL-intra-IC model compared to the ADL-intra model for SKU *i*.  Thus, in the updated results, a positive estimate of 0.21 indicates that one unit increase in "Randomness and growth" would cause a 0.21% more reduction in the MASE. We have now added a clearer description for the interpretation.  We have revised this section accordingly and make the description more explicit.  - p. 23, "All the results here indicate that we may pre-test these features for each SKU and then determine the optimal sales forecasting method specifically for that SKU." This is a very exploratory finding, not guided by prior hypotheses. Unless this can be verified on previously unseen data, please emphasize the tentative nature of this recommendation.  We thank the reviewer to point out the tentative nature of this indication. We have now revised the sentence: we only highlight the indications based on the current findings and we limit the scope of the finding by emphasizing that the determinants are only for the improved forecasting performance by the proposed methods compared to the model with similar specifications but overlook the structural change problem “Overall, we attempt to provide exploratory insights on the situations where our proposed methods may gain most benefits compared to the ADL-intra model..”  - p. 24, " The improved forecasting accuracy for product sales substantially contributes to retailers' profit". This is a very strong claim, and not backed by anything the authors did in their paper. Whether forecast accuracy improvements actually translate into better profits is not this clear-cut, since forecasts must still be translated into operational plans, which are constrained by logistics. In addition, supply chain operations rely far more on quantile forecasts than on point forecasts, since the total order includes safety amounts, and whether these quantile forecasts can be improved by the authors' proposals is not obvious. Please tone this claim down.  We thank the reviewer for this comment, and we tone down the claim of our contribution- we remove this sentence and add “Therefore, our study may provide retailers more effective forecasting methods” Minor points:  - p. 6, 1st equation: there is no intercept in the formula. If u\_t is assumed to have mean 0, this implies that a price of x=0 is associated with sales of y=0. I assume the authors meant to include an intercept parameter.  We have revised this section and we do not treat the variables to be price or price reductions (as pointed out by the reviewer, the sales will not be zero even there is no price reduction). We show the analytical evidence for a simple example where the model does not have an intercept. We highlight that more sophisticated scenarios (e.g., with an intercept and with endogenous variables) can be demonstrated using simulations. In the supplementary material, we demonstrate the impact of structural change on forecast bias and forecasting performance where the model has an intercept.  - Table 3: why do different models serve as benchmarks?  We compare the proposed ADL-intra-EWC model and the ADL-intra-IC model with the ADL-intra model because the ADL-intra model has similar specifications but overlook the problem of structural change), so that we know how much improvement is contributed by taking into account the problem of structural change. We compared the proposed methods the Base-lift method because it is still being widely used by industrial practitioners. We have now highlighted this in the revised manuscript.  - Figure 3: please provide more information in the figure caption instead of in the text or even in footnotes - the reader should not need to hunt through the text to understand the figure. Are the diamonds joined by lines group means? Are box widths \*proportional\* to numbers of SKUs in each category? "Are determined by" can be a log transform, a square root or anything else. Please ensure that the horizontal axes have the same extension so the plots are comparable, and that the whiskers are not cut off by the figure bounding box.    We have now described the Figure explicitly. We have provided the detailed information and we have also reproduced the Boxplots. The box widths are now proportional to the number of SKU’s in each product category. The diamonds represent the group means for each product category and joined by lines for illustration. The Boxplots are now with the same extension for the horizontal axes and without outliers being clipped.  - Table 7: please indicate in the table caption what "their counterparts" are. The entire table is unclear to me; how do the top and the bottom half differ? What does "Model with 5 factors and category dummy variables" as a caption to the bottom half (or is it?) refer to, in contrast to the top half? The text says on p. 23 that "the horizon is one to eight-week ahead", but the table says "Horizon = 8" - please clarify whether the horizon is eight weeks or \*up to\* eight weeks.  We have now completely revised Table 8 (the previous Table 7). It now shows the parameter estimates for the regression model with the five factors as independent variables. Previously We developed another regression model where the independent variables include the five factors and also category dummy variables. The parameter estimates of the five factors are consistent for these two models. The revised Table 8 now only includes the parameter estimates of the model which has the five factors as independent variables. This makes Table 8 more readable. Also, we put the following footnote:  “For robustness, we have developed an alternative regression model which also include dummy variables to capture potentially unobserved category effects, and we find the parameter estimate for the five factors to be consistent with those shown in Table 8.”  We have also clarified the description for the horizons.  - References: please provide full details for Loeb (2015) Revised  Typos:  - p. 3, 2nd para, l. 3: "The model which is subject to structural break" - either add "a" or change to "breaks"  Revised  - p. 3, 2nd para, l. 7: remove "in" after "including"  Revised  - p. 4, l. 2: "mention" -> "mentioned"  Revised  - p. 4, l. 3: "values" -> "value"  Revised  - p. 4, 2nd para, l. 2: "the change of" -> "changes in" Revised  - p. 5, l. 5: "forecast" -> "forecasts" Revised  - p. 6, 2nd para, l. 3: "structure" -> "structural" Revised  - p. 6, third equation block: italic and upright versions of beta, X and Q are mixed, which is painful to me (also on p. 7)). Please ensure proper mathematical typesetting. Lowercase x\_{T+h} and uppercase X\_{T+h} is used inconsistently here.  We thank the reviewer for this correction. We have now revised those inconsistent versions of letters and symbols. We have now discarded the symbol “Q” and use the lowercase consistently.  - p. 7, equation for the bias correction: lowercase omega is used both as the starting index of the summation (with the ending index indicated by W - why mix Greek and Latin?) and as the summation index itself.  Revised  - p. 7: "The estimated bias are" should be "is" Revised - p. 10, equation: can be slightly simplified by removing "=\eta, \eta"  We thank the reviewer- we think it might be more readable if we keep it.  - p. 19/20: Figure 3(c) is not "in the bottom-right corner", but in the bottom left.  We have now removed the previous Figure 3(c). This is because that we now focus on the ADL-intra-EWC method and the ADL-intra-IC method. We only consider the ADL-EWC-IC model to be exploratory and thus we do not show too much of its details for simplicity.    Reviewer #2: This is an interesting paper that is trying to investigate the forecasting performance of several ADL models over retailer product sales. The study is particularly focusing on the effects of structural breaks originating from marketing activities over the products.  Main Comments: 1.     I found the structure of this paper very confusing. For example, the introduction and literature review section are very poorly written with many overlaps and repetitions that are not at all informative for the reader. The contribution of the paper, as outlined, is very weak.  We have now completely revised the structure of the paper and we have taken out some of the arguments that may sound a bit repetitive. We have also highlighted explicitly our contributions.  2.     In the introduction the authors should clearly indicate what is the model they are introducing and why. On top of that, it should be made clear to the reader why the specific models seem appropriate for the retail forecasting exercise. It is quite astonishing that the reader does not get a gist of what he/she is going to see unless he/she reaches page 10.  We have now introduced what is new in our research earlier in the paper than we did in the previous version. We have also highlighted the value of the work regarding how it might impact inventory management practices similar to other related work in the field of forecasting.  3. The literature review is very short and several references are packed all together without any meaningful commentary (for example in page 5). I would really be interested to see what are the findings of studies assuming constant marketing activities, as this would highlight/clarify/validate potentially comparisons with the models at hand.    We have now completely revised the literature review. In section 2.1., we summarize the findings of previous studies which forecast retailer product sales at SKU level. e.g., their proposed methods and the rationales in more detail. In section 2.2., we summarize the (changing) effect of the marketing activities.    4.     The text is characterized by some generalizations that make the reader confused on what the authors are claiming. For example, in page 3 '… The data in retailer product sales… macroeconomics).' I would suggest a thorough read-through to the authors in order to make the text more to the point.  We take the reviewer’s suggestions and we have now streamlined the manuscript in line with this comment and taking into consideration other comments/suggestions by the other reviewer too.  5.     Section 3 and 4 seem a bit redundant the way they are presented. I would expect to see a methodology section, where these two sections could motivate/inform the selection of models in section 6. The ideal approach would be that section 3 and 4 are reduced substantially and included in the commentary of section 6 or if needed in an appendix.  We have now completely revised these sections.  6.     In section 5, it should be explained what the display and feature percentage is along with the motivation of the selection of these inputs. In my view, the data section should have been after the intro/literature review.  We have now added descriptions below Table 4. We put the section for the structural change after the literature review. It explains what would happen if we overlook the change in the effect of the marketing activities, which was introduced in the literature review section.  7.     In terms of the analysis during structural breaks, it is interesting that the authors do not make explicitly clear how they detect, test or analyse their results for structural breaks. It would be expected that in that type of paper where forecasting performance is evaluated within the presence of structural breaks that vast emphasis would be given on that issue. The general information provided in section 4 are not enough in my opinion. What tests have been done and what was the result? Except from the Chow test, have the authors investigated the Andrews approach (2003 , Econometrica) or Fixed Regressor Boostrap? The vague analysis on that aspect is a main shortcoming of this paper.  We previously described how we conduct the sequential Chow test in the Appendix in the supplementary material - we have now included this in section 6 of the paper. We conduct the sequential Chow test for up to 95% of the weeks in the estimation period. Suppose that we have an estimation period of 160 weeks. We conduct the Chow test for each of the 160\*0.95= 152 weeks (e.g., the 152 weeks in the centre of the 160 weeks, from week 5 to week 156). Each time we assume that there is a structural change occurring at one of these weeks. For example, we initially conduct the Chow test assuming a structural change occurring at week 5, and we obtain the p-value. We then conduct the Chow test for week 6, 7, and so forth until week 156 and each time we obtain the corresponding p-values. We reserve at least 5% of the weeks for the estimation of the test. Thus, we may obtain up to 152 p-values in total. The null hypothesis of no structural change will be rejected if any of these p-values is below the threshold. To mitigate the multiple comparison problem, we adopt a very small threshold, i.e., 0.001 instead of the commonly used 0.05.  In the revised manuscript, we highlight the fact that previous studies have proposed alternative tests (e.g., Donald W K Andrews, 1993; Donald W. K. Andrews & Ploberger, 1994; Bai & Perron, 1998, 2003; Brown, Durbin, & Evans, 1975). However, these tests have different focuses (e.g., the size and the location of the structural change) and more stringent assumptions (e.g., a known number of multiple changes as a priori knowledge). The estimation of the locations and the sizes using these tests were not satisfactory (Pesaran and Timmerman, 2005). In our study, the purpose is neither to detect the locations nor the number of structural changes, but focusing on investigating the presence of any structural change, so that we can estimate and then offset the bias (if using the IC method) or to accept a trade-off between the forecast bias and the forecast error variance (if using the EWC method). Therefore, we conduct a sequential Chow test which serves for this purpose and has the benefit of a simple implementation. The empirical results suggest that our models generate more accurate forecasts.  The Andrews’ approach published in *Econometrica* in 2003 (e.g., the end‐of‐sample instability test) has an advantage that it can be used when there are very limited data (e.g., even one observation) before or after structural change within the estimation sample. In comparison, to implement the sequential Chow test, we need to reserve some observations before and after the structural change (e.g., it is an F-test which compares the fit of the model before and after the structural change). However, as pointed out by the other reviewer, the effects of the marketing activities tend to change in a gradual way rather than an abrupt way. Thus, it is unlikely that there is a sudden and abrupt structural change occurring exclusively for a very few observations which are close to the forecast origin or the beginning of the estimation window (and it matters only if there are so few observations that a Chow test cannot be conducted). Empirically, we tried conducting the sequential Chow test with different settings so that we know if this situation potentially matters. For example, in the revised manuscript, we update the results by conducting the sequential Chow test for up to 95% of the weeks in the estimation period. This is compared to the previous results where we conduct the sequential Chow test for up to 70% of the weeks. Our proposed methods consistently have superior forecasting performance and the results for the two settings suggest little difference. Therefore, it is unclear if we could benefit from the advantage of the Andrews’ instability test as we do not see benefit when we push the assumed location for the structural change towards the edges of the estimation window. Also, the Andrews’ instability test assumes that explanatory variables must be strictly stationary. Otherwise the test will be associated with a distorted inference and mix the instability of the explanatory variables with the instability of the regression model. In the retailer context, the assumption of stationarity of the explanatory variables may not always be hold as product prices increase gradually during the two-to-three-year timeframe due to inflation, though intermittently associated with price reductions.    We thank the reviewer for pointing out the possibility of further improvement if a set of alternative tests can be evaluated and we leave the empirical question for future research.  8.     In the modelling part, I would expect to see a small description at least of the benchmark model, the base-lift model. Also, it is not well explained why Lasso is used two times. What motivates the authors to double apply Lasso shrinkage? Is there any similar study suggesting that? What are the expected benefits of this approach? Given that this paper comes down to a horse-racing application between ADL models constrained by Lasso operators, it is important to clarify why the ADL-raw models is combined with the ADL-own model? Wouldn't be a logical question from the reader as to why the Lasso operator is not used in a third stage, for example? A lot of discussion is provided before-hand (in cases unnecessary), but at section 6 that the reader wants to understand the reasoning and process of modelling, he/she struggles to do so because of the lack of explanations. Another important shortcoming for this work.  We have now added an explicit description for the Base-lift benchmark model.  We have now added explicit explanations for the modeling process of the ADL-intra-EWC method and ADL-intra-IC method. The LASSO procedure was initially used as a variable selection method to identify the important marketing variables (e.g., Huang et al., 2014). It was then used as a model simplification strategy following Ma et al. (2016). The general ADL model becomes the ADL-raw model after being simplified by the LASSO procedure, as shown in Figure 2.  We choose the LASSO procedure as a model simplification strategy because it proves to be effective, and it is also automatic. However, it runs the risk of missing important variables (e.g., potentially the price and promotions of the focal product). If we miss important variables, the final forecasts will be biased, and the forecast error variance will also rise (and this bias is not the bias incurred by the structural change but from the bias of the parameter estimate). Missing important variables is far more serious compared to the cost of efficiency (Davidson & MacKinnon, 2004).  Thus, we try to avoid this as much as we can. For example, if the price and promotion variables of the focal product are not included in the ADL-raw model, we try to bring them back, but only if they are retained in the ADL-own model by the LASSO procedure (as this suggests that they are useful/important). The supplementary parallel ADL model, i.e., in equation (8), by definition, has fewer explanatory variables compared to the general ADL model, i.e., in equation (7), and is less likely to suffer from multicollinearity compared to the latter. Thus, if the price and promotions of the focal product truly have effects on the product sales, it is less likely that they will be removed by both the ADL-raw model and the ADL-own model. However, if we further implement the LASSO procedure, we will bear the risk of missing important variables.  9.     In the experimental design section, the representation of the competing models is not clear. The authors should have included perhaps a table or a clear text on what is the differences between models in a concise and transparent way. Additionally, I am not sure why the roll-forward period is equal to two weeks, while the forecasts are done for 1, 4 and 8 weeks. The authors should explain why they used sMAPE over MAPE or why MASE is more informative. In terms of statistical accuracy, RMSE and Theil-U statistics are also quite often used. One final point is why are the authors inclined to use the adjustment of Cooper et al. (2009). This was not quite clear to me.  We have now added a clear text which highlights the feature and the difference between the models.  We implement the models with 18 rolling events, and for each time we roll the full estimation window forward for two weeks. Therefore, we will be able to evaluate our models for a longer time span (e.g., under this condition, the initial full estimation window is [1:160] and the last estimation window is [35:194], thus the data we used have a time span from week 1 to week 194, compared to a shorter time span from week 1 to week 167 if we only roll forward one week each time). This may potentially make our results more robust as we evaluate our models for a longer time span.  We include various traditional error measures in the evaluation. These error measures capture different aspects of the unobserved loss function for the retailer. We have explained why we report the results for the symmetric MAPE in a footnote. For example, the MAPE does not have an upper bound and vulnerable to outliers. We have added the descriptions for the advantages for the more recently developed error measures including the MASE and the RelAvgMAE. For example, they are more robust to outliers while the latter is readily interpretable as the percentage improvement (or worsening) of the focal method compared to a benchmark. These error measures need to be used as a whole to form the broad picture of the model comparison. We have now added the scaled MSE as an additional error measure. It conveys the information of the RMSE and the Theil’s U. Our methods consistently have the best forecasting performance.  We use the adjustment of Cooper et al. (2009) to mitigate the bias due to the logarithm transform (e.g., the expected value of the log-transformed variable does not equal to the log transform of the expected value). We have now added this in a footnote.  10.     The section 8 is quite extensive in terms of results in tables, but the comments are quite limited and badly written. For example, is there consistency in ranking across statistical measures? What exactly is the importance of Table 3? At such a forecasting exercise, a Diebold-Mariano test is needed to assess the significance of the differences in performances. The Wilcoxon Sign Rank is not enough. It is also not quite clear what Table 4 is offering. In Figure 3, is there a particular reason why these six product categories are selected for analysis?  We have now highlighted that the results are consistent for all the error measures.  We have now replaced the Wilcoxon SR test with Diebold-Mariano (DM) test in Table 3. We include Table 4 to demonstrate the performance of the various models depending on whether the focal product is being promoted. This is because the product sales exhibit very different characteristics (e.g., much higher variations) for the promoted period compared to the non-promoted period.  In Figure 3, we choose the six product categories for which the ADL-intra-EWC model and the ADL-intra-IC model have the highest advantage over the ADL-intra model (which has similar model specifications but overlook the problem of structural change), and we show the distribution of the improved forecasting performance using boxplots. We have highlighted this explicitly in the revised manuscript.  11.     Table 5 should probably be within the analysis of the determinants of the forecasting performance. I also find the explanation of the process confusing. Why is this factor selection followed? Have the authors explored another factor analysis?  We include the previous Table 5 (now Table 6) in section 7 to highlight the improved forecasting performance of the proposed methods for each category (i.e., at category level). We have now added explicit descriptions in section 7.  We have now completely revised section 8 which includes the factor selection. In section 8, we constructed factors from various measures for the data characteristics, and we explore if they are potentially the determinants of improved forecasting performance of the proposed methods (i.e., at SKU level)  We have now completely revised this section. We now construct the five factors using nine statistical measures because some previous measures are not informative (e.g., range, skewness, and kurtosis etc.) and they were allocated into different factors. This may make the interpretation of the factors difficult. The updated results have consistent findings. We have now revised Table 8 with descriptions and made it more readable.  Minor comments: 1.     The writing of the whole paper is not very good. A lot of proof-reading is required. Some examples: \*     'Under such a circumstance…activities' in the abstract.  Revised \*     '…the generated forecasts may potentially…' in the abstract  Revised \*     SKU abbreviation should be defined in the abstact and in text and then explained perhaps on a footnote.  Revised \*     '…proposed holistic methods to generate…' page 4, line 57  Revised \*     Footnote 7 should refer to figure 1.  Revised 2.     Sometimes writing lack of academic standard. For example, there is no numbering in equations, there is extensive use of bullet points within text, tables are not explained with footnotes, many equations appear within text, the tables style is not uniform etc.  We have now taken on board the comment on the writing style of the paper and we have now completely revised the corresponding sections.  3.     The paper is not well positioned in the OR forecasting literature.  We have now completely revised the manuscript. We have more explicitly and effectively positioned the work as one that can impact on the inventory management of retailers similar to related work in the domain of forecasting.  Overall, based on the above I am inclined to reject the paper due to its lack of solid contribution, convincing results, presentation and academic rigour. |
|  |

Bottom of Form