**A list of responses to the reviewers' comments**

We would like to thank the anonymous reviewers for their kind comments and valuable suggestions. We carefully read the reviewers’ reports and revised the manuscript according to their suggestions. All of the comments of the reviewers have been positively taken into account. The paper has improved substantially with their contributions.

In addition to the modifications according to referees’ suggestions, a major change in the revision is that we update the results by adding one-half the mean-squared error to the forecasts before transforming them back into sales units. Usually this adjustment may not be an issue in terms of accuracy so that it is often neglected, but when we try to response to one of referee’s bias measure suggestion, we realize it does have a relative large effect on the forecasting bias, which is potentially important in stocking decisions. We have also attempted to improve the clarity (and correct grammatical errors) in this version.

Our detailed responses to the comments of the reviewers are given below.

Reviewer #1:

This article reports different forecasting causal models capable of dealing with high dimensional data. This paper also analyses the impact of intra and inter category promotional information. This work employs weekly sales with promotional information at SKU level.

I think the paper is well-written (although I've found some minor mistakes) and well-structured. The topic is relevant and important, since many companies that rely on marketing campaigns could find this research useful. The work accomplished by the authors has been substantial and the results are promising.

Nonetheless, the paper in its present status has important weaknesses and until all of them are properly addressed, I would not recommend this paper to be published.

1. **Literature review:**

Although the authors have done a good literature review, in general. There is an important weakness in it, namely the judgmental forecasting part. It should be noted that many companies still use judgmental adjustments to forecast promotional sales. Therefore, I think it is important to review that stream of research. References that can help authors are:

[1] Fildes R and Goodwin P (2007). Against your better judgment? How organizations can improve their use of management judgment in forecasting. Interfaces 37(6): 70-576.

[2] Trapero JR, Pedregal DJ, Fildes R and Kourentzes N (2013). Analysis of judgmental adjustments in the presence of promotions. International Journal of Forecasting 29(2): 234-243

[3] Trapero, J.R., Kourentzes, N., Fildes, R. (2014). "On the identification of sales forecasting models in the presence of promotions", Journal of the Operational Research Society, In press.

It would have been very interesting to be able to compare your results with those obtained by experts (if available).

**Authors’ response:**

Thanks for this suggestion. We’ve adopted this suggestion by adding relevant references on judgmental adjustment in section 2.3, page 7. Unfortunately the IRI dataset we employed in this research doesn’t provide expert judgmental forecasts, so we can’t compare our forecasting results with judgmental adjustments.

1. **Model building**.

It is important that the model can be automatized. Nonetheless, I miss in the article some interpretation points. In particular:

2.1 Are all the variables included in the model statistically significant? Have you carried out any kind of analysis about that?

**Authors’ response:**

This is a very good question which we also considered during the conduction of this research. In our four steps process, no significance test is included to determine whether the selected variables by LASSO are statistical significant or not. LASSO is a regularization technique for simultaneous estimation and variable selection which continuously shrinks the coefficients toward 0 as the penalty increases. It is not based on traditional statistical significance test to select variables. The usual constructs like p-values, confidence intervals, etc., do not exist for LASSO estimates. Also we think it is unnecessary to do such a test because that doesn’t help in terms of improving forecasting accuracy. What is the purpose of such a test? To drop variables which are insignificant? LASSO has done good work in selecting variables and parameter estimation, how should we select variables and estimate the model again? Studies have shown that the models estimated by the LASSO procedure delivers forecasts significantly superior than the significance test based benchmarks models, e.g., Autometrics (Epprecht et al., 2013). After all, the objective of this research is not to select significant variables, but to improve forecasts.

In addition, there is still no answer on how to produce a p-value for each predictor given fitted lasso model at some *λ* (Lockhart et al.2014). The distribution of such a sparse estimator is non-Gaussian with point mass at zero, and this is the reason why standard bootstrap or subsampling techniques do not provide valid confidence regions or p-values.

References:

Lockhart R., Taylor J., Tibshirani R.J., Tibshirani R. (2014). Annals of Statistics, 42(2), 413-468.

Epprecht C., Dominique G. and Álvaro V., (2013), Comparing variable selection techniques for linear regression: LASSO and Autometrics, Documents de travail du Centre d'Economie de la Sorbonne, Université Panthéon-Sorbonne (Paris 1), Centre d'Economie de la Sorbonne, <http://EconPapers.repec.org/RePEc:mse:cesdoc:13080>.

2.2 Given the econometric nature of your models. Have you checked whether the residuals may contain any kind of correlation that can be modelled (by an ARIMA function, for instance)?

**Authors’ response:**

In the preliminary analysis, we carried out a Ljung-Box test with 30 lags on in-sample residual series from ADL-inter-all with three-stage LASSO. We found 761 residual series in 926 SKUs couldn’t meet the independently distributed assumption at 0.05 significant level. Considering that the number of significantly auto-correlated residual series are relative small, and some of these are just significant by chance (i.e., 50 out of 1000 white noise series would have chance to be significant at 0.05 level), we didn’t do further modeling efforts on those auto-correlated residuals. But still it would be interesting to investigate the potential for improving forecasts by modeling those residuals in the future studies.

2.3 There is a common and very simple forecasting algorithm entitled last-like forecasting algorithm that has not been used as a benchmark, unlike other references cited in your paper that use it. In particular, see section 3.3.1 (exponential smoothing with lift adjustment) in:

Gür Ali, Ö., SayIn, S., van Woensel, T., & Fransoo, J. (2009). SKU demand forecasting in the presence of promotions. Expert Systems with Applications, 36(10), 12340-12348.

In the end, if the models that authors proposed (quite complex) do not improved substantially such a simple algorithm, then the cost maybe is not worthwhile.

**Authors’ response:**

Thanks for this suggestion. We’ve adopted this by adding the base-lift as another benchmark model in the current manuscript. Please see section 4.2 and 4.5.2 in the revised paper for the details.

1. **Results**

3.1 Authors have been focused on magnitude error metrics, however there is a lack of bias error metrics as the Mean Error, for example.

**Authors’ response:**

Thanks for this suggestion. We’ve added the Mean percentage Error as a bias metric in the revised paper. Please see section 4.5 for the details.

3.2. I think it would be very interesting to look into the improvement percentage of the forecasting models proposed by the authors when there is a promotion on the SKU and when there is not any promotional activity in the SKU. In other words, up to what extent promotional activities of other SKUS can improve the forecasting accuracy of other SKUS that are not under any promotion. Furthermore, when there is a promotion on a determined SKU how much improvement can be expected if we use such own SKU promotional information.

**Authors’ response:**

Thanks for this suggestion. We’ve adopted this in the revised paper. We divide the forecasting periods as periods when the focal SKU is under promotion and periods when the focal SKU is non-promoted. We then compare the forecasting improvements in the two periods. Please see section 4.5.3 in the revised paper for the details.

**Reviewer #2:** The article is concerned with the relevant problem of SKU-level sales forecasting. The authors discuss state of the art and argue convincingly for the need to bring scalable, automated data mining approaches to the fore. Several approaches and methodologies mostly focused on the LASSO method are described and compared.

Several of the highlighted research advances can be useful for practitioners. The article is clearly written and easy to follow, with some areas for improvement.

Proposed areas of improvement:

1. I missed detail on the LASSO model selection. There's a "Complexity" meta (hyper) -parameter in the LASSO that controls the trade-off between model complexity and performance. It is important to describe how this parameter is determined.

**Authors’ response:**

Sorry about that. We’ve put a new sentence in section 3.3, page 15 as follows:

We use 10-fold cross-validation to obtain optimal value of penalty parameter in LASSO to give minimum cross-validated error.

2. LASSO encourages sparse models, but it is not the only alternative to reduce high variance in situations with more parameters than data points. LASSO assumes that the interdependency structure is sparse, but can we really assume that? A natural alternative is to try ridge regression (L2 penalty). L2 might perform as well or even better than the LASSO, when there are many, many tiny marketing effects combining to create a substantial response. The article would benefit from a discussion of ridge regression, and why it is not used as a natural benchmark for this study. Even better, add ridge regression to the model comparisons.

**Authors’ response:**

The “Bet on sparsity principle"(in the Elements of Statistical learning, Hastie, Tibshirani and Friedman, 2009) may give an answer about sparsity assumption when using LASSO dealing with high dimensionality. The L1 methods assume that the truth is sparse, in some basis. If the assumption holds true, then the parameters can be efficiently estimated using L1 penalties. If the assumption does not hold—so that the truth is dense—then no method will be able to recover the underlying model without a large amount of data per parameter.

Because of the form of the L1-penalty, the LASSO does variable selection and shrinkage, whereas ridge regression, in contrast, only shrinks. L1 penalties are convex and the assumed sparsity can lead to significant computational advantages. Ridge regression keeps all the variables in the model. If only a small part of them are useful, then most of the variables in the model will certainly generate more noise to worsen forecasts.

In the revision, we’ve added the ridge regression to the model comparisons. Please refer Table 2, page 23 for the results. With full set of information, the forecasting performance of ridge regression is worse than that of all the models estimated by LASSO. In fact, apart from the results reported in current paper, we also tested a series of other penalization based variable selection and model estimation algorithms, including ridge regression, as well as some newly proposed algorithms, such as the smoothly clipped absolute deviation (SCAD) in Fan (2001) and minimax concave penalty (MCP) in Zhang (2010). We found that LASSO is the most efficient method while obtain most accurate forecasts among others.

3. The article argues that the interrelationships are volatile and changeable from week to week. Accordingly the rolling scheme re-selects variables week-by-week. But Table 3 shows a fixed set of influential categories. I assume these are from the fixed scheme analysis. But this needs to be called out. I would also like to see results how the selected variables may change from week to week (do they?), to get a sense of the structural variation and to better understand "robustness" issues.

**Authors’ response:**

At the category level, we didn’t run a rolling scheme. So Table 3 reports the sets of influential categories only evaluated from calibration periods same as fixed scheme analysis. In the revision, we add a brief explanation about this in the section 4.4 of the paper.

At SKU level, both the structure and the value of selected variables could be changing during the rolling scheme. It is very hard to find a concise way to show how the selected variables changing from week to week. For each SKU, there are thousands candidate variables in each rolling period. By comparing the forecasting results, we can infer this structure changing.

4. The method to combine "diffusion factors" with LASSO selection should be made more clear. I was not able to understand how the clustering exactly works, and how this method would get around the principal PCA limitation (PCA ignoring the dependent variable).

**Authors’ response:**

Thanks for this advice. The clustering is explained in section 3.2 (page 12) as:

In the empirical study, each explanatory variable, i.e. sales lag, price, display and feature, across SKUs in the same category is regarded as a cluster. For each cluster, we conduct PCA dynamically and extract m Principle Components (PCs). So if we have *v* types of marketing instrument and c categories, then we extract *v*\**c*\**m* PCs.

The reason on how this method would get around principal PCA limitation is further explained in section 3.2 (page 13) as:

The PCA is an effective approach to lower the variable dimensionality, but it has a drawback in forecasting applications. Eigen-vectors corresponding to large eigenvalues are retained whereas those associated with small eigenvalues are discarded. Thus, the retained factors might not have any predictive power of the dependent variable whereas the discarded factors might be useful (Stock and Watson, 2002). Here we conduct PCA dynamically as the inputs to the proposed multistage LASSO. With the aid of LASSO, we can input more diffusion factors into the model (we use between 270 and 450 factors as candidate predictors in the empirical study). Thus the final retained factors are no longer only determined by their eigenvalues, but also by their predictive power. This combines the merit of PCA which is effective in dealing with collinearity and LASSO which is good at variable selection in high dimensional space while overcomes the drawbacks of each.

5. Pg. 22 has a typo: 'falls' when it should say 'fails'

**Authors’ response:**

Thanks for pointing this out. We’ve corrected them in the revision.