Response to reviewers

Optimal forecast reconciliation with time series selection 2024-08-23

We thank the reviewers for their careful reading of our paper; their comments have led to several improvements and corrections. In this revision, we have addressed all the comments raised by the reviewers, and we provide a point-to-point response to each comment of the review team. Reviewer comments are in black, our responses are in green.

Reviewer 2

The paper reformulates forecast reconciliation as a grouped variable selection problem. Regularisation and subset selection are carried out using state-of-the-art optimisation techniques, including Mixed Integer Programming. Recognising that reconciled forecasts $\tilde{\boldsymbol{y}}$ are simply linear combinations of base forecasts $\hat{\boldsymbol{y}}$ via $\tilde{\boldsymbol{y}} = \boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{y}}$, setting columns of $\boldsymbol{G}$ to zero can eliminate heavily misspecified forecasts from the combination. The simulation results and empirical examples are thorough. In so far as selection methods work, they do so when a diagonal matrix is plugged in to the objective function in place of the forecast covariance matrix.

Main Comments

- The discussion around Proposition 1 is confusing. The discussion mixes up the roles of 1) the constraint $\boldsymbol{G}\boldsymbol{S} = \boldsymbol{I}$ and 2) the property of preserving the unbiasedness of forecasts after reconciliation. The proposition as it is currently phrased states: *"Under the assumption of unbiasedness, the count of nonzero column entries of* $\boldsymbol{G}$*,… derived through solving equation 4 is at least equal to the number of time series at the bottom level."*. However, equation 4 is solved with the constraint $\boldsymbol{G}\boldsymbol{S} = \boldsymbol{I}$. This is what guarantees that unbiased base forecasts will remain unbiased after reconciliation. The wording implies that this reasoning operates the other way around.

  A further implication of imposing the constraint $\boldsymbol{G}\boldsymbol{S} = \boldsymbol{I}$ is that $\boldsymbol{G}$ must have no less than n_b non-zero columns, as is correctly argued in the proof to the proposition. However, an alternative and more precise wording of Proposition 1 would be *"If the assumption that forecast reconciliation preserves unbiasedness is imposed by enforcing* $\boldsymbol{G}\boldsymbol{S} = \boldsymbol{I}$*, then the number of nonzero column entries of* $\hat{\boldsymbol{G}}$ *will be no less than* n_b*"*.

  (Done) Improve the wording of the first part of Proposition 1.

- The second part of the Proposition 1 states *"In addition, we can restore the full hierarchal structure by aggregating/disaggregating the selected time series"*. This again is somewhat imprecise and possibly refers to a different issue to that proven here. I suspect that what is meant by 'restore' here, is the following. If the solution to Equation 4 yields a $\hat{\boldsymbol{G}}$ with exactly n_b non-zero columns, then these correspond to variables from which the full hierarchy can be obtained using nothing but the information embedded in the constraints. As the authors suggest, it may be possible that the zero columns correspond to series AA and BA, but not to series AA and AB, since in the later case, the aggregation constraints alone are insufficient for forecasts of AA and AB to be recovered. This is not a consequence of "assuming unbiasedness" or even that $\hat{\boldsymbol{G}}_{\cdot\mathbb{S}}$ has n_b columns. It is a consequence of enforcing the constraint $\boldsymbol{G}\boldsymbol{S} = \boldsymbol{I}$.

  To this I would add the following observations. First, solving equation (4) could lead to a solution for $\hat{\boldsymbol{G}}$ that has more than n_b non-zero columns. In this case there are in fact too many series and a while a coherent forecast can be 'restored', this cannot be done uniquely. Second, is the point that it is the constraint $\boldsymbol{G}\boldsymbol{S} = \boldsymbol{I}$ that enforces that the selected columns of $\hat{\boldsymbol{G}}$ will correspond to variables that can restore the hierarchy. The mechanism by which it does so, is not rigorously proven here, however, a proof should be possible by leaning on some of the arguments made in Zhang et al. (2023), which is cited by the authors.

  (Done) Improve the wording of the second part of Proposition 1, and add some explanations to include the observations the reviewer suggested.

(Done) Add proof for the second observation.

- As well as the statement of Proposition 1, the proof could be made tighter and clarified. In particular the same equation is essentially presented twice, once with $\hat{\boldsymbol{G}}$ and the second time with $\hat{\boldsymbol{G}}_{\cdot\mathbb{S}}$. Only the second of these is needed.

  (Done) Make the proof for Proposition 1 tighter and clarified.

- Finally, in terms of tightening up the mathematics on page 8, at the very bottom, the line $\text{vec}(\hat{\boldsymbol{y}}) = \hat{y}$, while correct, adds nothing since $\text{vec}(\hat{\boldsymbol{y}})$ is not used in equation (5). Also, on the second line, it would be worthwhile to make it explicit that $\boldsymbol{SG}\hat{\boldsymbol{y}} = \text{vec}(\boldsymbol{SG}\hat{\boldsymbol{y}}) = (\hat{\boldsymbol{y}} \otimes \boldsymbol{S})\,\text{vec}(\boldsymbol{G})$, currently the second and third terms are present but not the first.

  (Done) Tighten up the proof for Proposition 2.

- Some of the methods discussed in the literature review are 'in-sample' methods in the sense that $\hat{\boldsymbol{y}}_{t+h|t}$ are predictions in the form of fitted values ($\boldsymbol{y}_{t+h}$ is in training data when base forecasts are computed). Others (for example the RERM method) are 'out-of-sample' in the sense that $\hat{\boldsymbol{y}}$ are genuine forecasts. In principle all optimisation methods could use either an in-sample or out-of-sample approach. I believe that in this paper only 'in-sample' methods are considered. This is reasonable, however, this should be clearly stated at some point (and it would provide motivation for not using RERM in the simulation studies and empirical results).

  (Done) Point out the "in-sample" and "out-of-sample" methods when discussing methods in the literature review and introducing proposed methods.

  (Done) Provide reason for not using RERM and ERM in the simulation studies and empirical results. They demand extensive rounds of model training and significant computation time.

- More guidance could be given on the similarities between methods. For example Elasso seems to be the same as OLS-lasso with the important difference that only the latter retains the $\boldsymbol{GS} = \boldsymbol{I}$ constraint. This also begs two further questions. The first is why a $\boldsymbol{W}$ matrix is not used in the Elasso method. The second is why the $\boldsymbol{GS} = \boldsymbol{I}$ constraint is not dropped for other regularisation approaches.

  Clarify that the difference between Elasso and OLS-lasso lies in the presence of the constraint GS=I and also data sources.

  (Done) Discuss similarities between methods. (1) Similarities between three constrained "out-of-sample" reconciliation methods. (2) The difference between Elasso and the Subset, Parsimonious, and Lasso methods, which also explains two further questions raised by the reviewer.

  (Done) How to use a W matrix for Elasso? Discussion.

- Where the unbiasedness preserving property is dropped, the authors could also consider including an n_b-dimensional shift parameter $\boldsymbol{d}$, such that $\tilde{\boldsymbol{y}} = \boldsymbol{S}(\boldsymbol{d} + \boldsymbol{G}\tilde{\boldsymbol{y}})$. Then $\boldsymbol{d}$ can be trained alongside with $\text{vec}(\boldsymbol{G})$ and act as a bias correction. The problem should still be able to be written down as a least squares problem, optimising w.r.t. $(\boldsymbol{d}, \text{vec}(\boldsymbol{G}))$.

  (Done) Extend the Elasso method by including a shift parameter as a bias correction. Part of discussion.

- In the intuitive method it is on the one hand stated that '*when the jth diagonal element of* $\boldsymbol{A}$ *is zero, the jth column of* $\bar{\boldsymbol{G}}$ *becomes entirely composed of zeros'*. However, later it is stated that *"implementing grouped variable selection… can be challenging because it imposes restrictions of* $\bar{\boldsymbol{G}}$ *to ensure it adheres rigorously to the analytical solution of MinT while making the selection"*. The second, quite confusing statement, seems to contradict the first, if a zero element in $\boldsymbol{A}$ implies a zero column of $\bar{\boldsymbol{G}}$ then why is grouped selection even necessary? Also, when calling a method 'intuitive' it is important to discuss what makes it intuitive. Little intuition is given to motivate this method, rather an appeal is made

to reduce the number of parameters - perhaps 'parsimonious' method would be a more appropriate name.

(Done) Reword the second statement for the Intuitive method.

(Done) Rename the Intuitive method to the Parsimonious method.

(Done) Rename in all results.

- On page 24 the statement is made that *"the Elasso method consistently outperforms the others overall"*. Given that the Elasso performs poorly at short horizons and for some groups of bottom level variables, I think the use of 'consistently' is not warranted here.

  (Done) Improve the result explanation to make it more rigorous. Now "the Elasso method outperforms the others when evaluated on average results across the entire hierarchy and forecast horizon".

- The results for most methods are very close to one another. Some testing on whether the observed differences are significant should be added.

  (Done) Add MCB test for simulations.

- While it is valuable to report the series that tend to be selected more often, some additional context would be useful - in particular the forecast variance of each series (in order to determine whether series with high forecast error are dropped) and the forecast correlation (to determine whether uncorrelated series are selected).

  (Done) Report MASE of each series and correlation heatmap for one-step-ahead forecast errors. Hint: Higher-level series tends to have larger variance. The methods are preferred when the error correlation within the structure is negative.

Reviewer 3

This paper proposes novel forecast reconciliation methods incorporating time series selection. Two categories of such methods are proposed - one category is based on out-of-sample information while the other category of methods is based on in-sample information. These are illustrated via simulation studies and two empirical applications. The findings are argued to demonstrate improved forecast accuracy, especially at higher aggregation levels, longer forecast horizons and in situations with model misspecification.

The paper focuses on an important area (i.e., forecast reconciliation), is theoretically rigorous and provides a good summary of the relevant research. Incorporating time series selection to strengthen forecast reconciliation has both theoretical and practical significance. While the authors are to be commanded on their theoretical approach, the paper lacks any substantive discussion on the practical significance of the proposed methods. What are the precise implications of their proposed methods for analysts working in labour economics and tourism, for example (since their application data is from these two fields). Other users of such forecasts? What could be some potential implications for decision making? Policy making? These would be vital to enhancing the paper's reach and impact.

Done

- P.2 The authors claim "Through simulation experiments and two empirical applications, we demonstrate that our proposed methods guarantee coherent forecasts that outperform or match their respective benchmark methods" - how is such a guarantee provided? May wish to reconsider the wording.

  Done

- p.2-3 "A remarkable feature of the proposed methods is their ability to diminish disparities arising from using different estimates of the base forecast error covariance matrix, thereby mitigating challenges

associated with estimator selection, which is a prominent concern in the field of forecast reconciliation research." What about other concerns in forecast reconciliation research? It would be good to summarize these and address areas where the proposed methodology could support such concerns.

Done

- Conclusion section needs to be extended to discuss the practical repercussions of this work as well as the potential limitations it faces. The authors briefly touch upon one aspect in the final paragraph, but there would be other challenges and such an extended discussion would be critical in both highlighting the limitations as well as emphasizing the contributions of the current paper.

  Done

Reviewer 4

Summary

This manuscript aims to eliminate the negative effects of initially poor-performing base forecasts through time series selection. In the first group of methods, based on out-of-sample information, the authors formulate the problem as an optimization problem using diverse penalty functions. The second group of methods relax the unbiasedness assumption and introduces an additional reconciliation method with selection, utilizing in-sample observations and their fitted values. Both simulation and empirical studies show the great potential of the proposed methods.

Overall, this paper provides deep insights into forecast reconciliation methods and fits well with EJOR. However, a fair bit of work is required to get published in EJOR. Specifically, the following major and minor points could be considered to improve its exposition.

Major comments

1. While the proposed methodology opts to leave *poor* base forecasts unused in the creation of reconciled forecasts, the approach by Zhang et al. (2023) is primarily focused on preserving *good* base forecasts unchanged during the reconciliation process. The key difference lies in the handling of forecasts: the former method alters the forecasts of *poor* base forecasts, ensuring these do not influence other nodes, whereas the latter method keeps the forecasts of certain nodes immutable, which then impacts others. It is crucial for the authors to emphasize these distinctions and interconnections theoretically, empirically or through discussions.

   (Done) Highlight the differences and connections between the proposed methodology and the approach by Zhang et al. (2023). Discussion.

2. The authors' discussions on variable selection raise the question of whether they have considered a bi-level variable selection approach. Specifically, this entails allowing for both grouplevel and individual variable selection within those groups, an approach that could potentially enhance the precision and interpretability of the forecasting model. Such methodologies are well-documented in the literature, including the sparse group lasso (Simon et al., 2013), hierarchical Lasso (Zhou & Zhu, 2010), and the group bridge approach by Huang et al. (2009). Adopting a bi-level selection mechanism could provide deeper insights into the contribution of individual base forecasts, especially in terms of their significance when mapped to bottom-level disaggregated forecasts.

   Thank you for these helpful comments. In the context of forecast reconciliation, bi-level variable selection can be approached from two perspectives.

   First, following the idea of our methodology, we treat each time series in a given hierarchy as a variable. To achieve time series selection during reconciliation, we formulate an optimization problem that controls the number of nonzero column entries in the weighting matrix $\boldsymbol{G}$, which can be considered individual variable selection. For group-wise variable selection, we first have

to define groups of time series within the hierarchy. The task is challenging as both the grouping and the number of series within each group are unknown and can be determined subjectively. Thus, we choose not to consider bi-level variable selection from this perspective in our paper.

Second, by delving deeper into the optimization problem, we can treat each column of $\boldsymbol{G}$ as a group and each element as an individual. In this perspective, group-wise sparsity can be introduced by shrinking some columns of $\boldsymbol{G}$ towards zero, and within-group sparsity by shrinking some elements towards zero. This can be achieved by simply including an additional lasso penalty, as suggested in Simon et al. (2013), to address within-group sparsity. This provides insights into the contribution of individual base forecasts to the bottom-level reconciled forecasts. However, this would shift the focus away from our primary objective of time series selection, introducing additional hyperparameters and increasing computational complexity. Therefore, we decide to discuss this idea as a potential future research direction in Section 6.

3. Implementation of the GitHub repos provided by the authors tells that computation is somehow an issue in practice. Based on this, I suggest the authors consider the following.

   - Report the computational time would provide useful guides for the readers as well as the practitioners. Give discussions in terms of complexity and scalability, especially in the context of large-scale forecasting. This is crucial for applications in real-world scenarios.

     (Done) Report the computational time, discuss complexity and scalability.

   - Ida, Fujiwara & Kashima (2019) proposes a fast Block Coordinate Descent for Sparse Group Lasso, which efficiently skips the updates of the groups whose parameters must be zeros by using the parameters in one group. They claim their approach reduces the processing time by up to 97% from the standard approach. I suggest the authors check whether it is helpful in improving computation.

     (Done) Not helpful for Elasso.

4. The authors claim the proposed methods keep poor base forecasts unused in generating reconciled forecasts. They should check whether this is the case at the end of their experimental studies. Specifically, an analysis should be conducted to ascertain whether the forecasts that are not selected for reconciliation indeed correspond to suboptimal ones. This could involve a detailed examination of the performance metrics of excluded forecasts compared to those included, to ensure that the selection process aligns with the stated objective of excluding poor base forecasts. Such an investigation validates the method's effectiveness and strengthens the reliability of the proposed approach in practical forecasting scenarios.

   (Done) Check whether the series with poor forecasts are not selected, present results of each series. Consider scale-independent measures?

5. Given the variability in time series data—ranging from seasonal patterns, trend components, to noise levels—the authors could investigate how their methods perform across a diverse set of conditions. Examples are:
   - Examining the performance stability of the proposed reconciliation methods across time series with different levels of seasonality.
   - Assessing the impact of signal-to-noise ratios on the efficacy of the proposed methods.

   We agree on the importance of investigating the performance stability of the developed methods across diverse conditions. We believe that our simulations and applications encompass a wide range of data variations. First, within a given hierarchy, we account for variations in seasonality, trend, and noise levels across different series. Such variations reflect the inherent nature of hierarchical time series. For example, in the Australian labor force data, the total number of unemployed persons (the most aggregated series) shows much stronger seasonality and higher

signal-to-noise ratios compared to the unemployed persons data for individual states and territories. Second, we address variations in seasonality, trend, and noise levels across different hierarchies in the paper. We generate stationary data in Section 4.2, while examining quarterly data in Section 4.1 and analyzing monthly data in two applications in Section 5. Additionally, in Section 4.2, we explore the impact of correlation on the performance of the proposed reconciliation methods by controlling the error correlation in the simulated hierarchy. While the simulated data in Section 4.2 shows no trend, the real-world datasets in Section 5 display noticeable trends. Given the impracticality of considering all possible conditions and the constraints on the paper's length, we decide to retain the experiments as originally presented.

6. The current methodology section and experiments provide a foundational overview of the proposed forecast reconciliation approaches and how they are implemented. However, it would benefit significantly from a more detailed exposition on several fronts to enhance the reader's understanding. Specific areas include:

    - More detailed justification for the choice of penalty functions and the theoretical underpinnings that motivated these choices.

      (Done) Provide justification for the choice of penalty functions.

    - The inclusion of a sensitivity analysis regarding the hyperparameters associated with each method, such as penalty parameters in the optimization problem. Understanding how variations in these parameters affect the outcomes could provide valuable insights into the robustness and flexibility of the proposed approaches.

      (Done) Include a sensitivity analysis regarding penalty parameters.

Minor comments

1. I suggest the authors put *Variable selection* as one of the keywords.

   We have added "Variable selection" and removed "Grouped time series" from the keywords.

2. When cross-referencing equations, I suggest using \eqref from **amsmath** instead of \ref in the LaTeX Kernel to match the equation reference exactly.

   Thanks. Now fixed.

References

Huang, J, S Ma, H Xie & CH Zhang (2009). A group bridge approach for variable selection. *Biometrika* **96**(2), 339–355.

Ida, Y, Y Fujiwara & H Kashima (2019). Fast sparse group lasso. *Advances in Neural Information Processing Systems* **32**.

Simon, N, J Friedman, T Hastie & R Tibshirani (2013). A sparse-group lasso. *J Computational and Graphical Statistics* **22**(2), 231–245.

Zhang, B, Y Kang, A Panagiotelis & F Li (2023). Optimal reconciliation with immutable forecasts. *European J Operational Research* **308**(2), 650–660.

Zhou, N & J Zhu (2010). Group variable selection via a hierarchical lasso and its oracle property. *arXiv preprint arXiv:1006.2871*.