

European Journal of Operational Research

Probabilistic Forecast Reconciliation: Properties, Evaluation and Score Optimisation --Manuscript Draft--

Manuscript Number:	EJOR-D-21-02522R1
Article Type:	Theory and Methodology Paper
Section/Category:	(T) Forecasting
Keywords:	Forecasting; Scoring Rules; Hierarchical time series; Stochastic Gradient Descent
Corresponding Author:	George Athanasopoulos Monash University Caulfield East, Australia AUSTRALIA
First Author:	Anastasios Panagiotelis
Order of Authors:	Anastasios Panagiotelis Puwasala Gamakumara George Athanasopoulos Rob J Hyndman
Abstract:	<p>We develop a framework for forecasting multivariate data that follow known linear constraints. This is particularly common in forecasting where some variables are aggregates of others, commonly referred to as hierarchical time series, but also arises in other prediction settings. For point forecasting, an increasingly popular technique is reconciliation, whereby forecasts are made for all series (so-called base forecasts) and subsequently adjusted to cohere with the constraints. We extend reconciliation from point forecasting to probabilistic forecasting. A novel definition of reconciliation is developed and used to construct densities and draw samples from a reconciled probabilistic forecast. In the elliptical case, we prove that true predictive distributions can be recovered using reconciliation even when the location and scale of base predictions are chosen arbitrarily. Reconciliation weights are estimated to optimise energy or variogram score. The log score is not considered since it is improper when comparing unreconciled to reconciled forecasts, a result also proved in this paper. Due to randomness in the objective function, optimisation uses stochastic gradient descent. This method improves upon base forecasts in simulated and empirical data, particularly when the base forecasting models are severely misspecified. For milder misspecification, extending popular reconciliation methods for point forecasting results in similar performance to score optimisation.</p>
Suggested Reviewers:	<p>Feng Li feng.li@cufe.edu.cn Expert and thorough reviewer</p> <p>Nikolaos Kourentzes nikolaos@kourentzes.com Expert</p> <p>Fotios Petropoulos fotios@bath.edu Expert</p> <p>Tim Januschowski tjnsch@amazon.de Expert coming from a practitioner perspective.</p>
Opposed Reviewers:	

From: Professor
George Athanasopoulos
Econometrics and Business Statistics
Monash University
Caulfield, VIC, 3145, Australia

To: Professor Rudd H. Teunter
Editor, European Journal of Operational Research

May 18, 2022

Dear Ruud,

Please find attached our revised manuscript, EJOR-D-21-02522, entitled “Probabilistic Forecast Reconciliation: Properties, Evaluation and Score Optimisation” by Anastasios Panagiotelis, Puwasala Gamakumara, George Athanasopoulos, and Rob Hyndman for your consideration for publication in the *European Journal of Operational Research*. We also enclose a point by point response addressing the comments of the referees.

For the reviewers convenience, our responses to referees are provided in blue text and all new parts of the paper (in both the manuscript and response letter) are provided in red text.

Please note that in response to Reviewer 2, in terms of providing EJOR readers a non-technical pathway and intuition, and in order to not dilute the mathematical rigour of the paper important from a technical perspective, we have added a new section, Section 2, entitled “Outline of Main Results”. This section provides the reader less interested in technical details, with sufficient information in order to skip the technical sections that follow.

We look forward to hearing from you.

With kind regards and best wishes,

George Athanasopoulos

Editor

It has been decided that your paper should be reconsidered for publication in the EUROPEAN JOURNAL OF OPERATIONAL RESEARCH after a major revision taking into account the enclosed referees' comments. (You may wish to argue that some comments are invalid). Please also ensure that you have cited recent and relevant publications in EJOR and other OR journals.

We thank all reviewers for their invaluable/positive comments. All responses are in Blue font, any additions/changes to the paper are marked in Red. To the best of our knowledge, we have now added all recent and relevant publications in EJOR and other OR journals. This is particularly aided by the requests of the Reviewers.

Reviewer #1:

The manuscript provides a novel approach to probabilistic reconciliation using score optimization. It is concise, clearly structured and has a strong motivation, pointing out the necessary gaps in the literature. Following a detailed derivation of the SGD algorithm, the reconciled projections are evaluated using simulated and empirical data against a number of benchmark models. As the paper is very comprehensive, provides code as well as documentation, and appears quite mature, I have only a few minor comments.

- In the simulations, what are the assumptions on the covariance matrix in the joint Gaussian base forecasts?

In Section 6.2 we previously stated that we use “*the variance covariance matrix of the residuals*”. We now modify the statement to make sure that this is clear to now read: “*the variance covariance matrix of the residuals of the fitted models*”. In principle different choices could be used depending on assumptions made about how the covariance process should be modelled (e.g. time varying correlation), however such an issue lies beyond the scope of the simulation.

Why is it that the evaluated methods produce better scores for jointly distributed prediction errors?

In general, all methods score better for jointly distributed errors. This is due to the fact that assuming independence is likely to represent a (more severe) misspecification. The difference is substantial in the empirical application, especially with assuming Gaussian independent errors, as shown in Figure 8. We highlight this with the correlation heatmap we present in Figure 7. We also show some departures from normality in Figure 6. We comment on this stating “*Therefore, independent Gaussian probabilistic forecasts are likely to represent severe misspecification*” just before Figure 6.

In the simulations, all DGPs assume a joint error process and hence we expect that there is less misspecification error when probabilistic forecasts are generated using a joint distribution rather than assuming independence. As these differences are less pronounced we opt not to comment on these here and distract the reader from the main feature which is the performance evaluation across the reconciliation approaches.

- How well does the model scale to large, complex (grouped) hierarchies? How large is the computational demand compared to benchmark methods?

Relative to OLS and MinT, Score Optimisation is slower taking 2-3 minutes, mainly since it requires finding base forecasts over a rolling window (see line 2 of Algorithm 1). We note that implementing methods such as OLS and MinT for probabilistic rather than point forecasting using Theorem 3.5, is in itself novel and a significant contribution of the paper.

In an operational setting, where forecasts are made every period and kept, then these forecasts will already be available. The computational cost can also be mitigated by exploiting parallelisation (in our case we parallelised over different simulation/empirical scenarios rather than in the SGD itself). The stochastic gradient descent itself converges very quickly in most cases in less than 20 iterations.

- The authors might expand the discussion of the findings, in particular in section 7.2. How come the OLS approach performs so well given that it has been shown to lead to mediocre results in point forecasts?

A crucial point to make here is that OLS/MinT applied in the point forecasting setting are different and will thus have different properties to OLS/MinT applied in the probabilistic forecasting setting. In particular, for point forecasting, MinT will minimise the expected total mean squared error, although it should be noted that OLS also has some desirable properties for a loss function based on the L2 norm (see Panagiotelis et al 2021 for details). In the probabilistic setting we use scoring rules to evaluate forecast quality and these previous results do not necessarily apply. We now add the following statement to the paper (see Page 26, Lines 43-50):

“We suggest two possible reasons for the good performance of OLS in the probabilistic case. First, the energy score depends on the L2 norm of the difference between realizations and draws from the probabilistic forecast, which is similar to the setting for which OLS has optimal properties for point forecasts (see Panagiotelis et al, 2021). Second, for OLS there is less estimation uncertainty as fewer parameters need to be estimated.”

- In the electricity example, how is the joint distribution of the base forecasts estimated if they are generated using independent neural networks?

For both generating joint Gaussian and joint bootstrap probabilistic forecasts we use the residual covariance matrix of the fitted neural networks (fitted using the NNETAR function in the fable package, references are provided in the paper we do not repeat them here). This is identical to how the joint base forecasts are generated from the models in the simulations in Section 6. We have now added this information on Page 24, lines 40-42, by stating:

“Four situations were considered where base forecasts are assumed to be either Gaussian or bootstrapped from residuals, and either independent or dependent (we use the residual covariance matrix of the fitted neural networks, in a similar fashion as in Section 7).”

- A matter of preference, but the figures might look better when using a lighter theme than the out-of-the-box ggplot theme.

We acknowledge that this is a matter of taste. Having tried a few lighter themes our preference is to use the default ggplot theme.

Reviewer #2:

We thank the reviewer for the positive and constructive feedback on our paper. We respond below in detail to the comments that require a response.

Summary and General points

1. The paper is relevant to EJOR and continues a lineage of hierarchical forecasting papers recently published in the journal

No response required. Thank you for your positive view/comment.

2. The objective of the paper is to extend hierarchical forecasting methodology to the probabilistic setting.

No response required.

3. The paper is a difficult read due to the level of mathematical knowledge assumed. Whilst it is undoubtedly mathematically sophisticated, a good deal more effort could be made to explain the reasoning behind the linear algebra and borel sets. This sets a high bar for the empirical results - which in terms of the score optimisation algorithm in particular provide nothing by way of a reward for struggling through to the end. I'm sure the authors mathematical credentials are impeccable, but this is of little benefit to the journal if the reader is strongly inclined to give up reading on page 2.

We have now added a new section, after the introduction but before the main discussion of theoretical results (See Section 2 Outline of Main Results). The purpose of this section is to give a more intuitive and non-technical description of the main theoretical results. We

believe this new section provides the reader less interested in technical details sufficient information to skip the relevant sections.

Something that we now also emphasise in this new section (and in the introduction) is that the OLS and MinT methods are used in a novel way in this paper that relies on the way the theoretical results extend point reconciliation to probabilistic reconciliation. As a consequence, the assertion that the score optimisation algorithm offers little reward on the basis that it is outperformed by OLS and MinT in some settings, does not fairly represent the contributions of the paper. “OLS” and “MinT” as used here are not the existing point forecasting techniques, but probabilistic extensions of these techniques. These are introduced in this paper and are made possible by our definition of probabilistic coherence.

4. The paper makes significant and novel contributions in two main areas:
 - a. Firstly, in formally establishing how scoring rules can and should be applied in multivariate hierarchical contexts.

No response required.

- b. Secondly the paper posits a valuable and far-reaching result that sampling from underlying base forecast distributions, and reconciling the samples delivers valid probabilistic forecasts.

No response required.

5. These two contributions are important and worthwhile, and the paper should focus on explaining and exploring them in more detail, and making them accessible to the forecasting and OR community.

Both of these results are highlighted in the new section summarising the key theoretical results of the paper.

6. Additionally, the authors present a HF algorithm which seeks to produce probabilistic forecasts by optimising a scoring rule. This part of the paper is much less successful:
 - a. The empirical results are weak, with the somewhat complex methodology presented only outperforming simpler methods when the base forecast models are clearly and obviously very badly miss-specified. The reader is left somewhat short-changed in that the level of effort required to understand and deploy the algorithm is in no way justified by improved empirical results compared to simply sampling from the base forecast distributions.

We respectfully disagree with the referee here. Please allow us to make a few points which we now include in the paper in order to highlight the contributions of the paper.

1. The novel contribution of the paper is not only using score optimisation in forecast reconciliation. What is labelled OLS and MinT Shrink in the empirical application (and of course in the simulation setting) is also novel and goes beyond the point forecast setting (which has been previously explored by us or other authors). (Please see added text in the Introduction Page 4, lines 43-50)
 2. From our analysis we find that there are two possible sources of misspecification: (i) assuming independence and (ii) assuming Gaussian errors. Such misspecifications are not unusual in practice. For example, organisational sections or silos generating their forecasts independently, is common practice. We argue that one of the novel contributions of the hierarchical literature, including this paper, is to overcome such situations using forecast reconciliation. (Please see added text in the Introduction Page 3, lines 19-36).
 3. Despite an increasing recognition of the importance of probabilistic forecasts, it is not always the case in practice that organisations produce probabilistic forecasts. On the other hand it is more common for organisations to at least provide a predicted mean and a predicted variance. Having only a mean and variance available is also common with judgementally adjusted forecasts. Where only a mean and variance are available the logical parametric assumption is a normal distribution and bootstrapping (due to the lack of residuals) is not possible. Therefore the assumption of normality, while a misspecification for our empirical example, is not unrealistic in practice. We have now added discussion regarding the importance of the Gaussian setting in Section 4.2, Page 13, lines 8-15.
 4. Our results clearly show that score optimisation in forecast reconciliation is worth considering in such commonly observed misspecification settings. We believe this is a very strong, and also useful, result in practice.
- b. The empirical approach has a major conceptual weakness in that it fails to account for parameter uncertainty in the reconciliation process. HF processes are not parsimonious - most reconciliation models require the estimation of a vast dimensional covariance matrix. It is well known that such estimation exercises are fraught with difficulty, for example in the similar context of VAR estimation, it is well established that accounting for parameter uncertainty is critical, and the vast majority of published research in the area adopts a Bayesian approach. The fact that the author's own (full covariance) MinT approach only works well when shrinkage based estimation is adopted supports this viewpoint. The authors make this point in their conclusion, but it is of critical importance, and in my view undermines their empirical results here.

We agree with the referee to some extent here, but believe that much of what is being proposed goes beyond the scope of our paper. We agree with the importance of shrinkage whether it be used for estimating the parameters of a large variance covariance matrix (as required for MinT) or in regularising the reconciliation coefficients in the score optimisation algorithm. Regarding the latter, we certainly believe that this is a worthwhile avenue for research, and for this reason discuss it in the conclusion. Given that this paper has a

number of theoretical results, uses these results to extend OLS and MinT into the probabilistic setting (in a novel way) and proposes the score optimisation algorithm, we feel that considering shrinkage in the score optimisation setting is best handled in another paper. We note that we comment on this as “*A promising future research avenue*” in the conclusion of the paper (see Page 27, lines 58-61 and Page 28 lines 1-2).

- c. In my view the authors should focus on applying their own simple and robust idea of sampling from probabilistic forecasts and reconciling. The additional complexity of score optimisation without accounting for parameter uncertainty is certainly not justified by the results presented here.

We would contend that we do apply “*a simple and robust idea of sampling from probabilistic forecasts and reconciling*”. This method is used to produce joint probabilistic base forecasts either by assuming Gaussianity or via joint bootstrapping and then reconciling via popular point forecasting methods such as OLS and MinT. As argued elsewhere, these are novel extensions of existing point forecasting methods motivated by the definitions of probabilistic reconciliation proposed in this paper.

The novel score optimisation algorithm may not work in all cases. However, we have shown an important practical situation (as we argue above in our response to 6a) where it does work and works significantly better than the other reconciliation approaches. This is a novel contribution that one may choose to implement in a different setting and may possibly get better results. As discussed in our response to 6b we agree that shrinkage/regularisation is a potentially fruitful area of future research.

7. For publication in a Journal where real world application and decision making is important, the authors should make the effort to provide much more in the way of explanation and general reasoning behind their results.

We believe that this has been addressed in our response to the other points raised by the referee. In particular the new Section 2 as well as further discussion in the section on the empirical application provide further explanation.

Detailed Comments

1. Abstract... 'This method improves on base forecasts...' this is hardly a contribution... I think the same authors prove elsewhere that reconciliation always improves on base forecasts...

The results referred to here from our other work apply to the point forecast setting. The proposed manuscript explores the probabilistic setting, therefore the results here, even for MinT and OLS are all new. Even in the point forecasting setting, the statement made here by the referee is only partially true. In Panagiotelis et al. (2021) we do present general

results but for two objectives: (i) to guarantee that reconciled forecasts improve upon base forecasts and (ii) to find the reconciliation method that is best on average. To guarantee objective (i) the loss function has to be matched with the weights matrix used in the projection reconciliation approach.

Panagiotelis, A., G. Athanasopoulos, P. Gamakumara, and R. J. Hyndman (2021). Forecast reconciliation: A geometric view with new insights on bias correction. *International Journal of Forecasting* 37(1), 343–359.

2. Section 2.2 I think this definition will be completely meaningless to many readers. Surely there should be a way of setting this more clearly without recourse to borel sets etc? Of course the definition is important (and therefore merits more explanation) but In a journal focused on applications this would be better off in an appendix?

As well as the discussion added in the new Section 2, we have added further explanation to this definition (see Page 9 lines 5-15, following Definition 3.1). In particular we talk about assigning probabilities to “intervals”, “rectangles” or “regions”, since these are all cases of Borel sets when giving intuitive explanations. We continue to use Borel sets in the definitions only to ensure full rigour.

3. Section 2.3 - Again much more effort could be made to explain what is going on here

We now summarise this result more succinctly in the new Section 2. We believe that the additional discussion around Section 3.2 (what was previously 2.2) as well as the use of Figure 2, makes the rest of this section clear.

4. Theorem 3.5 - I think this is a really important result (and one of the key contributions of the paper). A proof is of course provided, and while the theorem makes intuitive sense, no attempt is made to clarify and explain the logic of what is going on, and as the proof builds on the earlier results, and many readers will have to take it on trust...

We have added additional explanation here (see the new text following Theorem 4.5 - previously Theorem 3.5). This proof also leans heavily on the concepts from Definitions 2.1 and 2.2 (now 3.1 and 3.2) which we believe are now more clearly explained.

5. Page 16, line 37. 'bottom level series have lower signal to noise than higher level'? It is a little surprising that this does not occur automatically, in practice the noise level in the higher level series would normally reduce via a diversification effect?

Yes we agree. The bottom-level series did automatically have a lower signal-to-noise ratio. The additional noise was added to the bottom-level to make this difference more pronounced and replicate a realistic/practical setting. We have revised the statement now to read:

“After simulating from the ARIMA models, additional noise is added to ensure bottom-level series have a considerably lower signal-to-noise ratio than upper-level series with details provided in Appendix D of the online supplement.”

Note we have also corrected a typo in the paper changing the second instance of the word “bottom” to “upper”.

6. Section 7.2 - As noted by the authors, the assumptions of independence and gaussian errors are clearly both badly violated in the data. The score optimisation algorithm generates 'statistically significant' improvements to base forecasts generated using naïve and inaccurate base forecasts! Is it really worth going to the trouble of reconciling such poor forecasts? When more sensible base forecast assumptions are made, simpler and more robust methods based on Theorem 3.5 perform more strongly.

As we argue in our response to comment 6a above, and we argue again here, it is absolutely worth reconciling such forecasts, or at least it is worth having the knowledge and a tool, such as score optimisation, that can successfully reconcile these:

1. We never know the quality of our forecasts a priori. Any evaluation of these will always be in sample and many times such evaluation is not possible, e.g., when these are judgmentally generated or in general not model based or the model is not made available.
2. In commonly observed organisational silos, independence is the only assumption possible.
3. A parametric assumption of Gaussianity for the errors is the usual alternative when only the mean and variance are available and/or when bootstrapping is not a possibility.

Reviewer #3:

This paper is concerned with the highly relevant problem of forecast reconciliation in a probabilistic setting. The authors (i) present a theoretical framework and theoretical results for this problem, (ii) propose an algorithm for probabilistic forecast reconciliation based on score optimization, and (iii) present results on both simulated and real data. In general, the paper is well written and of high quality. Considering all these elements, in my opinion, this paper is suitable for publication in EJOR after a minor revision.

We thank the reviewer for the positive and constructive feedback on our paper. We respond below in detail to the comments that require a response.

Please find below my comments.

General comments:

I very much like that the paper is concise and to the point. This is true for both the theoretical and empirical parts. The empirical sections are very good: nice presentation + insightful discussion for both the simulation and case study results. However, the introduction and the theoretical sections (2 to 5) are hard to follow at times, which I think is a consequence of these sections being somewhat too concise. Extra context is needed in some cases to make the text more comprehensible for a more general audience.

We have now added a new section (Section 2) giving a non-technical explanation of the main theoretical results. We have provided more context to some of the theorems at the request of Reviewer 2. We believe the changes have made the more challenging sections of the paper much more comprehensible to a general audience.

Some examples:

- P2L53: "Prior to the development of forecast reconciliation, the focus was on finding a subset of variables that could be subsequently aggregated or disaggregated to find forecasts for all series." Please clarify.

We have now revised the statement as follows:

"Prior to the development of forecast reconciliation, the focus was on forecasting a subset of variables at some selected level of aggregation, and subsequently aggregating or disaggregating these to generate forecasts for all series."

- P3L1: "These papers formulated reconciliation as a regression model, however subsequent work has formulated reconciliation as an optimisation problem where weights are chosen to minimise a loss." Please clarify.

Let us please note that this statement does not end where the reviewer indicates it ends. We reproduce below the full statement:

"These papers formulated reconciliation as a regression model, however subsequent work has formulated reconciliation as an optimisation problem where weights are chosen to minimise a loss, such as a weighted squared error (Van Erven and Cugliari, 2015; Nystrup et al., 2020), a penalised version thereof (Ben Taieb and Koo, 2019), or the trace of the forecast error covariance (Wickramasuriya et al., 2019)."

In order to further clarify we break the statement into two sentences and also add to it. The revised text now reads as follows:

"These papers formulated reconciliation as a regression model, reconciling the base forecasts by projecting them onto a subspace for which aggregation constraints hold. Subsequent work has formulated reconciliation as an optimisation problem where weights are chosen to minimise a loss, such as a weighted squared error (Van Erven and Cugliari,

2015; Nystrup et al., 2020), a penalised version thereof (Ben Taieb and Koo, 2019), or the trace of the forecast error covariance (Wickramasuriya et al., 2019)."

- P3L9: "The accuracy and popularity of forecast reconciliation methods can be attributed to a number of factors". However, only one factor is discussed, i.e., breaking down organizational silos. Please also discuss/mention other factors.

We have now revised the paragraph to read as follows:

"The popularity of forecast reconciliation methods can be attributed to a number of factors. Forecasts across different aggregation levels may be generated by different departments or 'silos' within an organisation, using different sets of predictors, modelling approaches, or expert judgement. Potentially, these are viewed as optimal within these divisions. Reconciliation represents a way to combine information via the sharing of forecasts, thus breaking down these silos. Although it may be difficult to share forecasting processes and associated information across different parts of a large organisation, the forecasts themselves are much easier to share and reconcile. In contrast to bottom-up and top-down approaches, which effectively discard the forecasts of all but one level, the combination of forecasts across all levels also leads to improved forecast accuracy."

- Ben Taieb et al. (2020) explanation: Either discuss it in more detail or in less detail, but now it is rather unclear what this method is exactly about. What is meant by 'reordering' base forecasts? To some extent this comment is also true for the discussion of Jeon et al. (2019): "ranking draws from independent base probabilistic forecasts before reconciliation is effective"?

To elaborate on "reordering", both JPP and BTTH obtain draws from the probabilistic forecast from each variable independently (let's think of these as vectors of length L). To reconcile these methods we take a single draw from each variable and combine these in an n -vector. In this case, the ordering of the draws within each vector of length L becomes critical and both BTTH and JPP attempt to do this in a way that captures dependence.

JPP do this in a way that is equivalent to reconciling quantiles (which is how we have described on Page 3 lines 56-57, Page 4 lines 21-23, and Page 19 lines 45-55). We have rewritten the sentence "*ranking draws...is effective*" since the connection between this and reconciling quantiles may not be clear. Regarding BTTH, we have expanded on our discussion of their method adding the following on Page 4 lines 1-6

"In particular, Ben Taieb et al.(2020) draw a sample of size L from the probabilistic forecasts of univariate models for the m bottom-level series and stack these in an $L \times m$ matrix. To induce dependence, the columns of this matrix are reordered so that the copula of the data matrix created, matches the empirical copula of the residuals. Samples of the aggregate series are obtained in a bottom-up fashion."

- P14L44: "These are typically high dimensional problems, deep neural networks handle millions of parameters, so this tool is well suited to our problem." What do you mean exactly by this and 'these'? Gamma is not necessarily high-dimensional, right?

The above sentence follows on from *"There is also a recent but growing literature on using SGD to optimise scoring rules (see Gasthaus et al., 2019; Janke and Steinke, 2020; Hofert et al., 2020, and references therein for examples)."* which highlights the recent literature using SGD to optimise scoring rules. *"These...problems"* refers to the problems tackled within these papers.

We have now rewritten the sentence to make it clear. The sentence now reads (see Page 16 lines 35-36):

"These papers typically deal with high dimensional problems, deep neural networks handle millions of parameters, so this tool is well suited to our problem."

Also although Gamma does not necessarily contain millions of parameters (although in some applications it may do so), we simply make this point to assuage any concerns a reader may have about the applicability of our methods to larger hierarchies.

- P14L61: Discuss in more detail how the proposed score optimization algorithm differs from the method proposed by Rangapuram et al. (2021). At first glance, an obvious one is that your approach allows using general techniques (including ets() for example) to generate probabilistic base forecasts, which is not the case for Rangapuram et al. (2021). However, how is your method different from the one presented in Rangapuram et al. (2021) as for the reconciliation procedure (inclusion of translation, orthogonality of projection(?)...), and what are the implications? Just to be clear: I do not expect you to add this method to the empirical results in Sections 6 and 7.

We have now extended the discussion of Rangapuram et al. (2021) as follows (see Page 16 lines 60 and Page 61 lines 38-43):

"Rangapuram et al. (2021) use a similar approach in their end-to-end forecasting process. Their method is more restrictive than what we propose here in that the projection must be orthogonal, base forecasts are not translated, and base forecasts must be generated by a DeepVAR."

The conclusion could be improved a bit by adding some nuance to the discussion:

- "Since the scores are approximated by Monte Carlo simulation, stochastic gradient descent is used for optimisation." One could also use batch gradient descent using all Monte Carlo simulations in one go, isn't it?

There is some subtlety to this point. In many contexts, (e.g. training neural networks) what makes stochastic gradient descent, *"stochastic"*, is that the training data are subsampled for

computational reasons. In our setting, the Energy (and Variogram) Scores cannot be computed in closed form and must be estimated using a Monte Carlo estimate. This is what requires us to use SGD even when all data are used. We have now added the following (see Page 16, lines 39-46):

"An important distinction is that the use of SGD, rather than gradient descent in these contexts, arises due to computational considerations as it is not efficient to use all data. In contrast we use all data and the 'stochastic' nature of our gradient descent arises since the score functions contain integrals that must be estimated by Monte Carlo."

- "This method is shown to lead to significant improvements over base forecasts, bottom-up methods and existing probabilistic reconciliation approaches across a wide variety of simulated and empirical examples." Refer also here to the degree of misspecification of the base forecasts.

-

We have now revised the statement and have added the suggestion. The statement now reads as follows (see Page 27 line 49):

This method is shown to lead to significant improvements over base forecasts, bottom-up methods and existing probabilistic reconciliation approaches across a wide variety of simulated and empirical examples, particularly when the base forecasting models are severely misspecified.

Minor comments:

- Add references to the first sentence of the Intro.

Thank you for the suggestion. We have now added references to these.

- P10L60: Add dot at the end of footnote 1.

Done.

- P12L18: Replace 'and' by 'are'?

Done.

- P14L60: Is 'projection' the correct term here (also see P14L2)?

Yes 'projection' is the correct term here.

- P17L35: Replace period by colon.

Done.

- P18L10: Variogram score is only introduced by referring to Scheuerer and Hamill (2015), but no definition or explanation is provided.

We have now added the definition and appropriate additional references and discussion on the discrimination ability of scoring rules (see Page 15 lines 1-13).

- P24L56: Remove dot after footnote 3 in text + add dot at the end of footnote 3.

Done.

- How many errors do you use for the construction of bootstrapped probabilistic forecasts? I think that this info is missing.

We have added the following on Page 19 lines 22-25.

“The number of bootstrap samples is set equal to the sample size both here and in Section 8.”

- Can you provide an explanation for the deviant performance of BTTH for the Gaussian DGP in terms of variogram scores? The fact that a definition of the variogram score (and a discussion of how this score is different from the energy score) is missing makes it impossible for the reader to reason about this observation.

We are unable to explain the deviant performance of BTTH in this setting. However, we have now added the definition of the variogram score. We also provide references that discuss scenarios in which the variogram score has good discriminatory power (see Page 15 lines 1-13).

- Forecast reconciliation in the probabilistic setting is rigorously developed.
- Point forecast reconciliation is extended in a novel way to the probabilistic setting.
- Results are derived for the Gaussian and non-Gaussian case.
- Theorems on scoring rules are derived with recommendations for forecast evaluation.
- A new reconciliation method based on score optimisation and stochastic gradient descent is proposed.
- The new methods are shown to improve forecast accuracy in a simulated and empirical setting.

Probabilistic Forecast Reconciliation: Properties, Evaluation and Score Optimisation

Anastasios Panagiotelis

Discipline of Business Analytics,

University of Sydney, NSW 2006, Australia.

Email: anastasios.panagiotelis@sydney.edu.au

and

Puwasala Gamakumara

Department of Econometrics and Business Statistics,

Monash University, VIC 3800, Australia.

Email: puwasala.gamakumara@monash.edu

and

George Athanasopoulos*

Department of Econometrics and Business Statistics,

Monash University, VIC 3145, Australia.

Email: george.athanasopoulos@monash.edu

and

Rob J Hyndman

Department of Econometrics and Business Statistics,

Monash University, VIC 3800, Australia.

Email: rob.hyndman@monash.edu

May 15, 2022

*The authors gratefully acknowledge the support of Australian Research Council Grant DP140103220. We also thank Professor Mervyn Silvapulle for valuable comments.

Abstract

We develop a framework for forecasting multivariate data that follow known linear constraints. This is particularly common in forecasting where some variables are aggregates of others, commonly referred to as hierarchical time series, but also arises in other prediction settings. For point forecasting, an increasingly popular technique is reconciliation, whereby forecasts are made for all series (so-called base forecasts) and subsequently adjusted to cohere with the constraints. We extend reconciliation from point forecasting to probabilistic forecasting. A novel definition of reconciliation is developed and used to construct densities and draw samples from a reconciled probabilistic forecast. In the elliptical case, we prove that true predictive distributions can be recovered using reconciliation even when the location and scale of base predictions are chosen arbitrarily. Reconciliation weights are estimated to optimise energy or variogram score. The log score is not considered since it is improper when comparing unreconciled to reconciled forecasts, a result also proved in this paper. Due to randomness in the objective function, optimisation uses stochastic gradient descent. This method improves upon base forecasts in simulated and empirical data, particularly when the base forecasting models are severely misspecified. For milder misspecification, extending popular reconciliation methods for point forecasting results in similar performance to score optimisation.

Keywords: Forecasting, Scoring Rules, Hierarchical Time Series, Stochastic Gradient Descent.

1 Introduction

Forecasting hierarchical time series arise in many decision making settings, including demand forecasting for supply chain management (Babai et al., 2021; Kourentzes and Athanasopoulos, 2021), forecasting electricity generation for planning infrastructure investment (Ben Taieb et al., 2020; Nystrup et al., 2020), forecasting mortality rates (Li and Hyndman, 2021), as well as applications in macroeconomics (Eckert et al., 2021; Athanasopoulos et al., 2020) and tourism management (Athanasopoulos et al., 2022; Kourentzes and Athanasopoulos, 2019). In recent years forecast reconciliation has become an increasingly popular method for handling such problems (see Hyndman and Athanasopoulos, 2021, for an overview). Reconciliation involves producing predictions for all variables and making a subsequent adjustment to ensure these adhere to known linear constraints. Despite the importance of having probabilistic forecasts available in a decision making setting, reconciliation methodology has primarily been developed with point prediction in mind. This paper develops a formal framework for probabilistic reconciliation, derives theoretical results that allow reconciled probabilistic predictions to be constructed and evaluated, and proposes an algorithm for optimal probabilistic reconciliation with respect to a proper scoring rule.

Before describing the need for probabilistic reconciliation we briefly review the literature on point forecast reconciliation. Prior to the development of forecast reconciliation, the focus was on forecasting a subset of variables at some selected level of aggregation, and subsequently

aggregating or disaggregating these to generate forecasts for all series. (see Dunn et al., 1976; Gross and Sohl, 1990, and references therein).

An alternative approach emerged with Athanasopoulos et al. (2009) and Hyndman et al. (2011) who recommended producing forecasts of all series (referred to as ‘base’ forecasts) and then adjusting, or ‘reconciling’, these forecasts to be ‘coherent’, i.e. adhere to the aggregation constraints. These papers formulated reconciliation as a regression model, **reconciling the base forecasts by projecting them onto a subspace for which aggregation constraints hold.**, however Subsequent work has formulated reconciliation as an optimisation problem where weights are chosen to minimise a loss, such as a weighted squared error (Van Erven and Cugliari, 2015; Nystrup et al., 2020), a penalised version thereof (Ben Taieb and Koo, 2019), or the trace of the forecast error covariance (Wickramasuriya et al., 2019).

The popularity of forecast reconciliation methods can be attributed to a number of factors. Forecasts across different aggregation levels may be generated by different departments or ‘silos’ within an organisation, using different sets of predictors, modelling approaches, or expert judgement. Potentially, these are viewed as optimal within these divisions. Reconciliation represents a way to combine information via the sharing of forecasts, thus breaking down these silos. Although it may be difficult to share forecasting processes and associated information across different parts of a large organisation, the forecasts themselves are much easier to share and reconcile. In contrast to bottom-up and top-down approaches, which effectively discard the forecasts of all but one level, the combination of forecasts across all levels also leads to improved forecast accuracy.

In contrast to point forecasts, the entire probability distribution of future values provides a full description of the uncertainty associated with the predictions (Gneiting and Katzfuss, 2014). The importance of probabilistic forecasts can be seen in decision making settings in risk management, when it is critical to quantify the probability of extreme events. Therefore probabilistic forecasting has become of great interest in many disciplines such as, economics (Rossi, 2014), meteorological studies (McLean Slaughter et al., 2013), energy forecasting (Ben Taieb et al., 2017) and retail forecasting (Böse et al., 2017). An early attempt towards probabilistic forecast reconciliation came from Shang and Hyndman (2017) who applied reconciliation to forecast quantiles, rather than to the point forecasts, in order to construct prediction intervals. This idea was extended to constructing a full probabilistic forecast by Jeon et al. (2019) who propose a number of algorithms, one of which is equivalent to reconciling a large number of forecast quantiles. Ben Taieb et al. (2020) also propose an algorithm to

obtain probabilistic forecasts that cohere to linear constraints. In particular, Ben Taieb et al. (2020) draw a sample of size L from the probabilistic forecasts of univariate models for the m bottom-level series and stack these in an $L \times m$ matrix. To induce dependence, the columns of this matrix are reordered so that the copula of the data matrix created, matches the empirical copula of the residuals. Samples of the aggregate series are obtained in a bottom-up fashion. The only sense in which top-level forecasts are used is in the mean, which is adjusted to match that obtained using the MinT reconciliation method (Wickramasuriya et al., 2019).

There are a number of shortcomings to Jeon et al. (2019) and Ben Taieb et al. (2020). First, little formal justification is provided for the algorithms, or for the sense that they generalise forecast reconciliation to the probabilistic domain. As such, both algorithms are based on sampling and neither can be used to obtain a reconciled density analytically. Both algorithms are tailored towards specific applications and conflate reconciliation with steps that reorder the base forecasts. For example, while Jeon et al. (2019) show that reconciling the quantiles of independent base probabilistic forecasts is effective, this may only be true due to the highly dependent time series considered in their application. A limitation of Ben Taieb et al. (2020) is that to ensure their sample from the base probabilistic forecast has the same empirical copula as the data, it must be of the same size as the training data. This will be problematic in applications with fewer observations than the smart meter data they consider. Further, Ben Taieb et al. (2020) only incorporate information from the forecast mean of aggregate variables, missing out on potentially valuable information in the probabilistic forecasts of aggregate data.

In this paper we seek to address a number of open issues in probabilistic forecast reconciliation. First, we develop in a formal way, definitions and a framework that generalise reconciliation from the point setting to the probabilistic setting. This is achieved by extending the geometric framework proposed by Panagiotelis et al. (2021) for point forecast reconciliation. An important feature of this definition is that it allows existing reconciliation methods such as OLS and MinT to be extended to the probabilistic setting. While OLS and MinT are not new methods in for point forecast reconciliation, their extension to probabilistic reconciliation in a way built upon the new definitions is novel in this paper. Second, we utilise these definitions to show how a reconciled forecast can be constructed from an arbitrary base forecast. Solutions are provided in the case where a density of the base probabilistic forecast is available and in the case where it is only possible to draw a sample from the base forecasting distribution. Third, we show that in the elliptical case, the correct predictive distribution can be recovered via linear reconciliation irrespective of the location and scale parameters of the base forecasts.

We also derive conditions for when this also holds for the special case of reconciliation via projection. Fourth, we derive theoretical results on the evaluation of reconciled probabilistic forecasts using multivariate scoring rules, including showing that the log score is improper when used to compare reconciled to unreconciled forecasts. Fifth, we propose an algorithm for choosing reconciliation weights by optimising a scoring rule. This algorithm exploits advances in stochastic gradient descent and is thus suited to scoring rules which are often only known up to an approximation. The algorithm and other methodological contributions described in this paper are implemented in the ProbReco package (Panagiotelis, 2020).

The remainder of the paper is structured as follows. **Section 2 provides a non-technical summary of the main theoretical results of the paper. We recommend that a reader who is notless concerned with the more technical and in particular probability theoretic issues, can safely proceed to Section 6 after reading this section.** In Section 3, after a brief review of point forecast reconciliation, novel definitions are provided for coherent forecasts and reconciliation in the probabilistic setting. In Section 4, we outline how reconciliation can be achieved in both the case where the density of the base probabilistic forecast is available, and in the case where a sample has been generated from the base probabilistic forecast. In Section 5, we consider the evaluation of probabilistic hierarchical forecasts via scoring rules, including theoretical results on the impropriety of the log score in the context of forecast reconciliation. The use of scoring rules motivates our algorithm for finding optimal reconciliation weights using stochastic gradient descent, which is described in Section 6 and evaluated in an extensive simulation study in Section 7. An empirical application on forecasting electricity generation from different sources is contained in Section 8. Finally Section 9 concludes with some discussion and thoughts on future research.

2 Outline of Main Results

Many results from the paper require some background in probability theory that may distract from readers more concerned with the practicalities of implementing probabilistic forecast reconciliation. In this section we briefly discuss the main theoretical results in a non-technical manner.

- First, we define the concept of *probabilistic coherence* in **Definition 3.1**. Loosely speaking, this is defined as any forecast assigns zero probability to events that do not meet the coherence condition (e.g. in a hierarchical setting, forecasts that do not correctly add up).

- This is distinct from *probabilistic forecast reconciliation*, which we define in **Definition 3.2**. In the same way that point forecast reconciliation begins with an incoherent forecast, in the probabilistic setting we begin with an incoherent probabilistic forecast. In the point forecasting setting we can consider a (usually linear) function that takes an incoherent point and maps it to a coherent points. In the probabilistic setting we consider the same types of functions, but think about them mapping *sets* of incoherent points to *sets* of coherent points. The probabilities assigned to these two sets are the same, giving us a general definition for probabilistic forecast reconciliation. The key implication of this definition is that any existing point reconciliation method (e.g. OLS or MinT) can be extended to the probabilistic setting.
- Using these definitions we can derive two practical ways of carrying out forecast reconciliation. The first is a method involving integration, but which in the case of *elliptical distributions* (including the Gaussian distribution) provides an elegant solution involving linear transformations of scale and location parameters (**Theorems 4.1 and 4.2**). We further prove that in the elliptical case, the true predictive distribution can be obtained via such a linear reconciliation method (**Theorem 4.3**).
- The second practical method for conducting forecast reconciliation relies on **Theorem 4.5**. This theorem states that a distribution can be reconciled by *simulating* from the base (incoherent) forecast and then reconciling each sampled vector as if it were a point forecast. This motivates the Score Optimal Reconciliation method introduced in Section 6.
- Finally, **Theorem 5.1** is an important results concerning the evaluation of reconciled probabilistic forecasts using the *log score*. This theorem implies that incoherent forecasts can even outperform the true predictive distribution when the log score is used for evaluation. This makes the log score ill suited to comparing the performance of incoherent probabilistic forecasts with reconciled forecasts.

A reader less concerned with the technical details of these results, can safely skip the next three sections and progress to the details of the Score Optimal Reconciliation algorithm proposed in Section 6.

3 Hierarchical probabilistic forecasts

Before extending coherence and reconciliation to the probabilistic setting, we briefly refresh these concepts for point forecasts. We follow the geometric interpretation introduced by Panagiotelis et al. (2021), since this formulation naturally generalises to probabilistic forecasting.

3.1 Point Forecasting

A hierarchical time series is a collection of time series adhering to linear constraints. Stacking the value of each series at time t into an n -vector \mathbf{y}_t , the constraints imply that \mathbf{y}_t lies in an m -dimensional linear subspace of \mathbb{R}^n for all t . This subspace is referred to as the coherent subspace and is denoted as \mathfrak{s} . A typical (and the original) motivating example is a collection of time series some of which are aggregates of other series. In this case $\mathbf{b}_t \in \mathbb{R}^m$ can be defined as the values of the most disaggregated or bottom-level series at time t and the aggregation constraints can be formulated as $\mathbf{y}_t = \mathbf{S}\mathbf{b}_t$, where \mathbf{S} is an $n \times m$ constant matrix for a given hierarchical structure.

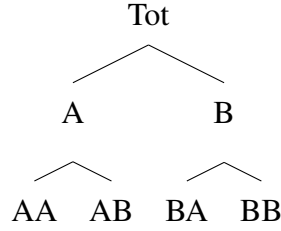


Figure 1: An example of a two-level hierarchical structure.

An example of a hierarchy is shown in Figure 1. There are $n = 7$ series of which $m = 4$ are bottom-level series. Also, $\mathbf{b}_t = [y_{AA,t}, y_{AB,t}, y_{BA,t}, y_{BB,t}]'$, $\mathbf{y}_t = [y_{Tot,t}, y_{A,t}, y_{B,t}, \mathbf{b}_t']'$, and

$$\mathbf{S} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ & & \mathbf{I}_4 \end{pmatrix},$$

where \mathbf{I}_4 is the 4×4 identity matrix.

The connection between this characterisation and the coherent subspace is that the columns of \mathbf{S} span \mathfrak{s} . Below, the notation $s : \mathbb{R}^m \rightarrow \mathbb{R}^n$ is used when premultiplication by \mathbf{S} is thought of as a mapping. Finally, while \mathbf{S} is defined in terms of m bottom-level series here, in general

any m series can be chosen with the S matrix redefined accordingly. The columns of all appropriately defined S matrices span the same coherent subspace \mathfrak{s} .

When forecasts of all n series are produced, they may not adhere to constraints. Such forecasts are called incoherent base forecasts and are denoted $\hat{\mathbf{y}}$. To exploit the fact that the target of the forecast adheres to known linear constraints, base forecasts can be adjusted in a process known as forecast reconciliation. This involves selecting a mapping $\psi : \mathbb{R}^n \rightarrow \mathfrak{s}$ and then setting $\tilde{\mathbf{y}} = \psi(\hat{\mathbf{y}})$, where $\tilde{\mathbf{y}} \in \mathfrak{s}$ is called the reconciled forecast. The mapping ψ may be considered as the composition of two mappings $\psi = s \circ g$. Here, $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ combines incoherent base forecasts of all series to produce new bottom-level forecasts, which are then aggregated via s . Many existing point forecasting approaches including the bottom-up (Dunn et al., 1976), OLS (Hyndman et al., 2011), WLS (Athanasopoulos et al., 2017) and MinT (Wickramasuriya et al., 2019) methods, are special cases where g is premultiplication by a matrix G and where SG is a projection matrix. These are summarised in Table 1.

Table 1: Summary of reconciliation methods for which SG is a projection matrix. Here W is some diagonal matrix, $\hat{\Sigma}_{sam}$ is a sample estimate of the residual covariance matrix and $\hat{\Sigma}_{shr}$ is a shrinkage estimator proposed by Schäfer and Strimmer (2005), given by $\tau \text{diag}(\hat{\Sigma}_{sam}) + (1 - \tau)\hat{\Sigma}_{sam}$ where $\tau = \frac{\sum_{i \neq j} \text{Var}(\hat{\sigma}_{ij})}{\sum_{i \neq j} \hat{\sigma}_{ij}^2}$ and σ_{ij} denotes the (i, j) th element of $\hat{\Sigma}_{sam}$.

Reconciliation method	G
OLS	$(S'S)^{-1}S'$
WLS	$(S'WS)^{-1}S'W$
MinT(Sample)	$(S'\hat{\Sigma}_{sam}^{-1}S)^{-1}S'\hat{\Sigma}_{sam}^{-1}$
MinT(Shrink)	$(S'\hat{\Sigma}_{shr}^{-1}S)^{-1}S'\hat{\Sigma}_{shr}^{-1}$

3.2 Coherent probabilistic forecasts

We now turn our attention towards a novel definition of coherence in a probabilistic setting. First let $(\mathbb{R}^m, \mathcal{F}_{\mathbb{R}^m}, \nu)$ be a probability triple, where $\mathcal{F}_{\mathbb{R}^m}$ is the usual Borel σ -algebra on \mathbb{R}^m . This triple can be thought of as a probabilistic forecast for the bottom-level series. A σ -algebra $\mathcal{F}_{\mathfrak{s}}$ can then be constructed as the collection of sets $s(\mathcal{B})$ for all $\mathcal{B} \in \mathcal{F}_{\mathbb{R}^m}$, where $s(\mathcal{B})$ denotes the image of \mathcal{B} under the mapping s .

Definition 3.1 (Coherent Probabilistic Forecasts). Given the triple, $(\mathbb{R}^m, \mathcal{F}_{\mathbb{R}^m}, \nu)$, a coherent probability triple $(\mathfrak{s}, \mathcal{F}_{\mathfrak{s}}, \check{\nu})$, is given by \mathfrak{s} , the σ -algebra $\mathcal{F}_{\mathfrak{s}}$ and a measure $\check{\nu}$, such that

$$\check{\nu}(s(\mathcal{B})) = \nu(\mathcal{B}) \quad \forall \mathcal{B} \in \mathcal{F}_{\mathbb{R}^m}.$$

To explain this definition simply, without recourse to Borel sets, consider a three variable hierarchy $A = B + C$ and let \mathcal{B} be a rectangle on the 2-dimensional space of B and C . The probability that an observation lies in this rectangle is $\nu(\mathcal{B})$. Then $s(\mathcal{B})$ will be some region in the coherent subspace \mathfrak{s} . The probability that a coherent forecast lies in this region is $\check{\nu}(s(\mathcal{B}))$. An alternative, but equivalent explanation is that coherent probabilistic forecasts assigns a probability of zero to any set of points that does not contain any coherent points.

To our best knowledge, the only other definition of coherent probabilistic forecasts is given by Ben Taieb et al. (2020) who define them in terms of convolutions. While these definitions do not contradict one another, our definition has two advantages. First it can more naturally be extended to problems with non-linear constraints with the coherent subspace \mathfrak{s} replaced with a manifold. Second, it facilitates a definition of probabilistic forecast reconciliation.

3.3 Probabilistic forecast reconciliation

Let $(\mathbb{R}^n, \mathcal{F}_{\mathbb{R}^n}, \hat{\nu})$ be a probability triple characterising a probabilistic forecast for all n series. The hat is used for $\hat{\nu}$ analogously with \hat{y} in the point forecasting case. The objective is to derive a reconciled measure $\tilde{\nu}$, assigning probability to each element of the σ -algebra $\mathcal{F}_{\mathfrak{s}}$.

Definition 3.2 (Reconciled Probabilistic Forecasts). The reconciled probability measure of $\hat{\nu}$ with respect to the mapping $\psi(\cdot)$ is a probability measure $\tilde{\nu}$ on \mathfrak{s} with σ -algebra $\mathcal{F}_{\mathfrak{s}}$ such that

$$\tilde{\nu}(\mathcal{A}) = \hat{\nu}(\psi^{-1}(\mathcal{A})) \quad \forall \mathcal{A} \in \mathcal{F}_{\mathfrak{s}},$$

where $\psi^{-1}(\mathcal{A}) := \{y \in \mathbb{R}^n : \psi(y) \in \mathcal{A}\}$ is the pre-image of \mathcal{A} , that is the set of all points in \mathbb{R}^n that $\psi(\cdot)$ maps to a point in \mathcal{A} .

This definition naturally extends forecast reconciliation to the probabilistic setting. In the point forecasting case, the reconciled forecast is obtained by transforming an incoherent forecast. For probabilistic forecasts, sets of points are transformed to sets of points, with the same probabilities assigned to these sets under the base and reconciled measures respectively. Also, since ψ can be expressed as a composition $s \circ g$, a reconciled probabilistic distribution ν can be obtained for m series such that $\nu(\mathcal{B}) = \hat{\nu}(g^{-1}(\mathcal{B}))$ for all $\mathcal{B} \in \mathcal{F}_{\mathbb{R}^m}$. A probabilistic forecast for the full hierarchy can then be obtained via Definition 3.1. This construction will be used in Section 4.

Definition 3.2 can use any continuous mapping ψ , where continuity is required to ensure that open sets in \mathbb{R}^n used to construct $\mathcal{F}_{\mathbb{R}^n}$ are mapped to open sets in \mathfrak{s} . However, hereafter, we restrict our attention to ψ as a linear mapping. This is depicted in Figure 2 when ψ is a projection. This figure is only a schematic, since most applications are high-dimensional. The arrow labelled \mathbf{S} spans an m -dimensional coherent subspace \mathfrak{s} , while the arrow labelled \mathbf{R} spans an $(n - m)$ -dimensional direction of projection. The mapping g collapses all points in the blue shaded region $g^{-1}(\mathcal{B})$, to the black interval \mathcal{B} . Under s , \mathcal{B} is mapped to $s(\mathcal{B})$ shown in red. Under our definition of reconciliation, the same probability is assigned to the red region under the reconciled measure as is assigned to the blue region under the incoherent measure.

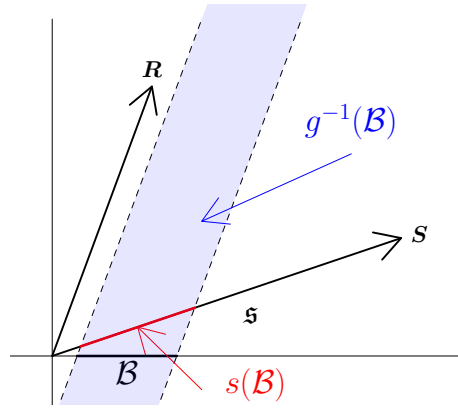


Figure 2: Summary of probabilistic forecast reconciliation. The probability that \mathbf{y}_{t+h} lies in the red segment under the reconciled probabilistic forecast equals the probability that \mathbf{y}_{t+h} lies in the shaded blue area under the unreconciled probabilistic forecast. Since most applications are high-dimensional, this figure is only a schematic.

4 Construction of Reconciled Distribution

In this section we derive theoretical results on how distributions on \mathbb{R}^n can be reconciled to a distribution on \mathfrak{s} . In Section 4.1 we show how this can be achieved analytically by a change of coordinates and marginalisation when the density is available. In Section 4.2 we explore this result further in the specific case of elliptical distributions. In Section 4.3 we consider reconciliation in the case where the density may be unavailable but it is possible to draw a

sample from the base probabilistic forecast distribution. Throughout we restrict our attention to linear reconciliation.

4.1 Analytical derivation of reconciled densities

The following theorem shows how a reconciled density can be derived from any base probabilistic forecast on \mathbb{R}^n .

Theorem 4.1 (Reconciled density of bottom-level). *Consider the case where reconciliation is carried out using a composition of linear mappings $s \circ g$ where g combines information from all levels of the base forecast into a new density for the bottom-level. The density of the bottom-level series under the reconciled distribution is*

$$\tilde{f}_b(\mathbf{b}) = |\mathbf{G}^*| \int \hat{f}(\mathbf{G}^- \mathbf{b} + \mathbf{G}_\perp \mathbf{a}) d\mathbf{a},$$

where \hat{f} is the density of the incoherent base probabilistic forecast, \mathbf{G}^- is an $n \times m$ generalised inverse of \mathbf{G} such that $\mathbf{G}\mathbf{G}^- = \mathbf{I}$, \mathbf{G}_\perp is an $n \times (n - m)$ orthogonal complement to \mathbf{G} such that $\mathbf{G}\mathbf{G}_\perp = \mathbf{0}$, $\mathbf{G}^* = \begin{pmatrix} \mathbf{G}^- & \mathbf{G}_\perp \end{pmatrix}$, and \mathbf{b} and \mathbf{a} are obtained via the change of variables

$$\mathbf{y} = \mathbf{G}^* \begin{pmatrix} \mathbf{b} \\ \mathbf{a} \end{pmatrix}.$$

Proof. See Appendix A.1. □

Theorem 4.2 (Reconciled density of full hierarchy). *Consider the case where a reconciled density for the bottom-level series has been obtained using Theorem 4.1. The density of the full hierarchy under the reconciled distribution is*

$$\tilde{f}_y(\mathbf{y}) = |\mathbf{S}^*| \tilde{f}_b(\mathbf{S}^- \mathbf{y}) \mathbb{1}\{\mathbf{y} \in \mathfrak{s}\},$$

where $\mathbb{1}\{\cdot\}$ equals 1 when the statement in braces is true and 0 otherwise,

$$\mathbf{S}^* = \begin{pmatrix} \mathbf{S}^- \\ \mathbf{S}'_\perp \end{pmatrix},$$

\mathbf{S}^- is an $m \times n$ generalised inverse of \mathbf{S} such that $\mathbf{S}^- \mathbf{S} = \mathbf{I}$, and \mathbf{S}_\perp is an $n \times (n - m)$ orthogonal complement to \mathbf{S} such that $\mathbf{S}'_\perp \mathbf{S} = \mathbf{0}$.

Proof. See Appendix A.1. □

Applying this result in the Gaussian case is shown in Appendix B in the online supplement.

4.2 Elliptical distributions

More generally, consider linear reconciliation of the form $\psi(\hat{y}) = S(d + G\hat{y})$. For an elliptical base probabilistic forecast, with location $\hat{\mu}$ and scale $\hat{\Sigma}$, the reconciled probabilistic forecast will also be elliptical with location $\tilde{\mu} = S(d + G\hat{\mu})$ and scale $\tilde{\Sigma} = SG\hat{\Sigma}G'S'$. This is a consequence of the fact that elliptical distributions are closed under linear transformations and marginalisation. While the base and reconciled distribution may be of a different form, they will both belong to the elliptical family. This leads to the following result.

Theorem 4.3 (Recovering the true density through reconciliation). *Assume the true predictive distribution is elliptical with location μ and scale Σ . Then for an elliptical base probabilistic forecast with arbitrary location $\hat{\mu}$ and scale $\hat{\Sigma}$, there exists d_{opt} and G_{opt} such that the true predictive distribution is recovered by reconciliation.*

Proof. First consider finding a G_{opt} for which the following holds,

$$\Sigma = SG_{\text{opt}}\hat{\Sigma}G'_{\text{opt}}S'. \quad (1)$$

Note that since the true data must be coherent Σ will not be full rank, while in contrast $\hat{\Sigma}$ usually will be (but need not be) full rank. Equation (1) can be solved as $G_{\text{opt}} = \Omega_0^{1/2}\hat{\Sigma}^{-1/2}$, where $\hat{\Sigma}^{1/2}$ is any matrix such that $\hat{\Sigma} = \hat{\Sigma}^{1/2}(\hat{\Sigma}^{1/2})'$.¹ Here, $\Omega_0^{1/2}(\Omega_0^{1/2})' = \Omega$ and Ω is the true scale matrix for the bottom-level series. To ensure conformability of matrix multiplication, $\Omega^{1/2}$ must be an $m \times n$ matrix; so it can be set to the Cholesky factor of Ω augmented with an additional $n - m$ columns of zeros. To reconcile the location, solve the following for d_{opt}

$$\mu = S(d_{\text{opt}} + G_{\text{opt}}\hat{\mu})$$

which is given by $d_{\text{opt}} = \beta - G_{\text{opt}}\hat{\mu}$, where β is defined so that $\mu = S\beta$. \square

The above theorem is not feasible in practice since exploiting the result requires knowledge of μ and Σ . However this result does have important consequences for the algorithm introduced in Section 6 motivating a linear form of reconciliation. In particular, SG_{opt} is not a projection matrix in general. This implies that in the probabilistic forecasting setting, it is advised to include a translation d in the reconciliation procedure. This holds even if the base forecasts are unbiased (i.e. $\hat{\mu} = \mu$) since in general $SG_{\text{opt}}\hat{\mu} \neq \mu$.

Although SG_{opt} is not a projection matrix in general, there are some conditions under which it will be. These are described by the following theorem.

¹For example a Cholesky factor which will be unique if $\hat{\Sigma}$ is full rank. If $\hat{\Sigma}$ is rank deficient, then although the Cholesky factor is no longer unique, any $\hat{\Sigma}^{1/2}$ such that $\hat{\Sigma} = \hat{\Sigma}^{1/2}(\hat{\Sigma}^{1/2})'$ can be used.

Theorem 4.4 (Optimal Projection for Reconciliation). *Let $\hat{\Sigma}$ be the scale matrix from an elliptical but incoherent base forecast and assume base forecasts are also unbiased. When the true predictive distribution is also elliptical, then this can be recovered via reconciliation using a projection if $\text{rank}(\hat{\Sigma} - \Sigma) \leq n - m$.*

Proof. See Appendix A.2. \square

Although results based on elliptical distribution may seem limited, we would note that in many operational settings, including where judgemental adjustments are made, a predicted mean and a predicted variance may be available instead of a full probabilistic forecast. In this case, a sensible parametric assumption would be to assume Gaussianity.

4.3 Simulation from a Reconciled Distribution

In practice it is often the case that samples are drawn from a probabilistic forecast since an analytical expression is either unavailable, or relies on unrealistic parametric assumptions. A useful result is the following.

Theorem 4.5 (Reconciled samples). *Suppose that $(\hat{\mathbf{y}}^{[1]}, \dots, \hat{\mathbf{y}}^{[L]})$ is a sample drawn from an incoherent probability measure $\hat{\nu}$. Then $(\tilde{\mathbf{y}}^{[1]}, \dots, \tilde{\mathbf{y}}^{[L]})$ where $\tilde{\mathbf{y}}^{[\ell]} := \psi(\hat{\mathbf{y}}^{[\ell]})$ for $\ell = 1, \dots, L$, is a sample drawn from the reconciled probability measure $\tilde{\nu}$ as defined in Definition 3.2.*

Proof. For any $\mathcal{A} \in \mathcal{F}_s$

$$\begin{aligned} \Pr(\hat{\mathbf{y}} \in \psi^{-1}(\mathcal{A})) &= \lim_{L \rightarrow \infty} L^{-1} \sum_{\ell=1}^L \mathbb{1}\{\hat{\mathbf{y}}^{[\ell]} \in \psi^{-1}(\mathcal{A})\} \\ &= \lim_{L \rightarrow \infty} L^{-1} \sum_{\ell=1}^L \mathbb{1}\{\psi(\hat{\mathbf{y}}^{[\ell]}) \in (\mathcal{A})\} \\ &= \Pr(\tilde{\mathbf{y}} \in (\mathcal{A})) \end{aligned}$$

\square

To say that a sample $\tilde{\mathbf{y}}$ is drawn from a probability distribution, then the proportion of points landing within a region (or strictly speaking Borel set) \mathcal{A} should equal the probability assigned to that region. As $L \rightarrow \infty$ these two quantities converge. Definition 3.2 establishes the connection between the probability assigned to $\psi^{-1}(\mathcal{A})$ under the base measure and probability assigned to \mathcal{A} under the reconciled measure

This result implies that reconciling each member of a sample drawn from an incoherent distribution provides a sample from the reconciled distribution. The schemes of Jeon et al.

(2019) and Rangapuram et al. (2021) are built upon this theorem. This result allows coherent forecasts to be built in a general and modular fashion, the mechanism for simulating base forecasts is separated from the question of reconciliation. This will become clear in the simulation study covered in Section 7.

5 Evaluation of Hierarchical Probabilistic Forecasts

An important issue in all forecasting problems is evaluating forecast accuracy. In the probabilistic setting, it is common to evaluate forecasts using proper scoring rules (see Gneiting and Raftery, 2007; Gneiting and Katzfuss, 2014, and references therein). Throughout, we follow the convention of negatively oriented scoring rules such that smaller values of the score indicate more accurate forecasts. In general, a scoring rule $K(., .)$, is a function taking a probability measure as the first argument and a realisation as the second argument (although for ease of notation we will at times replace the probability measure with its associated density in the first argument). A scoring rule is proper if $E_Q[K(Q, \omega)] \leq E_Q[K(P, \omega)]$ for all P , where P is any member of some class of probability measures (densities), Q is the true predictive and ω is a realisation. When this inequality is strict for all $P \neq Q$, the scoring rule is said to be strictly proper.

Since hierarchical forecasting is inherently a multivariate problem (the linear constraints affect all variables), our focus is on multivariate scoring rules. Arguably the simplest multivariate scoring rule is the log score. The log score simply evaluates the negative log density at the value of the realisation, $LS(P, \omega) = -\log f(\omega)$, where f is the density associated with a distribution P . The log score is more commonly used when a parametric form for the density is available.

Alternatively there are a number of other multivariate scoring rules that are difficult to compute using the probabilistic forecast density alone, but can be approximated using a sample drawn from that density. An example is the energy score (ES) (see Gneiting and Raftery, 2007, for details) which is a multivariate generalisation of the popular Cumulative Rank Probability Score (CRPS). The energy score is given by

$$ES(P, \omega) = E_P\|\mathbf{y} - \omega\|^\alpha - \frac{1}{2}E_P\|\mathbf{y} - \mathbf{y}^*\|^\alpha, \quad \alpha \in (0, 2], \quad (2)$$

where \mathbf{y} and \mathbf{y}^* are independent copies drawn from the distribution P . In the empirical results described later, we follow common convention by setting $\alpha = 1$. While the expectations in Equation (2) may have no closed form, they can be easily approximated via simulations using

a sample drawn from the probabilistic forecast. An alternative is the variogram score (VS) of order p (see Scheuerer and Hamill, 2015, for details) defined as

$$VS_p(P, \omega) = \sum_{i=1}^{n-1} \sum_{j=i+1}^n (|\omega_i - \omega_j|^p - E_P|y_i - y_j|^p)^2, \quad (3)$$

where ω_i, ω_j, y_i and y_j are the i -th and j -th components of ω and y as defined above. In the empirical results described later, we set $p = 0.5$. We refer readers to Alexander et al. (2021) and Bjerregård et al. (2021) for further discussion on the discrimination ability of these scoring rules.

5.1 The Log Score for Hierarchical Time Series

When an expression for the density of an incoherent base forecast is available, Section 4 describes how the density of a reconciled forecast can be recovered. With both densities available, the log score is a natural and straightforward scoring rule to use. However, the following theorem shows that the log score is improper in the setting of comparing incoherent to coherent forecasts.

Theorem 5.1 (Impropriety of log score). *When the true data generating process is coherent, then the log score is improper with respect to the class of incoherent measures.*

Proof. See Appendix A.3. □

As a result of Theorem 5.1 we recommend avoiding the log score when comparing reconciled and unreconciled probabilistic forecasts.

6 Score Optimal Reconciliation

We now propose an algorithm for finding reconciliation weights by optimising an objective function based on scores. For clarity of exposition, we consider the special case of the energy score. However, the algorithm can be generalised to any score that is computed by sampling from the probabilistic forecast. For example, in the simulations and the empirical application of Sections 7 and 8 we consider optimising with respect to both the energy and variogram scores. Motivated by Theorem 4.3, which shows that the true predictive density can be recovered (albeit only infeasibly) by linear reconciliation we consider reconciliation of the form $\tilde{y} = \psi_{\gamma}(\hat{y}) = S(d + G\hat{y})$, where $\gamma := (d, \text{vec}(G))$. This allows for more flexibility than a projection, which would imply the constraints $d = \mathbf{0}$ and $GS = I$. This added flexibility is also motivated by Theorem 4.3 which shows that projections in general are not guaranteed

to recover the true predictive distribution even in the elliptical case. When making an h -step-ahead forecast at time T , the objective used to determine an optimal value of γ is the total energy score based on in-sample information, given by

$$\mathcal{E}(\gamma) = \sum_{t=T}^{T+R-1} ES(\tilde{f}_{t+h|t}^{\gamma}, \mathbf{y}_{t+h}), \quad (4)$$

where $\tilde{f}_{t+h|t}^{\gamma}$ is a probabilistic forecast for \mathbf{y}_{t+h} made at time t and reconciled with respect to $\psi_{\gamma}(\cdot)$, and R is the number of score evaluations used in forming the objective function.

One of the challenges in optimising this objective function is that there is, in general, no closed form expression for the energy score. However, it can be easily approximated by simulation as

$$\hat{\mathcal{E}}(\gamma) = \sum_{t=T}^{T+R-1} \left[\frac{1}{Q} \left(\sum_{q=1}^Q \|\tilde{\mathbf{y}}_{t+h|t}^{[q]} - \mathbf{y}_{t+h}\| - \frac{1}{2} \|\tilde{\mathbf{y}}_{t+h|t}^{[q]} - \tilde{\mathbf{y}}_{t+h|t}^{*[q]}\| \right) \right], \quad (5)$$

where $\tilde{\mathbf{y}}_{t+h|t}^{[q]} = \mathbf{S}(\mathbf{d} + \mathbf{G}\hat{\mathbf{y}}_{t+h|t}^{[q]})$, $\tilde{\mathbf{y}}_{t+h|t}^{*[q]} = \mathbf{S}(\mathbf{d} + \mathbf{G}\hat{\mathbf{y}}_{t+h|t}^{*[q]})$ and $\hat{\mathbf{y}}_{t+h|t}^{[q]}, \hat{\mathbf{y}}_{t+h|t}^{*[q]} \stackrel{iid}{\sim} \hat{f}_{t+h|t}$ for $q = 1, \dots, Q$.

The objective function is optimised by Stochastic Gradient Descent (SGD). The SGD method has become increasingly popular in machine learning and statistics over the past decade having been applied to training neural networks (Bottou, 2010) and Variational Bayes (Kingma and Welling, 2013). There is also a recent but growing literature on using SGD to optimise scoring rules (see Gasthaus et al., 2019; Janke and Steinke, 2020; Hofert et al., 2020, and references therein for examples). These papers typically deal with high dimensional problems, deep neural networks handle millions of parameters, so this tool is well suited to our problem. An important distinction is that the use of SGD, rather than gradient descent in these contexts, arises due to computational considerations, as it is not efficient to use all data. In contrast we use all data and the ‘stochastic’ nature of our gradient descent arises since the score functions contain integrals that must be estimated by Monte Carlo.

It requires an estimate of the gradient $\partial \hat{\mathcal{E}} / \partial \gamma$ which is computed by automatically differentiating Equation (5) using the header only C++ library of the Stan project (Carpenter et al., 2015). The learning rates used for SGD are those of the Adam method (see Kingma and Ba, 2014, for details). Pseudo-code for the full procedure in the case where $h = 1$ is provided in Algorithm 1 and is implemented in the R package *ProbReco* (Panagiotelis, 2020).

While Algorithm 1 is not the first instance of calibrating parameters by optimising scoring rules (see Gneiting et al., 2005, for an earlier example), to the best of our knowledge it is the first instance of doing so to find a projection to be used in forecast reconciliation. Rangapuram

Algorithm 1 SGD with Adam for score optimal reconciliation (one-step-ahead forecasts). The initial value of γ is given by OLS reconciliation. Steps 9–14 are the standard steps for SGD with Adam. Squaring \mathbf{g}_j in Step 11 and division and addition in Step 14 are element-wise operations.

```

1: procedure SCOREOPT( $\mathbf{y}_1, \dots, \mathbf{y}_{T+R}, \beta_1, \beta_2, \epsilon, \eta$ ).
2:   for  $t = T : T + R - 1$  do
3:     Find base forecasts  $\hat{f}_{t+1|t}$  using  $t - T + 1, t - T + 2, \dots, t$  as training data.
4:   end for
5:   Initialise  $\mathbf{m}_0 = \mathbf{0}, \mathbf{v}_0 = \mathbf{0}$  and  $\gamma_0 = (\mathbf{0}, \text{vec}((\mathbf{S}'\mathbf{S})^{-1}\mathbf{S}'))$ 
6:   for  $j = 1, 2, 3, \dots$  up to convergence do
7:     Draw  $\hat{\mathbf{y}}_{t+1|t}^{[q]}, \hat{\mathbf{y}}_{t+1|t}^{*[q]} \sim \hat{f}_{t+1|t}$  for  $q = 1, \dots, Q, t = T, \dots, T + R - 1$ .
8:     Compute  $\tilde{\mathbf{y}}_{t+1|t}^{[q]}$  and  $\tilde{\mathbf{y}}_{t+1|t}^{*[q]}$  for  $q = 1, \dots, Q, t = T, \dots, T + R - 1$  using  $\gamma_{j-1}$ .
9:      $\mathbf{g}_j \leftarrow \partial \hat{\mathcal{E}} / \partial \gamma|_{\gamma=\gamma_{j-1}}$  ▷ Compute gradient
10:     $\mathbf{m}_j \leftarrow \beta_1 \mathbf{m}_{j-1} + (1 - \beta_1) \mathbf{g}_j$  ▷ Moving average of gradient
11:     $\mathbf{v}_j \leftarrow \beta_2 \mathbf{v}_{j-1} + (1 - \beta_2) \mathbf{g}_j^2$  ▷ Moving average of squared gradient
12:     $\hat{\mathbf{m}}_j \leftarrow \mathbf{m}_j / (1 - \beta_1^j)$  ▷ Bias correct
13:     $\hat{\mathbf{v}}_j \leftarrow \mathbf{v}_j / (1 - \beta_2^j)$  ▷ Bias correct
14:     $\gamma_j \leftarrow \gamma_{j-1} + \eta \frac{\hat{\mathbf{m}}_j}{(\hat{\mathbf{v}}_j + \epsilon)}$  ▷ Update weights
15:   end for
16:   Set the reconciled forecast as  $\hat{f}_{T+R+1|T+R}^{\gamma_{\text{opt}}}$  where  $\gamma_{\text{opt}}$  is the converged value of  $\gamma$ .
17: end procedure

```

et al. (2021) use a similar approach in their end-to-end forecasting process. Their method is more restrictive than what we propose here in that the projection must be orthogonal, base forecasts are not translated, and base forecasts must be generated by a DeepVAR.

Algorithm 1 is amenable to parallel computing architectures: the loop beginning at line 2 of the pseudo-code of Algorithm 1 can be done in parallel as can the computation of the gradient. Finally, the total score in Equation (4) can be replaced with a weighted sum where appropriate; for instance weights that decay for scores computed further in the past will favour choices of γ that produced better forecasting performance for more recent forecast windows.

7 Simulations

The aim of the simulations that follow is to demonstrate probabilistic forecast reconciliation including the algorithm discussed in Section 6. For all simulations, the tuning parameters

for the SGD are set as $\eta = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1 \times 10^{-8}$, which are the values recommended by Kingma and Ba (2014) and used in popular software packages such as TensorFlow, Keras and Torch amongst others. Convergence is achieved when the change in all gradients is less than 10% of the step size η . The number of sample periods used to construct the objective function is $R = 250$, while the number of draws used to estimate each score is $Q = 250$. All estimation of base models uses a sample size of $T = 500$. All forecast evaluations are carried out using a rolling window, also of size $W = 500$.

7.1 Data Generating Processes

The data generating process we consider corresponds to the 3-level hierarchical structure presented in Figure 1. Bottom-level series are first generated from $ARIMA(p, d, q)$ processes, which are in turn aggregated to form the middle and top-level series. The orders p and q are randomly selected from $\{1, 2\}$ for each bottom-level series. The AR and MA parameters are randomly and uniformly generated from $[0.3, 0.5]$ and $[0.3, 0.7]$ respectively, and only accepted if they belong to the stationary and invertible region. In addition a non-stationary case where d is randomly chosen for each bottom-level from $\{0, 1\}$ was considered, these results are omitted for brevity. A complete set of results are available at the github repository <https://git.io/JJwQB>.

We consider a multivariate Gaussian and a non-Gaussian setting for the errors driving the ARIMA processes. Specifically, the non-Gaussian errors are drawn from a meta-distribution of a Gumbel copula with Beta(1, 3) margins. After simulating from the ARIMA models, additional noise is added to ensure bottom-level series have a **considerably** lower signal-to-noise ratio than **upper**-level series with details provided in Appendix D of the online supplement. For each series the first 500 observations are ignored to avoid the impact of initial values.

7.2 Modelling and Base Forecasts

We fit univariate ARIMA models to each series using the `ARIMA()` function in the `fable` package (O'Hara-Wild et al., 2020) in R (R Core Team, 2018). Note that the order of the ARIMA models is not set to the true order but is chosen using the algorithm of Hyndman and Khandakar (2008), allowing for the possibility of misspecification. Indeed, an advantage of forecast reconciliation is the ability to down-weight the forecasts of series within the hierarchy that come from misspecified models. We also considered exponential smoothing (ETS)

models using the `ETS()` function in the `fable` package. These are omitted for brevity; please refer to <https://git.io/JJwQB> for a full set of results.

Let $\hat{\mathbf{y}}_{t+h|t} = (\hat{y}_{1,t+h|t}, \dots, \hat{y}_{n,t+h|t})'$, where $\hat{y}_{i,t+h|t}$ is the h -step-ahead point forecast for series i , and $\mathbf{E} := \{e_{i,t}\}_{i=1,\dots,n;t=1,\dots,T}$ is an $(n \times T)$ matrix of stacked residuals $e_{i,t}$. For each series and model, base probabilistic forecasts for $h = 1$ are constructed in the following four ways:

- **Independent Gaussian:** The base probabilistic forecast is made up of independent Gaussian distributions with the forecast mean and variance of variable i given by $\hat{y}_{i,t+h|t}$ and $\hat{\sigma}_{i,t+h|t}^2$, where $\hat{\sigma}_{i,t+h|t}^2$ is the sample variance of the residuals in the i th row of \mathbf{E} .
- **Joint Gaussian:** The base probabilistic forecast is a multivariate Gaussian distribution with the forecast mean $\hat{\mathbf{y}}_{t+h|t}$ and variance covariance matrix $\hat{\Sigma}$, where $\hat{\Sigma}$ is the variance covariance matrix of the residuals of the fitted models.
- **Independent Bootstrap:** Draw from the base probabilistic forecast independently for each variable as $\hat{y}_{i,t+h|t} + e_{i,\tau}$ with τ is drawn randomly (with replacement) from $1, 2, \dots, T$. The number of bootstrap samples is set equal to the sample size both here and in Section 8
- **Joint Bootstrap:** Draw from the joint probabilistic forecast with $\hat{\mathbf{y}}_{t+h|t} + \mathbf{e}_\tau$ where \mathbf{e}_τ is the τ th column of \mathbf{E} , where τ is drawn randomly (with replacement) from $1, 2, \dots, T$.

We restrict our attention to $h = 1$ although these methods can be generalised to larger h using the recursive method (Hyndman and Athanasopoulos, 2021). For multi-step-ahead forecasts, bootstraps should be block-wise to preserve serial dependence in the residuals.

7.3 Reconciliation

For each DGP, model and method for obtaining base forecasts, reconciled probabilistic forecasts are obtained using each of the following techniques:

- **Base:** The base forecasts with no reconciliation.
- **JPP:** The best method of Jeon et al. (2019). This is equivalent to reconciling quantiles. A sample is drawn from the base forecast, these are ranked, one variable at a time (so that the smallest value drawn from each variable are put together, etc.). These are then pre-multiplied by $\mathbf{S}(\mathbf{S}'\mathbf{W}\mathbf{S})^{-1}\mathbf{S}'\mathbf{W}$ where \mathbf{W} is a diagonal matrix with elements $(1/4^2, 1/2^2, 1/2^2, 1, 1, 1, 1)$. These are the squared reciprocals of the number of bottom-level series used to form an aggregate.
- **BTTH:** The method of Ben Taieb et al. (2020). This is a method whereby draws from the probabilistic forecasts of the bottom-level series are permuted so that they have the same empirical copula as the residuals. These are then aggregated to form a sample

from the distribution of all series. The mean is adjusted to be equivalent to the mean that would be obtained using the MinT method of Wickramasuriya et al. (2019) described in Table 1.

- **BottomUp:** Reconciliation via premultiplication by SG where $G = (\mathbf{0}_{m \times (n-m)}, \mathbf{I}_{m \times m})$.
- **OLS:** Reconciliation via pre-multiplication by $S(S'S)^{-1}S'$.
- **MinTShr:** Reconciliation via pre-multiplication using the shrinkage estimator of the covariance matrix used by Wickramasuriya et al. (2019) but applied to probabilistic rather than point forecasting.
- **ScoreOptE:** The algorithm described in Section 6 used to optimise energy score.
- **ScoreOptV:** The algorithm described in Section 6 used to optimise variogram score.

Note that JPP and BTTH previously exist in the literature. The methods BottomUp, OLS, and MinTShr have been used extensively for point forecasting but their application to probabilistic forecasting for general base forecasts is, to our best knowledge, novel.

In addition to these, two further reconciliation methods were considered; WLS, which reconciles via pre-multiplication by the matrix used in Jeon et al. (2019) but with no reordering of the draws, and MinTSam which uses a sample estimate of the covariance matrix rather than a shrinkage estimator. These methods were mostly dominated by OLS and MinTShr respectively and are omitted for brevity; please refer to <https://git.io/JJwQB> for a full set of results.

7.4 Energy score results for probabilistic forecasts

The left panel of Figure 3 shows the mean energy score for different reconciliation methods and different methods of generating base forecasts. When base probabilistic forecasts are generated independently, score optimisation with the energy score (ScoreOptE) performs best, while when base forecasts are generated jointly, the MinT method for reconciliation using the shrinkage estimator (MinTShr) yields the most accurate forecasts. The bottom-up method as well as BTTH and JPP fail to even improve upon base forecasts in all cases. As expected score optimisation using the variogram score does not perform as well as score optimisation using energy score, when evaluation is carried out with respect to the latter. However, the results are quite close suggesting that score optimisation is fairly robust to using an alternative proper score.

To assess significant differences between the reported results, we use post-hoc Nemenyi tests (Hollander et al., 2013). The Nemenyi test is a non-parametric test that identifies groups of forecasts which cannot be significantly distinguished from one another. We use the

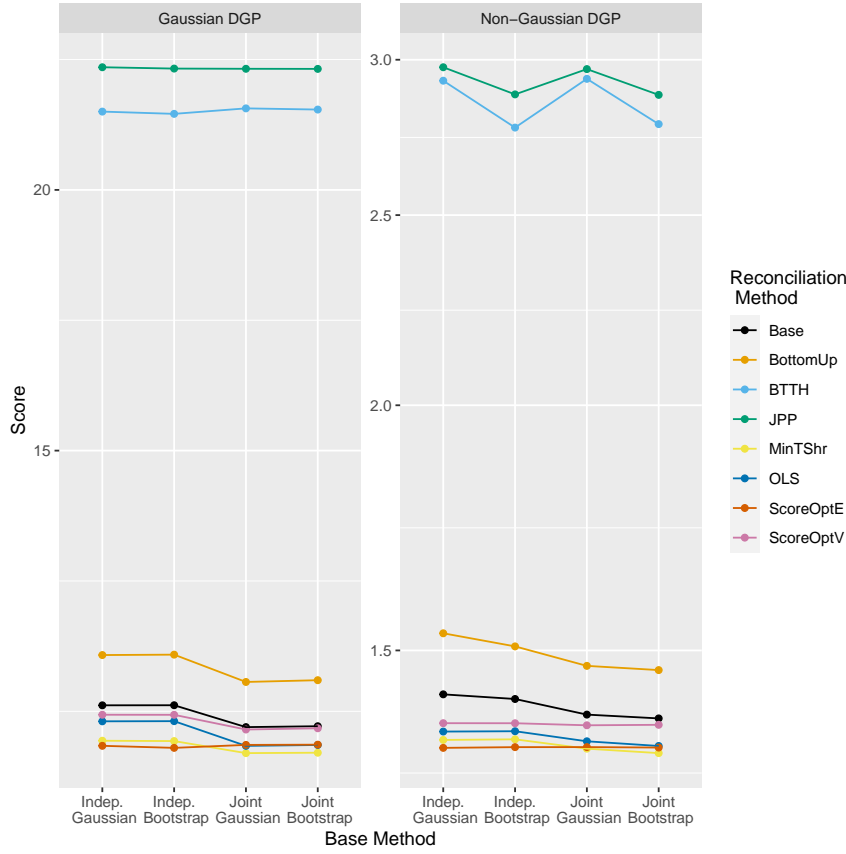


Figure 3: Mean energy scores using different base forecast and reconciliation methods. Left panel is the Gaussian data, right panel is the non-Gaussian data.

implementation of the tests available in the `tsutils` R package (Kourentzes, 2019). Figure 4 reports the results which should be looked at column-wise. A blue square indicates that the method in the corresponding row, is statistically indistinguishable from the method in that column. For all four methods of generating base forecasts, MinTShr, ScoreOptE and OLS significantly outperform base forecasts, bottom-up forecasts, BTTH and JPP.

The right panel of Figure 3 reports the mean energy score for the non-Gaussian DGP. Overall, the results are quite similar to the Gaussian DGP. The best performing reconciliation method is ScoreOptE when base probabilistic forecasts are independent, and MinTShr when base forecasts are dependent. The Nemenyi matrix is omitted for brevity; please refer to <https://git.io/JJwQB> for a full set of results. However, these lead to similar conclusion to Figure 4. The methods ScoreOptE, MinTShr and OLS are statistically indistinguishable from one another but are significantly better than base forecasts and the bottom-up method. The methods BTTH and JPP lead to a statistically significant deterioration in forecast quality relative to base forecasts.

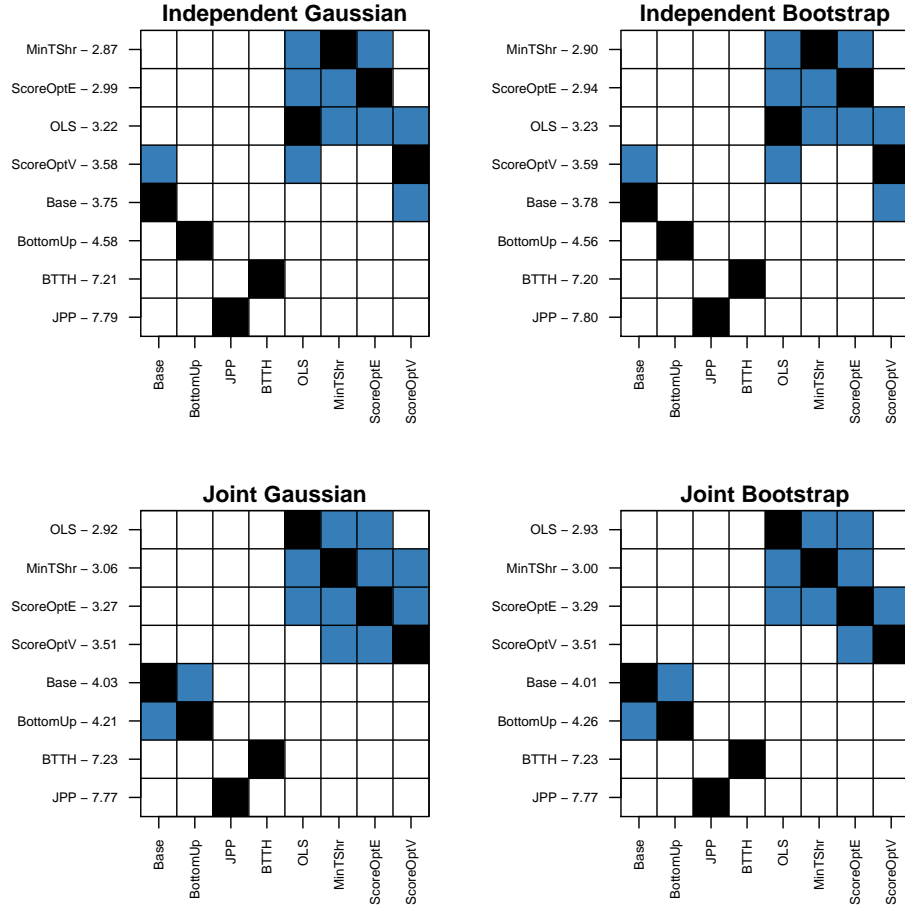


Figure 4: *Nemenyi matrix for Energy Score for Gaussian DGP.*

Similar conclusions can be drawn based on the results for both Gaussian and non-Gaussian probabilistic forecasts considering the mean variogram score. These are presented in Appendix E.

8 Forecasting Australian Electricity Generation

8.1 Data Description and Base Forecasts

To demonstrate the potential of the proposed methods, we consider an application to forecasting electricity generation from different sources of energy. Daily time series were obtained from opennem.org.au, a website that compiles publicly available data from the Australian Energy Market Operator (AEMO). Probabilistic day-ahead forecasts are crucial inputs into operational and planning decisions that ensure efficiency and stability of the power network. This has become a more challenging problem with growth in intermittent sources of generation such as wind and solar. The hierarchy comprises three levels of aggregation.

1. *Total* generation is the sum of generation from *Renewable* and *non-Renewable* sources.

2. *Renewable* generation is the sum of *Batteries*, *Hydro (inc. Pumps)*, *Solar*, *Wind* and *Biomass*. *Non-Renewable* is the sum of *Coal*, *Gas* and *Distillate*
3. *Battery* generation is given by *Battery (Discharging)* minus *Battery (Charging)*, *Hydro (inc. Pumps)* is *Hydro* generation minus *Pumps* (energy used to pump water upstream), while *Solar* generation is the sum of *Solar (Rooftop)* and *Solar (Utility)*. *Coal* generation is the sum of *Black Coal* and *Brown Coal*, while *Gas* is the sum of *Gas (OCGT)*, *Gas (CCGT)*, *Gas (Steam)*, *Gas (Reciprocating)*.

In total, there are $n = 23$ series of which $m = 15$ are bottom-level series.

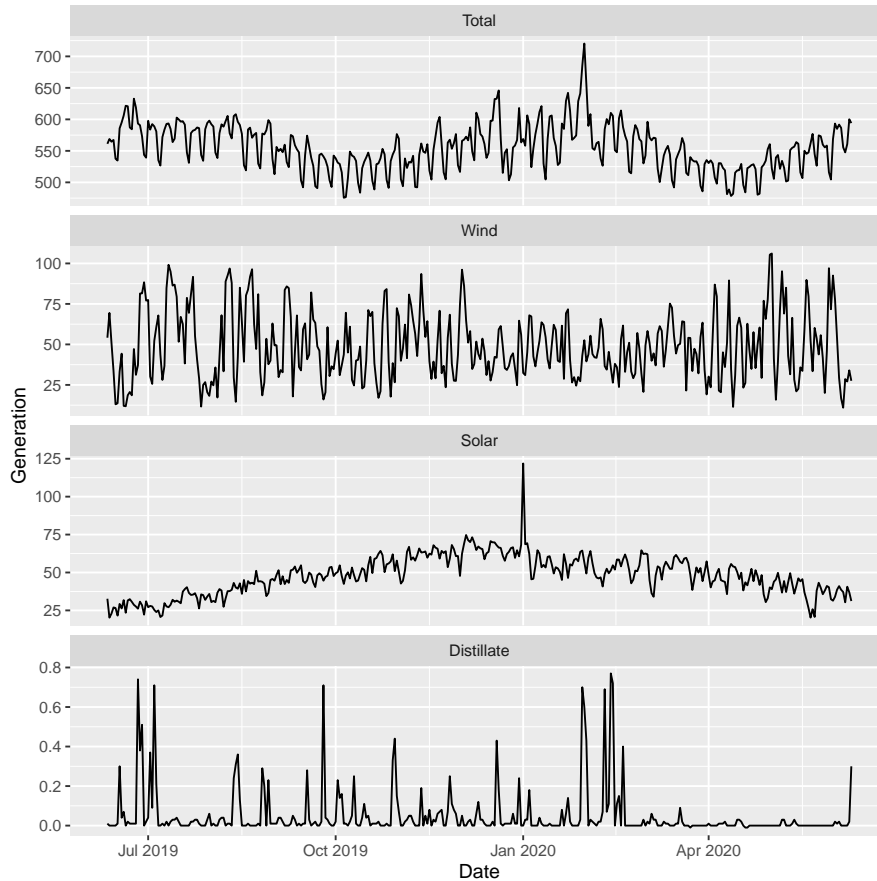


Figure 5: Time series plots for selected series from 11 June 2019 to 10 June 2020.

Figure 5 shows time plots for some selected series². The series exhibit some interesting and unique features. At the aggregate level, *Total* generation shows strong weekly seasonality, with troughs corresponding to weekends. An annual seasonal pattern is also displayed with peaks occurring during the months of June–August as well as December–February. These periods correspond to the winter and summer months in Australia for which electricity demand peaks for heating and cooling purposes respectively. As expected, generation from *Solar* peaks during the summer months of December–February. There are also some unusually

²Time plots for the remaining series area available from <https://git.io/JJwd0>.

large spikes observed in both the *Total* and *Solar* series during February and January 2020 respectively. *Wind* displays higher volatility (especially outside the summer months), while generation from *Distillate* exhibits aperiodic spikes. The diversity and prominence of the features in each series and each level of aggregation highlights the importance of modelling and forecasting each series on its own merits and then applying a reconciliation approach.

The forecast evaluation is based on a rolling window. Each training window consists of 140 days (20 weeks) of data. One-step-ahead forecasts were generated leading to 170 daily forecasts for evaluation. Each series was independently modelled using a one-layer feed-forward neural network with up to 28 lags of the target variable as inputs. This was implemented using the `NNETAR` function in the `fable` package. Neural networks are used to highlight the versatility of reconciliation to different forecasting approaches. While including more layers or meteorological variables as predictors will probably lead to improved base forecasts, the primary objective is to assess the effectiveness of different forecast reconciliation methods. For base forecasts, we also considered a multivariate model, namely a Vector Autoregression (VAR) with shrinkage, implemented using the R package `BigVAR` package (Nicholson et al., 2019). Since, for both base forecasts and reconciled forecasts the neural network outperformed the VAR, we focus only on the former. However, we note that even for the VAR results (summarised in Appendix F of the online supplement) reconciliation methods improve upon base forecasts. This highlights that forecast reconciliation should not be thought of as an alternative for forecasting from multivariate models, but rather as an option for improving both univariate and multivariate approaches.

Four situations were considered where base forecasts are assumed to be either Gaussian or bootstrapped from residuals, and either independent or dependent (we use the residual covariance matrix of the fitted neural networks, in a similar fashion as in Section 7). Figure 6 demonstrates departures from normality in the residuals of base forecasting models, while the correlation heatmap of these residuals in Figure 7 demonstrates departures from independence. Therefore, independent Gaussian probabilistic forecasts are likely to represent severe misspecification.

8.2 Reconciliation

The same reconciliation methods were used as in the simulation study with score optimisation based on an objective function with 56 days of score evaluations. For brevity, only the energy score results are presented; please refer to <https://git.io/JJMEH> for a full set of results.

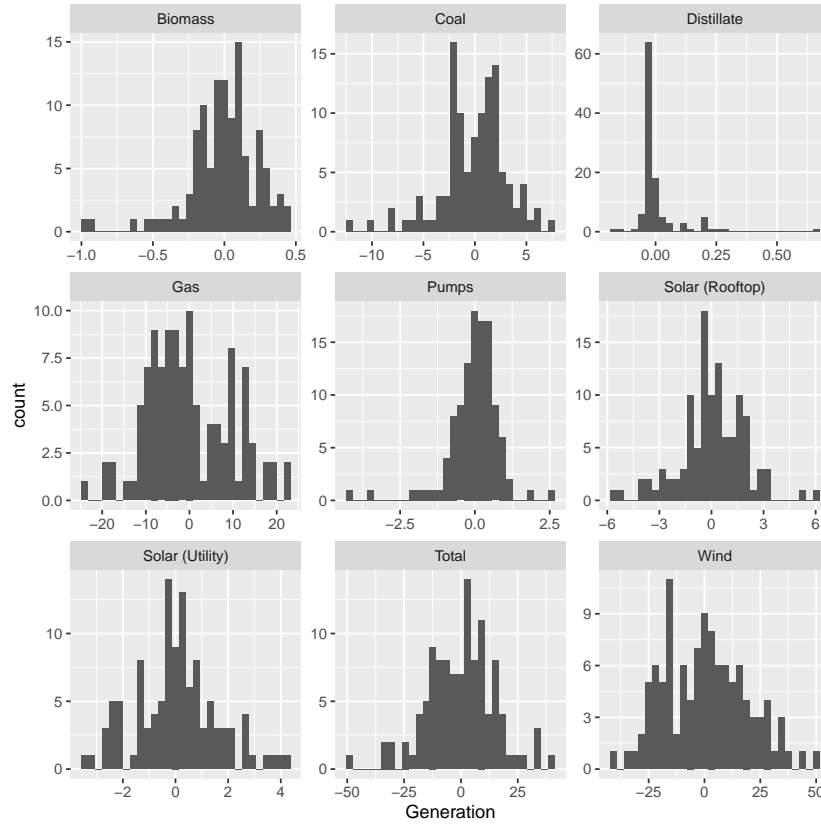


Figure 6: Densities of residuals for selected series from a typical training window of 2 October 2019 to 21 January 2020.

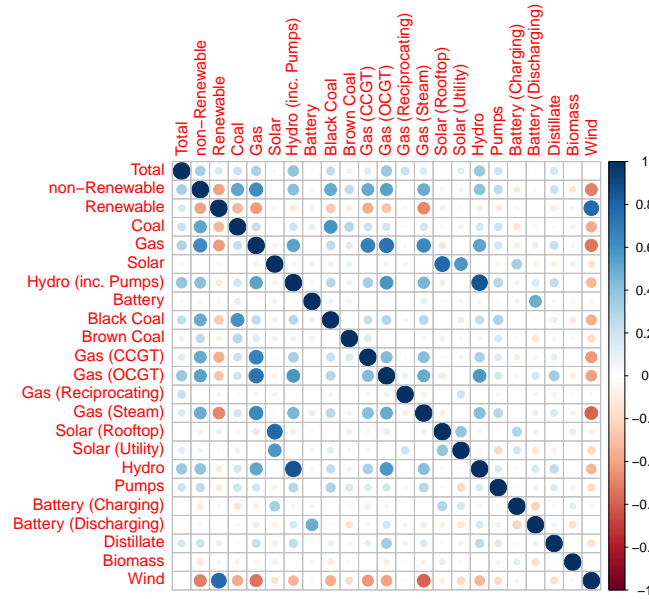


Figure 7: Correlation heatmap of residuals from a typical training window of 2 October 2019 to 21 January 2020. Blue circles indicate positive correlation, while red circles indicate negative correlation with larger circles indicating stronger correlations.

The mean energy score for all four base forecasting methods is summarised in Figure 8. When base forecasts are generated assuming both independence and a Gaussian distribution, score optimisation achieves a mean energy score that is considerably smaller than all other competing methods, with MinT providing the second smallest value. The superior forecasting performance of score optimisation is statistically significant, see the Nemenyi matrix in Figure 9 (left). This suggests that score optimisation is best for guarding against severe model misspecification.

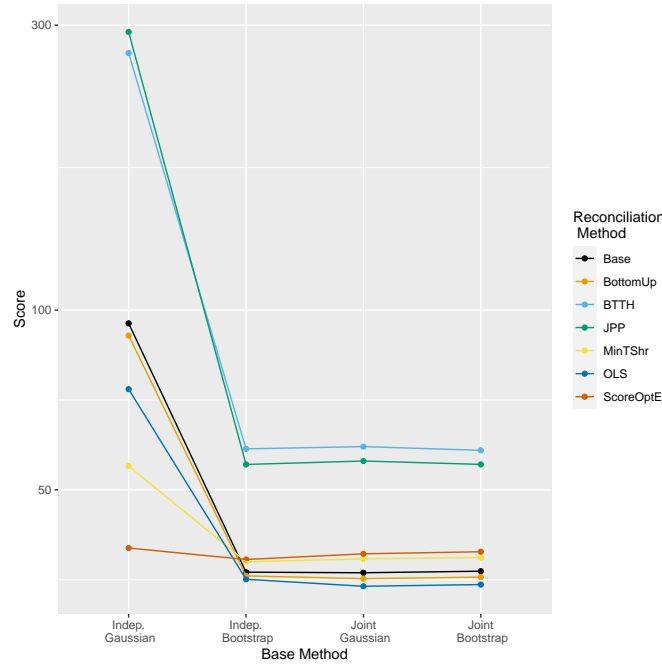


Figure 8: Mean Energy score for the electricity application for different base forecasting methods and different reconciliation methods.

For all other methods the best performing method is OLS. This difference is statistically significant.³ We suggest two possible reasons for the good performance of OLS in the probabilistic case. First, the energy score depends on the L2 norm of the difference between realizations and draws from the probabilistic forecast, which is similar to the setting for which OLS has optimal properties for point forecasts (see Panagiotelis et al., 2021). Second, for OLS there is less estimation uncertainty as fewer parameters need to be estimated. Although score optimisation does not improve upon base, the differences are not significant. For all base forecasts, both JPP and BTTH are significantly worse than base forecasts.

³See the Nemenyi matrix for jointly bootstrapped base forecasts in Figure 9 (right). The corresponding figures for joint Gaussian and independent bootstrap look mostly similar to the right panel of Figure 9; please refer to <https://git.io/JJMEH> for a full set of results.

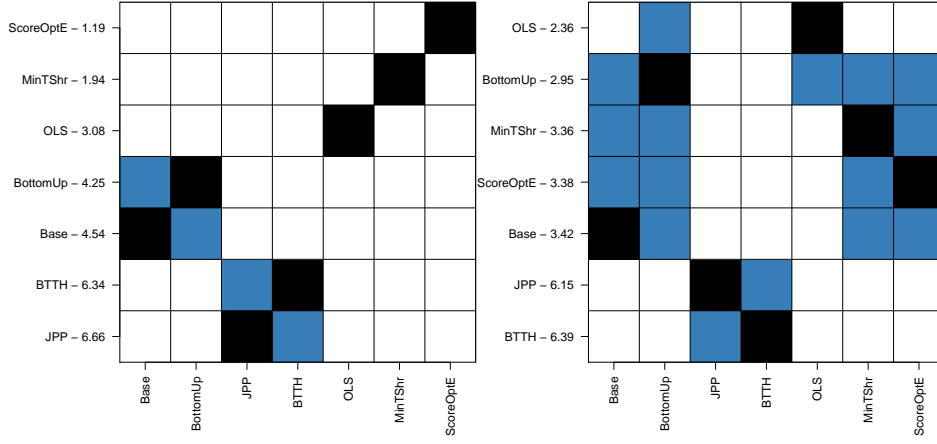


Figure 9: Nemenyi matrices for Energy score for the electricity application. Left: base forecasts are independent and Gaussian. Right: base forecasts are obtained by jointly bootstrapping residuals.

9 Conclusions

This paper introduces a rigorous formulation of forecast reconciliation in the probabilistic setting. It can be applied when the base forecast is either available as a density or when a sample has been drawn from the base forecast. In the elliptical case, we prove that reconciliation can recover the correct probability distribution if the base forecast is of the correct form, irrespective of the scale and location of the base forecast. Probably due to this reason, score optimisation works well in applications even when the base forecasts are assumed to be independent.

We also prove that the log score is not proper when comparing incoherent and coherent forecasts. Consequently, we introduce a new algorithm that trains reconciliation weights by minimising the energy score or variogram score. Since the scores are approximated by Monte Carlo simulation, stochastic gradient descent is used for optimisation. This method is shown to lead to significant improvements over base forecasts, bottom-up methods and existing probabilistic reconciliation approaches across a wide variety of simulated and empirical examples, **particularly when the base forecasting models are severely misspecified.**

An interesting result is that projection methods with certain optimality properties in the point forecasting setting, also work well when extended to the probabilistic case. In particular, a simple least squares projection is the best performing method in the high-dimensional empirical example, provided the base forecasts are not too badly misspecified. This may arise since projections implicitly provide constrained versions of the reconciliation weights. A promising future research avenue may involve regularised versions of score optimisation that

add an L_1 or L_2 penalty to the objective function. Alternatively, early stopping (Bühlmann and Yu, 2003) of the gradient descent may lead to a better bias-variance tradeoff in learning reconciliation weights.

A final important avenue of future research is the development of probabilistic forecast reconciliation for domains other than the real line. These may include domains constrained above zero, discrete domains, or domains that are a mixture of continuous distributions and discrete point masses. While such problems are challenging, the geometric interpretation of probabilistic forecast introduced in this paper, lays the foundation for this research agenda.

References

- Alexander, C., M. Coulon, Y. Han, and X. Meng (2021). Evaluating the Discrimination Ability of Proper Multivariate Scoring Rules. pp. 1–35.
- Athanasopoulos, G., R. A. Ahmed, and R. J. Hyndman (2009). Hierarchical forecasts for Australian domestic tourism. International Journal of Forecasting 25(1), 146 – 166.
- Athanasopoulos, G., P. Gamakumara, A. Panagiotelis, R. J. Hyndman, and M. Affan (2020). Hierarchical Forecasting. In Peter Fuleky (Ed.), Macroeconomic Forecasting in the Era of Big Data. Advanced Studies in Theoretical and Applied Econometrics. (vol 52 ed.), Chapter 21, pp. 689–719. Springer, Cham.
- Athanasopoulos, G., R. J. Hyndman, N. Kourentzes, and M. O’Hara-Wild (2022). Probabilistic forecasts using expert judgement: the road to recovery from COVID-19. Journal of Travel Research forthcoming, 1–64.
- Athanasopoulos, G., R. J. Hyndman, N. Kourentzes, and F. Petropoulos (2017). Forecasting with temporal hierarchies. European Journal of Operational Research 262(1), 60–74.
- Azzalini, A. (2020). sn: The Skew-Normal and Related Distributions such as the Skew- t . Università di Padova, Italia. R package version 1.6-1.
- Babai, M. Z., J. E. Boylan, and B. Rostami-Tabar (2021). Demand forecasting in supply chains: a review of aggregation and hierarchical approaches. International Journal of Production Research forthcoming, 1–25.

- Ben Taieb, S., R. Huser, R. J. Hyndman, and M. G. Genton (2017). Forecasting uncertainty in electricity smart meter data by boosting additive quantile regression. IEEE Transactions on Smart Grid 7(5), 2448–2455.
- Ben Taieb, S. and B. Koo (2019, 07). Regularized regression for hierarchical forecasting without unbiasedness conditions. In KDD '19: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1337–1347.
- Ben Taieb, S., J. W. Taylor, and R. J. Hyndman (2020). Hierarchical probabilistic forecasting of electricity demand with smart meter data. Journal of the American Statistical Association. in press.
- Bjerregård, M. B., J. K. Møller, and H. Madsen (2021). An introduction to multivariate probabilistic forecast evaluation. Energy and AI 4, 100058.
- Böse, J.-H., V. Flunkert, J. Gasthaus, T. Januschowski, D. Lange, D. Salinas, S. Schelter, M. Seeger, and Y. Wang (2017). Probabilistic demand forecasting at scale. Proceedings of the VLDB Endowment 10(12), 1694–1705.
- Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In Y. Lechevallier and G. Saporta (Eds.), Proceedings of COMPSTAT'2010, pp. 177–186. Physica-Verlag HD.
- Bühlmann, P. and B. Yu (2003). Boosting with the L_2 loss: regression and classification. Journal of the American Statistical Association 98(462), 324–339.
- Carpenter, B., M. D. Hoffman, M. Brubaker, D. Lee, P. Li, and M. Betancourt (2015). The Stan math library: Reverse-mode automatic differentiation in C++.
- Dunn, D. M., W. H. Williams, and T. L. Dechaine (1976). Aggregate versus subaggregate models in local area forecasting. Journal of the American Statistical Association 71(353), 68–71.
- Eckert, F., R. J. Hyndman, and A. Panagiotelis (2021). Forecasting Swiss exports using Bayesian forecast reconciliation. European Journal of Operational Research 291(2), 693–710.
- Gasthaus, J., K. Benidis, Y. Wang, S. S. Rangapuram, D. Salinas, V. Flunkert, and T. Januschowski (2019). Probabilistic forecasting with spline quantile function rnns. In

The 22nd international conference on artificial intelligence and statistics, pp. 1901–1910.
PMLR.

Gneiting, T. and M. Katzfuss (2014). Probabilistic forecasting. Annual Review of Statistics and Its Application 1, 125–151.

Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102(477), 359–378.

Gneiting, T., A. E. Raftery, A. H. Westveld III, and T. Goldman (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Monthly Weather Review 133(5), 1098–1118.

Gross, C. W. and J. E. Sohl (1990). Disaggregation methods to expedite product line forecasting. Journal of Forecasting 9(3), 233–254.

Hofert, M., A. Prasad, and M. Zhu (2020). Applications of multivariate quasi-random sampling with neural networks.

Hollander, M., D. A. Wolfe, and E. Chicken (2013). Nonparametric statistical methods. John Wiley & Sons.

Hyndman, R. J., R. A. Ahmed, G. Athanasopoulos, and H. L. Shang (2011). Optimal combination forecasts for hierarchical time series. Computational Statistics and Data Analysis 55(9), 2579–2589.

Hyndman, R. J. and G. Athanasopoulos (2021). Forecasting: Principles and Practice (3rd ed.). Melbourne, Australia: OTexts.

Hyndman, R. J. and Y. Khandakar (2008). Automatic time series forecasting: The forecast package for R. Journal of Statistical Software 26(3), 1–22.

Janke, T. and F. Steinke (2020, Aug). Probabilistic multivariate electricity price forecasting using implicit generative ensemble post-processing. 2020 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS).

Jeon, J., A. Panagiotelis, and F. Petropoulos (2019). Probabilistic forecast reconciliation with applications to wind power and electric load. European Journal of Operational Research 279(2), 364–379.

Kingma, D. P. and J. Ba (2014). Adam: A method for stochastic optimization.

Kingma, D. P. and M. Welling (2013). Auto-encoding variational Bayes.

Kourentzes, N. (2019). tsutils: Time Series Exploration, Modelling and Forecasting. R package version 0.9.0.

Kourentzes, N. and G. Athanasopoulos (2019). Cross-temporal coherent forecasts for Australian tourism. Annals of Tourism Research 75, 393–409.

Kourentzes, N. and G. Athanasopoulos (2021). Elucidate structure in intermittent demand series. European Journal of Operational Research 288(1), 141–152.

Li, H. and R. J. Hyndman (2021). Assessing mortality inequality in the U.S.: What can be said about the future? Insurance: Mathematics and Economics 99, 152–162.

McLean Sloughter, J., T. Gneiting, and A. E. Raftery (2013). Probabilistic wind vector forecasting using ensembles and Bayesian model averaging. Monthly Weather Review 141(6), 2107–2119.

Nicholson, W., D. Matteson, and J. Bien (2019). BigVAR: Dimension Reduction Methods for Multivariate Time Series. R package version 1.0.6.

Nystrup, P., E. Lindstrom, P. Pinson, and H. Madsen (2020). Temporal hierarchies with autocorrelation for load forecasting. European Journal of Operational Research 280(3), 876 – 888.

O’Hara-Wild, M., R. Hyndman, and E. Wang (2020). fable: Forecasting Models for Tidy Time Series. R package version 0.2.0.

Panagiotelis, A. (2020). ProbReco: Score Optimal Probabilistic Forecast Reconciliation. R package version 0.1.0.

Panagiotelis, A., G. Athanasopoulos, P. Gamakumara, and R. J. Hyndman (2021). Forecast reconciliation: A geometric view with new insights on bias correction. International Journal of Forecasting 37(1), 343–359.

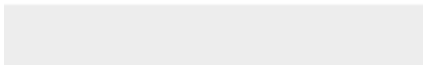
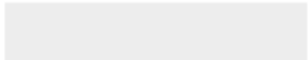
R Core Team (2018). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

- Rangapuram, S. S., L. D. Werner, K. Benidis, P. Mercado, J. Gasthaus, and T. Januschowski (2021). End-to-End Learning of Coherent Probabilistic Forecasts for Hierarchical Time Series. In International Conference on Machine Learning, pp. 8832–8843.
- Rossi, B. (2014). Density forecasts in economics, forecasting and policymaking. Technical report, Els Opuscles del CREI.
- Schäfer, J. and K. Strimmer (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical Applications in Genetics and Molecular Biology 4(1).
- Scheuerer, M. and T. M. Hamill (2015). Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. Monthly Weather Review 143(4), 1321–1334.
- Shang, H. L. and R. J. Hyndman (2017). Grouped functional time series forecasting: An application to age-specific mortality rates. Journal of Computational and Graphical Statistics 26(2), 330–343.
- Székely, G. J. and M. L. Rizzo (2013). Energy statistics: A class of statistics based on distances. Journal of Statistical Planning and Inference 143(8), 1249–1272.
- Van Erven, T. and J. Cugliari (2015). Game-theoretically optimal reconciliation of contemporaneous hierarchical time series forecasts. In Modeling and Stochastic Learning for Forecasting in High Dimensions, pp. 297–317. Springer.
- Wickramasuriya, S. L., G. Athanasopoulos, and R. J. Hyndman (2019). Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. Journal of the American Statistical Association 114(526), 804–819.



[Click here to access/download](#)

LaTeX Source Files
all_files.zip



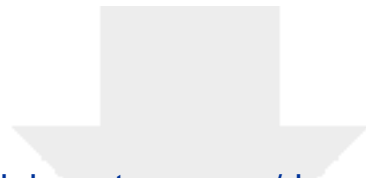
Dear Ruud

We have now addressed the following technical problem.

1. Please remove the PDF version of manuscript and upload the editable source file (TEX or DOC).

[We have now done that.](#)

Cheers,
George



[Click here to access/download](#)

Supplementary Material

ProbabilisticReconciliationR1-Appendices.pdf

