**6.4 Score results for Gaussian probabilistic forecasts**

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.

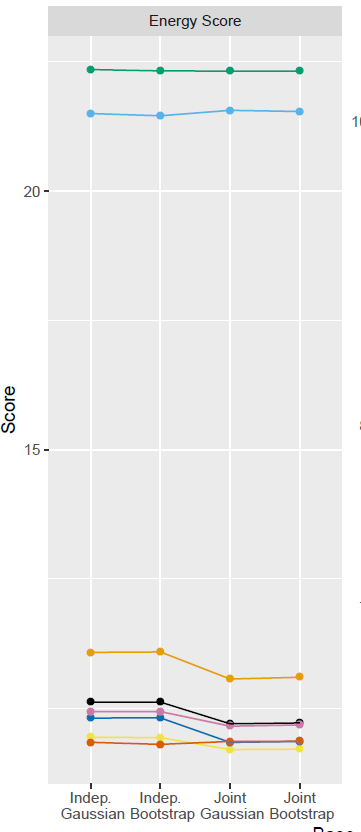
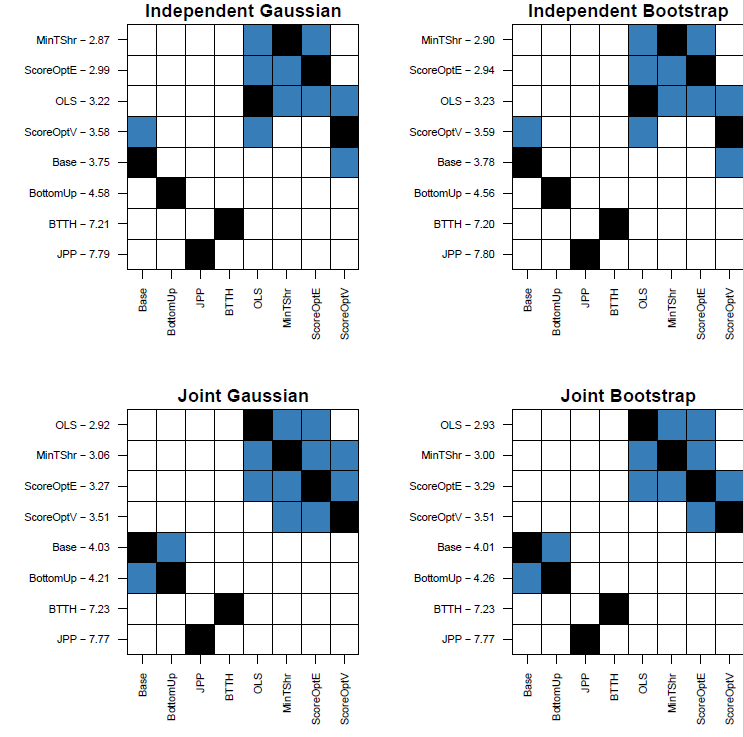


Figure 3: *Mean Energy scores for Gaussian DGP using different base forecast and reconciliation methods.*

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 implementation of the tests available in the tsutils R package (Kourentzes, 2019). Figure 5 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.

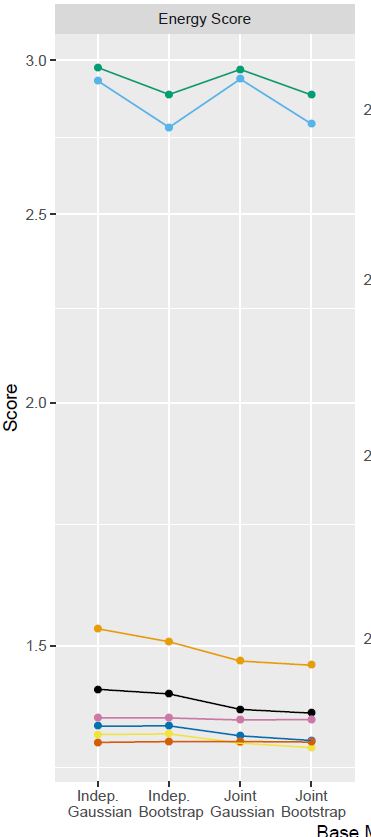


**Figure 5:** *Nemenyi matrix for Energy Score for Gaussian DGP.*

**6.5 Results for non-Gaussian probabilistic forecasts**

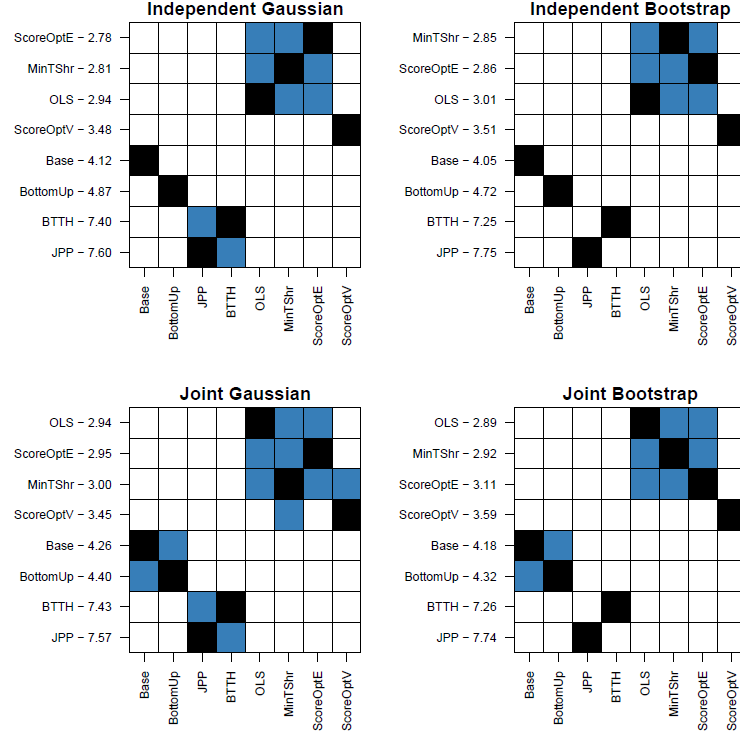
Figure 4 reports the mean energy score for the non-Gaussian DGP. Overall, the results are quite similar to the Gaussian DGP. The best performing reconciliation method is ScoreOptE when base probabilistic forecasts are independent, and MinTShr when base forecasts are dependent. The Nemenyi matrix is presented in Figure 6. These lead to similar conclusion to Figure 5. 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.

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 XXX.



**Figure 4:** *Mean scores for non-Gaussian DGP using different base forecast and reconciliation*

*methods. Left panel is energy score, right panel is variogram score.*



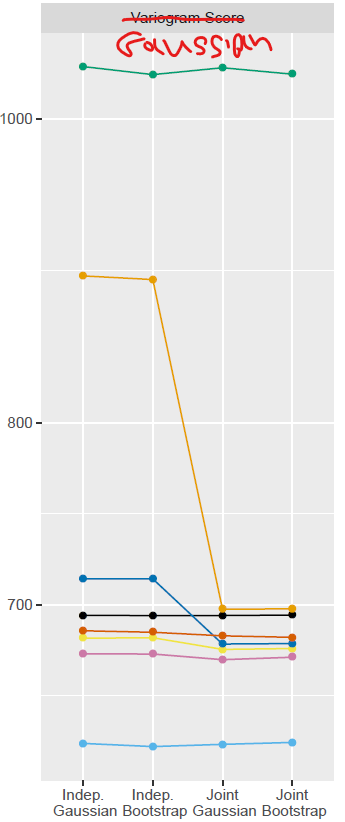
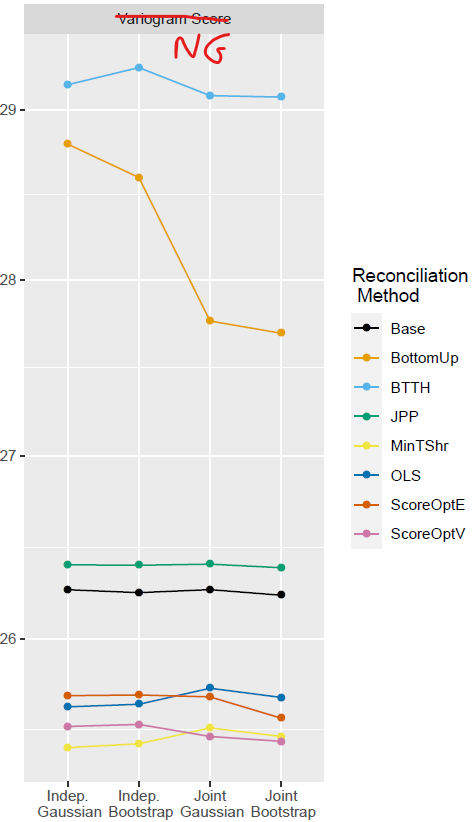
**Figure 6:** *Nemenyi matrix for Energy Score for non-Gaussian DGP.*

**Appendix – Simulation results for Variogram score**

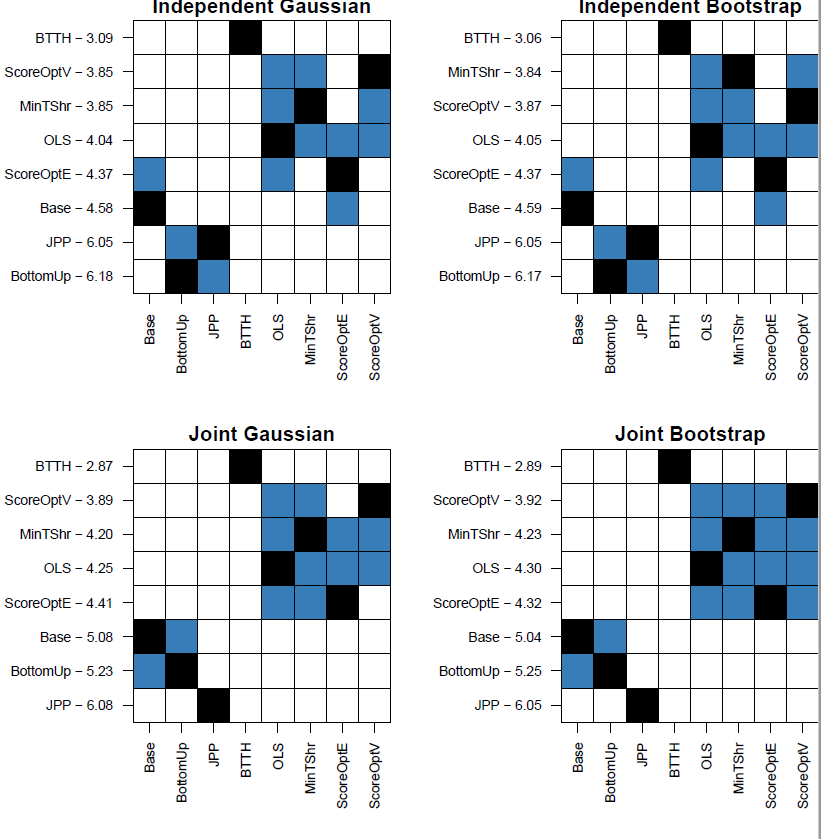
Figure XX shows the mean variogram score for different reconciliation methods and different methods of generating base forecasts. The results on the left panel are for a Gaussian DGP while the results on the right panel are for a non-Gaussian DGP.

For this specific Gaussian DGP, base model and score, BTTH significantly outperforms all other methods. However, this result was not observed when using BTTH for any other simulation scenario, including those discussed earlier in the paper as well as the results for the non-Gaussian DGP shown in the right panel. Excluding this result, score optimisation with respect to the variogram score is the best performing method with MinTShr and OLS also performing well. Score optimisation, OLS, MinTShr and BTTH all lead to significant improvements relative to base, bottom-up and JPP as shown in Figure Nemenyi.

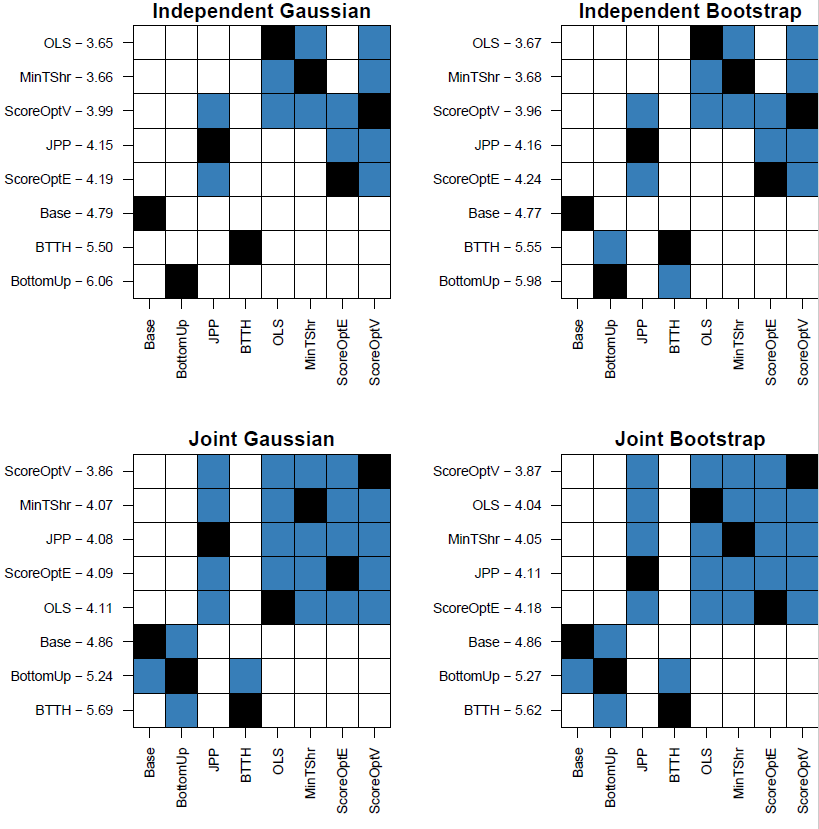
For the non-Gaussian DGP score optimisation with respect to the variogram score yields the best performance when base forecasts are dependent, while MinTShr yields the best performance when base forecasts are independent. In contrast to the Gaussian DGP, the JPP method leads to significant improvements over base forecasts, while the BTTH method leads to a significantly worse performance than base forecasts. The Nemenyi matrix is presented in Figure 10.

**Figure XX:** *Variogram**mean scores for Gaussian (left panel) and non-Gaussian (right panel) DGPs using different base forecast and reconciliation methods.*



**Figure 8:** *Nemenyi matrix for Variogram score with a Gaussian DGP.*



**Figure 9:** *Nemenyi matrix for Variogram score with a non-Gaussian DGP.*