A different view of hierarchical forecasting and connections with known modelling problem classes

Nikolaos Kourentzes

Professor of Predictive Analytics

Lancaster University Management School, UK

October 4, 2019

LUMS

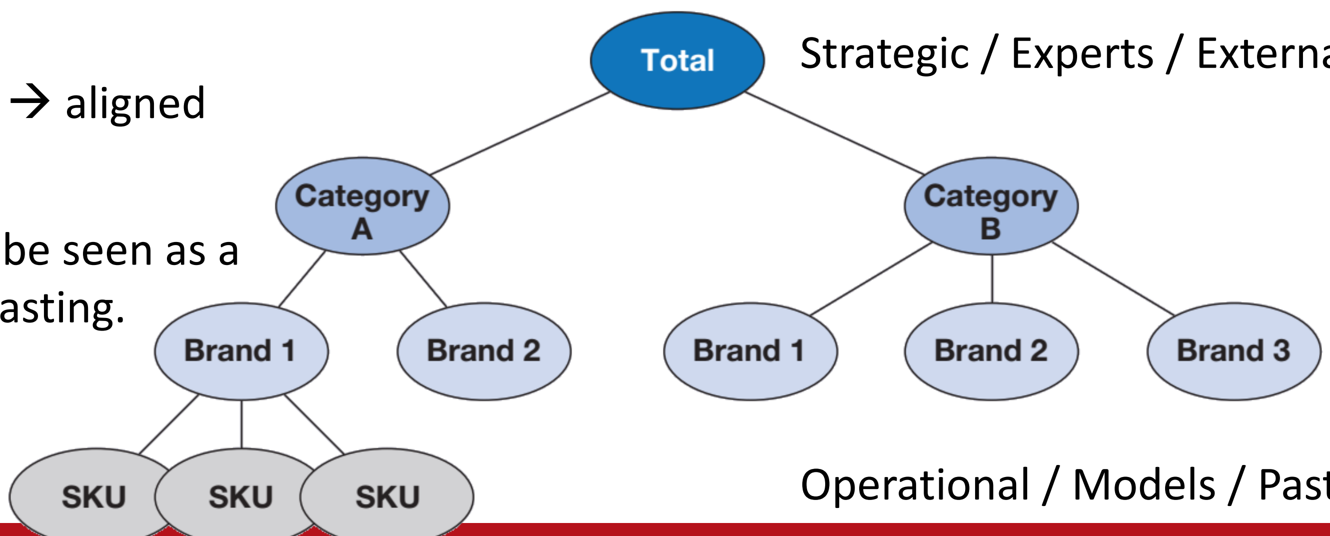
Marketing Analytics and Forecasting

A classic business problem

- Companies rely on forecasts to support decision making at different levels and functions.

Level	Horizon	Scope	Forecasts	Methods	Information
Operational	Short	Local	Way too many	Statistical	Univariate/Hard
Tactical	Medium	Regional	↕	↕	↕
Strategic	Long	Global	Few expensive	Experts	Multivariate/Soft

- The challenge: Forecasts must be aligned.
- Aligned forecasts → aligned decisions.
- The problem can be seen as a hierarchical forecasting.

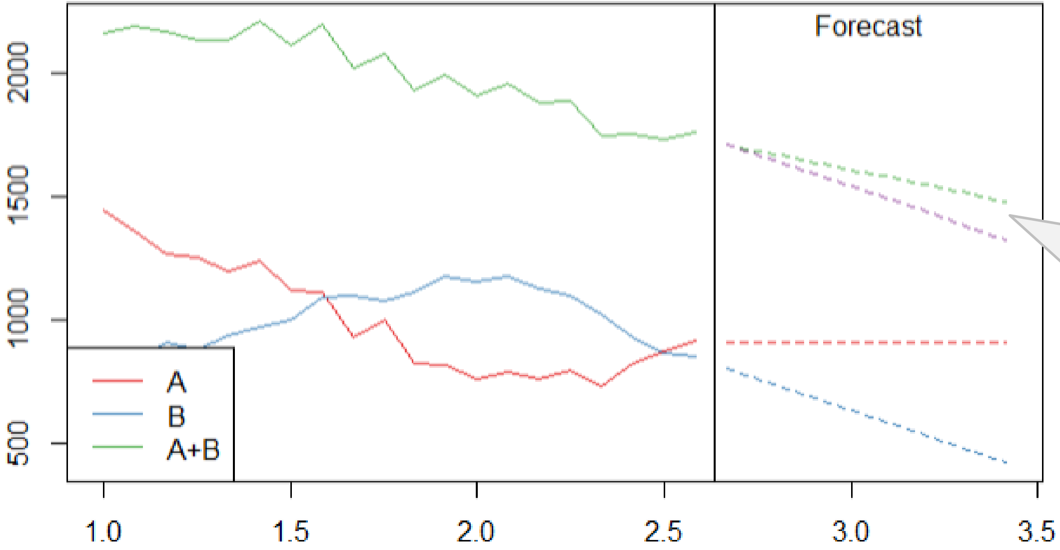


Strategic / Experts / External Info

Operational / Models / Past sales

Coherent forecasts

- As we aggregate data, some structures become more prominent (trends, seasonality), while others become less obvious (promotional activity) and noise is filtered.
- Although all series are based on the same information, this does not mean that the same information is useable → different models/parameters/forecasts.
- Example: forecasting A and B separately or forecasting their sum does not lead to the same result!



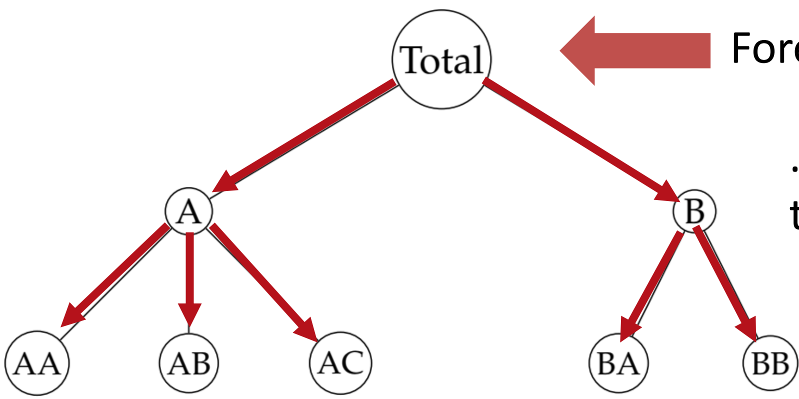
F(A+B) and F(A)+F(B) will typically be different, we need to impose equality (coherency of forecasts).

F(A+B) or F(A)+F(B) is correct? Coherency avoids this question

Hierarchical Forecasting

The two principal approaches to achieve hierarchical forecasts have been the **Top-Down** and **Bottom-Up** (Ord et al., 2017).

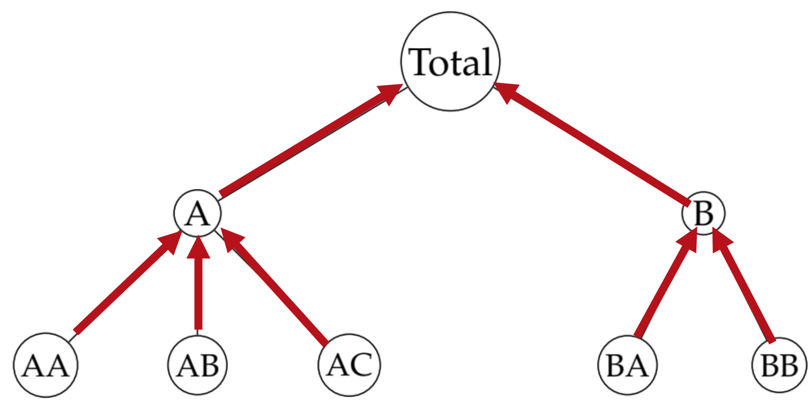
Top-Down



Forecast here!

... and disaggregate to lower levels

Bottom-Up



... and aggregate to higher levels

Forecast here!

Optimal Combinations

The recently proposed **optimal combinations** recast the hierarchical problem in the following way – we recast the problem as a forecast reconciliation model (Hyndman et al., 2011).

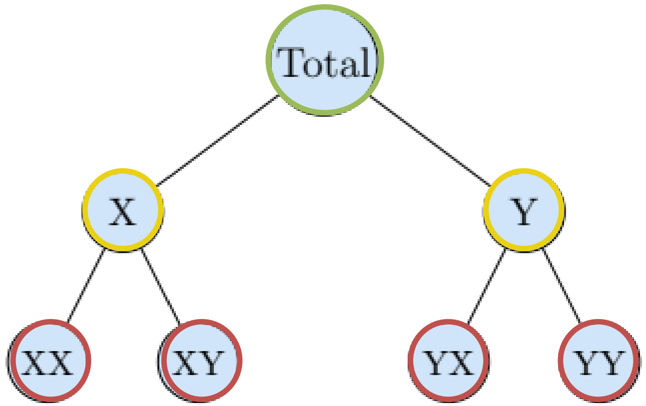
- $\boldsymbol{b} = (y_{xx}, y_{xy}, y_{yx}, y_{yy})'$ Lower level series
- $\boldsymbol{y} = (y_{tot}, y_x, y_y, \boldsymbol{b}')'$ All series
- $\boldsymbol{y} = \boldsymbol{S}\boldsymbol{b}$ Mapping of lower to all

$$\boldsymbol{S} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ & \boldsymbol{I}_m & & \end{bmatrix}$$

Top level

Middle level(s)

Bottom level



- $\hat{\boldsymbol{y}}_h$ h-step ahead forecasts for $\boldsymbol{y}$, i.e. all series.

Then we can write: $\tilde{\boldsymbol{y}}_h = \boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{y}}_h$

- where $\boldsymbol{G}$ projects (somehow!) linearly the forecasts to the lowest levels, so as to minimise $\tilde{\boldsymbol{y}}_h$- $\hat{\boldsymbol{y}}_h$, and $\tilde{\boldsymbol{y}}_h$ the coherent forecasts.

Optimal Combinations

The conventional **top-down** and **bottom-up** can be written as $\tilde{\boldsymbol{y}}_h = \boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{y}}_h$, and in these cases $\boldsymbol{G}$ uses a single level, ignoring all other information available.

With optimal combinations:

- $\boldsymbol{G} = \left(\boldsymbol{S}'\boldsymbol{W}_h^{-1}\boldsymbol{S}\right)^{-1}\boldsymbol{S}'\boldsymbol{W}_h^{-1}$, where $\boldsymbol{W}_h^{-1}$ is the variance-covariance matrix of h-step ahead errors.
- Given some forecasts $\hat{\boldsymbol{y}}_h$, obtained in any way, the only unknown to achieve coherent forecasts is $\boldsymbol{W}_h$.
- In principle, it should be the variance-covariance matrix of the reconciliation errors, but that poses a “chicken and the egg problem”. Wickramasuriya et al. (2019) showed that we can use the forecast errors instead.

Optimal Combinations

The exact estimation of $\boldsymbol{W}_h$ is problematic:

- Obtaining h-step ahead forecast errors is computationally demanding and at times, depending on the available sample and forecasting approach, not feasible.
- The dimension of $\boldsymbol{W}_h$ can easily become very large, causing estimation problems.

We typically assume that $\boldsymbol{W}_h = k\boldsymbol{W}_1$, i.e. that it is proportional to the 1-step ahead errors of the variance-covariance matrix.

- Estimation of $\boldsymbol{W}_1^{-1}$ is still non-trivial, but there are several approximation methodologies that perform well.

Optimal Combinations

Some of the more successful attempts (Athanasopoulos et al., 2017; Wickramasuriya et al., 2019):

- Assume homoscedasticity across everything: $W_{OLS} = I_m$
- Assume proportional increase in variance: Structural scaling.
- Assume no cross-effects: Variance scaling.
- Let the data speak using the full covariance matrix, with shrunk off-diagonals.

Structural scaling

$$\begin{bmatrix} 4 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Variance scaling

$$\begin{bmatrix} \hat{\sigma}^2_{Tot} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \hat{\sigma}^2_{X} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \hat{\sigma}^2_{Y} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \hat{\sigma}^2_{XX} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \hat{\sigma}^2_{XY} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \hat{\sigma}^2_{YX} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \hat{\sigma}^2_{YY} \end{bmatrix}$$

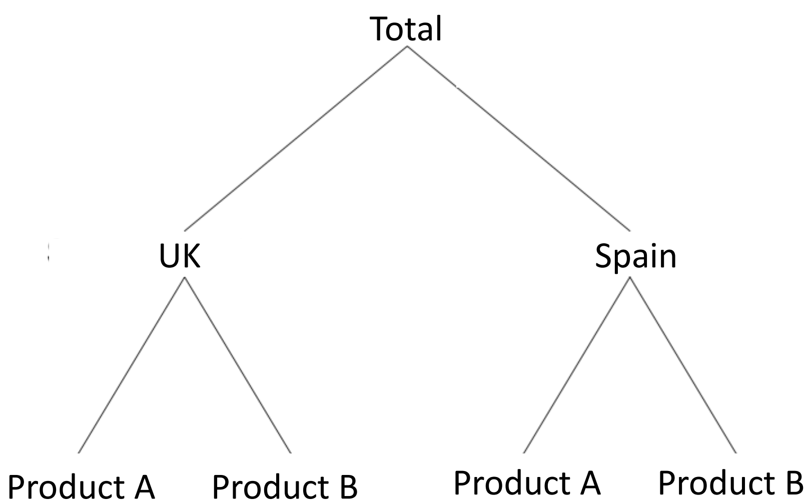
MinT shrinkage ($\rho_{i,j} \to 0$)

$$\begin{bmatrix} \hat{\sigma}^2_{Tot} & \hat{\rho}_{Tot,X} & \hat{\rho}_{Tot,Y} & \hat{\rho}_{Tot,XX} & \hat{\rho}_{Tot,XY} & \hat{\rho}_{Tot,YX} & \hat{\rho}_{Tot,YY} \\ \hat{\rho}_{X,Tot} & \hat{\sigma}^2_{X} & \hat{\rho}_{X,Y} & \hat{\rho}_{X,XX} & \hat{\rho}_{X,XY} & \hat{\rho}_{X,YX} & \hat{\rho}_{X,YY} \\ \hat{\rho}_{Y,Tot} & \hat{\rho}_{Y,X} & \hat{\sigma}^2_{Y} & \hat{\rho}_{Y,XX} & \hat{\rho}_{Y,XY} & \hat{\rho}_{Y,YX} & \hat{\rho}_{Y,YY} \\ \hat{\rho}_{XX,Tot} & \hat{\rho}_{XX,X} & \hat{\rho}_{XX,Y} & \hat{\sigma}^2_{XX} & \hat{\rho}_{XX,XY} & \hat{\rho}_{XX,YX} & \hat{\rho}_{XX,YY} \\ \hat{\rho}_{XY,Tot} & \hat{\rho}_{XY,X} & \hat{\rho}_{XY,Y} & \hat{\rho}_{XY,XX} & \hat{\sigma}^2_{XY} & \hat{\rho}_{XY,YX} & \hat{\rho}_{XY,YY} \\ \hat{\rho}_{YX,Tot} & \hat{\rho}_{YX,X} & \hat{\rho}_{YX,Y} & \hat{\rho}_{YX,XX} & \hat{\rho}_{YX,XY} & \hat{\sigma}^2_{YX} & \hat{\rho}_{YX,YY} \\ \hat{\rho}_{YY,Tot} & \hat{\rho}_{YY,X} & \hat{\rho}_{YY,Y} & \hat{\rho}_{YY,XX} & \hat{\rho}_{YY,XY} & \hat{\rho}_{YY,YX} & \hat{\sigma}^2_{YY} \end{bmatrix}$$

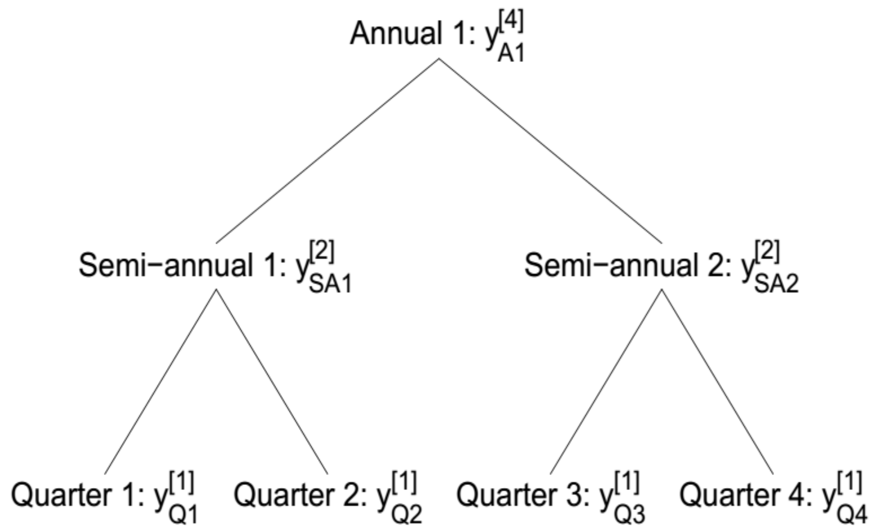
Temporal Hierarchies

Kourentzes et al. (2014) and Athanasopoulos et al. (2017) proposed the temporal analogue to hierarchical forecasting. The objective now is to join short-term and long-term forecasting.

Cross-sectional hierarchy



Temporal hierarchy



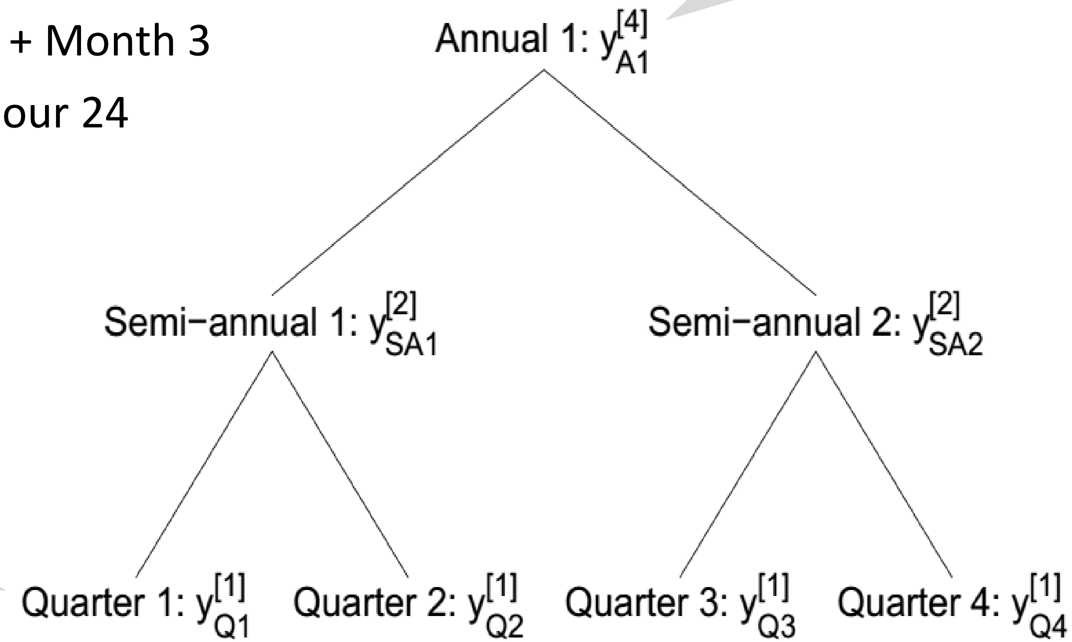
Temporal Hierarchies

Operational planning is done at detailed daily/weekly/monthly series, while tactical planning is done at monthly/quarterly/yearly series.

These series are naturally connected, for example:

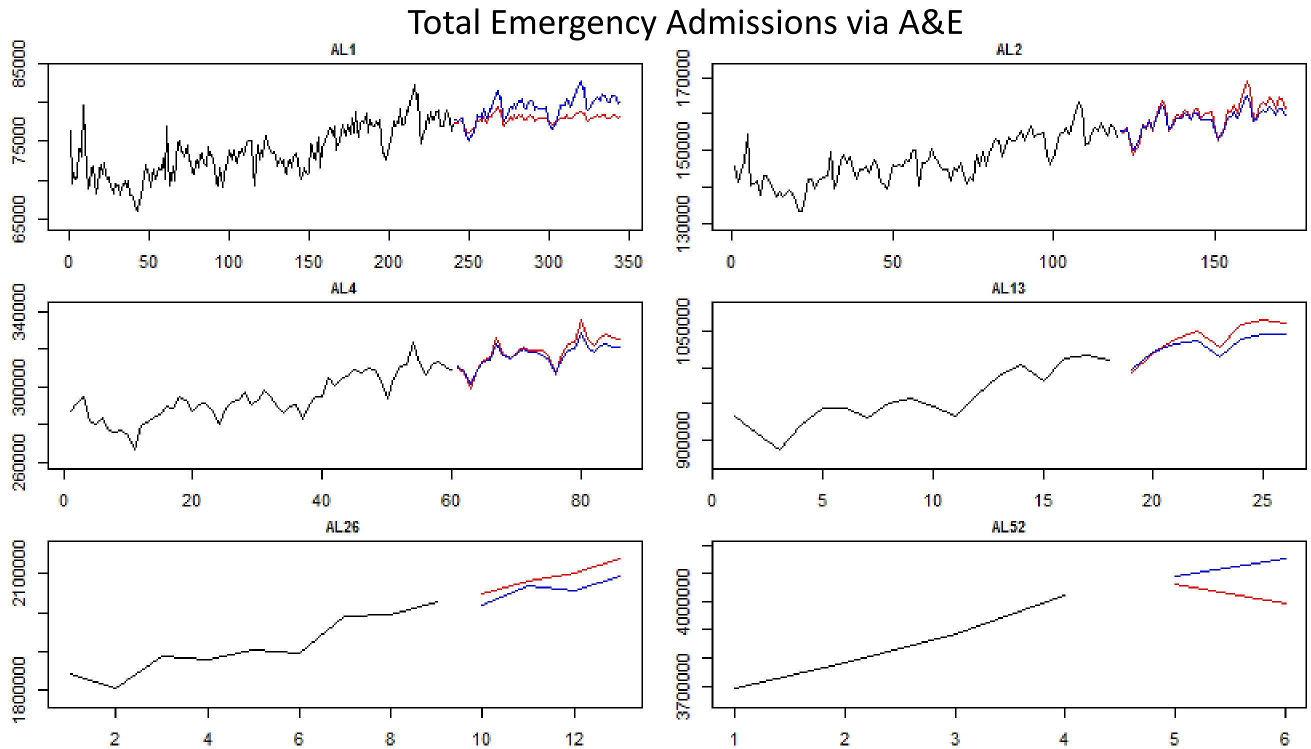
- Year 1 = Quarter 1 + Quarter 2 + Quarter 3 + Quarter 4
- Quarter 1 = Month 1 + Month 2 + Month 3
- Day 1 = Hour 1 + Hour 2 + … + Hour 24
- etc.

Aggregate external information: e.g. macroeconomic



Disaggregate internal information: e.g. promotions

Example: Predicting A&E admissions



Red is the prediction of the base model – at each level separately
Blue is the temporal hierarchy forecasts

Observe how information is `borrowed' between temporal levels. Base models for instance provide very poor weekly and annual forecasts

Example: Predicting A&E admissions

Collect weekly data for UK A&E wards.
13 time series: covering different types of emergencies and different severities (measured as time to treatment)
Span from week 45 2010 (7^{th} Nov 2010) to week 24 2015 (7^{th} June 2015)

Accurately predict to support staffing and training decisions.
Aligning the short and long term forecasts is important for consistency of planning and budgeting.

- Test set: 52 weeks.
- Rolling origin evaluation.
- Forecast horizons of interest: t+1, t+4, t+52 (1 week, 1 month, 1 year).
- ARIMA as base forecasting model.
- Evaluation on MASE (Mean Absolute Scaled Error).

Example: Predicting A&E admissions

Aggr. Level	h	Base	Reconciled	Change
Weekly	1	1.6	1.3	-17.2%
Weekly	4	1.9	1.5	-18.6%
Weekly	13	2.3	1.9	-16.2%
Weekly	1-52	2.0	1.9	-5.0%
Annual	1	3.4	1.9	-42.9%

Red is the prediction of the base model – at each level separately
Blue is the temporal hierarchy forecasts

- Accuracy gains at all planning horizons
- Crucially, forecasts are reconciled leading to aligned plans

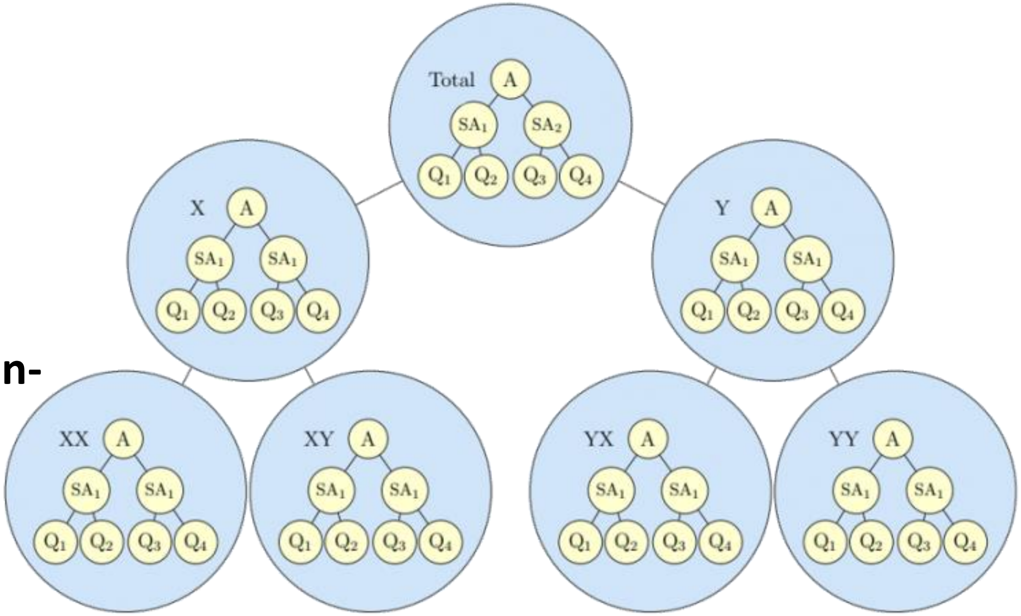
Cross-Temporal Hierarchies

The two sides of hierarchical forecasting have their limitations as well:

- Cross-sectional: we forecast a SKU at individual (lowest) and total sales at top level, i.e. do we need total sales at short term forecasting, which is fine for SKU level?
- Temporal: we forecast a SKU at short-term and long-term, i.e. sales of X at size Y for short-term (operational) and long-term (strategic) horizons. What decision would long-term forecasting of sales of individual SKU serve?

What we need is to combine both using **cross-temporal hierarchies**. The same formulation applies.

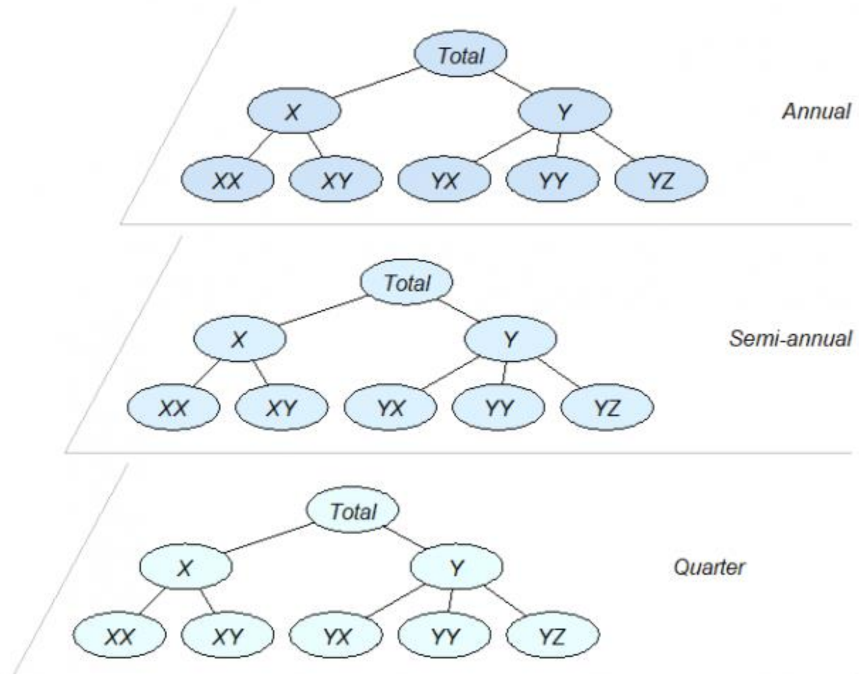
- But now constructing $\boldsymbol{S}$ and $\boldsymbol{G}$ is non-trivial and computationally demanding, because of their **dimensionality** and **non-unique mapping**.



Cross-Temporal Hierarchies

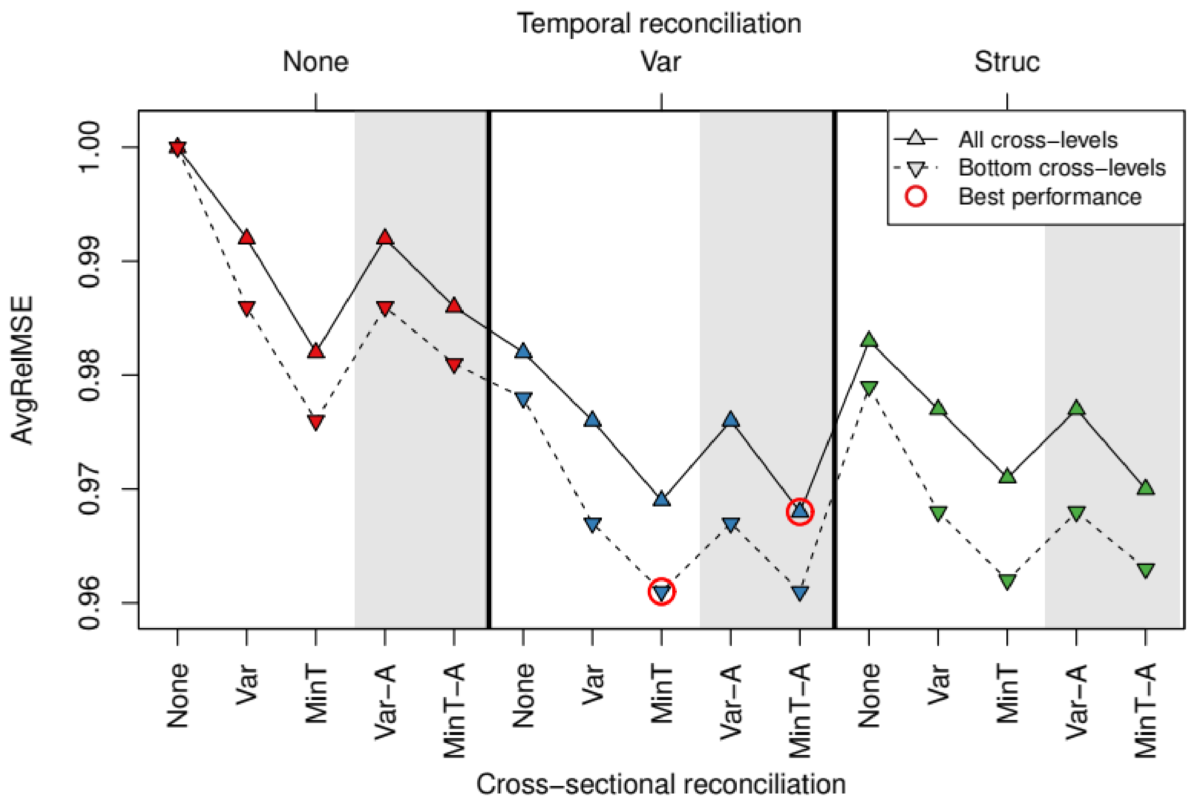
Kourentzes & Athanasopoulos (2019) proposed a methodology to split the estimation, reducing the size of the problem and still achieving cross-temporally coherent forecasts.

1. Produce temporal hierarchy forecasts for all time series in the cross-sectional hierarchy → **coherent in the time dimension**.
2. Estimate $\boldsymbol{G}$ at all temporal levels (as in figure!)
3. Calculate common $\overline{\boldsymbol{G}} = k^{-1} \sum_{=1}^{k} \boldsymbol{G}_k$, where k is the number of temporal levels.
4. Reconcile using $\widetilde{\boldsymbol{y}}_h = \boldsymbol{S}\overline{\boldsymbol{G}}\widehat{\boldsymbol{y}}_h$.



Empirical evaluation

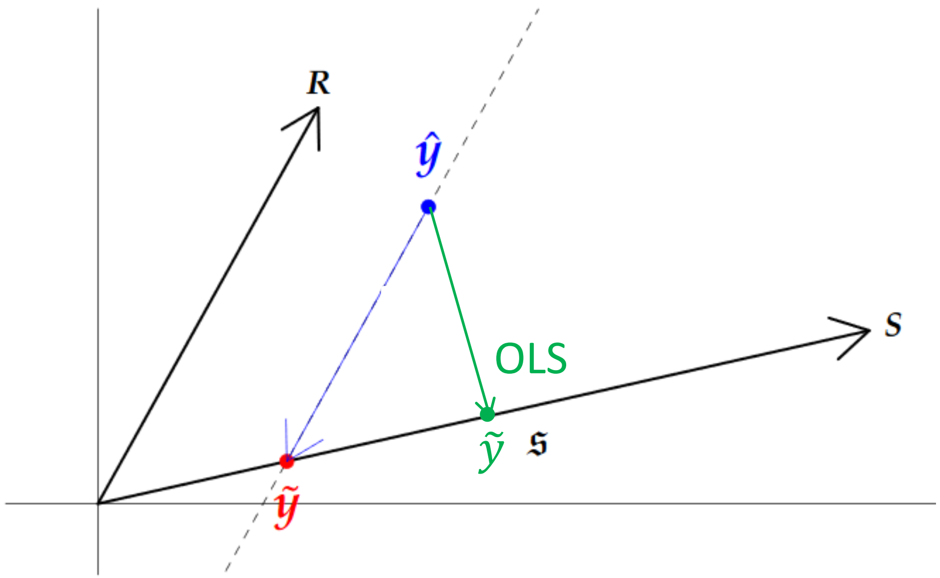
- Total to regional monthly tourism flows for Australia. 111 series, spanning 10 years.
- Test set 6 years, with rolling origin evaluation. Relative RMSE (<1 better) to base forecast.
- Forecast using exponential smoothing. Results with ARIMA similar.



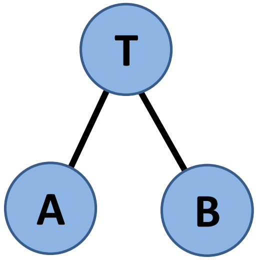
Figures in grey are cross-temporally coherent

A geometric view of the problem

- Gamakumara et al., 2018 and Athanasopoulos et al., 2019 propose an elegant geometric interpretation of hierarchical forecasting

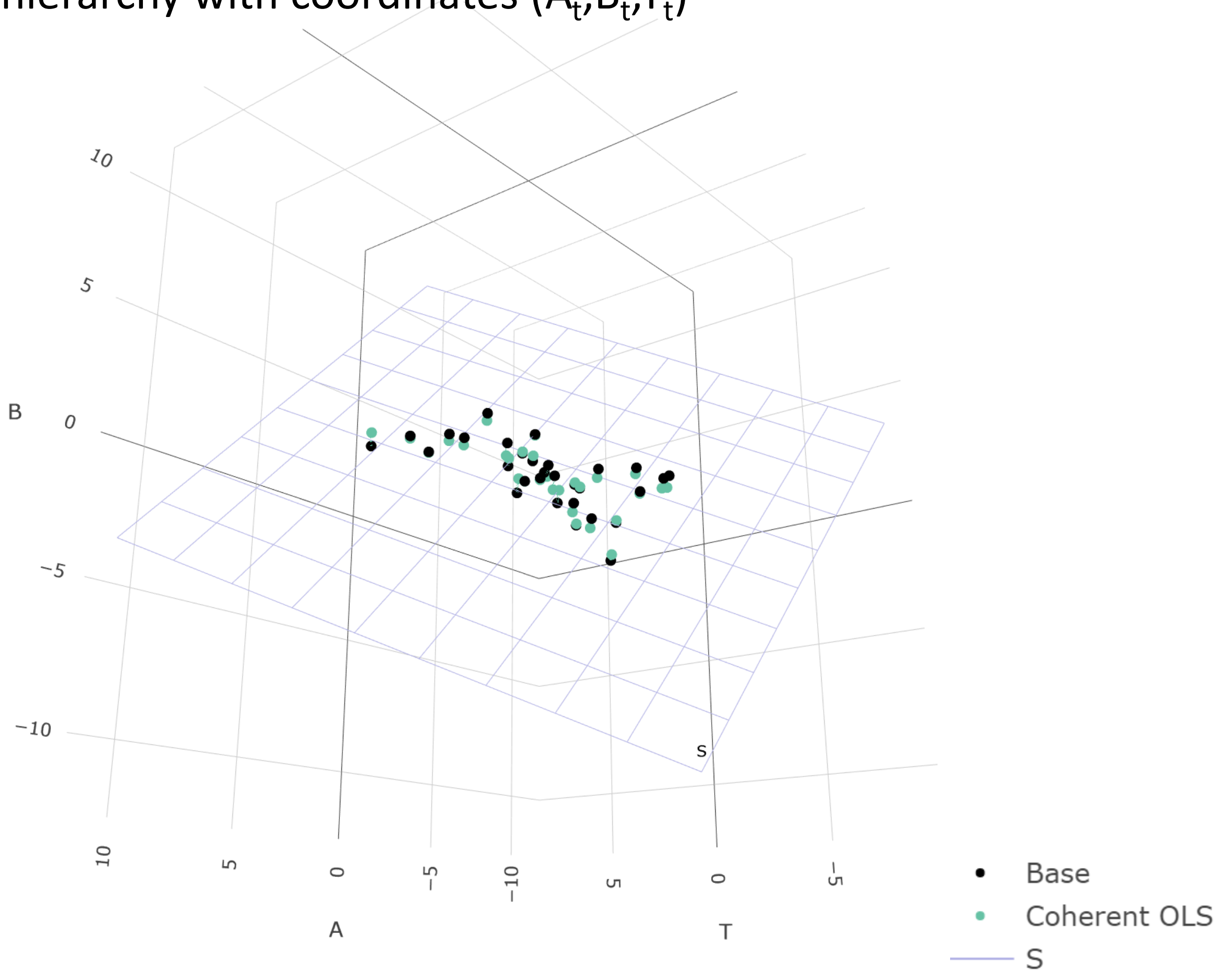


- … which I still don't fully get, so we started exploring further.
- This is work in progress, so we keep it simple and use this hierarchy
  - Number of series: m = 2 (lowest level), n = 3 (all levels)



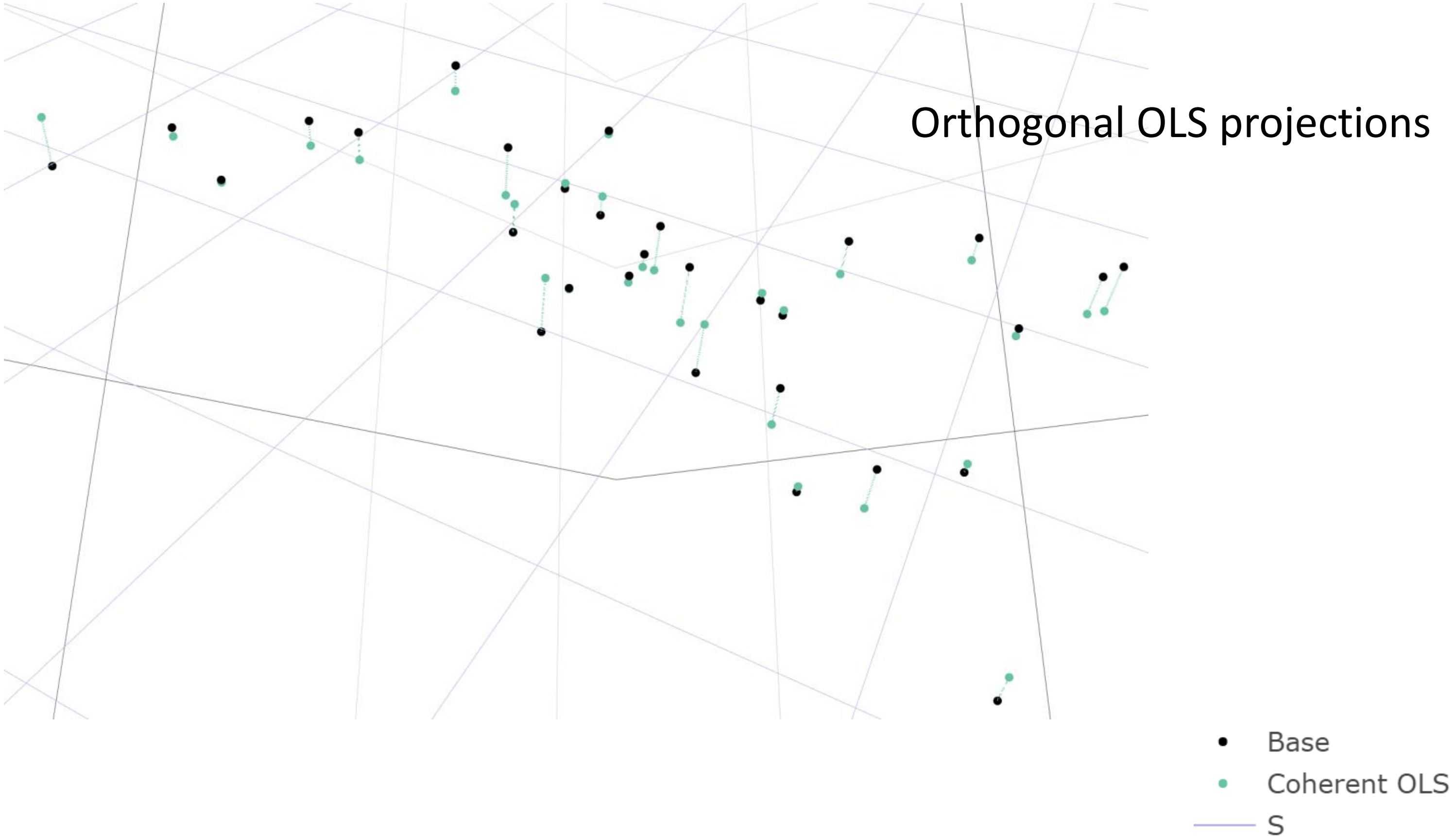
A geometric view of the problem

Each point is a whole hierarchy with coordinates (A_t,B_t,T_t)



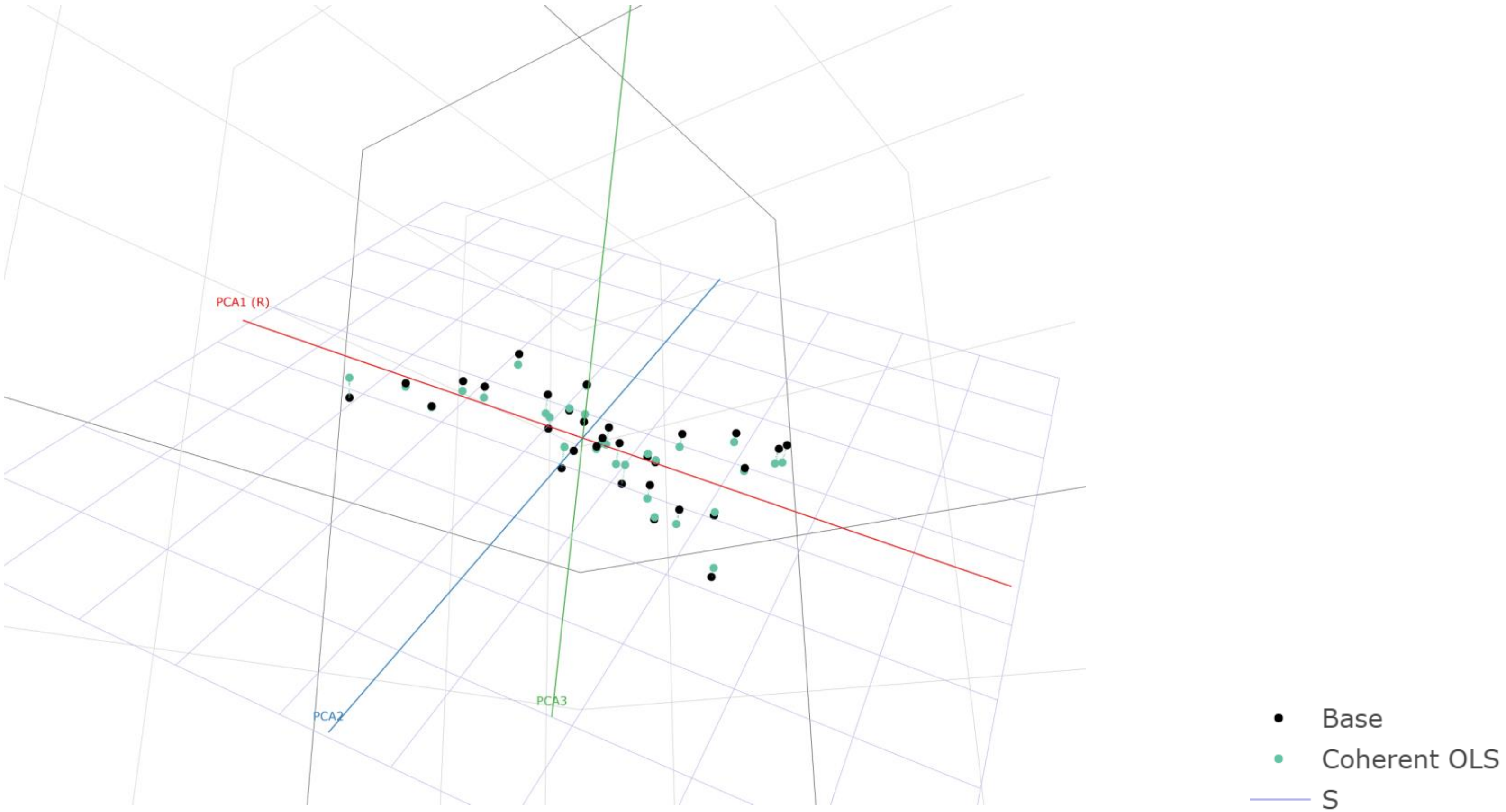
A geometric view of the problem

Each point is a whole hierarchy with coordinates (A_t, B_t, T_t)



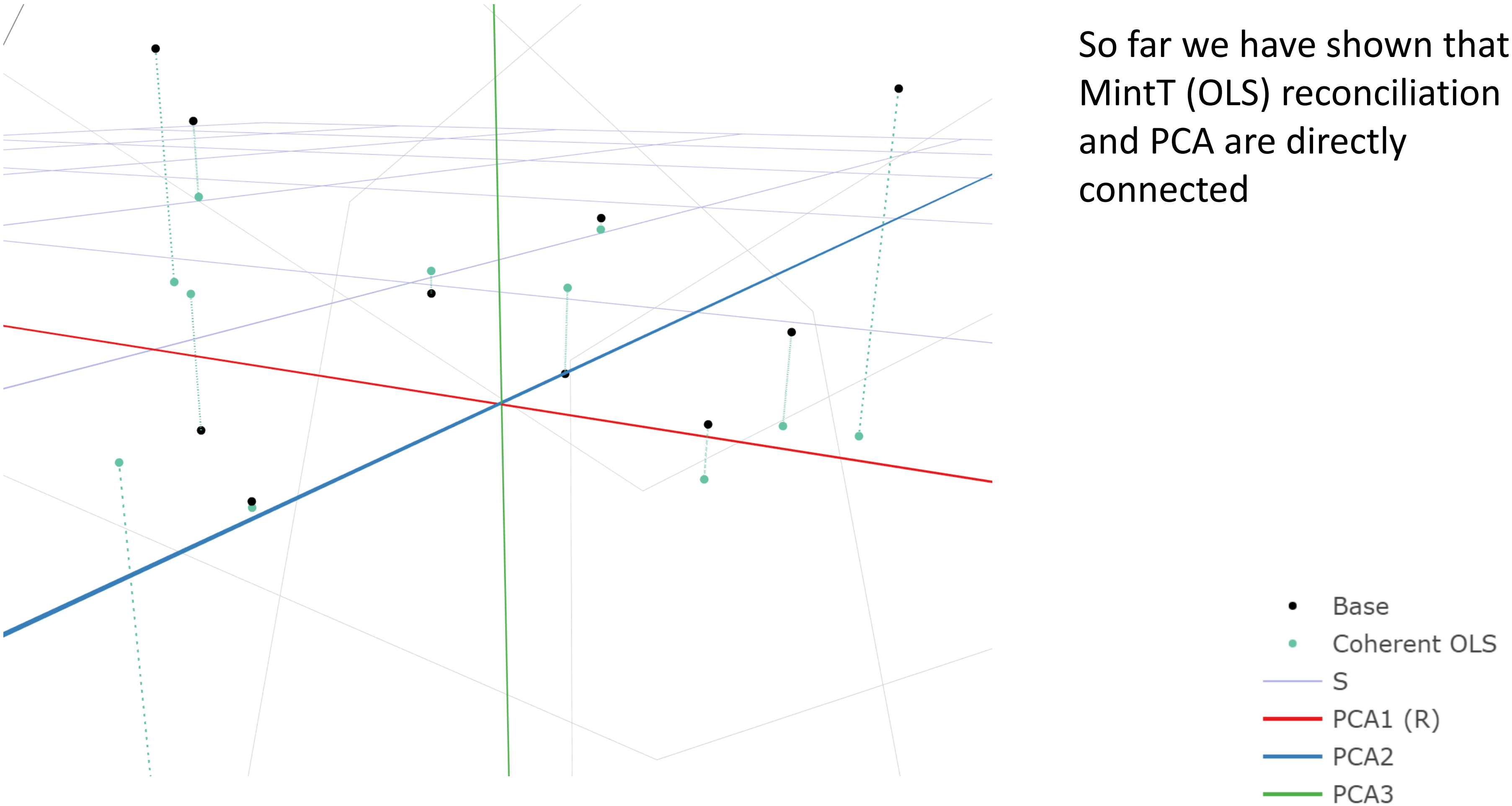
A geometric view of the problem

- The coherent space is defined by the highest variance principal components



A geometric view of the problem

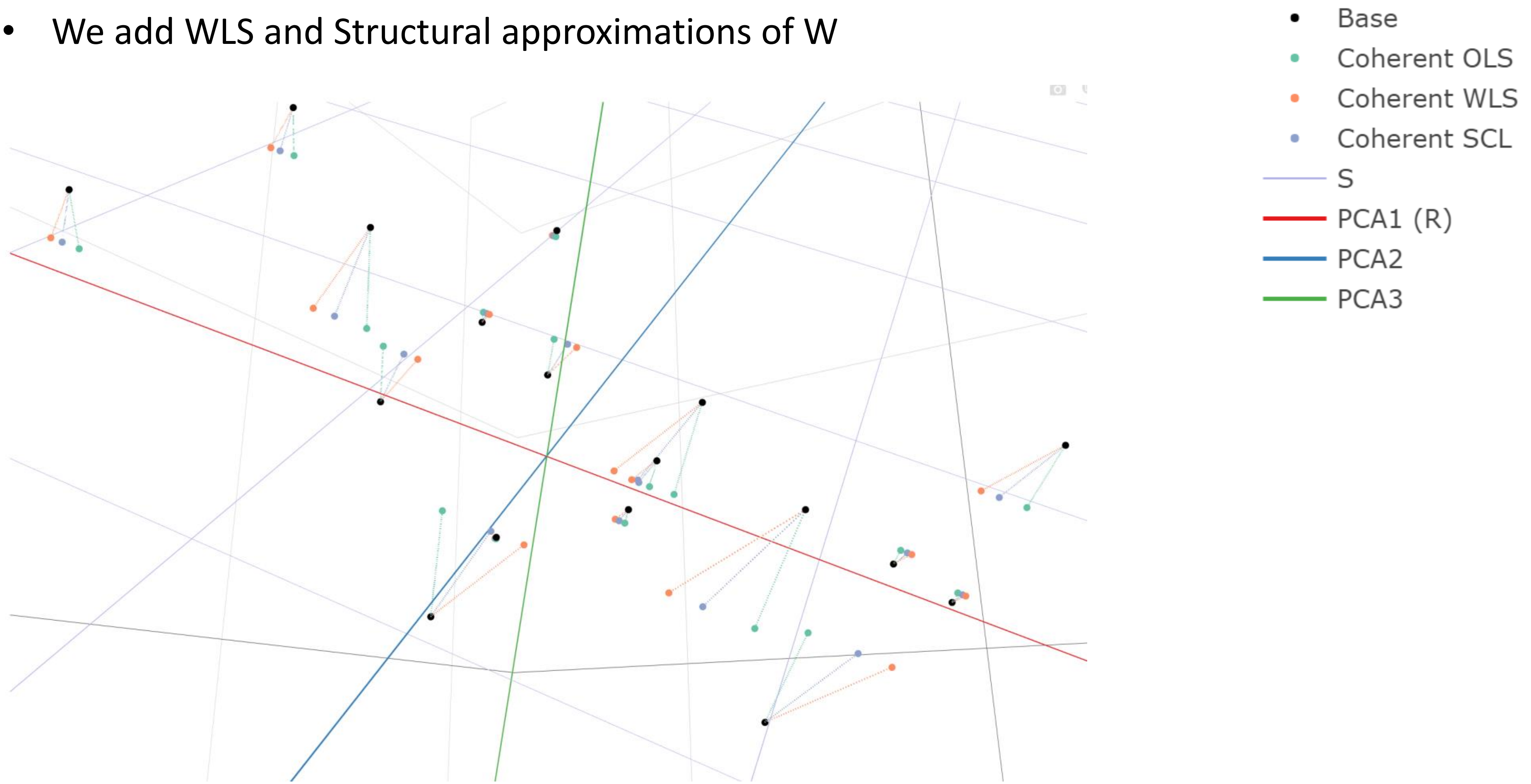
- The OLS projection is along the lowest variance components



So far we have shown that MintT (OLS) reconciliation and PCA are directly connected

A geometric view of the problem

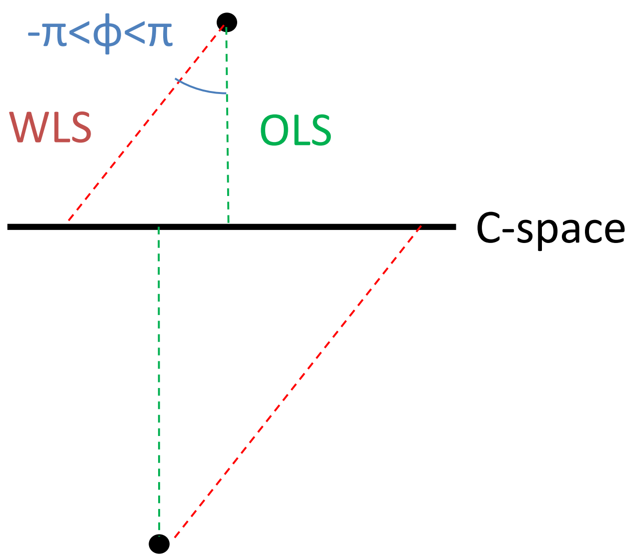
- We add WLS and Structural approximations of W

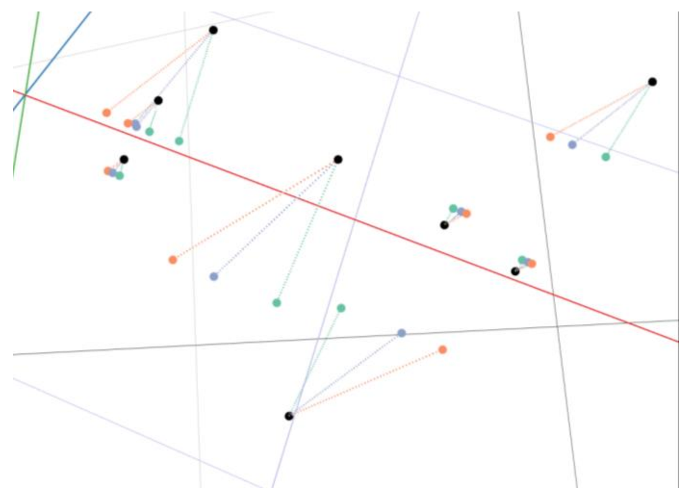


- It appears that all they do is all projections are parallel to PCA1
- If so, WLS is inefficient, as it is just OLS with a single rotation angle!

A geometric view of the problem

- Let us see what this means:
  - OLS is the orthogonal projection.
  - To get to WLS (or Structural) we just need a simple rotation.

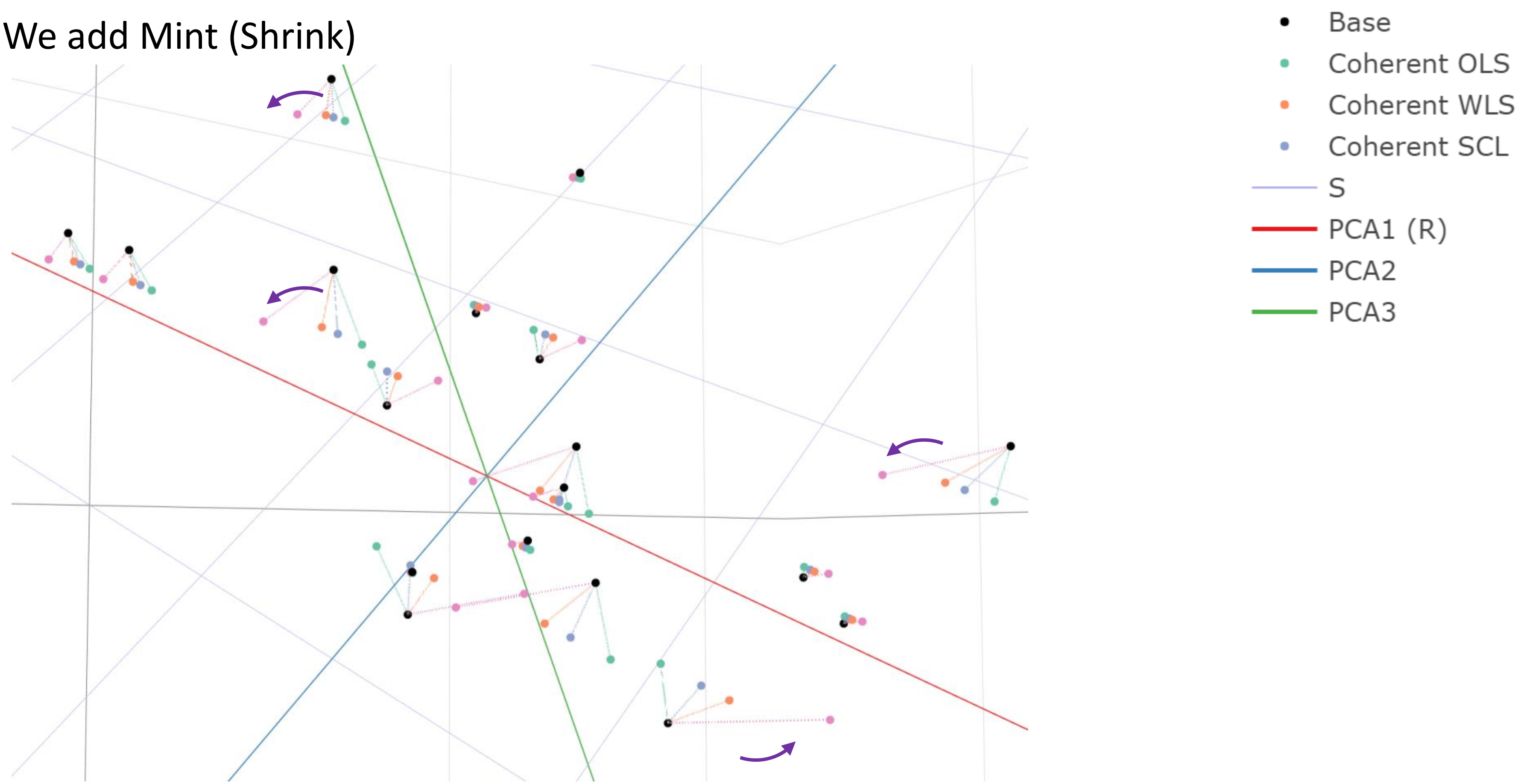




- If a point is over or under the C-space defines the direction of projection.
- Note that a rotation of the projection direction (rotation from PCA3) is a translation (+x,+y) in the C-space.
- Rotation requires 1 parameter, translation requires 2 parameters, WLS requires 3.

A geometric view of the problem

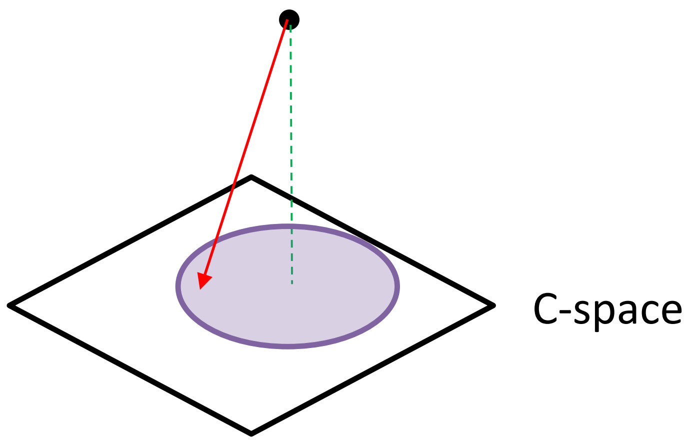
- We add Mint (Shrink)



- This is no longer parallel along the PCA1 line, the off-diagonals add rotation!
- Again the direction of rotation is defined by the norm of the C-space, itself defined by PCA3.

A geometric view of the problem

- So the general solution is:



This needs either

- Two rotation angles
- Or one rotation angle and a spin
- Or two translation directions
- … anyway, it is 2 parameters, not 3 (WLS) or 6 (Shrink) ← These are inefficient
- More generally we need m parameters (number of lowest level dimensions) and not estimating a covariance matrix

A geometric view of the problem

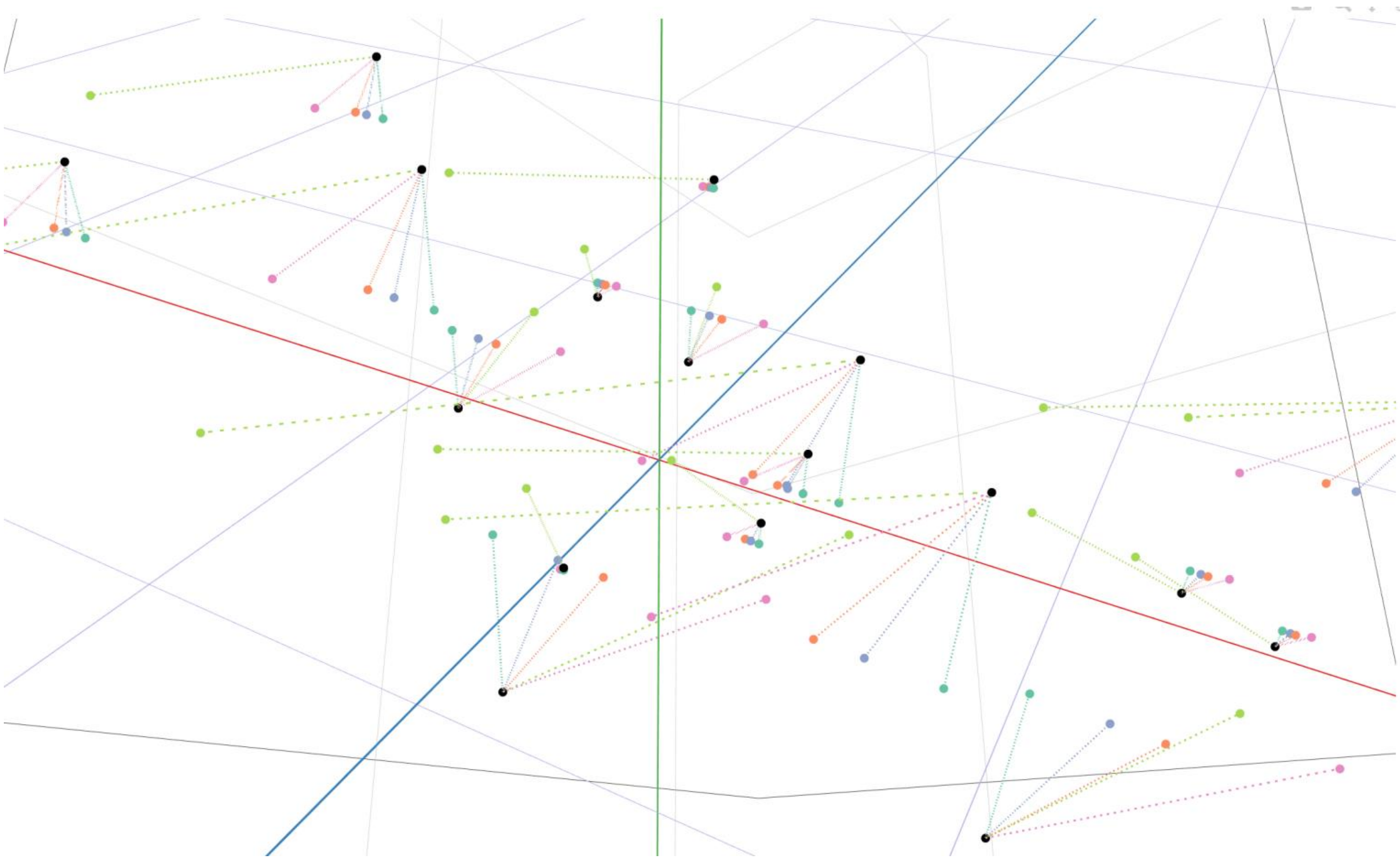
- There are three operators we can rely on:
  - Rotation
  - Translation
  - Scaling

- You may have already observed that a higher dimensional rotation is a lower dimensional translation in our problem.

- Here is the question: we are in 3D space, what does a 3D rotation mean for MintT?
  - More generally we can apply these operators on the C-space or on the B-space.
  - So far we only worked on the C-space.

A geometric view of the problem

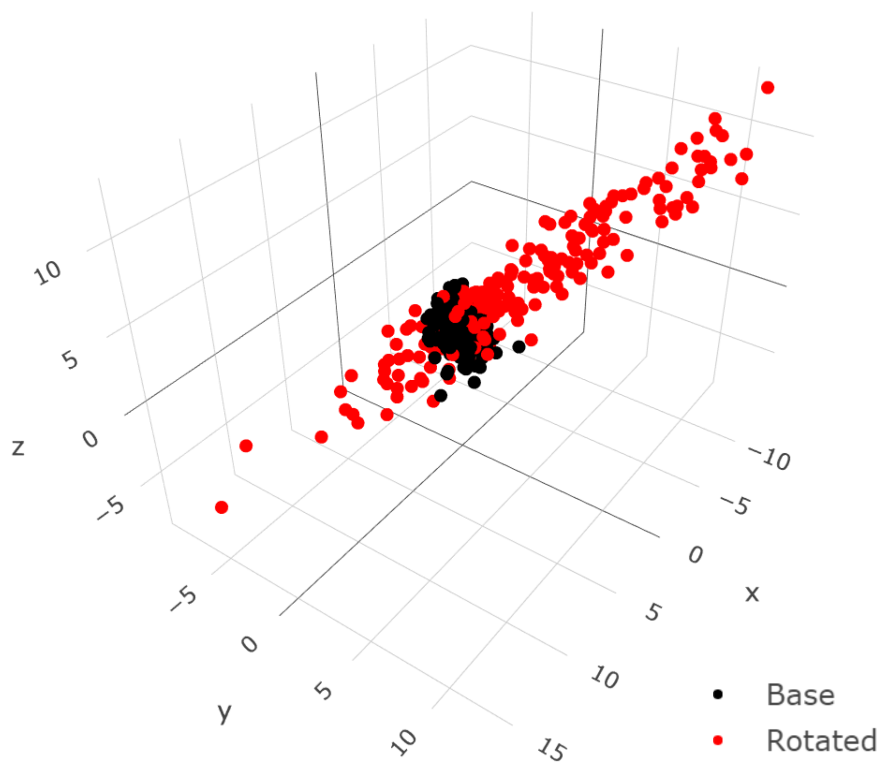
- We allow for operators in the B-space (the important one is rotation)



- We get the green lines… they are for each point different, why?

A geometric view of the problem

- We allow for operators in the B-space (the important one is rotation)



- A (n-parameter) B-space rotation with OLS projection is equivalent to a nonlinear reconciliation on the C-space.
- A (m-parameter) C-space rotation with OLS projection is equivalent to a linear reconciliation on the C-space. This case encompasses all existing approximations of W.

Some results

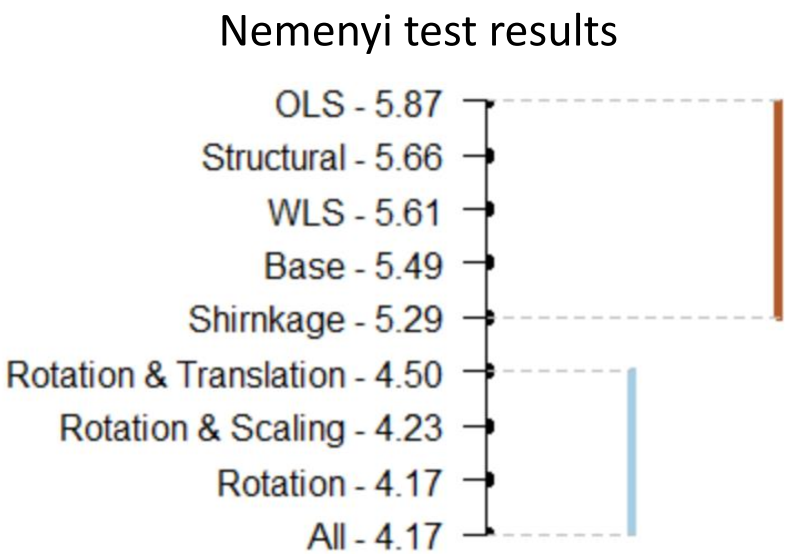
- Results over 500 simulated hierarchies.
- We start with appropriate base forecasts.
- AvgRelRMSE – The geometric mean of the ratio of RMSE over RMSE of base forecasts.

Reconciliation	AvgRelRMSE
Base	1
OLS	1.003
WLS	0.999
Structural	1.001
Shrinkage	0.994
Rotation	0.974
Rotation & Translation	0.979
Rotation & Scaling	0.973
All	0.976

Some results

- Results over 500 simulated hierarchies.
- AvgRelRMSE – The geometric mean of the ratio of RMSE over RMSE of base forecasts.

6 (n*2) parameters
6 (n*2) parameters
3 (n) parameters
9 (n*3) parameters



That’s not all!

- The connection with PCA makes it apparent that the hierarchical problem (with the three basic operators on B-space and C-space) is connected to a number of classes of modelling problems:
  - PCA → apparent
  - Multivariate models → The natural generalisation of MinT
  - Regression (needs some imagination, but essentially what you want is a multivariate system and focus only on one row)
  - Model combination (we know MinT does this, but we always see MinT as encompassed by model combination, but if I can do regression, I can do model combination)
- Why is this exciting? If we can show that:
  - We solve these inefficiently
  - A class of nonlinearities is simply a B-space rotation

  … then we have something fairly powerful.

An example of forecast combination

- We need to redefine the S matrix, but the rest works fine.

$$S = \begin{array}{|cc|l} 0.5 & 0.5 & \text{Some initial forecast (target)} \\ 1 & 0 & \text{Forecast A} \\ 0 & 1 & \text{Forecast B} \end{array}$$

- An example: use the Airline passengers time series and forecast t+12 using all ETS models.
  - Model select using AICc
  - Model combination using AICc
  - Hierarchical starting from model selection
  - Hierarchical starting from model combination

Method	Select AICc	Combine AICc	Hierarchical Select	Hierarchical Combine
RMSE	26.77	26.95	**24.66**	24.73

Conclusions

- MinT provides good hierarchical forecasting results. There is a cross-section and a temporal problem. The temporal provides higher accuracy gains.
- Cross-temporal hierarchy forecasts provide a single view of the future across market demarcations and planning horizons → “one number forecast”.
- The geometric interpretation of MinT reconciliation shows that the way we currently solve it is inefficient. We need less parameters in such a highly structured problem.
- Once seen as a rotation (or translation) then we can use these operators in either B-space or C-space. The former implies nonlinear projections.
- We conjecture that the hierarchical problem has equivalences with some of the fundamental time series modelling challenges:
  - If it holds it may be able to sow us ways to solve classic problems more efficiently/better.

Resources

Annals of Tourism Research 75 (2019) 393–409

ELSEVIER

Contents lists available at ScienceDirect

Annals of Tourism Research

journal homepage: www.elsevier.com/locate/annals

ANNALS

Cross-temporal coherent forecasts for Australian tourism

Nikolaos Kourentzes[a,*], George Athanasopoulos[b]

[a] Lancaster University Management School, Department of Management Science, Lancaster LA1 4YX, UK
[b] Department of Econometrics and Business Statistics, Monash University, Australia

Check for updates

ARTICLE INFO

Associate editor: Haiyan Song

Keywords:
Cross-sectional aggregation
Temporal aggregation
Forecast combinations
Spatial correlations

ABSTRACT

Key to ensuring a successful tourism sector is timely policy making and detailed planning. National policy formulation and strategic planning requires long-term forecasts at an aggregate level, while regional operational decisions require short-term forecasts, relevant to local tourism operators. For aligned decisions at all levels, supporting forecasts must be 'coherent', that is they should add up appropriately, across relevant demarcations (e.g., geographical divisions or market segments) and also across time. We propose an approach for generating coherent forecasts across both cross-sections and planning horizons for Australia. This results in significant improvements in forecast accuracy with substantial decision making benefits. Coherent forecasts help break intra- and inter-organisational information and planning silos, in a data driven fashion, blending information from different sources.

- References within the published paper.
- Useful R packages for cross-temporally coherent forecasts
  - thief – Temporal hierarchies;
  - hts – Cross-sectional hierarchies;
  - MAPA - alternative for temporally coherent forecasts.

Thank you for your attention! Questions?

Nikolaos Kourentzes

email: nikolaos@kourentzes.com
twitter @nkourentz
Blog: http://nikolaos.kourentzes.com

LUMS Marketing Analytics and Forecasting