A geometry inspired hierarchical forecasting methodology

Nikolaos Kourentzes
Skövde Artificial Intelligence Lab
Skövde University, Sweden

George Athanasopoulos
Department of Econometrics and Business Statistics, Monash University, Australia

Anastasios Panagiotelis
Discipline of Business Analytics, University of Sydney, Australia

International Symposium on Forecasting 2020

26/10/2020

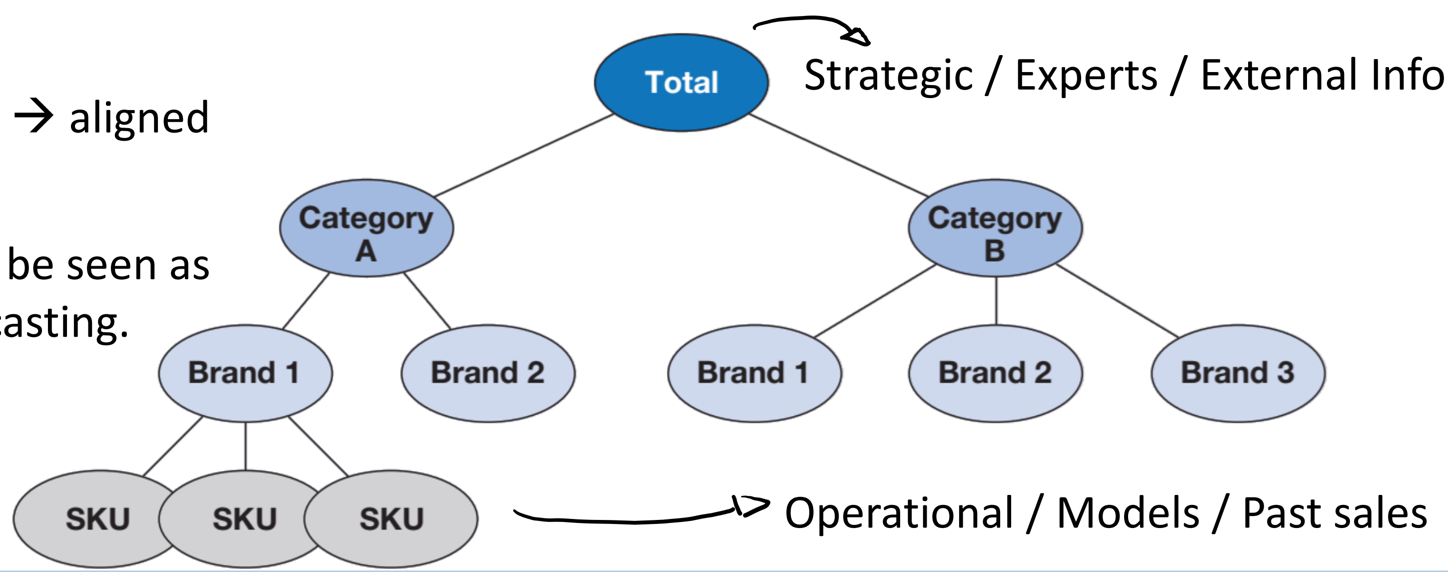
HÖGSKOLAN I SKÖVDE · MONASH University · THE UNIVERSITY OF SYDNEY

Hierarchical forecasting in a nutshell

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

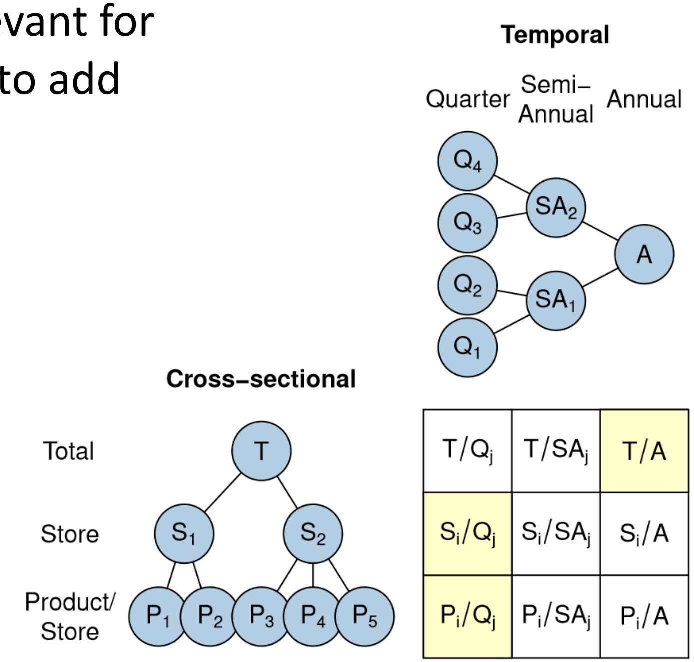
Level	Horizon	Scope	Forecasts	Methods	Information
Operational	Short	Local	Way too many	Statistical	Univariate/Hard
Tactical	Medium	Regional	↕	↕	↕
Strategic	Long	Global	Few expensive	Experts	Multivariate/Soft

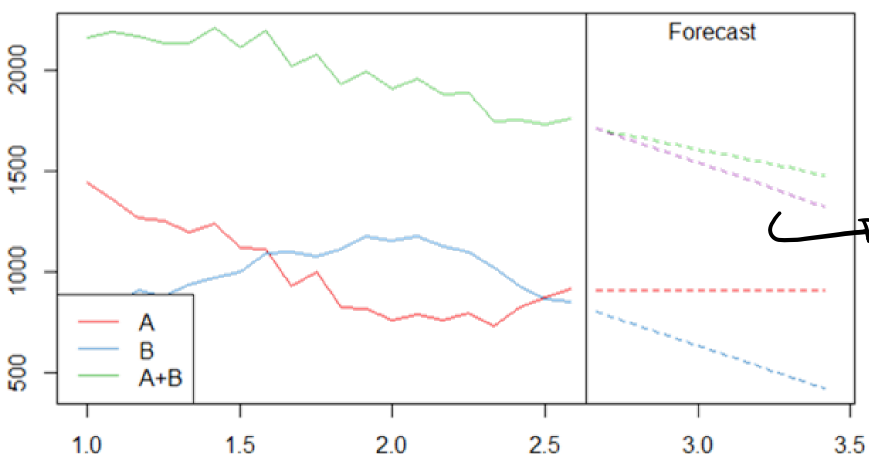
- The challenge: Forecasts must be aligned.
- Aligned forecasts → aligned decisions.
- The problem can be seen as hierarchical forecasting.



Hierarchical forecasting in a nutshell

- But not all forecasts or levels in the hierarchy are relevant for decision makers → still useful as “statistical devices” to add information to the hierarchical forecast
- It (perhaps!) is more helpful to think hierarchical forecasting as a multivariable (multivariate or univariate) problem.
- The different variables (nodes/levels of the hierarchy) are connect through the coherency constraints.





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

Coherency: $F(A+B) = F(A)+F(B)$

Hierarchical forecasting in a nutshell

- One way to manage this is to use the MinT reconciliation approach

Reconciled **coherent** forecasts ← $\tilde{\boldsymbol{y}}_h = \boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{y}}_h$ → Matrix of **base** forecasts of all variables

Summing matrix, i.e. the map of the hierarchy

Magic

- Observe that base forecasts are linearly combined to give us the reconciled forecasts
- $\boldsymbol{G}$ tells us how the information from the different forecasts is combined

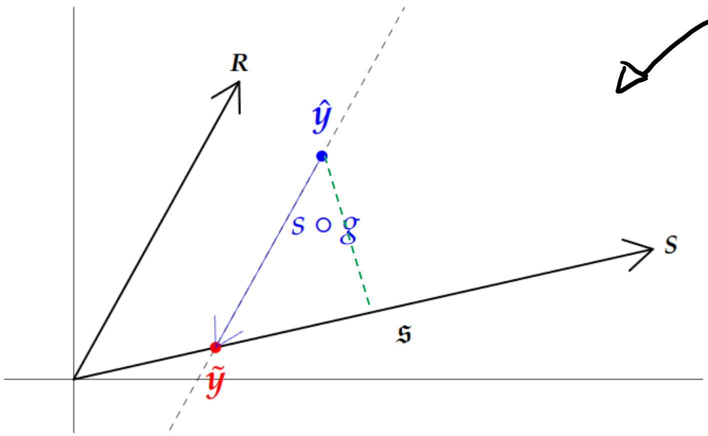
$$\boldsymbol{G} = \left(\boldsymbol{S}'\boldsymbol{W}^{-1}\boldsymbol{S}\right)^{-1}\boldsymbol{S}'\boldsymbol{W}^{-1}$$

An approximation of the covariance matrix of the **relevant** forecast errors

- Different $\boldsymbol{W}$ gives us a variety of approximations, with varying degrees of simplifications (e.g. independence) or estimation tricks (e.g. restrictions and shrinkage).

A geometric interpretation

- Instead of perceiving the problem algebraically (regression/combination), we can look at it from a geometric view



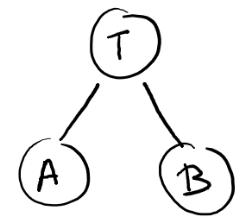
What this figure wants to say is that **base forecasts are projected on a coherent space**. If W is approximated using OLS then we get an **orthogonal** projection, otherwise an **oblique**.

- This is great because it tells us two things:
    - The coherent multivariable object has always lowest error than the base counterpart.
    - All coherent objects live on the same coherent space (let's call it **C-space**).

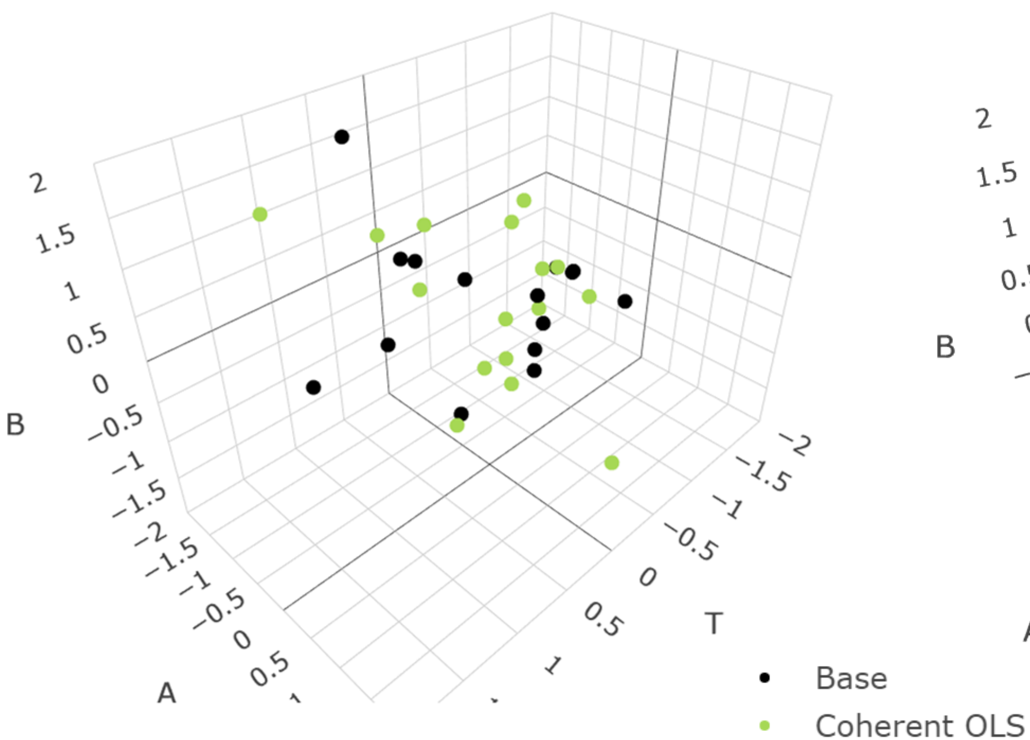
- But this figure is a bit arcane… so let's explore more.

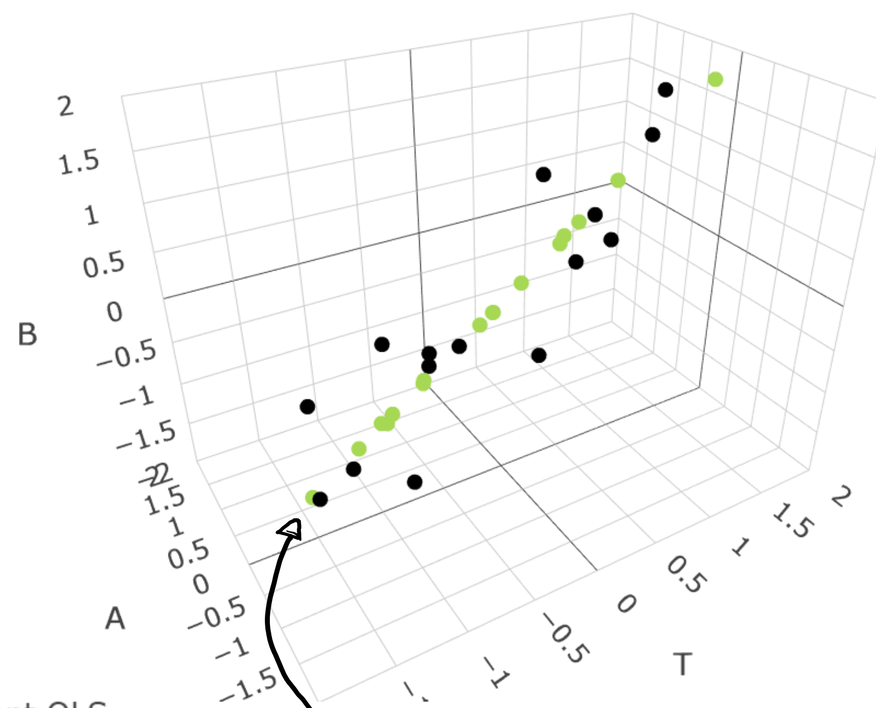
A simplified geometric interpretation

- We stick to a small hierarchy with 3 nodes, so that we can visualise both the **B-space** (where the base forecasts live) and the **C-space** fully.



- We simulate a problem and reconcile it using the OLS approximation ($\boldsymbol{W} = diag([1\ 1\ 1])$ – this needs no estimation, both $\boldsymbol{S}$ and $\boldsymbol{W}$ are known.

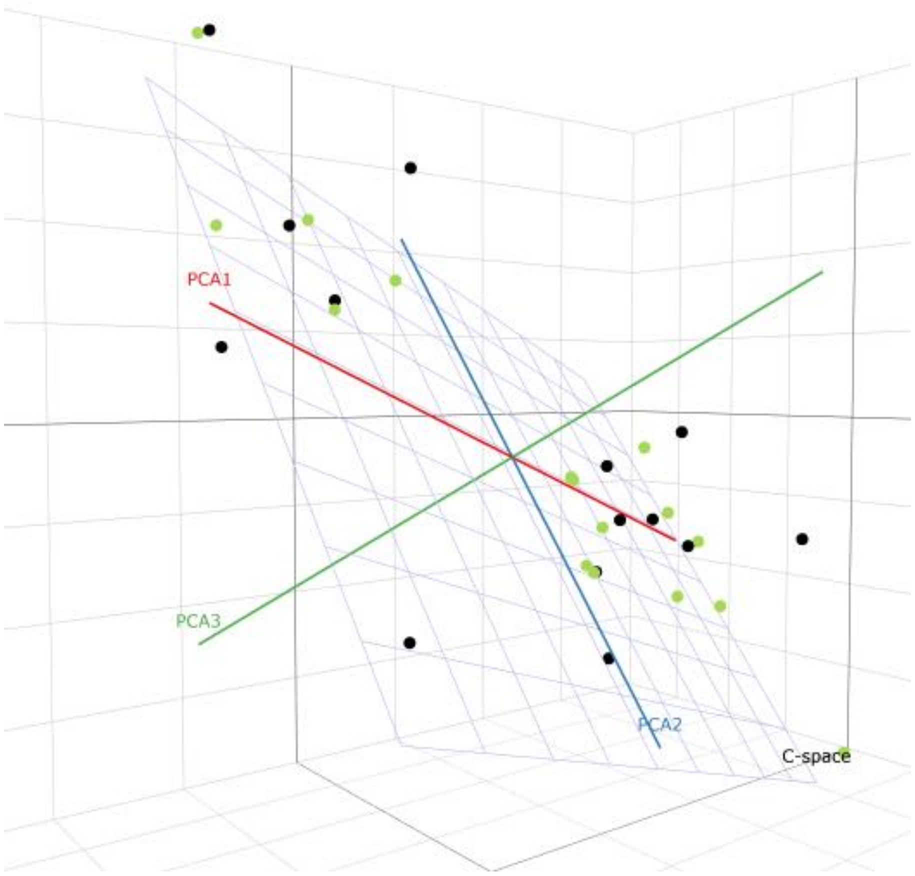




This has to be a plane!

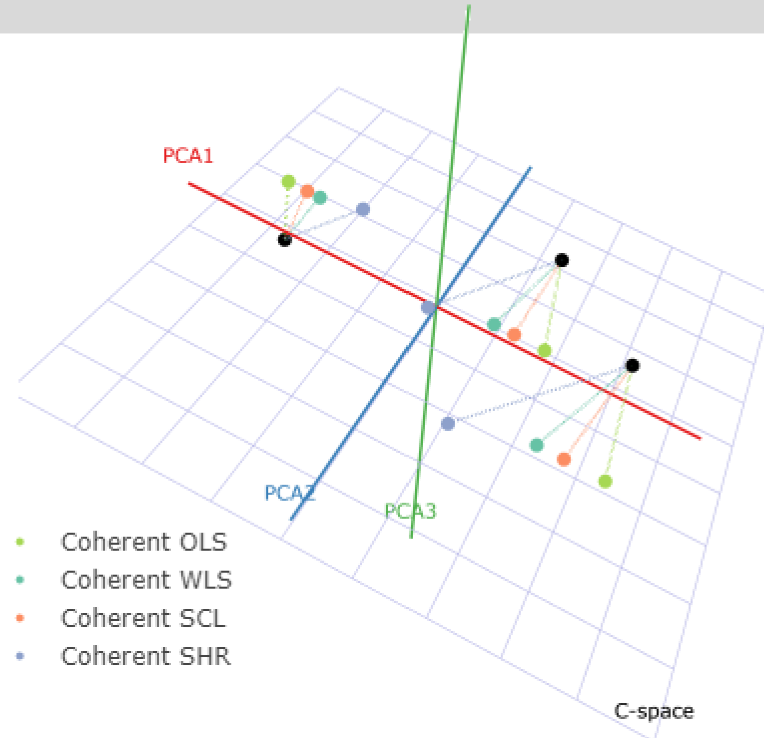
A simplified geometric interpretation

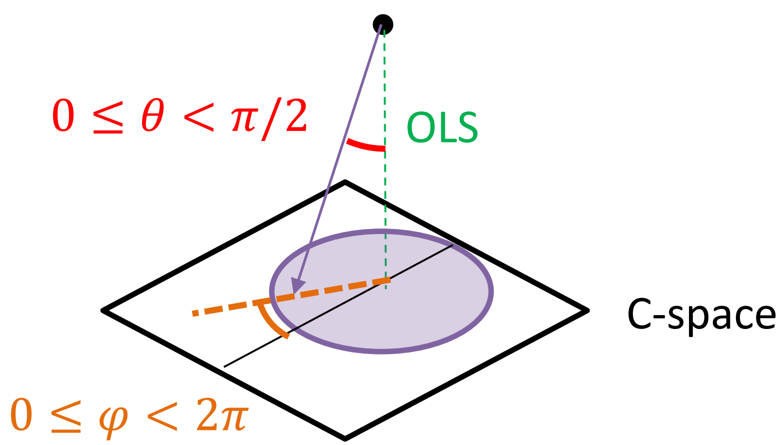
- It is a plane because the bottom level has only two nodes, this defines the dimensionality of C-space, while the dimensionality of B-space is the number of nodes prescribed by $\boldsymbol{S}$.
- There are a few ways to define a 3D plane, we pick two convenient:
  - Two intersecting lines on the plane;
  - A point and a normal vector of the plane (a vector perpendicular to the place).
- It turns out that the principal components of the coherent forecasts (or coherent forecast residuals) do exactly these representations.
- Generally, we need the m-first PC to describe the C-space, where m is the number of bottom-level nodes.



A simplified geometric interpretation

- Now let us add more $\boldsymbol{W}$ approximations:
  - OLS is parallel to the 3rd principal component;
  - WLS (only variance estimates in the diagonal of $\boldsymbol{W}$), SCL (scaling approximation, assume only additivity of variance), SHR (full covariance with shrinkage) are oblique projections.





- $\boldsymbol{W}$ is a 3x3 matrix, but the geometric view requires only 2 parameters to produce all $\boldsymbol{W}$ approximations, angles θ and φ, which are also bounded.
  - We get an efficiency bonus!

A geometry inspired reconciliation

- We can ask an optimizer to directly find the two angles, subject to our constraints, by minimising directly:

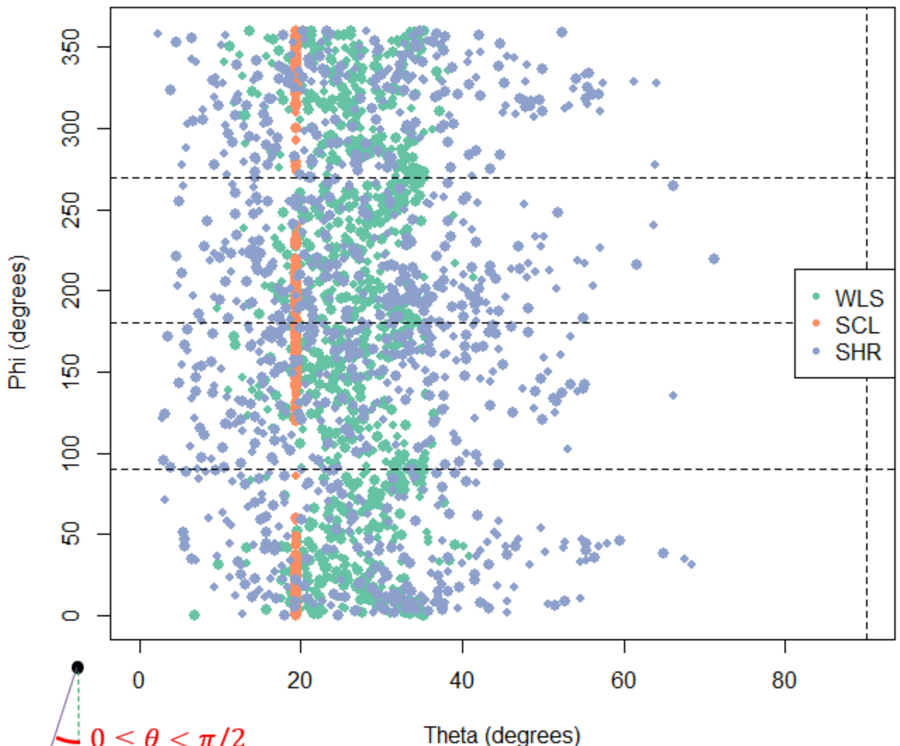
$$(\boldsymbol{y}_t - \tilde{\boldsymbol{y}}_{\boldsymbol{t}})^2$$

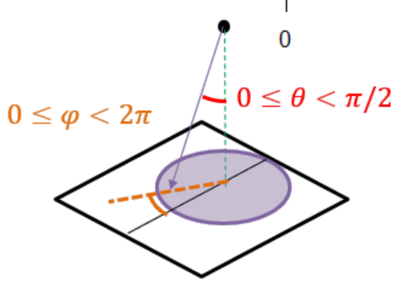
- As the solution by construction will be on the C-space, it is also coherent.
- Observe that we do not need the MinT framework anymore, as we do not need to estimate $\boldsymbol{W}$ or $\boldsymbol{G}$, but rather how to rotate the OLS projection vectors from each point.
- Nonetheless, it is easy to translate between angles and MinT solutions.

- Does it work?
  - Well… not really. Two issues:
    - The optimization is quite difficult and needs many tricks to make it work;
    - Even so, it is a 3D solution, so practically of little interest;
    - But there is more to it..! (spoiler: it is still a more efficient solution!)

A geometry inspired reconciliation

- Let us translate the various $\boldsymbol{W}$ approximations to angles from a 1000 simulated hierarchies.

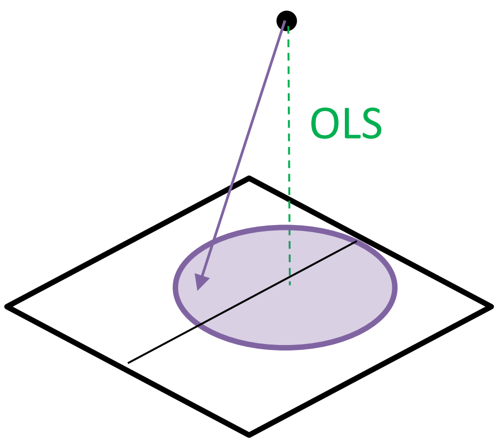


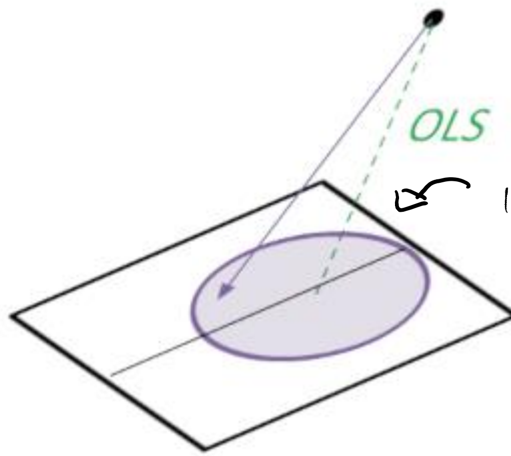


- Not all options are used!
- SCL (Structural scaling, no estimation) results in a fixed θ (19.47);
- WLS (variance in the diagonal of $\boldsymbol{W}$) varies around 27.16;
- SHR (full covariance with shrinkage) varies more with a mean of 28.28;
- Less assumptions of the approximation → more θ!
- Angle φ is denser for some regions (see SCL), but overall independent of θ.

A geometry abused reconciliation

- We can approximate a rotation by sheering the projection vectors of the OLS.





It's oblique!

This can in principle cover the complete $\theta - \varphi$ plot

- To do this we only need to multiply the B-space by a vector with as many elements as the dimensionality of the B-space and "back-multiply" to get things back to the original C-space.

$$\boldsymbol{G} = \left(\boldsymbol{S}'\boldsymbol{W}^{-1}\boldsymbol{S}\right)^{-1}\boldsymbol{S}'\boldsymbol{W}^{-1}$$

Completely different from MinT ☺

$$\boldsymbol{W} = \begin{bmatrix} w_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_n \end{bmatrix}$$

All non-diagonals are zero

We optimise on $(\boldsymbol{y}_t - \tilde{\boldsymbol{y}}_t)^2$

Less efficient by $n - m$ compared to angles

A geometry abused reconciliation

- Does it work?
  - In toy 3D examples yes → of little practical relevant → try on larger hierarchies.
  - Two cases, base forecasts are ETS, rolling origin evaluation.

	Frequency	Number of bottom level series	Number of All series	Sample	Test set	Horizon
Case 1	Quarterly	56	89	36	20	4
Case 2	Monthly	76	111	240	120	12

Scale MSE per series

Total error across series

Relative error	Base	OLS	SCL	WLS	SHR	RAX
	Case 1					
MSE	1	0.992	0.987	0.991	0.983	**0.977**
TSE	1	0.984	0.989	0.991	0.984	**0.976**
	Case 2					
MSE	1	0.983	0.965	0.956	0.928	**0.923**
TSE	1	0.940	0.955	0.936	0.914	**0.901**

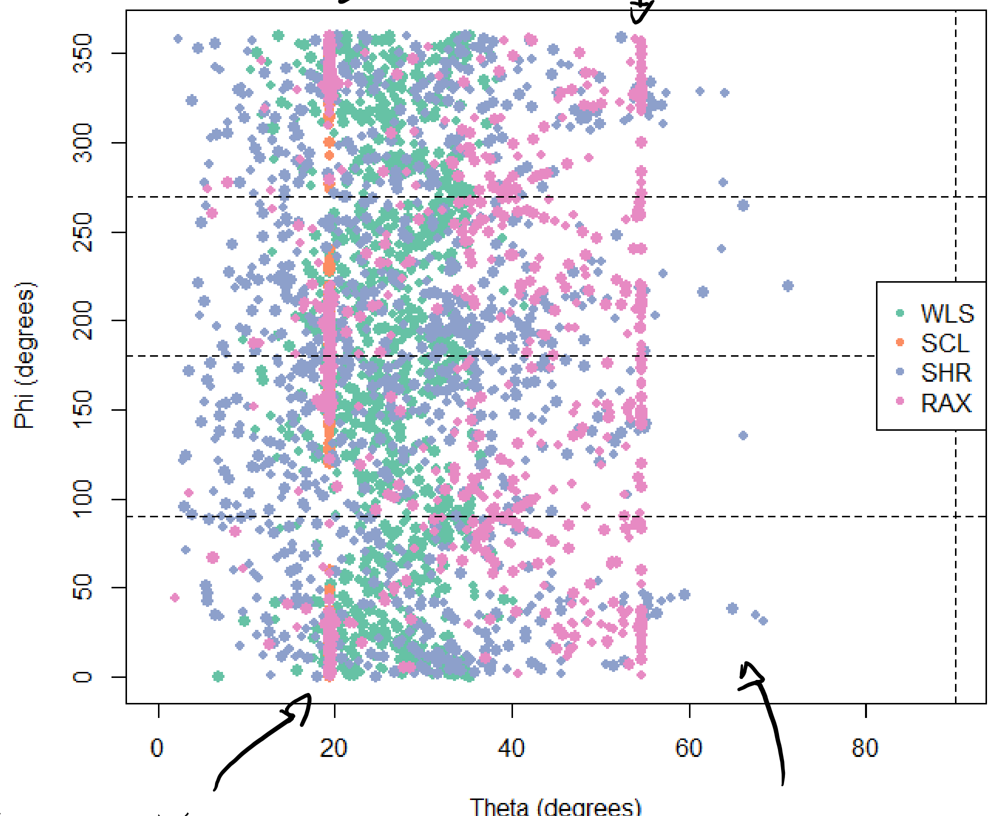
"Rotation approximation"

A geometry abused reconciliation

- Is it doing the same thing though?
- Back to 3D examples

Note it has a soft preference on θ.

It seems to be doing something different!



There is something to the angle of SCL

Very few (bad) SHR solutions go beyond RAX

A heuristic geometry abused reconciliation

- Going from a vector rotation solution to its approximation (RAX) we lost on efficiency. Perhaps we can gain back the lost efficiency by using some heuristics:
  - Estimate only the weights for the lowest level of the hierarchy and then use $\boldsymbol{W} = \boldsymbol{S}\boldsymbol{W}_b$ → HRAX: some efficiency as the rotation approach.
  - Estimate the weights for all levels except the lowest. Use HRAX (or WLS) weights for the lowest level → HRAXc
  - We can also modify WLS as HRAX to get HWLS. Note tha SCL is the equivalent for OLS.

Relative error	Base	OLS	SCL	WLS	SHR	RAX	HRAX	HRAXc	HWLS
	Case 1								
MSE	1	0.992	0.987	0.991	0.983	**0.977**	0.983	0.985	0.989
TSE	1	0.984	0.989	0.991	0.984	**0.976**	0.985	0.988	0.988
	Case 2								
MSE	1	0.983	0.965	0.956	0.928	**0.923**	0.945	0.933	0.959
TSE	1	0.940	0.955	0.936	0.914	**0.901**	0.926	0.915	0.940

Conclusions

- All MinT and MinT like solutions can be described efficiently using rotations and there seems to be a preference in the obliqueness of the solutions.
- Estimation errors of $\boldsymbol{W}$ are easy to spot when looking at the angle representation. Same for the restricting effect of the assumptions in the various approximations.
- Rotation reconciliation is difficult to optimise, and does not scale up easily, but it encompasses existing frameworks.
- The rotation approximation (RAX) can overcome these and seems to perform better than other good approximations.
- Next steps:
  - Observe that the loss function of RAX can be anything, that gives it a lot of flexibility.
  - Although it is inspired by rotations, it is merely a projection from B-space to C-space.
    - This we can solve analytically, by using directly the loss of RAX, instead of restricting us to MinT or similar.
    - We will show you this next time!

Thank you for your attention!
Questions?

Nikolaos Kourentzes

email: nikolaos@kourentzes.com

twitter @nkourentz

Blog: http://nikolaos.kourentzes.com

Slides available here

Special issue @IJF

Innovations in Hierarchical Forecasting

http://bit.ly/ijfhierarchical

HÖGSKOLAN I SKÖVDE