MONASH University

Forecast reconciliation: Geometry, optimization and beyond

Anastasios Panagiotelis

2 July 2025

Outline

Outline

Hierarchical Time Series

- At its most general, **multivariate** data bound together by some constraints.

Hierarchical Time Series

- At its most general, **multivariate** data bound together by some constraints.
- Typically these constraints are **linear**, although later I will present new work for non-linear constraints.

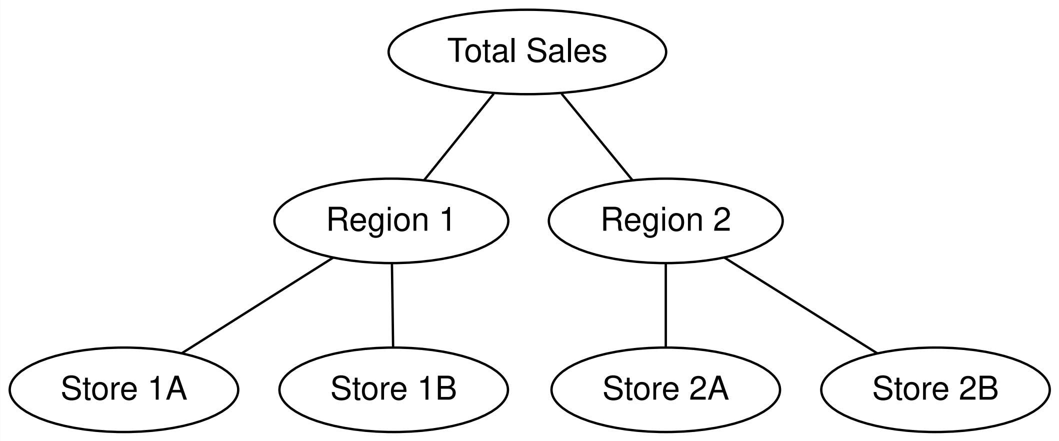
Hierarchical Time Series

- At its most general, **multivariate** data bound together by some constraints.
- Typically these constraints are **linear**, although later I will present new work for non-linear constraints.
- Most commonly arise due to an **aggregation** structure, hence the name 'hierarchical'.

Hierarchical Time Series

- At its most general, **multivariate** data bound together by some constraints.
- Typically these constraints are **linear**, although later I will present new work for non-linear constraints.
- Most commonly arise due to an **aggregation** structure, hence the name 'hierarchical'.
- Need not be hierarchical, alternative structures are grouped (or crossed) aggregation, or temporal aggregation.

Hierarchy


Total Sales
Region 1
Region 2
Store 1A
Store 1B
Store 2A
Store 2B

One representation

- For the simple hierarchy shown earlier:

$$\begin{pmatrix} y_{\mathrm{Tot}} \\ y_1 \\ y_2 \\ y_{1A} \\ y_{1B} \\ y_{2A} \\ y_{2B} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \times \begin{pmatrix} y_{1A} \\ y_{1B} \\ y_{2A} \\ y_{2B} \end{pmatrix}$$

$$\mathbf{y} = \mathbf{S} \times \mathbf{b}$$

Forecasting

- Observations will always **cohere** to the constraints.

Forecasting

- Observations will always **cohere** to the constraints.
- Forecasts generally will not.

Forecasting

- Observations will always **cohere** to the constraints.
- Forecasts generally will not.
  - Different forecasts are made by different agents.

Forecasting

- Observations will always **cohere** to the constraints.
- Forecasts generally will not.
  - Different forecasts are made by different agents.
  - Hard to construct a method that guarantees coherence.

Forecasting

- Observations will always **cohere** to the constraints.
- Forecasts generally will not.
  - Different forecasts are made by different agents.
  - Hard to construct a method that guarantees coherence.
- This talk is about **two-stage** processes whereby incoherent **base** forecasts are adjusted to be coherent.

Forecasting

- Observations will always **cohere** to the constraints.
- Forecasts generally will not.
  - Different forecasts are made by different agents.
  - Hard to construct a method that guarantees coherence.
- This talk is about **two-stage** processes whereby incoherent **base** forecasts are adjusted to be coherent.
- Note there is also work on **end-to-end** forecasting (e.g. Rangapuram et al. 2021).

Forecast reconciliation

- Start with a vector of incoherent forecasts $\hat{\mathbf{y}}$.

Forecast reconciliation

- Start with a vector of incoherent forecasts $\hat{\mathbf{y}}$.
- Recall the characterization earlier with $\mathbf{S}$

Forecast reconciliation

- Start with a vector of incoherent forecasts $\hat{\mathbf{y}}$.
- Recall the characterization earlier with $\mathbf{S}$
- 'Regress' $\hat{\mathbf{y}}$ on $\mathbf{S}$.

Forecast reconciliation

- Start with a vector of incoherent forecasts $\hat{\mathbf{y}}$.
- Recall the characterization earlier with $\mathbf{S}$
- 'Regress' $\hat{\mathbf{y}}$ on $\mathbf{S}$.
- Use prediction as reconciled forecasts $\tilde{\mathbf{y}}$, i.e.

Forecast reconciliation

- Start with a vector of incoherent forecasts $\hat{\mathbf{y}}$.
- Recall the characterization earlier with $\mathbf{S}$
- 'Regress' $\hat{\mathbf{y}}$ on $\mathbf{S}$.
- Use prediction as reconciled forecasts $\tilde{\mathbf{y}}$, i.e.

$$\tilde{\mathbf{y}} = \mathbf{S}(\mathbf{S}'\mathbf{S})^{-1}\mathbf{S}'\hat{\mathbf{y}}$$

- This is called OLS reconciliation.

Generalizations

- Where OLS works, it makes sense to consider GLS

$$\tilde{\mathbf{y}} = \mathbf{S}(\mathbf{S}'\mathbf{W}\mathbf{S})^{-1}\mathbf{S}'\mathbf{W}\hat{\mathbf{y}}$$

Generalizations

- Where OLS works, it makes sense to consider GLS

$$\tilde{\mathbf{y}} = \mathbf{S}(\mathbf{S}'\mathbf{W}\mathbf{S})^{-1}\mathbf{S}'\mathbf{W}\hat{\mathbf{y}}$$

- Setting $\mathbf{W}^{-1}$ to the covariance matrix of $\mathbf{y} - \hat{\mathbf{y}}$ optimizes expected squared error loss.

Generalizations

- Where OLS works, it makes sense to consider GLS

$$\tilde{\mathbf{y}} = \mathbf{S}(\mathbf{S}'\mathbf{W}\mathbf{S})^{-1}\mathbf{S}'\mathbf{W}\hat{\mathbf{y}}$$

- Setting $\mathbf{W}^{-1}$ to the covariance matrix of $\mathbf{y} - \hat{\mathbf{y}}$ optimizes expected squared error loss.
- This is the well-known **MinT method** of Wickramasuriya, Athanasopoulos, and Hyndman (2019).

A geometric view

- Another alternative is a geometric view.

A geometric view

- Another alternative is a geometric view.
- The constraints define a linear subspace $\mathcal{S} \subset \mathbb{R}^n$.

A geometric view

- Another alternative is a geometric view.
- The constraints define a linear subspace $\mathcal{S} \subset \mathbb{R}^n$.
  - Sometimes $\mathfrak{s}$ notation is used.

A geometric view

- Another alternative is a geometric view.
- The constraints define a linear subspace $\mathcal{S} \subset \mathbb{R}^n$.
  - Sometimes $\mathfrak{s}$ notation is used.
- This can naturally be extended to other settings

A geometric view

- Another alternative is a geometric view.
- The constraints define a linear subspace $\mathcal{S} \subset \mathbb{R}^n$.
  - Sometimes $\mathfrak{s}$ notation is used.
- This can naturally be extended to other settings
  - For discrete data $\mathcal{S}$ is a set of points (Zhang, Panagiotelis, and Kang 2023).

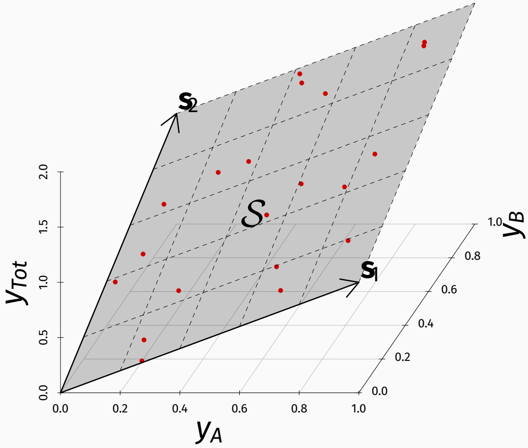
A geometric view

- Another alternative is a geometric view.
- The constraints define a linear subspace $\mathcal{S} \subset \mathbb{R}^n$.
  - Sometimes $\mathfrak{s}$ notation is used.
- This can naturally be extended to other settings
  - For discrete data $\mathcal{S}$ is a set of points (Zhang, Panagiotelis, and Kang 2023).
  - Non linear constraints will be covered later.

A geometric view

- Another alternative is a geometric view.
- The constraints define a linear subspace $\mathcal{S} \subset \mathbb{R}^n$.
  - Sometimes $\mathfrak{s}$ notation is used.
- This can naturally be extended to other settings
  - For discrete data $\mathcal{S}$ is a set of points (Zhang, Panagiotelis, and Kang 2023).
  - Non linear constraints will be covered later.
- The simplest three-variable hierarchy $y_{\mathrm{Tot}} = y_A + y_B$ for real-valued data is depicted on the next slide.

Coherent subspace



Geometry of reconciliation

- Reconciliation is a **map** from $\psi : \mathbb{R}^n \to \mathcal{S}$.

Geometry of reconciliation

- Reconciliation is a **map** from $\psi : \mathbb{R}^n \rightarrow \mathcal{S}$.
- The OLS and MinT method are **projections**, i.e. specific cases of ψ.

Geometry of reconciliation

- Reconciliation is a **map** from $\psi : \mathbb{R}^n \rightarrow \mathcal{S}$.
- The OLS and MinT method are **projections**, i.e. specific cases of ψ.
- Projections solve an optimisation problem

Geometry of reconciliation

- Reconciliation is a **map** from $\psi : \mathbb{R}^n \rightarrow \mathcal{S}$.
- The OLS and MinT method are **projections**, i.e. specific cases of ψ.
- Projections solve an optimisation problem

$$\tilde{\mathbf{y}} = \underset{\mathbf{y}}{argmin} (\mathbf{y} - \hat{\mathbf{y}})' \mathbf{W} (\mathbf{y} - \hat{\mathbf{y}})$$

- Subject to $\mathbf{y} \in \mathcal{S}$

Why projections?

- The reconciliation maps can be more general than projections.

Why projections?

- The reconciliation maps can be more general than projections.
- However projections have some nice properties

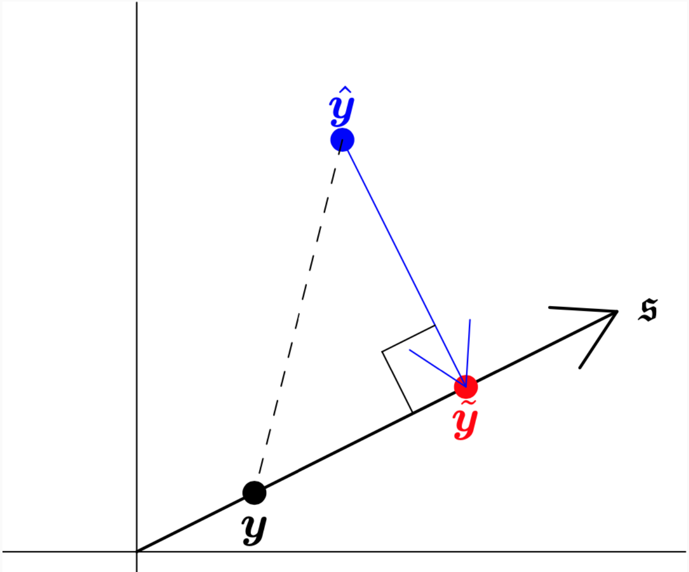
Why projections?

- The reconciliation maps can be more general than projections.
- However projections have some nice properties
  - OLS always reduces forecast error.

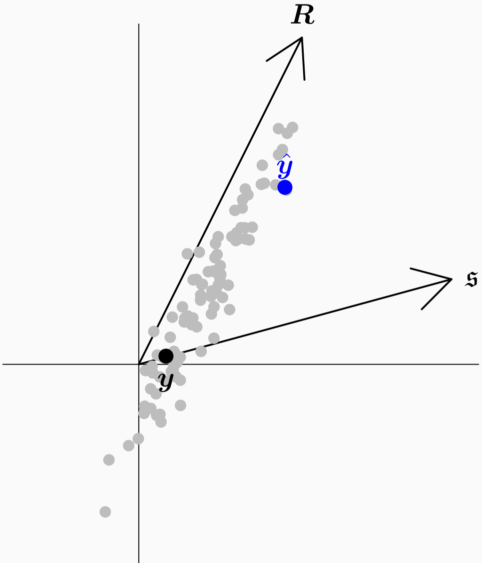
Why projections?

- The reconciliation maps can be more general than projections.
- However projections have some nice properties
  - OLS always reduces forecast error.
  - MinT minimizes forecast error in expectation.

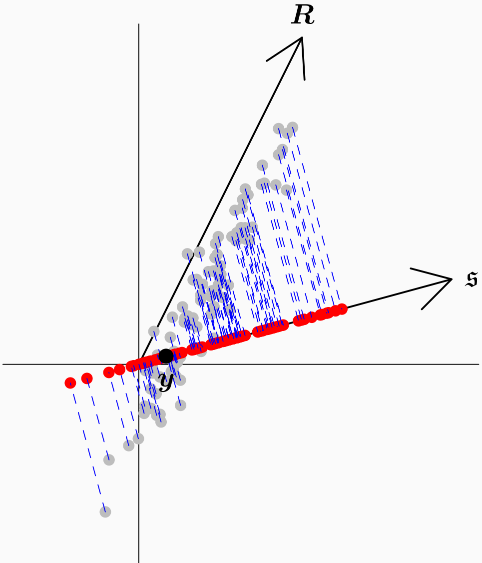
OLS Reconciliation



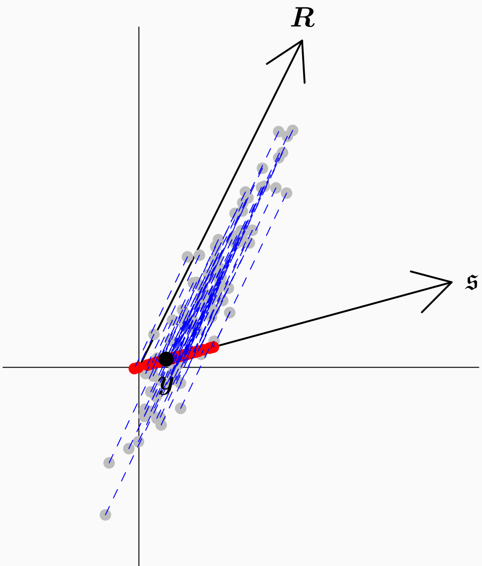
Why MinT?



Why MinT?



Why MinT?



Outline

Problem

- The regression interpretation does not naturally lend itself to be extended to probabilistic forecasting.

Problem

- The regression interpretation does not naturally lend itself to be extended to probabilistic forecasting.
- Alternative approaches define the reconciled distribution using *copulas* (Ben Taieb, Taylor, and Hyndman 2021) or by *conditioning* (see Corani et al. 2021).

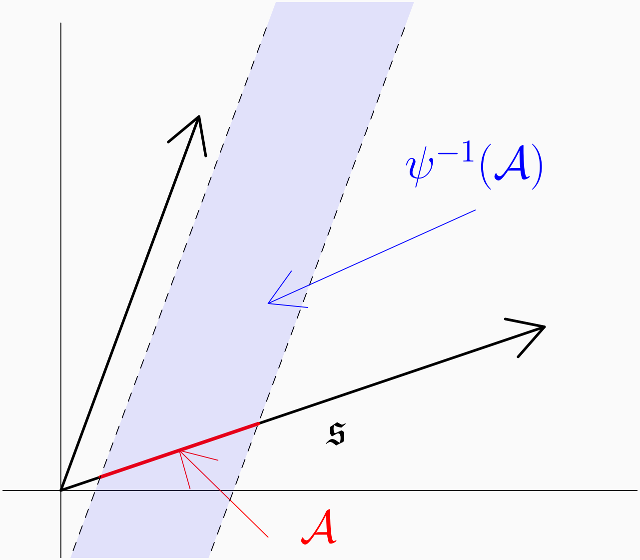
Problem

- The regression interpretation does not naturally lend itself to be extended to probabilistic forecasting.
- Alternative approaches define the reconciled distribution using *copulas* (Ben Taieb, Taylor, and Hyndman 2021) or by *conditioning* (see Corani et al. 2021).
- Some notions of reconciling draws from probabilistic distributions (Jeon, Panagiotelis, and Petropoulos 2019).

Problem

- The regression interpretation does not naturally lend itself to be extended to probabilistic forecasting.
- Alternative approaches define the reconciled distribution using *copulas* (Ben Taieb, Taylor, and Hyndman 2021) or by *conditioning* (see Corani et al. 2021).
- Some notions of reconciling draws from probabilistic distributions (Jeon, Panagiotelis, and Petropoulos 2019).
- Later formalized reconciliation as a **pushforward** (Panagiotelis et al. 2023).

Probabilistic Reconciliation



Probabilistic Reconciliation

- Let $\hat{\mu}$ be a measure on $\mathbb{R}^n$.

Probabilistic Reconciliation

- Let $\hat{\mu}$ be a measure on $\mathbb{R}^n$.
- Let $\mathcal{A}$ be some region entirely within $\mathcal{S}$ and $\psi : \mathbb{R}^n \rightarrow \mathcal{S}$.

Probabilistic Reconciliation

- Let $\hat{\mu}$ be a measure on $\mathbb{R}^n$.
- Let $\mathcal{A}$ be some region entirely within $\mathcal{S}$ and $\psi : \mathbb{R}^n \rightarrow \mathcal{S}$.
- Let $\psi^{-1}(\mathcal{A})$ be the pre-image of $\mathcal{A}$

Probabilistic Reconciliation

- Let $\hat{\mu}$ be a measure on $\mathbb{R}^n$.
- Let $\mathcal{A}$ be some region entirely within $\mathcal{S}$ and $\psi : \mathbb{R}^n \rightarrow \mathcal{S}$.
- Let $\psi^{-1}(\mathcal{A})$ be the pre-image of $\mathcal{A}$
  - $\forall \mathbf{x} \in \psi^{-1}(\mathcal{A}), \psi(\mathbf{x}) \in \mathcal{A}$.

Probabilistic Reconciliation

- Let $\hat{\mu}$ be a measure on $\mathbb{R}^n$.
- Let $\mathcal{A}$ be some region entirely within $\mathcal{S}$ and $\psi : \mathbb{R}^n \rightarrow \mathcal{S}$.
- Let $\psi^{-1}(\mathcal{A})$ be the pre-image of $\mathcal{A}$
  - $\forall \mathbf{x} \in \psi^{-1}(\mathcal{A}), \psi(\mathbf{x}) \in \mathcal{A}$.
- The reconciled measure $\tilde{\mu}$ is defined as

Probabilistic Reconciliation

- Let $\hat{\mu}$ be a measure on $\mathbb{R}^n$.
- Let $\mathcal{A}$ be some region entirely within $\mathcal{S}$ and $\psi : \mathbb{R}^n \rightarrow \mathcal{S}$.
- Let $\psi^{-1}(\mathcal{A})$ be the pre-image of $\mathcal{A}$
  - $\forall \mathbf{x} \in \psi^{-1}(\mathcal{A}), \psi(\mathbf{x}) \in \mathcal{A}$.
- The reconciled measure $\tilde{\mu}$ is defined as

$$\tilde{\mu}(\mathcal{A}) = \hat{\mu}(\psi^{-1}(\mathcal{A}))$$

- $\tilde{\mu}$ is the **pushforward** of $\hat{\mu}$ by ψ, denoted as $\psi\#\hat{\mu}$

Optimality

- This merely defines a way of getting a reconciled distribution from some incoherent base distribution.

Optimality

- This merely defines a way of getting a reconciled distribution from some incoherent base distribution.
- What is the optimal ψ?

Optimality

- This merely defines a way of getting a reconciled distribution from some incoherent base distribution.
- What is the optimal ψ?
- In the point forecasting world the ψ given by MinT is optimal for squared loss.

Optimality

- This merely defines a way of getting a reconciled distribution from some incoherent base distribution.
- What is the optimal ψ?
- In the point forecasting world the ψ given by MinT is optimal for squared loss.
- What does optimality even mean for distributional forecasts?

Scoring rules

- A scoring rule $S : \mathcal{P} \times \mathbb{R}^n \to \mathbb{R}$ takes a distributional forecast from a family $\mathcal{P}$ and an observation and assigns a *score* that measures forecast quality.

Scoring rules

- A scoring rule $S : \mathcal{P} \times \mathbb{R}^n \to \mathbb{R}$ takes a distributional forecast from a family $\mathcal{P}$ and an observation and assigns a *score* that measures forecast quality.
- In the context of probabilistic forecasting, it has been proven that choosing ψ to be the same projection as MinT is optimal for log score when probabilistic forecasts are Gaussian (Wickramasuriya 2023)

Scoring rules

- A scoring rule $S : \mathcal{P} \times \mathbb{R}^n \rightarrow \mathbb{R}$ takes a distributional forecast from a family $\mathcal{P}$ and an observation and assigns a *score* that measures forecast quality.
- In the context of probabilistic forecasting, it has been proven that choosing ψ to be the same projection as MinT is optimal for log score when probabilistic forecasts are Gaussian (Wickramasuriya 2023)
- In other cases we can optimize using a data driven approach.

Score Optimization

- Obtain a sequence of base forecasts $\hat{\mu}_t$ and corresponding realizations $\mathbf{y}_t$.

Score Optimization

- Obtain a sequence of base forecasts $\hat{\mu}_t$ and corresponding realizations $\mathbf{y}_t$.
- Parameterize ψ by some vector $\boldsymbol{\theta}$ (denoted $\psi_{\boldsymbol{\theta}}$)

Score Optimization

- Obtain a sequence of base forecasts $\hat{\mu}_t$ and corresponding realizations $\mathbf{y}_t$.
- Parameterize ψ by some vector $\boldsymbol{\theta}$ (denoted $\psi_{\boldsymbol{\theta}}$)
- For example, $\psi_{\boldsymbol{\theta}}(\mathbf{y}) = \mathbf{S}(\mathbf{d} + \mathbf{G}\mathbf{y})$, $\boldsymbol{\theta} = \left(\mathbf{d}', vec(\mathbf{G})'\right)'$

Score Optimization

- Obtain a sequence of base forecasts $\hat{\mu}_t$ and corresponding realizations $\mathbf{y}_t$.
- Parameterize ψ by some vector $\boldsymbol{\theta}$ (denoted $\psi_{\boldsymbol{\theta}}$)
- For example, $\psi_{\boldsymbol{\theta}}(\mathbf{y}) = \mathbf{S}(\mathbf{d} + \mathbf{G}\mathbf{y})$, $\boldsymbol{\theta} = \left(\mathbf{d}', vec(\mathbf{G})'\right)'$
- Optimize the following

Score Optimization

- Obtain a sequence of base forecasts $\hat{\mu}_t$ and corresponding realizations $\mathbf{y}_t$.
- Parameterize ψ by some vector $\boldsymbol{\theta}$ (denoted $\psi_{\boldsymbol{\theta}}$)
- For example, $\psi_{\boldsymbol{\theta}}(\mathbf{y}) = \mathbf{S}(\mathbf{d} + \mathbf{G}\mathbf{y})$, $\boldsymbol{\theta} = \left(\mathbf{d}', vec(\mathbf{G})'\right)'$
- Optimize the following

$$\underset{\boldsymbol{\theta}}{argmin} \sum_t S(\psi_{\boldsymbol{\theta}} \# \hat{\mu}_t, \mathbf{y}_t)$$

Some practical points

- Scoring rules that have been used include log score, energy score and variogram score.

Some practical points

- Scoring rules that have been used include log score, energy score and variogram score.
- Paired forecasts and observations can be obtained using rolling or expanding window schemes.

Some practical points

- Scoring rules that have been used include log score, energy score and variogram score.
- Paired forecasts and observations can be obtained using rolling or expanding window schemes.
- Different optimal values of θ can be obtained for different forecast horizons.

Some practical points

- Scoring rules that have been used include log score, energy score and variogram score.
- Paired forecasts and observations can be obtained using rolling or expanding window schemes.
- Different optimal values of θ can be obtained for different forecast horizons.
- Often draw a sample from $\hat{\mu}$ rather than work with the distribution itself.

Some practical points

- Scoring rules that have been used include log score, energy score and variogram score.
- Paired forecasts and observations can be obtained using rolling or expanding window schemes.
- Different optimal values of θ can be obtained for different forecast horizons.
- Often draw a sample from $\hat{\mu}$ rather than work with the distribution itself.
- Optimization by first order methods (e.g. SGD).

Main takeaways from empirical work

- For heavily misspecified model (independent and Gaussian base forecasts)

Main takeaways from empirical work

- For heavily misspecified model (independent and Gaussian base forecasts)
  - Score Optimisation outperforms benchmarks including projections.

Main takeaways from empirical work

- For heavily misspecified model (independent and Gaussian base forecasts)
  - Score Optimisation outperforms benchmarks including projections.
- For less misspecified models (jointly bootstraped base forecasts)

Main takeaways from empirical work

- For heavily misspecified model (independent and Gaussian base forecasts)
  - Score Optimisation outperforms benchmarks including projections.
- For less misspecified models (jointly bootstraped base forecasts)
  - Projections perform similarly to score optimization.

Main takeaways from empirical work

- For heavily misspecified model (independent and Gaussian base forecasts)
  - Score Optimisation outperforms benchmarks including projections.
- For less misspecified models (jointly bootstraped base forecasts)
  - Projections perform similarly to score optimization.
- Reconciliation gives forecaster a ‘second chance’.

Main takeaways from empirical work

- For heavily misspecified model (independent and Gaussian base forecasts)
  - Score Optimisation outperforms benchmarks including projections.
- For less misspecified models (jointly bootstraped base forecasts)
  - Projections perform similarly to score optimization.
- Reconciliation gives forecaster a ‘second chance’.
- Target reconciliation towards ‘end goal’.

Outline

Pinball loss

- Many forecasting problems involve optimizing pinball loss.

$$L_\alpha(y, q) = \alpha(y_i - q)I(y_i \geq q) + (1 - \alpha)(q - y_i)I(y_i < q)$$

Pinball loss

- Many forecasting problems involve optimizing pinball loss.

$$L_\alpha(y, q) = \alpha(y_i - q)I(y_i \geq q) + (1 - \alpha)(q - y_i)I(y_i < q)$$

- Here, I(.) equals 1 when the statement in parentheses is true, 0 otherwise.

Pinball loss

- Many forecasting problems involve optimizing pinball loss.

$$L_\alpha(y, q) = \alpha(y_i - q)I(y_i \geq q) + (1 - \alpha)(q - y_i)I(y_i < q)$$

- Here, I(.) equals 1 when the statement in parentheses is true, 0 otherwise.
- Quantiles minimize expected pinball loss $E_Y\left[L_\alpha(y, q)\right]$

In reconciliation

- To target quantiles we optimize.

$$\underset{\theta}{argmin} \sum_{i=1}^{n} \sum_{t \in \mathcal{T}_{\text{train}}} L_\alpha(y_{i,t}, \tilde{q}_{i,t})$$

In reconciliation

- To target quantiles we optimize.

$$\underset{\theta}{argmin} \sum_{i=1}^{n} \sum_{t \in \mathcal{T}_{\text{train}}} L_\alpha(y_{i,t}, \tilde{q}_{i,t})$$

- Subject to the constraints

$$\tilde{q}_{i,t} = \underset{q}{argmin} E_{\tilde{Y}_{i,t}} \left[L_\alpha(\tilde{y}_{i,t}, q) \right]$$

In reconciliation

- To target quantiles we optimize.

$$\underset{\theta}{argmin} \sum_{i=1}^{n} \sum_{t \in \mathcal{T}_{\mathrm{train}}} L_{\alpha}(y_{i,t}, \tilde{q}_{i,t})$$

- Subject to the constraints

$$\tilde{q}_{i,t} = \underset{q}{argmin} E_{\tilde{Y}_{i,t}} \left[L_{\alpha}(\tilde{y}_{i,t}, q) \right]$$

Note $\tilde{y}_{i,t}$ depend on θ

Optimization

- This is an example of **bi-level optimization**.

Optimization

- This is an example of **bi-level optimization**.
- It is further complicated by the fact that pinball loss is not smooth.

Optimization

- This is an example of **bi-level optimization**.
- It is further complicated by the fact that pinball loss is not smooth.
- It is also complicated by the need to approximate expectations with sample equivalents.

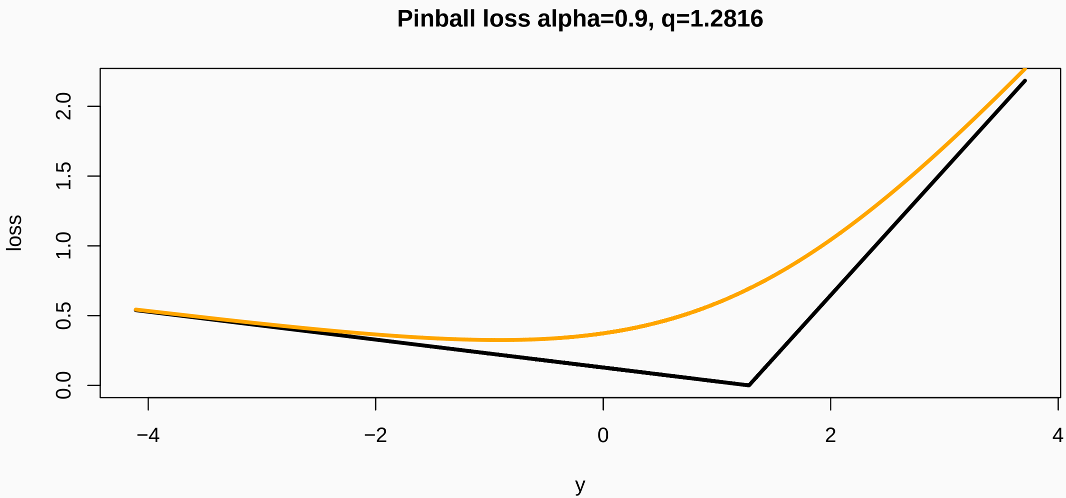
Smooth pinball loss

Use approximation converging to pinball loss as $\beta \to \infty$

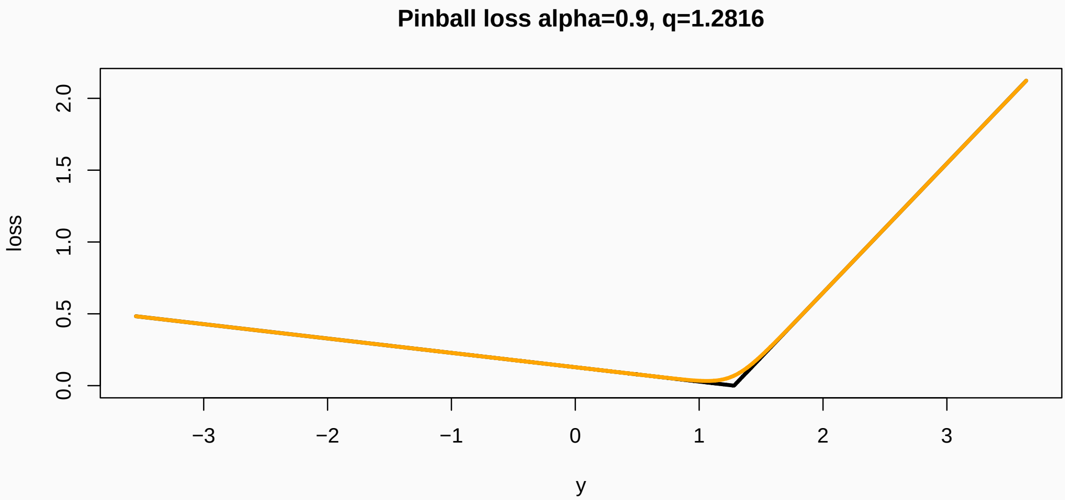
$$L_\alpha^\beta(y, q) = \frac{1}{\beta} \log\left(e^{\beta\alpha(y-q)} + e^{\beta(1-\alpha)(q-y)}\right)$$

Unlike the pinball function it is smooth, meaning we can use first order methods (like Stochastic Gradient Descent).

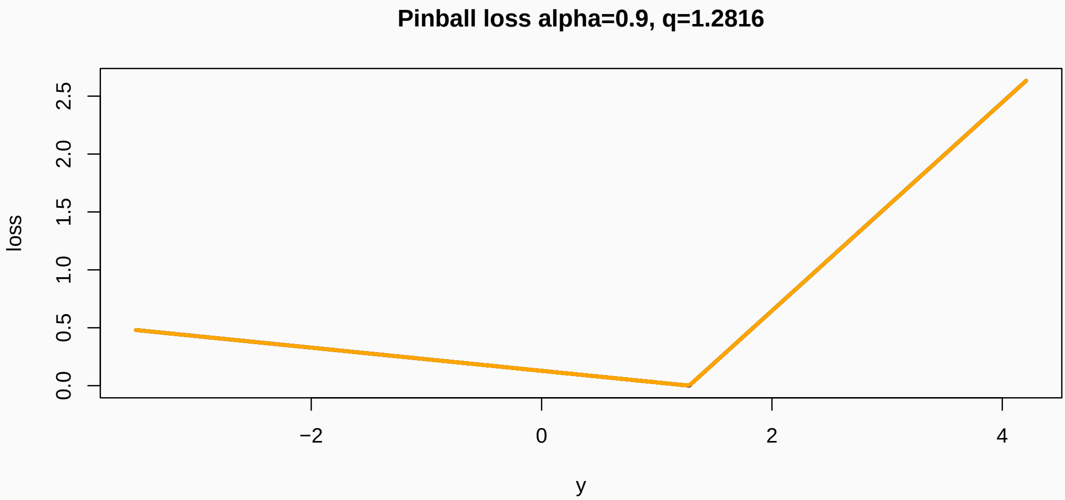
Smoothed pinball loss (β = 1)



Smoothed pinball loss (β = 10)



Smoothed pinball loss (β = 100)



What we want to solve

$$\underset{\theta}{argmin} \sum_{i=1}^{n} \sum_{t \in \mathcal{T}_{\text{train}}} L_\alpha(y_{i,t}, \tilde{q}_{i,t})$$

- Subject to the constraints

$$\tilde{q}_{i,t} = \underset{q}{argmin} E_{\tilde{Y}_{i,t}} \left[L_\alpha(\tilde{y}_{i,t}, q) \right]$$

What we can solve

$$\underset{\theta}{argmin} \sum_{i=1}^{n} \sum_{t \in \mathcal{T}_{\text{train}}} L_\alpha^\beta(y_{i,t}, \tilde{q}_{i,t})$$

- Subject to the constraints

$$\tilde{q}_{i,t|t-1} = \underset{q}{argmin} \sum_{j} L_\alpha^\beta(\tilde{y}_{i,t}^{(j)}, q)$$

What we can solve

$$\underset{\theta}{argmin} \sum_{i=1}^{n} \sum_{t \in \mathcal{T}_{\mathrm{train}}} L_{\alpha}^{\beta}(y_{i,t}, \tilde{q}_{i,t})$$

- Subject to the constraints

$$\tilde{q}_{i,t|t-1} = \underset{q}{argmin} \sum_{j} L_{\alpha}^{\beta}(\tilde{y}_{i,t}^{(j)}, q)$$

where $\tilde{y}_{i,t}^{(j)} = \psi_{\theta}\left(\hat{y}_{i,t}^{(j)}\right)$ and $\hat{y}_{i,t}^{(j)} \sim \hat{\mu}_t$ for $j = 1, \ldots, J$

What has been proven

- Let $f(\theta)$ be the problem we want to solve

What has been proven

- Let $f(\theta)$ be the problem we want to solve
- Let $f^{\beta}(\theta)$ be the smooth approximation of pinball loss

What has been proven

- Let $f(\theta)$ be the problem we want to solve
- Let $f^{\beta}(\theta)$ be the smooth approximation of pinball loss
- Let $f^{(J)}$ be the approx. from using pinball loss and J draws.

What has been proven

- Let $f(\theta)$ be the problem we want to solve
- Let $f^{\beta}(\theta)$ be the smooth approximation of pinball loss
- Let $f^{(J)}$ be the approx. from using pinball loss and J draws.
- Prove minimizer of approximate problem converges to minimizer of the ‘true’ problem by proving

What has been proven

- Let $f(\theta)$ be the problem we want to solve
- Let $f^{\beta}(\theta)$ be the smooth approximation of pinball loss
- Let $f^{(J)}$ be the approx. from using pinball loss and J draws.
- Prove minimizer of approximate problem converges to minimizer of the 'true' problem by proving

$$\sup_{\theta \in \boldsymbol{\Theta}} \left| f^{\beta}(\theta) - f(\theta) \right| \to 0 \text{ as } \beta \to \infty$$

$$\sup_{\theta \in \boldsymbol{\Theta}} \left| f^{(J)}(\theta) - f^{\beta}(\theta) \right| \to 0 \text{ as } J \to \infty.$$

Convergence of SGD

- Optimization via SGD, taking care to pass gradient through argmin in lower level.

Convergence of SGD

- Optimization via SGD, taking care to pass gradient through argmin in lower level.
- Note that the approximation of the expectation in the constraint means that the gradient is a biased estimate.

Convergence of SGD

- Optimization via SGD, taking care to pass gradient through argmin in lower level.
- Note that the approximation of the expectation in the constraint means that the gradient is a biased estimate.
- However the variant of SGD we use will converge if

Convergence of SGD

- Optimization via SGD, taking care to pass gradient through argmin in lower level.
- Note that the approximation of the expectation in the constraint means that the gradient is a biased estimate.
- However the variant of SGD we use will converge if
  - Bounded second moment of the stochastic gradient.

Convergence of SGD

- Optimization via SGD, taking care to pass gradient through argmin in lower level.
- Note that the approximation of the expectation in the constraint means that the gradient is a biased estimate.
- However the variant of SGD we use will converge if
  - Bounded second moment of the stochastic gradient.
  - L-smoothness.

Convergence of SGD

- Optimization via SGD, taking care to pass gradient through argmin in lower level.
- Note that the approximation of the expectation in the constraint means that the gradient is a biased estimate.
- However the variant of SGD we use will converge if
  - Bounded second moment of the stochastic gradient.
  - L-smoothness.
- Both are proven to hold for the functions we consider.

Convergence of SGD

- Optimization via SGD, taking care to pass gradient through argmin in lower level.
- Note that the approximation of the expectation in the constraint means that the gradient is a biased estimate.
- However the variant of SGD we use will converge if
  - Bounded second moment of the stochastic gradient.
  - L-smoothness.
- Both are proven to hold for the functions we consider.
- Important to check convergence of SGD.

Empirical study

- Use Australian tourism data.

Empirical study

- Use Australian tourism data.
- Grouped hierarchy of states and purpose of travel.

Empirical study

- Use Australian tourism data.
- Grouped hierarchy of states and purpose of travel.
- Dimension of $\mathbf{S}$ is 40×28.

Empirical study

- Use Australian tourism data.
- Grouped hierarchy of states and purpose of travel.
- Dimension of $\mathbf{S}$ is 40×28.
- Seasonal ARIMA used for base forecasts. Distributional forecasts assume Gaussian errors and skew t errors.

Empirical study

- Use Australian tourism data.
- Grouped hierarchy of states and purpose of travel.
- Dimension of $\mathbf{S}$ is 40×28.
- Seasonal ARIMA used for base forecasts. Distributional forecasts assume Gaussian errors and skew t errors.
- Train on 10 years (120 observations), evaluation on 7 years (84 observations).

Pinball Loss - Out of Sample (Normal errors)

	Quantile Level			
Method	0.05	0.2	0.8	0.95
Base	32*	85*	101	46
OLS	32*	84*	104	51
WLS	**31***	**82***	112	65
MinT	**31***	**82***	111	65
QOpt	35*	85*	100*	**41***

Bold denotes best performing method, asterisk(*) denotes inclusion in model confidence set (Hansen et. al., 2011).

Outline

The problem

- What if the constraints are non-linear?

The problem

- What if the constraints are non-linear?
- For example ratios are common quantities of interest.

The problem

- What if the constraints are non-linear?
- For example ratios are common quantities of interest.
  - Mortality rates are Deaths divided by Exposure.

The problem

- What if the constraints are non-linear?
- For example ratios are common quantities of interest.
  - Mortality rates are Deaths divided by Exposure.
  - Unemployment rates are number of Unemployed divided by Labor Force.

The problem

- What if the constraints are non-linear?
- For example ratios are common quantities of interest.
  - Mortality rates are Deaths divided by Exposure.
  - Unemployment rates are number of Unemployed divided by Labor Force.
- Both of these examples are also be subject to aggregation.

Problem formulation

- In general there are C constraints $g_1(\mathbf{y}) = 0, \ldots g_C(\mathbf{y}) = 0$, or more compactly $g(\mathbf{y}) = \mathbf{0}$.

Problem formulation

- In general there are C constraints $g_1(\mathbf{y}) = 0, \ldots g_C(\mathbf{y}) = 0$, or more compactly $g(\mathbf{y}) = \mathbf{0}$.
- The level set of points $\mathbf{y} : g(\mathbf{y}) = \mathbf{0}$ defines a coherent *surface* or *manifold* continue to be denoted as $\mathcal{S}$.

Problem formulation

- In general there are C constraints $g_1(\mathbf{y}) = 0, \ldots g_C(\mathbf{y}) = 0$, or more compactly $g(\mathbf{y}) = \mathbf{0}$.
- The level set of points $\mathbf{y} : g(\mathbf{y}) = \mathbf{0}$ defines a coherent *surface* or *manifold* continue to be denoted as $\mathcal{S}$.
- Non-linear reconciliation solves the following problem:

Problem formulation

- In general there are C constraints $g_1(\mathbf{y}) = 0, \ldots g_C(\mathbf{y}) = 0$, or more compactly $g(\mathbf{y}) = \mathbf{0}$.
- The level set of points $\mathbf{y} : g(\mathbf{y}) = \mathbf{0}$ defines a coherent *surface* or *manifold* continue to be denoted as $\mathcal{S}$.
- Non-linear reconciliation solves the following problem:

$$\tilde{\mathbf{y}} = \underset{\mathbf{y}}{argmin}(\mathbf{y} - \hat{\mathbf{y}})'\mathbf{W}(\mathbf{y} - \hat{\mathbf{y}})$$

- Subject to $\mathbf{y} \in \mathcal{S}$.

Towards Theory

- Note focus is still on point forecasts and projections.

Towards Theory

- Note focus is still on point forecasts and projections.
- First consider case of convex constraints.

Towards Theory

- Note focus is still on point forecasts and projections.
- First consider case of convex constraints.
  - Reconciliation guaranteed to improve base forecast, but only in hypograph.

Towards Theory

- Note focus is still on point forecasts and projections.
- First consider case of convex constraints.
  - Reconciliation guaranteed to improve base forecast, but only in hypograph.
- For more general constraints use following strategy

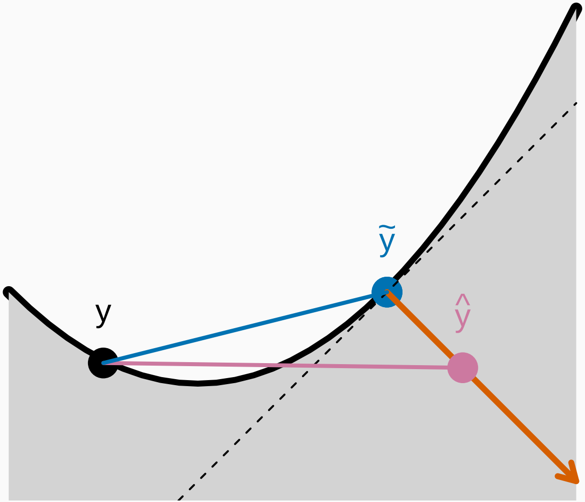
Towards Theory

- Note focus is still on point forecasts and projections.
- First consider case of convex constraints.
  - Reconciliation guaranteed to improve base forecast, but only in hypograph.
- For more general constraints use following strategy
  - Find closest point on the coherent manifold equidistant from the base and reconciled forecast.

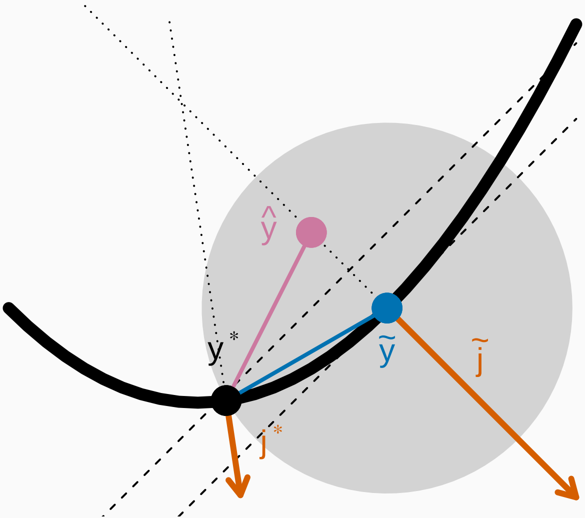
Towards Theory

- Note focus is still on point forecasts and projections.
- First consider case of convex constraints.
  - Reconciliation guaranteed to improve base forecast, but only in hypograph.
- For more general constraints use following strategy
  - Find closest point on the coherent manifold equidistant from the base and reconciled forecast.
  - This defines a ball in which reconciliation always outperforms base forecasts.

Convex Function: Hypograph



Any function



Radius of the Ball

- The radius of the ball on the previous slide is given by

$$r = \sqrt{\boldsymbol{\kappa}'\mathbf{J}^{*\prime}\mathbf{J}^{*}\boldsymbol{\kappa} + \mu\boldsymbol{\kappa}'\mathbf{J}^{*\prime}\tilde{\mathbf{J}}\boldsymbol{\lambda} + \frac{\mu^2}{4}\boldsymbol{\lambda}\tilde{\mathbf{J}}'\tilde{\mathbf{J}}\boldsymbol{\lambda}}$$

Radius of the Ball

- The radius of the ball on the previous slide is given by

$$r = \sqrt{\boldsymbol{\kappa}'\mathbf{J}^{*\prime}\mathbf{J}^{*}\boldsymbol{\kappa} + \mu\boldsymbol{\kappa}'\mathbf{J}^{*\prime}\tilde{\mathbf{J}}\boldsymbol{\lambda} + \frac{\mu^2}{4}\boldsymbol{\lambda}\tilde{\mathbf{J}}'\tilde{\mathbf{J}}\boldsymbol{\lambda}}$$

- $\mathbf{J}^*$ and $\tilde{\mathbf{J}}$ are gradients of the constraint evaluated at $\mathbf{y}^*$ and $\tilde{\mathbf{y}}$ respectively.

Radius of the Ball

- The radius of the ball on the previous slide is given by

$$r = \sqrt{\boldsymbol{\kappa}'\mathbf{J}^{*\prime}\mathbf{J}^{*}\boldsymbol{\kappa} + \mu\boldsymbol{\kappa}'\mathbf{J}^{*\prime}\tilde{\mathbf{J}}\boldsymbol{\lambda} + \frac{\mu^2}{4}\boldsymbol{\lambda}\tilde{\mathbf{J}}'\tilde{\mathbf{J}}\boldsymbol{\lambda}}$$

- $\mathbf{J}^*$ and $\tilde{\mathbf{J}}$ are gradients of the constraint evaluated at $\mathbf{y}^*$ and $\tilde{\mathbf{y}}$ respectively.
- λ and $\boldsymbol{\kappa}$ are Lagrange multipliers associated with certain optimization problems.

How is this useful?

- This theory tells us that non-linear forecast reconciliation is more likely to succeed when

How is this useful?

- This theory tells us that non-linear forecast reconciliation is more likely to succeed when
  - Base forecast is far from coherent manifold.

How is this useful?

- This theory tells us that non-linear forecast reconciliation is more likely to succeed when
  - Base forecast is far from coherent manifold.
  - The constraint function has lower curvature.

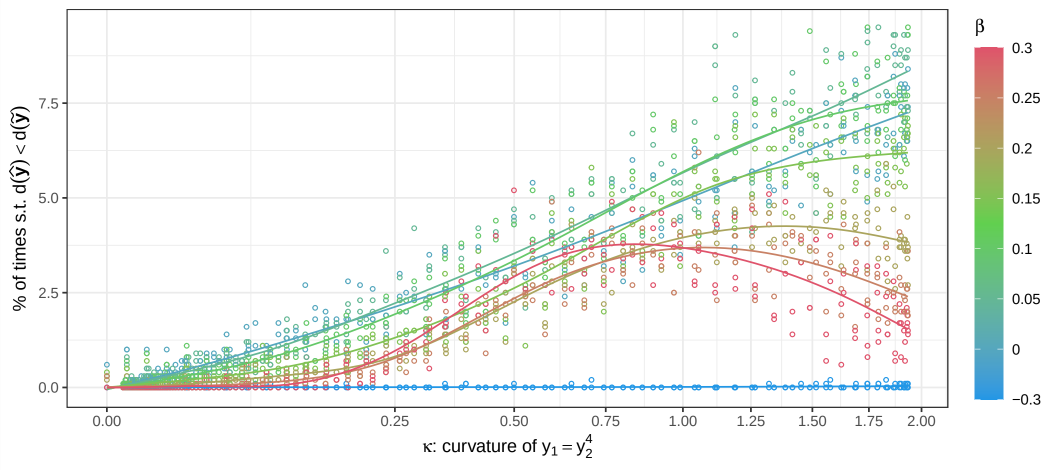
How is this useful?

- This theory tells us that non-linear forecast reconciliation is more likely to succeed when
  - Base forecast is far from coherent manifold.
  - The constraint function has lower curvature.
  - When reconciled forecast is in a high probability region of the true DGP.

How is this useful?

- This theory tells us that non-linear forecast reconciliation is more likely to succeed when
  - Base forecast is far from coherent manifold.
  - The constraint function has lower curvature.
  - When reconciled forecast is in a high probability region of the true DGP.
  - When some constraints are convex and the bast forecast is more likely to lie in the hypographs of these constraints.

Simulation results



Mortality Data

- Annual (1969-2019) data on

Mortality Data

- Annual (1969-2019) data on
  - Exposure (E)

Mortality Data

- Annual (1969-2019) data on
    - Exposure (E)
    - Deaths (D)

Mortality Data

- Annual (1969-2019) data on
  - Exposure (E)
  - Deaths (D)
  - Mortality rates (M)

Mortality Data

- Annual (1969-2019) data on
  - Exposure (E)
  - Deaths (D)
  - Mortality rates (M)
- For US as a whole and 9 census regions.

Mortality Data

- Annual (1969-2019) data on
  - Exposure (E)
  - Deaths (D)
  - Mortality rates (M)
- For US as a whole and 9 census regions.
- E and D respect aggregation constraints.

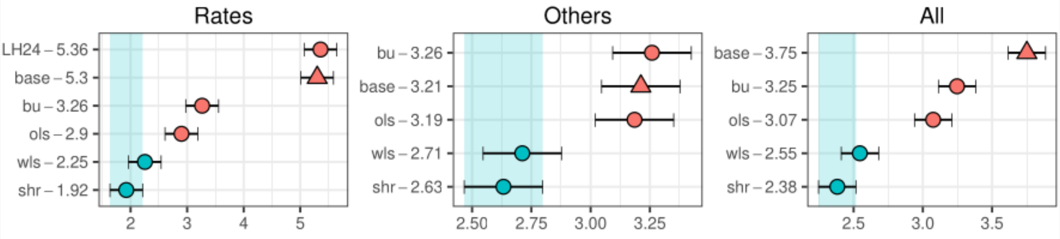
Mortality Data

- Annual (1969-2019) data on
  - Exposure (E)
  - Deaths (D)
  - Mortality rates (M)
- For US as a whole and 9 census regions.
- E and D respect aggregation constraints.
- M need not respect hierarchical constraints.

Mortality Data

- Annual (1969-2019) data on
  - Exposure (E)
  - Deaths (D)
  - Mortality rates (M)
- For US as a whole and 9 census regions.
- E and D respect aggregation constraints.
- M need not respect hierarchical constraints.
- However $M = D/E$ for each region.

Mortality Data



Outline

Setup

- Suppose we are interested in multivariate forecasting but do not have linear (or non-linear) constraints.

Setup

- Suppose we are interested in multivariate forecasting but do not have linear (or non-linear) constraints.
- Is there anything interesting about forecast reconciliation?

Setup

- Suppose we are interested in multivariate forecasting but do not have linear (or non-linear) constraints.
- Is there anything interesting about forecast reconciliation?
- Surprisingly... Yes!

Setup

- Suppose we are interested in multivariate forecasting but do not have linear (or non-linear) constraints.
- Is there anything interesting about forecast reconciliation?
- Surprisingly... Yes!
- New work on Forecast Linear Augmented Projection (FLAP)

What is FLAP?

- Suppose the target is to forecast $\mathbf{y} \in \mathbb{R}^m$.

What is FLAP?

- Suppose the target is to forecast $\mathbf{y} \in \mathbb{R}^m$.
- We construct new synthetic series $\mathbf{c} \in \mathbb{R}^p$ where $\mathbf{c} = \mathbf{\Phi}\mathbf{y}$.

What is FLAP?

- Suppose the target is to forecast $\mathbf{y} \in \mathbb{R}^m$.
- We construct new synthetic series $\mathbf{c} \in \mathbb{R}^p$ where $\mathbf{c} = \mathbf{\Phi y}$.
  - The choice of $\mathbf{\Phi}$ is arbitrary.

What is FLAP?

- Suppose the target is to forecast $\mathbf{y} \in \mathbb{R}^m$.
- We construct new synthetic series $\mathbf{c} \in \mathbb{R}^p$ where $\mathbf{c} = \mathbf{\Phi y}$.
  - The choice of $\mathbf{\Phi}$ is arbitrary.
- The augmented vector $(\mathbf{c}', \mathbf{y}')'$ coheres to known linear constraints.

What is FLAP?

- Suppose the target is to forecast $\mathbf{y} \in \mathbb{R}^m$.
- We construct new synthetic series $\mathbf{c} \in \mathbb{R}^p$ where $\mathbf{c} = \boldsymbol{\Phi}\mathbf{y}$.
  - The choice of $\boldsymbol{\Phi}$ is arbitrary.
- The augmented vector $(\mathbf{c}', \mathbf{y}')'$ coheres to known linear constraints.
- Forecast all components of $(\mathbf{c}', \mathbf{y}')'$

What is FLAP?

- Suppose the target is to forecast $\mathbf{y} \in \mathbb{R}^m$.
- We construct new synthetic series $\mathbf{c} \in \mathbb{R}^p$ where $\mathbf{c} = \mathbf{\Phi y}$.
  - The choice of $\mathbf{\Phi}$ is arbitrary.
- The augmented vector $(\mathbf{c}', \mathbf{y}')'$ coheres to known linear constraints.
- Forecast all components of $(\mathbf{c}', \mathbf{y}')'$
- Reconcile using MinT.

The key idea

- It is known from the properties of MinT that we will reduce forecast variance for $(\mathbf{c}', \mathbf{y}')'$.

The key idea

- It is known from the properties of MinT that we will reduce forecast variance for $(\mathbf{c}', \mathbf{y}')'$.
- It has been proven that the same is true when only looking at a subset of the hierarchy (in our case, $\mathbf{y}$).

The key idea

- It is known from the properties of MinT that we will reduce forecast variance for $(\mathbf{c}', \mathbf{y}')'$.
- It has been proven that the same is true when only looking at a subset of the hierarchy (in our case, $\mathbf{y}$).
- This result allows the benefits of reconciliation to be applied to problems where there are no constraints at all!

The key idea

- It is known from the properties of MinT that we will reduce forecast variance for $(\mathbf{c}', \mathbf{y}')'$.
- It has been proven that the same is true when only looking at a subset of the hierarchy (in our case, $\mathbf{y}$).
- This result allows the benefits of reconciliation to be applied to problems where there are no constraints at all!
- We also prove that the forecast variance in non-increasing as more synthetic components are added.

No free lunch

- Originally FLAP stood for ‘Free Lunch’ augmented projection.

No free lunch

- Originally FLAP stood for ‘Free Lunch’ augmented projection.
- All proofs assume error covariance matrix used in MinT is **known**. In practice it is estimated.

No free lunch

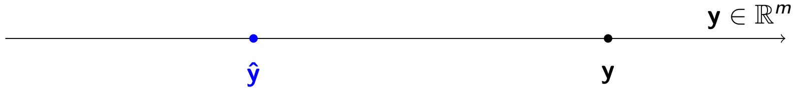
- Originally FLAP stood for ‘Free Lunch’ augmented projection.
- All proofs assume error covariance matrix used in MinT is **known**. In practice it is estimated.
- The quality of covariance matrix estimates deteriorate with higher dimension.

No free lunch

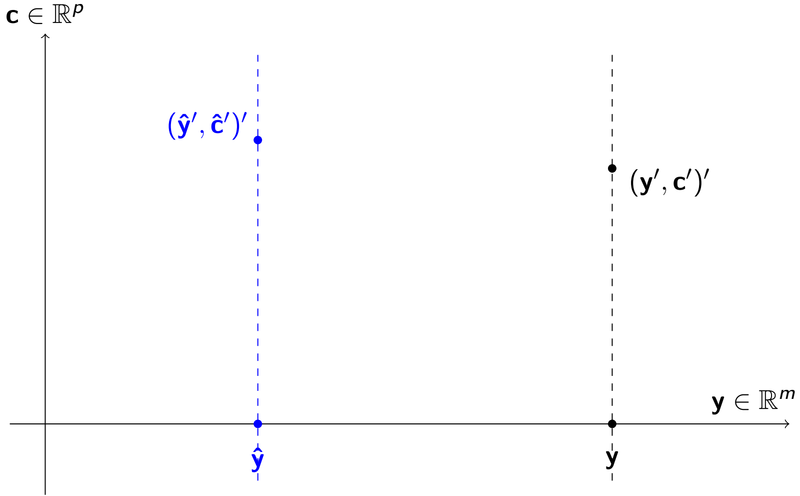
- Originally FLAP stood for ‘Free Lunch’ augmented projection.
- All proofs assume error covariance matrix used in MinT is **known**. In practice it is estimated.
- The quality of covariance matrix estimates deteriorate with higher dimension.
- However for finite dimension, the benefit of FLAP outweighs errors in estimating covariance matrix.

Geometry of FLAP

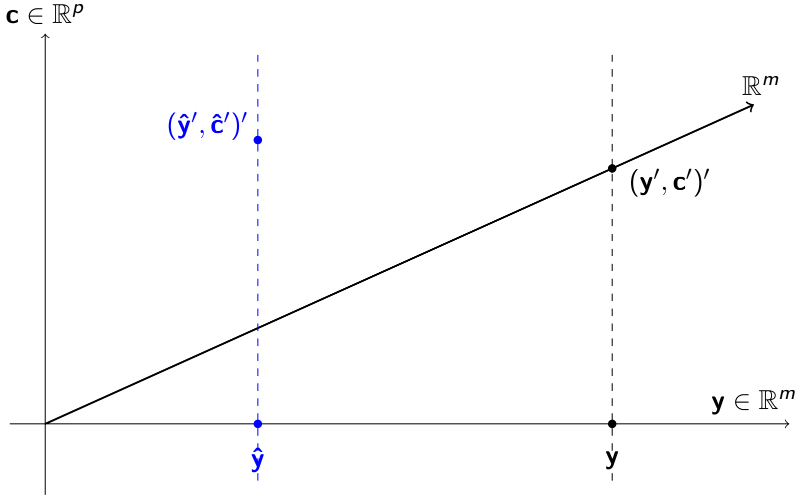
Geometry of FLAP



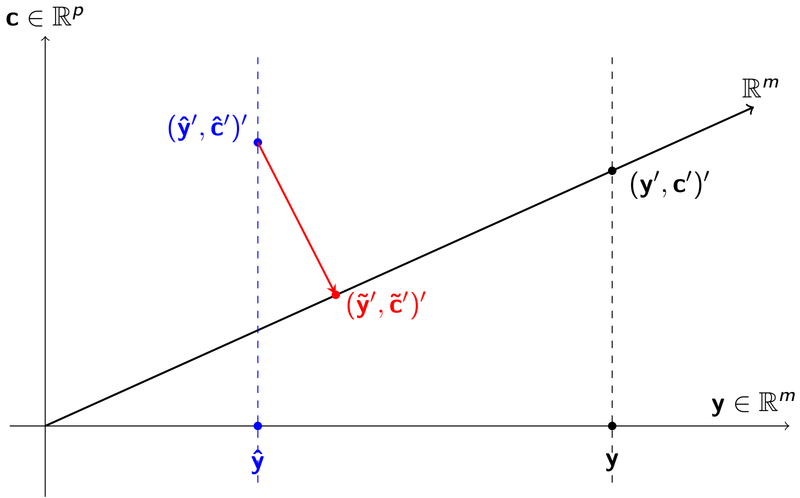
Geometry of FLAP



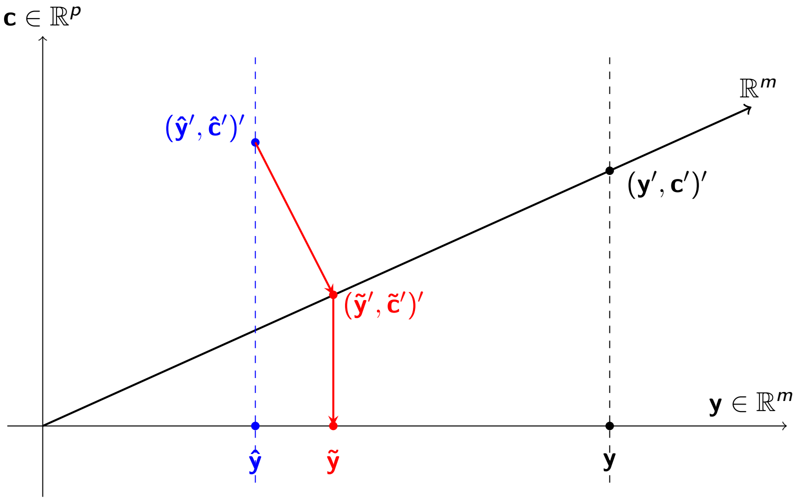
Geometry of FLAP



Geometry of FLAP



Geometry of FLAP



FRED-MD

- Monthly data of macroeconomic variables (McCracken and Ng, 2016).

FRED-MD

- Monthly data of macroeconomic variables (McCracken and Ng, 2016).
- Data from Jan 1959 – Sep 2023. 777 observations on 122 series.

FRED-MD

- Monthly data of macroeconomic variables (McCracken and Ng, 2016).
- Data from Jan 1959 – Sep 2023. 777 observations on 122 series.
- Expanding time series validation with

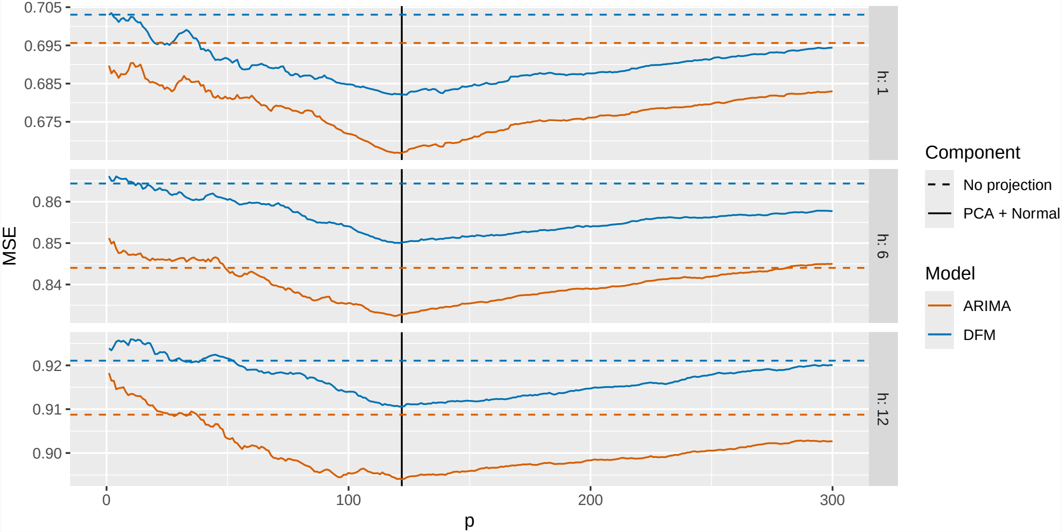
FRED-MD

- Monthly data of macroeconomic variables (McCracken and Ng, 2016).
- Data from Jan 1959 – Sep 2023. 777 observations on 122 series.
- Expanding time series validation with
    - Initial sample size of 25 years and

FRED-MD

- Monthly data of macroeconomic variables (McCracken and Ng, 2016).
- Data from Jan 1959 – Sep 2023. 777 observations on 122 series.
- Expanding time series validation with
  - Initial sample size of 25 years and
  - Forecast horizon up to 12 months.

FRED-MD



0.705
0.695
0.685
0.675
0.86
0.85
0.84
0.92
0.91
0.90
MSE
h: 1
h: 6
h: 12
0
100
200
300
p
Component
No projection
PCA + Normal
Model
ARIMA
DFM

Working Paper and R Package

YF Yang, G Athanasopoulos, RJ Hyndman, and A Panagiotelis (2024). “Forecast Linear Augmented Projection (FLAP): A free lunch to reduce forecast error variance”.

Department of Econometrics and Business Statistics, Monash University, Working Paper Series 13/24.

You can install the stable version from CRAN

```r
## CRAN.R-project.org/package=flap
install.packages("flap")
```

or the development version from Github

```r
## github.com/FinYang/flap
# install.packages("remotes")
remotes::install_github("FinYang/flap")
```

Outline

Final thoughts

- Sometimes understanding the same problem in a different way opens new doors in research.

Final thoughts

- Sometimes understanding the same problem in a different way opens new doors in research.
- Theory, methodology and application all matter. The connections and feedback loops between them are important.

Final thoughts

- Sometimes understanding the same problem in a different way opens new doors in research.
- Theory, methodology and application all matter. The connections and feedback loops between them are important.
  - Do not neglect any of these!

Final thoughts

- Sometimes understanding the same problem in a different way opens new doors in research.
- Theory, methodology and application all matter. The connections and feedback loops between them are important.
  - Do not neglect any of these!
- Work with the right people.

The right people









Forecast reconciliation

- Forecast reconciliation is a practical and interesting problem with many open questions.

Forecast reconciliation

- Forecast reconciliation is a practical and interesting problem with many open questions.
  - How can we guarantee that reconciled probabilistic forecasts are correctly calibrated?

Forecast reconciliation

- Forecast reconciliation is a practical and interesting problem with many open questions.
  - How can we guarantee that reconciled probabilistic forecasts are correctly calibrated?
  - Multi-objective optimization aspects of reconciliation problem.

Forecast reconciliation

- Forecast reconciliation is a practical and interesting problem with many open questions.
  - How can we guarantee that reconciled probabilistic forecasts are correctly calibrated?
  - Multi-objective optimization aspects of reconciliation problem.
  - Other loss functions?

Forecast reconciliation

- Forecast reconciliation is a practical and interesting problem with many open questions.
  - How can we guarantee that reconciled probabilistic forecasts are correctly calibrated?
  - Multi-objective optimization aspects of reconciliation problem.
  - Other loss functions?
- Jump on the bandwagon!

Links

Postdoc opportunity

Link to slides