Probabilistic reconciliation of hierarchical forecasts

Lorenzo Zambon

PhD Spring Workshop 2022

16/03/2022

Università degli Studi di Pavia

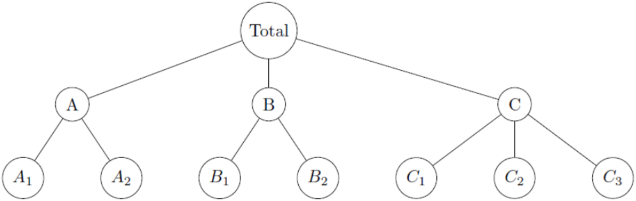
Università della Svizzera Italiana

Index

Reconciliation of hierarchical forecasts

Hierarchical forecasts

We want to predict several quantities organized in a certain structure:



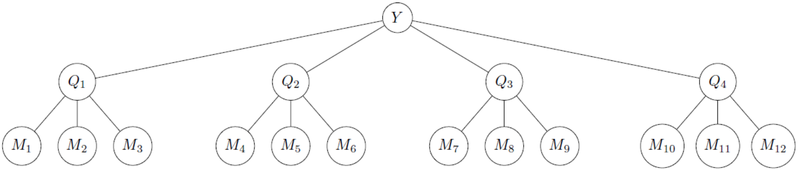
Often: **Time Series** forecasting

Time Series: sequence of data taken at equally spaced points in time
For instance: daily sales of a given product, monthly rainfall, yearly GDP of a country...

- geographical hierarchies
- temporal hierarchies

Temporal Hierarchy

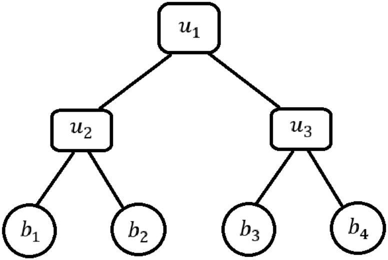
Consider a **monthly** time series: we want to predict the next 12 values
We may consider the **quarterly** or **yearly** time series obtained by aggregating monthly values!



- Use some model to obtain base forecasts for monthly, quarterly, and yearly TS
- Forecasts should be **coherent**: e.g., the sum of the forecasts for the first 3 months should be equal to the forecast for the first quarter
- **Reconciliation** is used to get coherent forecasts
- Reconciliation methods have been shown to improve the accuracy over base forecasts (Athanasopoulos et al. [2017])

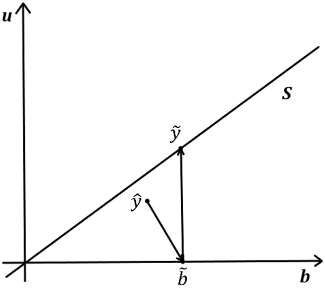
Notation

- **Bottom** observations: $b = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix} \in \mathbb{R}^m$
- **Upper** observations: $u = \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix} \in \mathbb{R}^n$
- **Hierarchy**: $\boldsymbol{y} = \boldsymbol{S}\boldsymbol{b}$, where $y = \begin{bmatrix} u \\ b \end{bmatrix}$, $S = \begin{bmatrix} A \\ \hdashline I_m \end{bmatrix} \in \mathbb{R}^{(n+m)\times m}$



$$S = \left[\begin{array}{cccc} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ \hdashline 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right]$$

Point reconciliation

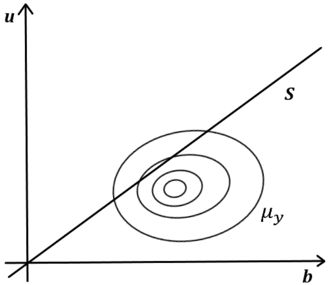


- Point forecast: $\hat{y} = \begin{bmatrix} \hat{u} \\ \hat{b} \end{bmatrix}$
- If $\hat{y} \notin \mathcal{S} := \{y : \; y = Sb\}$
  $\rightarrow$ the prediction is **incoherent**
- $\tilde{b} = G\hat{y}$, for some $G \in R^{m \times (n+m)}$
- $\tilde{y} = S\tilde{b}$

How to choose G?

- Bottom-Up: $G = \left[\underline{0} \;\vdots\; I_m\right]$
- MinT: $G = \left(S^T W^{-1} S\right)^{-1} \left(S^T W^{-1}\right)$, where $W = \text{cov}(y - \hat{y})$

Probabilistic Reconciliation



- Forecasts are given by a probability distribution $\mu_y \in \mathcal{P}(\mathbb{R}^{n+m})$, called **base distribution**, rather then a point
- The aim is to find a **reconciled distribution** $\tilde{\mu}_y \in \mathcal{P}(\mathcal{S})$

- Panagiotelis et al. [2020]: given a map $\psi : \mathbb{R}^{n+m} \to \mathcal{S}$, e.g. $\psi(\hat{y}) = SG\hat{y}$, define the reconciled distribution $\tilde{\mu}_y \in \mathcal{P}(\mathcal{S})$ as $\tilde{\mu}_y = \psi_{\#}\mu_y$
- Rangapuram et al. [2021]: coherence is imposed during training by L^2-projecting samples on the subspace $\mathcal{S}$
- Corani et al. [2021]: analytically compute the reconciled distr. in the Gaussian case

Reconciled distribution

Suppose that μ_y is absolutely continuous w.r.t. Lebesgue $\rightarrow$ π_y **density** of μ_y
i.e. $\mu_y(B) = \int_B \pi_y(y)dy$ for any $B \in \mathcal{B}(\mathbb{R}^{n+m})$
If μ_y is discrete: use probability mass function instead of density

Intuitively: I should only look at the probabilities of the points on $\mathcal{S}$
Since $\mathcal{S} \sim \mathbb{R}^m$ through the map $b \rightarrow Sb \implies$ we only focus on b

Define $\tilde{\pi}(b) = \pi(b \mid u = Ab) = \frac{\pi_y(Ab,b)}{\int_{\mathbb{R}^m} \pi_y(Ax,x)dx} \propto \pi_y(Ab, b)$

If u and b independent $\implies \tilde{\pi}(b) \propto \pi_u(Ab)\pi_b(b)$

How to sample from $\tilde{\pi}(b)$?

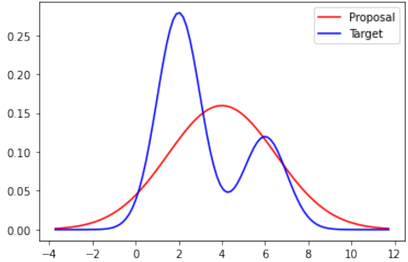
- Markov Chain Monte Carlo
- Importance Sampling

Reconciliation using Importance Sampling

Main idea of IS

Given a **target** distribution q and a function g, I want to compute $\mathbb{E}_{Y\sim q}\left[g(Y)\right]$

If I was able to draw $y_1, \ldots, y_N \overset{\text{IID}}{\sim} q \implies \mathbb{E}\left[g(Y)\right] \approx \frac{1}{N}\sum_{i=1}^N g(y_i)$



- Fix a **proposal** distribution p
- Draw $z_1, \ldots, z_N \overset{\text{IID}}{\sim} p$ and $\forall i = 1, \ldots, N$ compute $w_i := \frac{\pi_q(z_i)}{\pi_p(z_i)}$
- Then: $\mathbb{E}\left[g(Y)\right] \approx \frac{1}{N}\sum_{i=1}^N g(z_i)\cdot w_i$

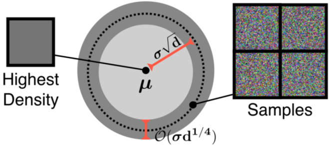
Indeed:

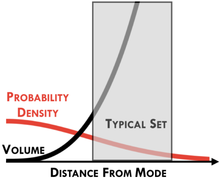
$$\mathbb{E}_{Y\sim q}\left[g(Y)\right] = \int g(y)\pi_q(y)dy = \int g(y)\frac{\pi_q(y)}{\pi_p(y)}\pi_p(y)dy = \mathbb{E}_{Z\sim p}\left[g(Z)w(Z)\right],$$

where $w(z) := \frac{\pi_q(z)}{\pi_p(z)}$

Curse of dimensionality

Target distribution q must be absolutely continuous w.r.t. the proposal distribution p, i.e. $\pi_p > 0$ where $\pi_q > 0$. It is crucial to choose the right proposal!



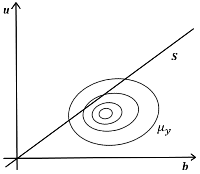


Curse of dimensionality:

- As the dimension of the space increases, it gets harder to find good proposals, i.e. good approximations of the target distribution
- The performance of IS typically decreases exponentially
- In high dimensions, the mass is concentrated in a small proportion of the space!
- The effective sample size, defined as
$$\text{ESS} := \frac{\left(\sum_i w_i\right)^2}{\sum_i w_i^2},$$ drops to 1

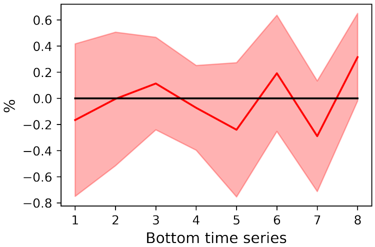
Reconciliation using IS

- Target distribution: $\tilde{\pi}(b) \propto \pi_u(Ab)\pi_b(b)$
- Proposal distribution: $\pi_b(b)$
- (Unnormalized) weights: $w_i := \pi_u(Ab_i)$



If all the distributions are Gaussian
$\implies$ $\tilde{\pi}(b)$ can be analytically computed
Corani et al. [2021]

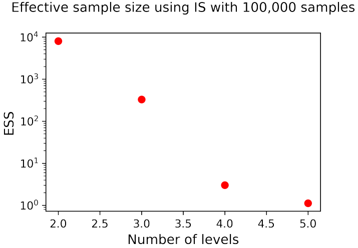
Percentage error using IS on the posterior means

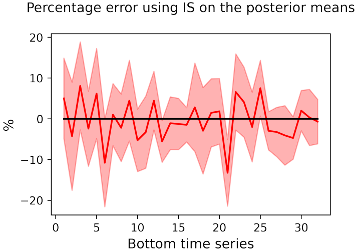


- 3 levels, 8 bottom nodes, 7 upper nodes
- 100,000 samples drawn, repeated for 30 times to compute 95% C.I.
- less than 0.1 seconds for 100,000 samples

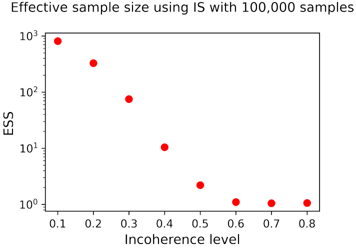
Reconciliation using IS

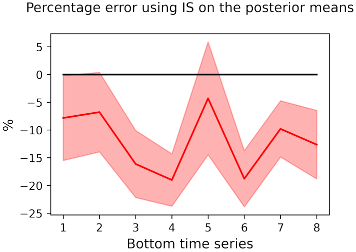
If the **size of the hierarchy** grows:





If the **incoherence level** grows:





Bottom-Up Importance Sampling

Bottom-Up Importance Sampling

Hence, when we have big hierarchies or a large incoherence, IS is practically unusable!

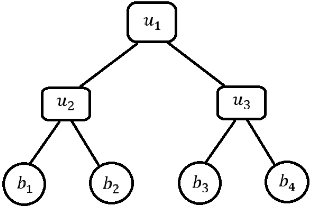
Suppose that:

- The hierarchy is given by a **tree**: each node only has one parent
- All the forecasts are independent: $\pi_y(y) = \pi_y(u, b) = \pi_{u_1}(u_1) \cdot \dots \cdot \pi_{b_m}(b_m)$

In this framework, we propose the **Bottom-Up Importance Sampling** algorithm:

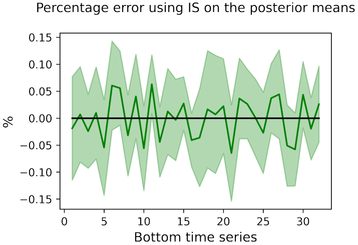
- The idea is to split a single $n-$dimensional Importance Sampling task into n one-dimensional IS tasks
- Reconciliation is performed by iteratively condition on each upper observation, from the bottom to the top
- Curse of dimensionality is deeply mitigated
- Works even if the distributions are only available through samples

BUIS algorithm

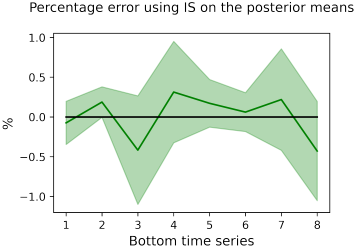


- Sample $\left(b_1^{(i)}, b_2^{(i)}, b_3^{(i)}, b_3^{(i)}\right)_{i=1,\dots,N}$ from π_b
- Compute weights: $w_i^{(u_2)} = \pi_{u_2}(b_1^{(i)} + b_2^{(i)})$, $w_i^{(u_3)} = \pi_{u_3}(b_3^{(i)} + b_4^{(i)})$
- Resample with replacement from the weighted sample $\left((b_1^{(i)}, b_2^{(i)}), w_i^{(u_2)}\right)_i$ to get $\left(b_1^{(j)}, b_2^{(j)}\right)_{j=1,\dots,N}$ (same for b_3 and b_4)
- Compute weights $w_j^{(u_1)} = \pi_{u_1}(b_1^{(j)} + b_2^{(j)} + b_3^{(j)} + b_4^{(j)})$
- Resample with replacement from $\left((b_1^{(j)}, b_2^{(j)}, b_3^{(j)}, b_4^{(j)}), w_j^{(u_1)}\right)_j$ to get an unweighted sample from $\tilde{\pi}(b)$

Gaussian distributions



Hierarchy levels: 5, incoherence: 0.2



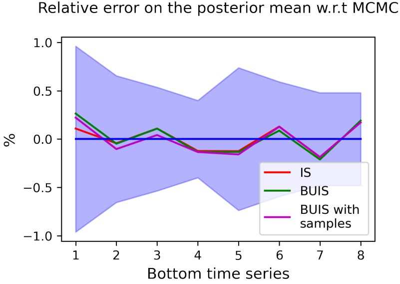
Hierarchy levels: 3, incoherence: 0.8

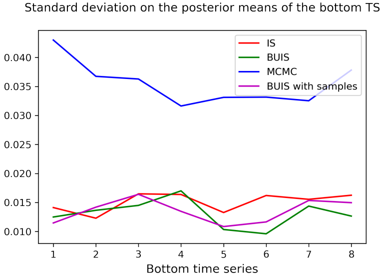
		Average absolute error on the bottom means	
N. levels	inc.	IS	BUIS
3	0.2	0.17%	0.04%
3	0.8	11.92%	0.23%
5	0.2	3.98%	0.03%

Poisson distributions

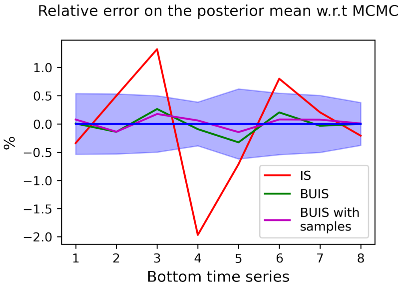
We run 30 experiments in the discrete setting, with two levels of incoherence:

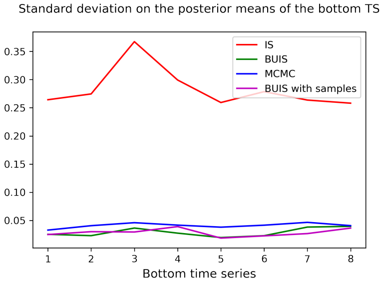
- All the distributions are Poisson
- Hierarchy: 3 levels, 8 bottom nodes, 7 upper nodes; incoherence: 0.2, 0.8
- We test IS and BUIS, using both the analytic pmf and samples, with $100,000$ samples
- No analytical solutions $\rightarrow$ we compare with the results obtained using MCMC (4 chains, 5000 samples each)





Poisson distributions





		Average time			
N. levels	inc.	IS	BUIS	BUIS w/ samples	MCMC
3	0.2	0.16 s	0.31 s	1.63 s	138.6 s
3	0.8	0.16 s	0.31 s	1.63 s	114.1 s

Temporal reconciliation using BUIS

Dataset: 22 monthly time series from the *Campy* dataset
Model: GLM (*tscount* package); forecasts are in the form of samples
Hierarchy:

- 12 monthly observations
- 6 bi-monthly observations, 4 quarterly observations, 3 four-monthly observations, 2 biannual observations, 1 annual observation

Not a tree! We use BUIS on the largest tree, then IS on the remaining constraints

Given $\alpha \in (0, 1)$, we denote by l and u the lower and upper bounds of the $(1 - \alpha)$ interval of the forecast distribution. The Mean Interval Score is defined as

$$\mathrm{MIS}(l, u, y, \alpha) := (u - l) + \frac{2}{\alpha}(l - y)\mathbb{1}_{\{y<l\}} + \frac{2}{\alpha}(y - u)\mathbb{1}_{\{y>u\}}$$

	Average Scaled MIS			
α	base	Gauss + NegBin	NegBin	samples
0.05	8.09	7.70	8.09	8.32
0.1	6.73	5.72	5.99	6.19
0.33	4.18	3.42	3.56	3.77

References

G. Athanasopoulos, R. J Hyndman, N. Kourentzes, and F. Petropoulos. Forecasting with temporal hierarchies. *European Journal of Operational Research*, 262(1):60–74, 2017. ISSN 0377-2217.

G. Corani, D. Azzimonti, J. P. S. C. Augusto, and M. Zaffalon. Probabilistic reconciliation of hierarchical forecast via bayes' rule. *Lecture Notes in Computer Science*, page 211–226, 2021.

A. Panagiotelis, P. Gamakumara, G. Athanasopoulos, and R. J Hyndman. Probabilistic forecast reconciliation: Properties, evaluation and score optimisation. *Working Paper No. 26/20*, 2020. URL `https://www.monash.edu/business/ebs/research/publications/ebs/wp26-2020.pdf`.

S. S. Rangapuram, L. D. Werner, K. Benidis, P. Mercado, J. Gasthaus, and T. Januschowski. End-to-end learning of coherent probabilistic forecasts for hierarchical time series. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Thank you for your attention!