Università degli Studi di Pavia
Università della Svizzera Italiana

Joint PhD program in Computational Mathematics and
Decision Sciences
XXXV cycle

Probabilistic forecast reconciliation: theory, algorithm, and applications

Advisors:
Prof. Stefano GUALANDI
Prof. Giorgio CORANI

PhD Dissertation of:
Lorenzo Zambon

Academic year 2021-2022

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Often time series are organized into a hierarchy. For example, the total visitors of a country can be divided into regions and the visitors of each region can be further divided into sub-regions. Hierarchical time series are common in several fields, such as retail sales (Makridakis et al. 2021) or electricity demand (Taieb et al. 2021). Hierarchical forecasts should be *coherent*. For instance, the sum of the forecasts for the sub-regions should match the forecast for the entire region. However, the forecasts produced independently for each time series (*base forecasts*) do not generally satisfy the summing constraints; they are hence *incoherent*.

Reconciliation algorithms (Hyndman et al. 2011; Wickramasuriya et al. 2019) adjust the incoherent base forecasts, making them coherent. In the process, they generally improve the accuracy compared to the base forecasts (Athanasopoulos et al. 2020). Indeed, forecast reconciliation has recently been reinterpreted as forecast combination (Hollyman et al. 2021; Di Fonzo and Girolimetto 2022). Reconciliation is particularly important for temporal hierarchies (Athanasopoulos et al. 2017; Kourentzes and Athanasopoulos 2021), in which forecasts are produced for the same variable at different temporal scales. For instance, reconciliation can be used to enforce coherence between monthly, quarterly, and yearly forecasts.

Most literature focuses on the reconciliation of the *point forecasts* (Hyndman et al. 2011; Wickramasuriya et al. 2019; Wickramasuriya et al. 2020; Di Fonzo and Girolimetto 2023). However, to support decision making we need to provide the entire reconciled predictive distribution, not only the reconciled point forecasts. Two algorithms for probabilistic reconciliation are proposed by Jeon et al. 2019 and Taieb et al. 2021; however, they have a number of shortcomings, as explained by Panagiotelis et al. 2022. In particular, little formal justification is provided for the algorithms, which are tailored towards specific applications. Recent applications of probabilistic

reconciliation include solar energy forecasting (Yang 2020). In Corani et al. 2020, the reconciled distribution is obtained through a Bayesian approach, but only under the Gaussian assumption. Wickramasuriya 2021 also focuses on the Gaussian case. Panagiotelis et al. 2022 formally defines probabilistic reconciliation as a projection. The parameters of the projection are optimized through Stochastic Gradient Descent in order to minimize a chosen scoring rule. A limit of this approach is that it does not scale to large hierarchies. Moreover, it does not deal with discrete distributions, and thus it cannot treat count time series, which are very common (Kolassa 2016).

In Chapter 2, we provide a general definition of coherence for probabilistic forecasts, which applies to discrete and continuous distributions. Then, we propose a concept of probabilistic reconciliation based on conditioning. It is rather general, as it can be applied to continuous and count time series. A similar approach has been independently developed by Corani et al. 2022. However, the underlying Markov Chain Monte Carlo (MCMC) algorithm does not scale well to large hierarchies. Our approach to sample from the reconciled distribution is based on importance sampling (IS). Yet, vanilla IS is not effective to sample from high dimensional distributions; this prevents using it to reconcile large hierarchies. Moreover, we show that a large incoherence is connected to a low performance of IS. The numerical experiments confirm that IS is not robust with respect to the hierarchy size and the incoherence level.

In Chapter 3, we thus propose a new algorithm, which we call Bottom-Up Importance Sampling (BUIS). This algorithm allows to efficiently sample from the reconciled distribution, with a speedup of up to three orders of magnitude compared to the method of Corani et al. 2022. This is possible since BUIS is based on importance sampling (IS), which works in parallel and not sequentially as MCMC. Our algorithm is able to overcome the drawbacks of vanilla IS. Moreover, BUIS can be used even if the base forecast distribution is only available through samples. We also provide a formal proof of convergence of BUIS to the actual distribution. A current limit of this algorithm is that it assumes the conditional independence of the base forecasts. We leave for future work the extension of the algorithm to manage also the correlation between forecasts. We run several experiments on synthetic data, which show that BUIS is able to efficiently sample from the reconciled distribution, even in the case of big hierarchies or large incoherence levels. Finally, we test our method exhaustively on time series extracted from different data sets, providing a clear improvement in the performance of the probabilistic forecasts.

In Chapter 4, we study the effects of reconciliation on the forecast dis-

tribution. In the Gaussian case, where the reconciled distribution can be obtained analytically, the reconciled mean of a variable is a compromise between the base mean of that variable and a linear combination, according to the hierarchy, of the base means of the other variables. For instance, the reconciled mean of the upper variable is a convex combination of the base and the bottom-up mean. Moreover, the variance of the forecast distribution of each variable decreases after reconciliation. However, in general, especially for count distributions, this may not be true: if there is a large incoherence, the variance may increase. On the other hand, when we deal with asymmetric distributions, a small incoherence may lead to a negative shift on the mean of both the bottom and the upper variables. We illustrate this point both from a theoretical viewpoint, and using some examples with Bernoulli and Poisson distributions. Then, we present an application to count time series of extreme events on the Credit Default Swap (CDS) market. The probabilistic forecasts are computed using a multivariate negative binomial score-driven model, proposed by Agosto 2022. We efficiently reconcile all the 3508 daily probabilistic forecasts, achieving a clear improvement in the performance of the forecasts, and observing the effects of the reconciliation discussed above.

In Chapter 5, we introduce and study the p-Fourier Discrepancy Functions, a new family of metrics for comparing discrete probability measures. Discrepancies are important tools for every task that requires the comparison of probability measures, such as assessing the performance of probabilistic forecasts. The performance of probabilistic forecasts is typically evaluated using a scoring rule, which is a function that takes as arguments a probability measure and a realization. A scoring rule K should be proper, i.e. $\mathbb{E}_{x\sim Q}[K(Q,x)] \leq \mathbb{E}_{x\sim Q}[K(P,x)]$, where Q is the forecast distribution, x the realization, and P any other probability measure. Proper scoring rules provide performance measures that address calibration and sharpness simultaneously (Gneiting et al. 2008). In the univariate setting, the most common scoring rule is arguably the rank probability score (RPS), which is defined as the l^2-distance between the predictive cumulative distribution function (CDF) and the true CDF, i.e. the step function at the true value of the time series (Kolassa 2016). Hierarchical forecasting is, however, a multivariate problem. A generalization of the RPS to the multivariate setting is the energy score (ES), which is a proper scoring rule (Gneiting et al. 2008), unlike other common multivariate scoring rules as the logarithmic score (Panagiotelis et al. 2022). ES is based on the energy distance, which can be expressed as a Fourier-based metric (Székely and Rizzo 2013). In this chapter, we introduce the Fourier Discrepancies, a discretized version of the χ_r-metrics (Rachev 1991). We show that the Fourier Discrepancies can be

expressed as the square root of a bilinear form induced by a positive definite matrix, hence they are 1-homogeneous and convex. We also prove that the squared Fourier Discrepancy is twice differentiable and that both its gradient and Hessian have an explicit formula. Finally, we study the lower and upper tight bounds of the Fourier Discrepancy in terms of the Total Variation distance.

Conclusions and future work are placed in Chapter 6. For the sake of clarity, we only report the essential proofs in the body of the thesis and leave the others in the appendix.

Chapter 2

Probabilistic reconciliation

This chapter is organized as follows. In Section 2.1, we set our notation and briefly recall temporal hierarchies and point reconciliation. In Section 2.2, we provide a general definition of coherence for probabilistic forecasts. We then obtain the expression of the reconciled distribution through conditioning, and we compare it to the existing literature. For the sake of clarity, in Section 2.3 we analytically derive the reconciled distribution in the Gaussian case. Our approach to sample from the reconciled distribution, based on importance sampling, is shown in Section 2.4, where we also highlight the issues that arise when dealing with large hierarchies or high incoherence between the base forecasts.

2.1 Notation

Consider the hierarchy of Figure 2.1. We denote by $\mathbf{b} = [b_1, \dots, b_m]^T$ the vector of bottom variables, and by $\mathbf{u} = [u_1, \dots, u_{n-m}]^T$ the vector of upper variables. We then denote by

$$\mathbf{y} = \begin{bmatrix} \mathbf{u} \\ \mathbf{b} \end{bmatrix} \in \mathbb{R}^n$$

the vector of all the variables. The hierarchy may be expressed as a set of linear constraints:

$$\mathbf{y} = \mathbf{S}\mathbf{b}, \text{ where } \mathbf{S} = \begin{bmatrix} \mathbf{A} \\ \hdashline \mathbf{I} \end{bmatrix}. \tag{2.1}$$

Here, $\mathbf{I} \in \mathbb{R}^{m \times m}$ is the identity matrix. $\mathbf{S} \in \mathbb{R}^{n \times m}$ is called *summing matrix*, while $\mathbf{A} \in \mathbb{R}^{(n-m) \times m}$ is called *aggregating matrix*. The constraints can thus be written as $\mathbf{u} = \mathbf{A}\mathbf{b}$. We use a graphical tree-like representation for the hierarchy, where each node is the sum of its children, as in Hyndman and

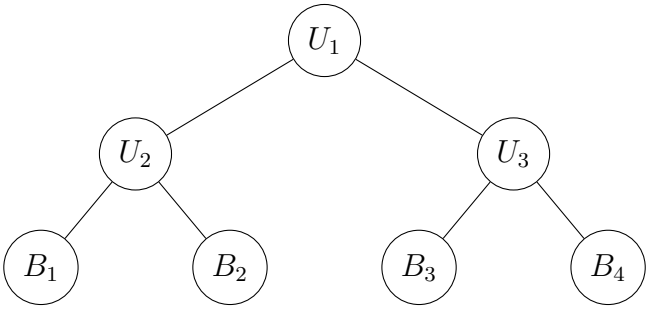


Figure 2.1: A simple hierarchy with 4 *bottom and* 3 *upper variables*

Athanasopoulos 2021, Chapter 11. For example, the aggregating matrix of the hierarchy in Figure 2.1 is given by

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}.$$

A point $\mathbf{y} \in \mathbb{R}^n$ is said to be *coherent* if it satisfies the constraints given by the hierarchy. We denote by $\mathcal{S}$ the set of coherent points, which is a linear subspace of $\mathbb{R}^n$:

$$\mathcal{S} := \{\mathbf{y} \in \mathbb{R}^n : \mathbf{y} = \mathbf{S}\mathbf{b}\}. \tag{2.2}$$

2.1.1 Temporal hierarchies

In temporal hierarchies (Athanasopoulos et al. 2017; Kourentzes and Athanasopoulos 2021), forecasts are generated for the same variable at different temporal scales. For instance, a quarterly time series may be aggregated to obtain semi-annual and annual series. If we are interested in predictions up to one year ahead, we compute the four quarterly forecasts $\widehat{q}_1, \widehat{q}_2, \widehat{q}_3, \widehat{q}_4$, the two semi-annual forecasts $\widehat{s}_1, \widehat{s}_2$, and the annual forecast $\widehat{a}_1$. We then obtain the hierarchy in Figure 2.1, with $\mathbf{b} = [\widehat{q}_1, \widehat{q}_2, \widehat{q}_3, \widehat{q}_4]^T$ and $\mathbf{u} = [\widehat{a}_1, \widehat{s}_1, \widehat{s}_2]^T$. The base forecasts independently computed at different frequencies are incoherent: for example, the quarterly predictions do not sum up to the annual prediction. Reconciliation adjusts the base forecasts, enforcing coherence.

2.1.2 Point forecasts reconciliation

Let us now denote by $\widehat{\mathbf{y}} = \left[\widehat{\mathbf{u}}^T \,|\, \widehat{\mathbf{b}}^T\right]^T$ the vector of the base (incoherent) forecasts. In the literature, point reconciliation is typically presented as a two-step process (Hyndman et al. 2011; Panagiotelis et al. 2021). First, the

reconciled bottom forecasts are computed by combining the base forecasts of the whole hierarchy:

$$\widetilde{\mathbf{b}} = \mathbf{G}\widehat{\mathbf{y}},$$

for some matrix $\mathbf{G} \in \mathbb{R}^{m \times n}$. Then, the reconciled forecasts are obtained as

$$\widetilde{\mathbf{y}} = \mathbf{S}\widetilde{\mathbf{b}},$$

where $\mathbf{S}$ is the summing matrix. Hence, $\widetilde{\mathbf{y}}$ is coherent by design. For example, if we set $\mathbf{G} = [\mathbf{I} \mid \mathbf{0}]$, we have the bottom-up approach, which sums up the bottom forecasts, ignoring the base forecasts of the upper variables (Hyndman and Athanasopoulos 2021, Chapter 11.2). In the minT method (Wickramasuriya et al. 2019), $\mathbf{G}$ is defined as

$$\mathbf{G} = (\mathbf{S}^T\mathbf{W}^{-1}\mathbf{S})^{-1}\mathbf{S}^T\mathbf{W}^{-1},$$

where $\mathbf{W}$ is the covariance matrix of the errors of the base forecasts. This method minimizes the trace of the covariance matrix of the reconciled forecasts, under the assumption that the base forecasts are unbiased.

2.1.3 Probabilistic framework

In many cases, decision making requires an estimate of the uncertainty of the predictions (Gneiting and Katzfuss 2014). This requires a probabilistic framework, in which forecasts are in the form of probability distributions. We denote by $\widehat{\nu} \in \mathcal{P}(\mathbb{R}^n)$ the forecast distribution for $\mathbf{y}$, where $\mathcal{P}(\mathbb{R}^n)$ is the space of probability measures on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, and $\mathcal{B}(\mathbb{R}^n)$ is the Borel σ-algebra on $\mathbb{R}^n$. Moreover, we denote by $\widehat{\nu}_u$ and $\widehat{\nu}_b$ the marginal base forecast distributions of, respectively, the upper and the bottom components of $\mathbf{y}$.

The forecast distribution $\widehat{\nu}$ may be either discrete or absolutely continuous. In the following, if there is no ambiguity, we will use $\widehat{\pi}$ to denote either its probability mass function, in the former case, or its density, in the latter. Therefore, if $\widehat{\nu}$ is discrete, we have

$$\widehat{\nu}(F) = \sum_{x \in F} \widehat{\pi}(x),$$

for any $F \in \mathcal{B}(\mathbb{R}^n)$. Note that the sum is well-defined as $\widehat{\pi}(x) > 0$ for at most countably many x's. On the contrary, if $\widehat{\nu}$ is absolutely continuous, for any $F \in \mathcal{B}(\mathbb{R}^n)$ we have

$$\widehat{\nu}(F) = \int_F \widehat{\pi}(x)\, dx.$$

2.2 Probabilistic Reconciliation

In this section, we discuss the notion of coherence in the probabilistic framework and our approach to probabilistic reconciliation.

In the non-probabilistic framework, a point forecast is incoherent if it does not belong to the set $\mathcal{S}$, defined as in (2.2). Let $\widehat{\nu} \in \mathcal{P}(\mathbb{R}^n)$ be a forecast distribution. Intuitively, we say that $\widehat{\nu}$ is incoherent if there exists a set T of incoherent points, i.e. $T \cap \mathcal{S} = \emptyset$, such that $\widehat{\nu}(T) > 0$. Or, equivalently, if $supp(\widehat{\nu}) \nsubseteq \mathcal{S}$, where $supp(\widehat{\nu}) := \{\mathbf{x} : \widehat{\pi}(\mathbf{x}) > 0\}$ is the support of $\widehat{\nu}$. We now define the summing map $s : \mathbb{R}^m \to \mathbb{R}^n$ as

$$s(\mathbf{b}) = \mathbf{S}\mathbf{b}. \tag{2.3}$$

The image of s is given by $\mathcal{S}$. Moreover, from (2.3) and (2.1), s is injective. Hence, s is a bijective map between $\mathbb{R}^m$ and $\mathcal{S}$, with inverse given by $s^{-1}(\mathbf{y}) = \mathbf{b}$, where $\mathbf{y} = \left[\mathbf{u}^T, \mathbf{b}^T\right]^T \in \mathcal{S}$. As explained in Panagiotelis et al. 2022, for any $\nu \in \mathcal{P}(\mathbb{R}^m)$ we may obtain a distribution $\widetilde{\nu} \in \mathcal{P}(\mathcal{S})$ as $\widetilde{\nu} = s_{\#}\nu$, namely the pushforward of ν using s:

$$\widetilde{\nu}(F) = \nu(s^{-1}(F)), \qquad \forall\, F \in \mathcal{B}(\mathcal{S}),$$

where $s^{-1}(F) := \{\mathbf{b} \in \mathbb{R}^m : s(\mathbf{b}) \in F\}$ is the preimage of F. In other words, $s_{\#}$ builds a probability distribution for $\mathbf{y}$ supported on the coherent subspace $\mathcal{S}$ from a distribution on the bottom variables $\mathbf{b}$. Since s is a measurable bijective map, $s_{\#}$ is a bijection between $\mathcal{P}(\mathbb{R}^m)$ and $\mathcal{P}(\mathcal{S})$, with inverse given by $(s^{-1})_{\#}$ (see Appendix A.1). We thus propose the following definition.

Definition 1. *We call coherent distribution any distribution $\nu \in \mathcal{P}(\mathbb{R}^m)$.*

This definition works with any type of distribution. Moreover, since it does not require s to be a linear map, it can be applied to any problem where non-linear constraints are involved.

2.2.1 Probabilistic reconciliation

In the probabilistic framework, the aim of reconciliation is to obtain a coherent reconciled distribution $\widetilde{\nu} \in \mathcal{P}(\mathbb{R}^m)$ from the base forecast distribution $\widehat{\nu} \in \mathcal{P}(\mathbb{R}^n)$.

A naive approach could be to simply set $\widetilde{\nu} = \widehat{\nu}_b$; this may be considered a probabilistic bottom-up, which ignores any probabilistic information about the upper series.

Panagiotelis et al. 2022 proposes a reconciliation method based on projections. Given a continuous map $\psi : \mathbb{R}^n \to \mathcal{S}$, the reconciled distribution is defined as $\widetilde{\nu} = \psi_{\#}\widehat{\nu} \in \mathcal{P}(\mathcal{S})$, i.e. $\widetilde{\nu}(F) = \widehat{\nu}(\psi^{-1}(F))$, for any

$F \in \mathcal{B}(\mathbb{R}^n)$. Since the map ψ is expressed as $\psi = s \circ g$, with $g : \mathbb{R}^n \to \mathbb{R}^m$ and $s : \mathbb{R}^m \to \mathcal{S}$, the reconciled distribution may be equivalently defined as $\widetilde{\nu} = g_{\#}\widehat{\nu} \in \mathcal{P}(\mathbb{R}^m)$. Note that, if $\widehat{\mathbf{y}}_1, \dots, \widehat{\mathbf{y}}_N$ are independent samples from the base incoherent forecast distribution $\widehat{\nu}$, then $\widetilde{\mathbf{y}}, \dots, \widetilde{\mathbf{y}}_N$, defined as $\widetilde{\mathbf{y}}_i := g(\mathbf{y}_i)$ for $i = 1, \dots, N$, are independent samples from the reconciled distribution $\widetilde{\nu}$. The function $g : \mathbb{R}^n \to \mathbb{R}^m$ combines information from all the levels by projecting on the bottom level the incoherent forecasts. For instance, if we define g as $g(\mathbf{y}) = \mathbf{b}$, where $\mathbf{y} = \left[\mathbf{u}^T, \mathbf{b}^T\right]^T$, we obtain the probabilistic bottom-up, as $\widetilde{\nu} = \widehat{\nu}_b$. In Panagiotelis et al. 2022, the map g is assumed to be in the form $g(\mathbf{y}) = \mathbf{d} + \mathbf{G}\mathbf{y}$, with $\mathbf{d} \in \mathbb{R}^m$ and $\mathbf{G} \in \mathbb{R}^{m \times n}$, and the parameter $\gamma := (\mathbf{d}, vec(\mathbf{G}))$ is optimized through stochastic gradient descent (SGD) to minimize a chosen scoring rule.

2.2.2 Probabilistic Reconciliation through conditioning

We now present our approach to probabilistic reconciliation, based on conditioning on the hierarchy constraints. Let $\widehat{\mathbf{Y}} = (\widehat{\mathbf{U}}, \widehat{\mathbf{B}})$ be a random vector with law given by $\widehat{\nu}$, so that $\widehat{\nu}_u$ and $\widehat{\nu}_b$ are the laws of, respectively, $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{B}}$.

Let us first suppose that the base forecast distribution $\widehat{\nu} \in \mathcal{P}(\mathbb{R}^n)$ is discrete. We define $\widetilde{\nu}$ by conditioning on the coherent subspace $\mathcal{S}$:

$$\begin{aligned}
\widetilde{\nu}(F) &= \mathbb{P}(\widehat{\mathbf{B}} \in F \mid \widehat{\mathbf{Y}} \in \mathcal{S}) \\
&= \frac{\mathbb{P}(\widehat{\mathbf{B}} \in F,\, \widehat{\mathbf{Y}} \in \mathcal{S})}{\mathbb{P}(\widehat{\mathbf{Y}} \in \mathcal{S})} \\
&= \frac{\mathbb{P}(\widehat{\mathbf{B}} \in F,\, \widehat{\mathbf{U}} = \mathbf{A}\widehat{\mathbf{B}})}{\mathbb{P}(\widehat{\mathbf{U}} = \mathbf{A}\widehat{\mathbf{B}})} \\
&= \frac{\sum_{\mathbf{b} \in F} \widehat{\pi}(\mathbf{A}\mathbf{b}, \mathbf{b})}{\sum_{\mathbf{x} \in \mathbb{R}^m} \widehat{\pi}(\mathbf{A}\mathbf{x}, \mathbf{x})},
\end{aligned} \tag{2.4}$$

for any $F \in \mathcal{B}(\mathbb{R}^m)$, provided that $\mathbb{P}(\widehat{\mathbf{Y}} \in \mathcal{S}) > 0$. The sums in (2.4) are well-defined, as $\widehat{\pi}(\mathbf{y}) > 0$ for at most countably many $\mathbf{y}$'s. Hence, $\widetilde{\nu}$ is a discrete probability distribution with pmf given by

$$\widetilde{\pi}(\mathbf{b}) = \frac{\widehat{\pi}(\mathbf{A}\mathbf{b}, \mathbf{b})}{\sum_{\mathbf{x} \in \mathbb{R}^m} \widehat{\pi}(\mathbf{A}\mathbf{x}, \mathbf{x})} \propto \widehat{\pi}(\mathbf{A}\mathbf{b}, \mathbf{b}). \tag{2.5}$$

Note that, if $\widehat{\nu}$ is absolutely continuous, we have that $\widehat{\nu}(\mathcal{S}) = 0$, since the Lebesgue measure of $\mathcal{S}$ is zero. Hence, $\mathbb{P}(\widehat{\mathbf{B}} \in F \mid \widehat{\mathbf{Y}} \in \mathcal{S})$ is not well-

defined. However, if we denote by $\widehat{\pi}$ the density of $\widehat{\nu}$, the last expression is still well-posed. We thus give the following definition.

Definition 2. *Let $\widehat{\nu} \in \mathcal{P}(\mathbb{R}^n)$ be a base forecast distribution. The reconciled distribution through conditioning is defined as the probability distribution $\widetilde{\nu} \in \mathcal{P}(\mathbb{R}^m)$ such that*

$$\widetilde{\pi}(\mathbf{b}) \propto \widehat{\pi}(\mathbf{A}\mathbf{b}, \mathbf{b}), \tag{2.6}$$

where $\widehat{\pi}$ and $\widetilde{\pi}$ are the densities of (respectively) $\widehat{\nu}$ and $\widetilde{\nu}$, if $\widehat{\nu}$ is absolutely continuous, or the probability mass functions otherwise.

To rigorously derive (2.6) in the continuous case, we proceed as follows. Let us define the random vector $\mathbf{Z} := \widehat{\mathbf{U}} - \mathbf{A}\widehat{\mathbf{B}}$. Note that the event $\{\widehat{\mathbf{Y}} \in \mathcal{S}\}$ coincides with $\{\mathbf{Z} = \mathbf{0}\}$. The joint density of $(\mathbf{Z}, \widehat{\mathbf{B}})$ can be easily computed (Appendix A.1):

$$\pi_{(\mathbf{Z},\widehat{\mathbf{B}})}(\mathbf{z}, \mathbf{b}) = \widehat{\pi}(\mathbf{z} + \mathbf{A}\mathbf{b}, \mathbf{b}).$$

Then, the conditional density of $\widehat{\mathbf{B}}$ given $\mathbf{Z} = \mathbf{0}$ is given by (Çinlar 2011, Chapter 4):

$$\begin{aligned}
\widetilde{\pi}(\mathbf{b}) &= \frac{\pi_{(Z,B)}(\mathbf{0}, \mathbf{b})}{\int_{\mathbb{R}^m} \pi_{(Z,B)}(\mathbf{0}, \mathbf{x})\, d\mathbf{x}} \\
&= \frac{\widehat{\pi}(\mathbf{A}\mathbf{b}, \mathbf{b})}{\int_{\mathbb{R}^m} \widehat{\pi}(\mathbf{A}\mathbf{x}, \mathbf{x})\, d\mathbf{x}} \\
&\propto \widehat{\pi}(\mathbf{A}\mathbf{b}, \mathbf{b}),
\end{aligned}$$

provided that $\int_{\mathbb{R}^m} \widehat{\pi}(\mathbf{A}\mathbf{x}, \mathbf{x})\, d\mathbf{x} > 0$. Finally, note that, if $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{B}}$ are independent, (2.6) may be rewritten as

$$\widetilde{\pi}(\mathbf{b}) \propto \widehat{\pi}_u(\mathbf{A}\mathbf{b}) \cdot \widehat{\pi}_b(\mathbf{b}), \tag{2.7}$$

where $\widehat{\pi}_u$ and $\widehat{\pi}_b$ are the densities of (respectively) $\widehat{\nu}_u$ and $\widehat{\nu}_b$.

From a Bayesian perspective, the reconciliation process can be interpreted as a generalization of the Bayes' rule. Indeed, the base distribution on the bottom variables may be interpreted as the prior:

$$\mathbf{b} \sim \widehat{\nu}_b, \tag{2.8}$$

while the likelihood expresses the hierarchy constraints:

$$\pi(\mathbf{u} \mid \mathbf{b}) = \delta_{\{\mathbf{u}=\mathbf{A}\mathbf{b}\}}. \tag{2.9}$$

Thus, the evidence is not given by a single observation, but rather by a probability distribution, i.e., the base conditional distribution of $\widehat{\mathbf{U}}$ given

$\widehat{\mathbf{B}}$. In the area of Bayesian networks, this approach is known as updating using soft evidence (Darwiche 2009, Chapter 3.6); it is at the core of the reconciliation approach by Corani et al. 2022.

While in Panagiotelis et al. 2022 the reconciled distribution was obtained by projecting the base distribution $\widehat{\nu}$ on $\mathcal{S}$, in this work $\widetilde{\nu}$ is obtained by conditioning $\widehat{\nu}$ on the constraints given by the hierarchy. Our approach can be applied to both continuous and discrete distributions. On the other hand, the approach based on projection optimizes the parameters through stochastic gradient descent, which is computationally expensive and not applicable for discrete distributions.

In our approach, the behaviour of the base distribution outside the coherent subspace is ignored: intuitively, we do not take into account the probability of incoherent points, since they are not "admissible". Indeed, (2.6) clearly shows that $\widetilde{\nu}$ only depends on the values of $\widehat{\nu}$ on $\mathcal{S}$. The reconciled distribution through conditioning satisfies the following property: for each pair of coherent points $\mathbf{y}_1, \mathbf{y}_2 \in \mathcal{S}$, we have

$$\frac{\widetilde{\pi}(\mathbf{y}_1)}{\widetilde{\pi}(\mathbf{y}_2)} = \frac{\widehat{\pi}(\mathbf{y}_1)}{\widehat{\pi}(\mathbf{y}_2)} \tag{2.10}$$

if $\pi(\mathbf{y}_2) \neq 0$, and $\widetilde{\pi}(\mathbf{y}_2) = 0$ if $\pi(\mathbf{y}_2) = 0$. The reconciliation thus preserves the odds ratio: if, for example, $\mathbf{y}_1$ is three times more likely than $\mathbf{y}_2$ according to the base distribution, then it is the same also for the reconciled distribution.

2.3 Gaussian case

When the base forecast distribution is a multivariate Gaussian, the reconciled distribution is also Gaussian, and its mean and covariance matrix can be analytically computed (Corani et al. 2020). In this case, reconciliation through conditioning coincides with minT, which has been proven to minimize the log score (Wickramasuriya et al. 2019; Wickramasuriya 2021). Let

$$\widehat{\mathbf{Y}} = \begin{bmatrix} \widehat{\mathbf{U}} \\ \widehat{\mathbf{B}} \end{bmatrix} \sim \mathcal{N}\left(\widehat{\mathbf{y}}, \widehat{\mathbf{\Sigma}}_Y\right) \tag{2.11}$$

be the base forecast distribution for the entire hierarchy, where

$$\widehat{\mathbf{y}} = \begin{bmatrix} \widehat{\mathbf{u}} \\ \widehat{\mathbf{b}} \end{bmatrix}, \quad \widehat{\mathbf{\Sigma}}_Y = \begin{bmatrix} \widehat{\mathbf{\Sigma}}_U & \widehat{\mathbf{\Sigma}}_{UB} \\ \widehat{\mathbf{\Sigma}}_{UB}^T & \widehat{\mathbf{\Sigma}}_B \end{bmatrix}.$$

Let us define $\mathbf{T} \in \mathbb{R}^{n \times n}$ as

$$\mathbf{T} = \begin{bmatrix} \mathbf{0} & \mathbf{I}_m \\ \mathbf{I}_{n-m} & -\mathbf{A} \end{bmatrix},$$

and let $\mathbf{Z} := \mathbf{T}\widehat{\mathbf{Y}}$. Hence, $\mathbf{Z}$ is Gaussian:

$$\mathbf{Z} \sim \mathcal{N}\left(\mathbf{T}\hat{\mathbf{y}},\, \mathbf{T}\widehat{\boldsymbol{\Sigma}}_Y\mathbf{T}^T\right). \tag{2.12}$$

We have

$$\begin{aligned} \mathbf{T}\hat{\mathbf{y}} &= \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} - \mathbf{A}\hat{\mathbf{b}} \end{bmatrix}, \\ \mathbf{T}\widehat{\boldsymbol{\Sigma}}_Y\mathbf{T}^T &= \begin{bmatrix} \widehat{\boldsymbol{\Sigma}}_B & \widehat{\boldsymbol{\Sigma}}_{UB}^T - \widehat{\boldsymbol{\Sigma}}_B\mathbf{A}^T \\ \widehat{\boldsymbol{\Sigma}}_{UB} - \mathbf{A}\widehat{\boldsymbol{\Sigma}}_B & \mathbf{Q} \end{bmatrix}, \end{aligned} \tag{2.13}$$

where $\mathbf{Q} = \widehat{\boldsymbol{\Sigma}}_U - \widehat{\boldsymbol{\Sigma}}_{UB}\mathbf{A}^T - \mathbf{A}\widehat{\boldsymbol{\Sigma}}_{UB}^T + \mathbf{A}\widehat{\boldsymbol{\Sigma}}_B\mathbf{A}^T$. Since

$$\mathbf{Z} = \begin{bmatrix} \hat{\mathbf{B}} \\ \hat{\mathbf{U}} - \mathbf{A}\hat{\mathbf{B}} \end{bmatrix} =: \begin{bmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \end{bmatrix},$$

the reconciled bottom distribution is given by the conditional law of $\mathbf{Z}_1$ given $\mathbf{Z}_2 = 0$, which is a multivariate Gaussian. Assuming that the covariance matrix of $\mathbf{Z}_2$ is positive definite, we have

$$\mathbf{Z}_1 \,|\, \mathbf{Z}_2 = 0 \sim \mathcal{N}\left(\widetilde{\mathbf{b}},\, \widetilde{\boldsymbol{\Sigma}}_B\right),$$

where

$$\begin{aligned} \widetilde{\mathbf{b}} &= \hat{\mathbf{b}} + \left(\widehat{\boldsymbol{\Sigma}}_{UB}^T - \widehat{\boldsymbol{\Sigma}}_B\mathbf{A}^T\right)\mathbf{Q}^{-1}(\mathbf{A}\hat{\mathbf{b}} - \hat{\mathbf{u}}), \\ \widetilde{\boldsymbol{\Sigma}}_B &= \widehat{\boldsymbol{\Sigma}}_B - \left(\widehat{\boldsymbol{\Sigma}}_{UB}^T - \widehat{\boldsymbol{\Sigma}}_B\mathbf{A}^T\right)\mathbf{Q}^{-1}\left(\widehat{\boldsymbol{\Sigma}}_{UB}^T - \widehat{\boldsymbol{\Sigma}}_B\mathbf{A}^T\right)^T. \end{aligned}$$

Since $\widetilde{\mathbf{U}} = \mathbf{A}\widetilde{\mathbf{B}}$ and $\widetilde{\mathbf{B}} \sim \mathcal{N}\left(\widetilde{\mathbf{b}},\, \widetilde{\boldsymbol{\Sigma}}_B\right)$, the reconciled upper distribution is also Gaussian: $\widetilde{\mathbf{U}} \sim \mathcal{N}\left(\widetilde{\mathbf{u}},\, \widetilde{\boldsymbol{\Sigma}}_U\right)$, with

$$\widetilde{\mathbf{u}} = \mathbf{A}\widetilde{\mathbf{b}}, \qquad \widetilde{\boldsymbol{\Sigma}}_U = \mathbf{A}\widetilde{\boldsymbol{\Sigma}}_B\mathbf{A}^T. \tag{2.14}$$

If we define $\mathbf{D} := \widehat{\boldsymbol{\Sigma}}_U - \widehat{\boldsymbol{\Sigma}}_{UB}\mathbf{A}^T$, from the above equations we have

$$\begin{aligned}
\widetilde{\mathbf{u}} &= \mathbf{A}\widehat{\mathbf{b}} + \left(\mathbf{A}\widehat{\boldsymbol{\Sigma}}_{UB}^T - \mathbf{A}\widehat{\boldsymbol{\Sigma}}_B\mathbf{A}^T\right)\mathbf{Q}^{-1}(\mathbf{A}\widehat{\mathbf{b}} - \widehat{\mathbf{u}}) \\
&= \mathbf{A}\widehat{\mathbf{b}} + (\mathbf{D} - \mathbf{Q})\,\mathbf{Q}^{-1}(\mathbf{A}\widehat{\mathbf{b}} - \widehat{\mathbf{u}}) \\
&= \mathbf{A}\widehat{\mathbf{b}} + \mathbf{D}\mathbf{Q}^{-1}(\mathbf{A}\widehat{\mathbf{b}} - \widehat{\mathbf{u}}) - (\mathbf{A}\widehat{\mathbf{b}} - \widehat{\mathbf{u}}) \\
&= \widehat{\mathbf{u}} + \left(\widehat{\boldsymbol{\Sigma}}_U - \widehat{\boldsymbol{\Sigma}}_{UB}\mathbf{A}^T\right)\mathbf{Q}^{-1}(\mathbf{A}\widehat{\mathbf{b}} - \widehat{\mathbf{u}}).
\end{aligned}$$

Moreover, we have

$$\begin{aligned}
\widetilde{\boldsymbol{\Sigma}}_U &= \mathbf{A}\widehat{\boldsymbol{\Sigma}}_B\mathbf{A}^T - \left(\mathbf{A}\widehat{\boldsymbol{\Sigma}}_{UB}^T - \mathbf{A}\widehat{\boldsymbol{\Sigma}}_B\mathbf{A}^T\right)\mathbf{Q}^{-1}\left(\mathbf{A}\widehat{\boldsymbol{\Sigma}}_{UB}^T - \mathbf{A}\widehat{\boldsymbol{\Sigma}}_B\mathbf{A}^T\right)^T \\
&= \mathbf{A}\widehat{\boldsymbol{\Sigma}}_B\mathbf{A}^T - (\mathbf{D} - \mathbf{Q})\,\mathbf{Q}^{-1}\left(\mathbf{D}^T - \mathbf{Q}\right) \\
&= \mathbf{A}\widehat{\boldsymbol{\Sigma}}_B\mathbf{A}^T - \mathbf{D}\mathbf{Q}^{-1}\mathbf{D}^T + \mathbf{D} + \mathbf{D}^T - \mathbf{Q} \\
&= \widehat{\boldsymbol{\Sigma}}_U - \left(\widehat{\boldsymbol{\Sigma}}_U - \widehat{\boldsymbol{\Sigma}}_{UB}\mathbf{A}^T\right)\mathbf{Q}^{-1}\left(\widehat{\boldsymbol{\Sigma}}_U - \widehat{\boldsymbol{\Sigma}}_{UB}\mathbf{A}^T\right)^T.
\end{aligned}$$

2.4 Sampling from the reconciled distribution

The reconciled distribution $\widetilde{\nu}$ is not, in general, a known distribution. In a non-Gaussian framework, we generally need to resort to sampling approaches. Our method is based on Importance Sampling (IS), which we briefly recall in the next subsection. For a complete discussion, we refer to Elvira and Martino 2021. From now on, we will use the term density to denote either the probability mass function (for discrete distributions) or the density with respect to the Lebesgue measure (for absolutely continuous distributions).

2.4.1 Importance Sampling

Importance Sampling is a popular technique used to approximate expectations with respect to a target distribution by sampling from another distribution. It was first introduced in the 50s in statistical physics (Kahn and Marshall 1953; Hammersley and Morton 1954), and since then it has been extensively used and developed.

Let X be a random variable with density p. Suppose we want to compute the expected value $m = \mathbb{E}[f(X)]$, for some function f. If we are able to draw independent samples $x_1, \dots, x_N$ from p, we can use the standard Monte Carlo

estimate:

$$\mathbb{E}[f(X)] \approx \frac{1}{N}\sum_{i=1}^{N} f(x_i). \tag{2.15}$$

In many cases, however, sampling from p could be impractical, or it could lead to a very high variance of the Monte Carlo estimator (2.15).

Now, let q be another density such that $q(x) > 0$ if $f(x)p(x) \neq 0$. We have that

$$\mathbb{E}[f(X)] = \int f(x)p(x)\,dx = \int f(x)\frac{p(x)}{q(x)}q(x)\,dx = \mathbb{E}[f(Y)w(Y)],$$

where w is defined as $w(y) = \frac{p(y)}{q(y)}$ and Y is a random variable with density q. Hence, if $y_1, \ldots, y_N$ are independent samples drawn from q, the importance sampling estimate is given by

$$\mathbb{E}[f(X)] \approx \frac{1}{N}\sum_{i=1}^{N} w(y_i)f(y_i). \tag{2.16}$$

In most practical cases, the density p is known only up to a normalizing constant: we can then write $w(y) = c\bar{w}(y)$, where $\bar{w}$ is known but the constant c is unknown. In such cases, we may replace (2.16) with the self-normalized importance sampling estimate

$$\mathbb{E}[f(X)] \approx \frac{\sum_{i=1}^{N} \bar{w}(y_i)f(y_i)}{\sum_{i=1}^{N} \bar{w}(y_i)}. \tag{2.17}$$

A crucial role for the efficiency of IS is played by the choice of the proposal distribution q. In most applications, a good proposal should be a good approximation of the target distribution p. A common diagnostic to assess the efficiency of IS is the Effective Sample Size. The ESS of a weighted sample represents the number of independent samples from the target distribution that yields the same efficiency in the estimation. The efficiency is usually interpreted in terms of the variance of the Monte Carlo estimator. A popular approximation of the ESS (Elvira et al. 2022) is given by

$$\widehat{ESS} = \frac{\left(\sum_{i=1}^{N} w(y_i)\right)^2}{\sum_{i=1}^{N} w(y_i)^2}. \tag{2.18}$$

Notice that $\widehat{ESS}$ is a number between 1 and N. If we are able to sample directly from the target distribution, i.e. $q = p$, then $w(y) = 1$ for any y,

hence $\widehat{ESS} = N$. On the contrary, when we use a bad proposal distribution, typically few weights are much larger than the others, leading to a very low ESS. In particular, when the dimension of the space grows, it gets harder to find a good proposal, and the performance of IS, in terms of ESS, drops dramatically (Agapiou et al. 2017). This phenomenon is usually referred to as curse of dimensionality.

2.4.2 Probabilistic reconciliation via IS

Let $\widetilde{\nu}$, as in Definition 2), be the target distribution. We set $\widehat{\nu}_b$ as proposal distribution. Given a sample $\mathbf{b}_1, \dots, \mathbf{b}_N$ drawn form $\widehat{\nu}_b$, the weights are computed as

$$w_i := \frac{\widehat{\pi}(\mathbf{A}\mathbf{b}_i, \mathbf{b}_i)}{\widehat{\pi}_b(\mathbf{b}_i)}. \tag{2.19}$$

Then, $(\mathbf{b}_i, \tilde{w}_i)_{i=1,\dots,N}$ is a weighted sample from $\widetilde{\nu}$, where $\tilde{w}_i := {}^{w_i}/_{\sum_{j=1}^N w_j}$ are the normalized weights. Note that (2.19) may be interpreted as the conditional density of $\widehat{\mathbf{U}}$ at the point $\mathbf{A}\mathbf{b}_i$, given that $\widehat{\mathbf{B}} = \mathbf{b}_i$. Loosely speaking, we draw samples $(\mathbf{b}_i)_i$ from the base bottom distributions, and then weight how likely they are using the base upper distributions. We thus combine the information contained in the distributions of both the bottom and the upper variables. From a Bayesian perspective, we sample from the prior, and then we assign weights to the samples by using the soft evidence. Finally, note that, under the assumption of independence between $\widehat{\mathbf{B}}$ and $\widehat{\mathbf{U}}$, the density of $\widetilde{\nu}$ may be factorized as in (2.7). In this case,

$$w_i = \widehat{\pi}_u(\mathbf{A}\mathbf{b}_i). \tag{2.20}$$

2.4.3 Limitations of IS

It is well-known that importance sampling is not effective to sample from high dimensional distributions; this prevents using it to reconcile large hierarchies. We also expect low performance when the proposal distribution $\widehat{\nu}_b$ is not a good approximation of the target distribution $\widetilde{\nu}$. The following result relates the Kullback-Leibler divergence (Kullback and Leibler 1951) between the base and reconciled distribution to the efficiency of IS.

Proposition 1. *Let $\widehat{\mathbf{B}}$ be a random vector distributed as $\widehat{\nu}_b$, and let $W := \widehat{\pi}(\mathbf{A}\widehat{\mathbf{B}}, \widehat{\mathbf{B}})/\widehat{\pi}_b(\widehat{\mathbf{B}})$. Then, the Kullback-Leibler divergence of the base bottom distribution from the reconciled bottom distribution is given by*

$$KL(\widehat{\nu}_b \,\|\, \widetilde{\nu}) = \log\left(\mathbb{E}[W]\right) - \mathbb{E}[\log(W)]. \tag{2.21}$$

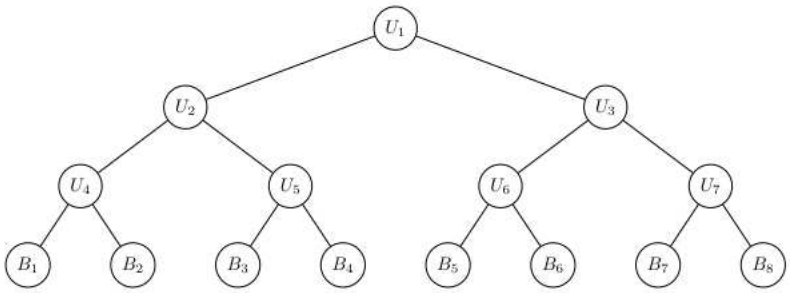


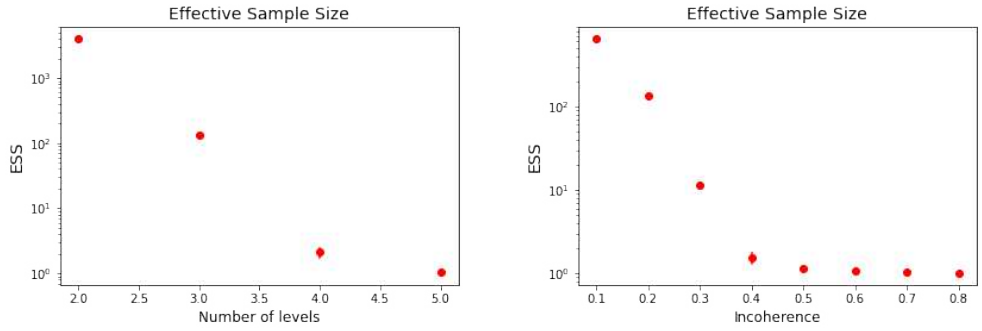
Figure 2.2: A binary hierarchy

In Appendix A.2, we recall the definition of the KL divergence and we report the proof of Proposition 1. The right-hand side of (2.21) is a measure of the dispersion of the random variable W. Indeed, by the Jensen's inequality, it is always non-negative, and it is zero when W is constant a.s.; it gets larger as W becomes more dispersed. In the context of the measures of inequality, it usually referred to as Mean Logarithm Deviation (Haughton and Khandker 2009). Note that, from (2.19), the importance sampling weights are IID copies of W. Hence, the more distant are the base and the reconciled distribution, in terms of the Kullback-Leibler divergence, the more dispersed are the IS weights. As explained above, a large dispersion of the weights leads to a low effective sample size, and thus to a poor performance of importance sampling.

The incoherence of the forecasts is often defined as the difference between the bottom-up mean and the base upper mean, i.e. $\mathbf{A}\widehat{\mathbf{b}} - \widehat{\mathbf{u}}$. Intuitively, we expect that the distance between the distributions of $\mathbf{A}\widehat{\mathbf{B}}$ and $\widehat{\mathbf{U}}$ grows as the incoherence grows, and therefore also the distance between $\widehat{\nu}_b$ and $\widetilde{\nu}$, as the latter merges the information coming from the bottom and the upper variables.

We run some experiments to test the dependence of the efficiency of IS on the size of the hierarchy and on the percentage level of incoherence. We set a binary hierarchy, described by a tree where each node has 1 parent and 2 children (except for the top and the bottom nodes). If k is the number of levels of the hierarchy, then there are $m = 2^k$ bottom nodes and $1 + 2 + \cdots + 2^{k-1}$ upper nodes. An examples for $k = 3$ levels is reported in Figure 2.2.

The base distribution is defined by setting an independent Poisson distribution on each node of the hierarchy. We fix a vector $\boldsymbol{\lambda}_b \in \mathbb{R}_+^m$ of the bottom means. Then, we fix an incoherence level $\epsilon > 0$, and we set the means of the upper nodes as $\boldsymbol{\lambda}_u = (1 + \epsilon)\mathbf{A}\boldsymbol{\lambda}_b$. Hence, for example, an incoherence level of $\epsilon = 0.5$ means that the base upper means are 50% greater than the sum



(a) *(b)*

Figure 2.3: Effective sample size as the dimension of the hierarchy (left) or the incoherence level (right) grows. The y axis is logarithmic

of the corresponding base bottom means.

We set $\epsilon = 0.2$, and for each hierarchy size $k \in \{2, 3, 4, 5\}$ we draw $100,000$ weighted samples from $\widetilde{\nu}$ and we compute the effective sample size. We repeat 30 times and take the average. The results are reported in Figure 2.3a. Then, we set $k = 3$ and we do the same for $\epsilon \in \{0.1, 0.2, \ldots, 0.8\}$. The results are reported in Figure 2.3b. As expected, the effective sample size dramatically drops as the hierarchy size or the incoherence level grows. Note that, in Figure 2.3, the y axis is logarithmic.

These experiments, along with Proposition 1, justify the need of a more robust algorithm. In the next chapter, we introduce the Bottom-Up Importance Sampling algorithm.

Chapter 3

Algorithm and experiments

This chapter is organized as follows. In Section 3.1, we introduce the Bottom-Up Importance Sampling algorithm. We then show how to adapt it in order to deal with distributions in the form of samples, or in case of grouped non-hierarchical time series. In Section 3.2, we run several experiments on synthetic data to show the efficiency of our algorithm. Finally, in Section 3.3, we test our method exhaustively on time series extracted from different data sets, providing a significant improvement in the quality of the probabilistic forecasts.

3.1 Bottom-Up Importance Sampling algorithm

First, we state the main assumption of our algorithm:

Assumption 1. *The base forecasts of each variable are independent.*

For instance, consider the Gaussian base forecast distribution defined by (2.11): it satisfies Assumption 1 only if the covariance matrix $\widehat{\mathbf{\Sigma}}_Y$ is diagonal. Of course, a coherent distribution on the entire hierarchy does not satisfy this assumption, because of the constraints on the variables. However, this is a common assumption when forecasts are produced independently for each time series, as their source of errors are assumed to be independent. For example, this is typically the case with temporal hierarchies (Athanasopoulos et al. 2017). For the sake of simplicity, we present our algorithm making also the following assumption:

Assumption 2. *The data structure is hierarchical.*

This means that the data structure disaggregates in a unique hierarchical manner (Hyndman and Athanasopoulos 2021, Chapter 11.1). Hence, it is

represented by a tree, in which every node only has one parent. An example is given by the binary hierarchy in Figure 2.2. This assumption will be relaxed in Section 3.1.2, so that the algorithm can deal also with non-hierarchical structures.

Under Assumption 1 and Assumption 2, we develop a new algorithm, called Bottom-Up Importance Sampling (BUIS). The core idea is to split a single $(n-m)$-dimensional importance sampling problem into $n-m$ one-dimensional problems. To do so, we start by drawing a sample from the base distribution $\widehat{\nu}_b$. Then, for each level of the hierarchy, from bottom to top, we update the sample through an importance sampling step. At each step, the "partially" reconciled distribution is used as proposal. In this way, we encapsulate the information contained in the base distributions of the upper time series, as explained in Section 2.4.2. The advantage of this algorithm is that we independently perform importance sampling for each upper variable. This deeply alleviates the curse of dimensionality.

For each level $l = 1, \dots, L$ of the hierarchy, we denote the upper variables at level l by $u_{1,l}, \dots, u_{k_l,l}$. Moreover, for any upper variable $u_{j,l}$, we denote by $b_{1,(j,l)}, \dots, b_{q_{j,l},(j,l)}$ the bottom variables that sum up to $u_{j,l}$. In this way, we have that $\sum_{l=1}^{L} k_l = n-m$, the number of upper variables, while $\sum_{j=1}^{k_l} q_{j,l} = m$, the number of bottom variables, for each level l.

Let us consider, for example, the hierarchy in Figure 2.1. For the first level $l = 1$, we have $k_1 = 2$, $u_{1,1} = U_2$, and $u_{2,1} = U_3$. Moreover, $q_{1,1} = q_{2,1} = 2$, and $b_{1,(1,1)} = B_1$, $b_{2,(1,1)} = B_2$, $b_{1,(2,1)} = B_3$, $b_{2,(2,1)} = B_4$. For the last level $l = 2$, we have $k_2 = 1$, $u_{1,2} = U_1$, $q_{1,2} = 4$, $b_{1,(1,2)} = B_1$, $b_{2,(1,2)} = B_2$, $b_{3,(1,2)} = B_3$, $b_{4,(1,2)} = B_4$.

Algorithm 1 Bottom-Up Importance Sampling

Sample $\left(\mathbf{b}^{(i)}\right)_{i=1,\dots,N}$ from $\widehat{\pi}_b$
for l in levels **do**
  for $j = 1, \dots, k_l$ **do**
    $\breve{w}^{(i)} \leftarrow \widehat{\pi}_{u_{j,l}}\left(\sum_{t=1}^{q_{j,l}} b^{(i)}_{t,(j,l)}\right)$ for $i = 1, \dots, N$
    $w^{(i)} \leftarrow \frac{\breve{w}^{(i)}}{\sum_h \breve{w}^{(h)}}$ for $i = 1, \dots, N$
    $\left(\bar{\mathbf{b}}^{(i)}_j\right)_i \leftarrow$ **Resample** $\left(\left(b^{(i)}_{1,(j,l)}, \dots, b^{(i)}_{q_{j,l},(j,l)}\right), w^{(i)}\right)_i$
  end for
  $\mathbf{b}^{(i)} \leftarrow \left[\bar{\mathbf{b}}^{(i)}_1, \dots, \bar{\mathbf{b}}^{(i)}_{k_l}\right]$ for $i = 1, \dots, N$
end for
return $\left(\mathbf{b}^{(i)}\right)_i$

The BUIS algorithm is presented above (Alg. 1). The "Resample" step is

performed by sampling with replacement from the discrete distribution given by

$$\mathbb{P}\left(\mathbf{b} = \left(b^{(i)}_{1,(j,l)}, \dots, b^{(i)}_{q_{j,l},(j,l)}\right)\right) = w^{(i)}, \tag{3.1}$$

for all $i = 1, \dots, N$. We explicit the BUIS algorithm on the simple hierarchy in Figure 2.1:

1. Sample $(b_j^{(i)})_{i=1,\dots,N}$ from $\widehat{\pi}_{B_j}$, for $j = 1, 2, 3, 4$
2. Compute the weights $(w^{(i)})_{i=1,\dots,N}$ with respect to U_2 as
$$w^{(i)} = \widehat{\pi}_{U_2}\left(b_1^{(i)} + b_2^{(i)}\right)$$
3. Sample $\left(\bar{b}_1^{(i)}, \bar{b}_2^{(i)}\right)_i$ with replacement from $\left((b_1^{(i)}, b_2^{(i)}), w^{(i)}\right)_{i=1,\dots,N}$
4. Repeat step 2 and 3 using B_3, B_4 and U_3 to get $\left(\bar{b}_3^{(i)}, \bar{b}_4^{(i)}\right)_i$
5. Set $\left(b_1^{(i)}, b_2^{(i)}, b_3^{(i)}, b_4^{(i)}\right)_i = \left(\bar{b}_1^{(i)}, \bar{b}_2^{(i)}, \bar{b}_3^{(i)}, \bar{b}_4^{(i)}\right)_i$ and move to the next level
6. Compute the weights $(w^{(i)})_{i=1,\dots,N}$ with respect to U_1 as
$$w^{(i)} = \widehat{\pi}_{U_1}\left(b_1^{(i)} + b_2^{(i)} + b_3^{(i)} + b_4^{(i)}\right)$$
7. Sample $\left(\bar{b}_1^{(i)}, \bar{b}_2^{(i)}, \bar{b}_3^{(i)}, \bar{b}_4^{(i)}\right)_i$ with replacement from $\left((b_1^{(i)}, b_2^{(i)}, b_3^{(i)}, b_4^{(i)}), w^{(i)}\right)_i$

Proposition 2. *The output of the BUIS algorithm is approximately a sample drawn from the reconciled distribution $\widetilde{\nu}$.*

The proof is reported in Appendix A.3.

3.1.1 Sample-based BUIS

The densities of the forecast distributions are not always available in analytical form. For instance, probabilistic forecasts on count time series are typically given as samples (Liboschik et al. 2017). However, we are able to perform reconciliation even without the analytical form of the densities. Since we only deal with one-dimensional densities to compute the weights, we may effectively use approximations based on samples. For discrete distributions, we use the empirical distribution. As for the continuous setting,

several methods are available to approximate the true density, such as kernel density estimation (Chen 2017). Therefore, we only need to replace line 4 in Algorithm 1 with:

$$\begin{aligned}
&\textbf{Sample } \left(u_{j,l}^{(i)}\right)_{i=1,\dots,N} \text{ from } \widehat{\pi}_{u_{j,l}} \\
&\breve{\pi} \leftarrow \textbf{Density Estimation}\left(\left(u_{j,l}^{(i)}\right)_{i=1,\dots,N}\right) \\
&\breve{w}^{(i)} \leftarrow \breve{\pi}\left(\textstyle\sum_{t=1}^{q_{j,l}} b_{t,(j,l)}^{(i)}\right) \quad \text{for } i = 1, \dots, N
\end{aligned}$$

From a computational perspective, the sample-based algorithm is slower due to the density estimation step.

3.1.2 More complex hierarchies: grouped time series

Time series with a data structure that does not disaggregate in a unique hierarchical manner are referred to as grouped time series (Hyndman and Athanasopoulos 2021, Chapter 11). For instance, consider a weekly time series, for which we compute the following temporal aggregates: 2-weeks, 4-weeks, 13-weeks, 26-weeks, 52-weeks. If we deal with one year forecasts, we have 52 bottom variables and $26 + 13 + 4 + 2 + 1 = 46$ upper variables. Clearly, this structure cannot be represented as a tree.

Since Assumption 2 is not satisfied, the BUIS algorithm, as described in Section 3.1, cannot be used. Indeed, as highlighted in the proof, we need the independence of $\bar{\mathbf{b}}_1, \dots, \bar{\mathbf{b}}_{k_l}$ to multiply their densities. If Assumption 2 does not hold, correlations between bottom variables are created when conditioning on the upper levels.

To overcome this problem, we proceed as follows. First, we find the largest sub-hierarchy within the group structure. For instance, in the example above, we consider the sub-hierarchy given by the bottom variables and by the 2-weeks, 4-weeks and 52-weeks aggregates. All the other upper variables are then regarded as additional constraints. We use the BUIS algorithm on the sub-hierarchy, obtaining a sample $\mathbf{b}$. Then, we compute the weights on $\mathbf{b}$ using the base distributions of the additional constraints. This is equivalent to performing a plain IS, where we use the output of BUIS on the hierarchical part as proposal distribution. In this way, we reduce the dimension of the IS task from $n - m$, the total number of upper constraints, to the number of constraints that are not included in the sub-hierarchy: in the above example, from 46 to 6. We highlight that the distribution we sample from would be the same even with different choices of sub-hierarchies. However, picking the largest one is the best choice from a computational perspective. We still

refer to this extended version of the algorithm as Bottom-Up Importance Sampling.

3.2 Experiments on synthetic data

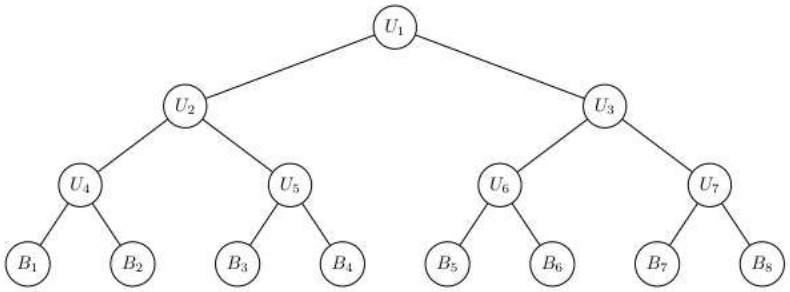


Figure 3.1: A binary hierarchy

We present experiments on synthetic data, aimed at checking the correctness of our algorithm. We compare the accuracy of IS, BUIS, and the method by Corani et al. 2022, which we implement as in their paper using the library *PyMC3* (Salvatier et al. 2016). *PyMC3* adopts an adaptive Metropolis-Hastings algorithm (Haario et al. 2001) for discrete distributions and the No-U-Turn Sampler (NUTS, Hoffman, Gelman, et al. 2014) for continuous distributions. We run 4 chains with $5,000$ samples each. We use $100,000$ samples for IS and BUIS, which are faster. We perform experiments on two different hierarchies: the binary tree of Figure 3.1 and the weekly hierarchy described in Section 3.1.2. We implemented the BUIS algorithm in Python. We make available a notebook to reproduce our synthetic experiments at the url `https://drive.google.com/file/d/1dUThfSfWv9Qij6-slwtYwMd-Prw2KagR/view?usp=sharing`.

3.2.1 Reconciling Gaussian forecasts

We start by considering Gaussian base forecasts, for which the reconciled distribution can be analytically computed (Corani et al. 2020). For the binary hierarchy, we set on each bottom node a Gaussian distribution with mean randomly chosen in the interval $[5, 10]$, and standard deviation $\widehat{\sigma}_b = 2$. We denote by $\widehat{\mathbf{m}}_b \in \mathbb{R}^8_+$ the vector of the base bottom means. We introduce incoherence by setting the means of the base forecast of the upper variables as $\widehat{\mathbf{m}}_u = (1+\epsilon)\mathbf{A}\widehat{\mathbf{m}}_b$, where $\mathbf{A}$ is the aggregating matrix and ϵ is the incoherence level. We consider the incoherence levels $\epsilon \in \{0.1,\ 0.3,\ 0.5\}$. Hence, an

incoherence level of 0.5 means that the base upper means are 50% greater than the sum of the corresponding base bottom means. We set $\widehat{\sigma}_u = 3$ as standard deviation for the base forecast of each upper variable.

We compare the reconciled mean computed via sampling ($\bar{\mathbf{y}}$) with the analytically reconciled mean ($\mathbf{y}^a$) using the mean absolute percentage error (MAPE):

$$\text{MAPE}(\bar{\mathbf{y}}, \mathbf{y}^a) = \frac{1}{n} \sum_{i=1}^{n} \frac{|\bar{y}_i - y_i^a|}{y_i^a} \cdot 100.$$

We use the MAPE as it is an intuitive way for comparing the relative error between $\bar{\mathbf{y}}$ and $\mathbf{y}^a$, although it is generally not recommended for evaluating the forecast accuracy (Kolassa 2016).

We repeat each experiment 30 times using the same parameters, and we report the average errors in Table 3.1. Remarkably BUIS reduces the error with respect to IS, dealing robustly also with large incoherence. The results are graphically represented in Figure 3.2, where we show the boxplot of the reconciled mean of a bottom variable. The blue line represents the exact value. We complete our analysis by reporting the 2-Wasserstein distance (Panaretos and Zemel 2019) between the true reconciled distribution and the empirical distribution obtained via sampling. The results, shown in Table 3.2, are similar to those discussed for the mean. Hence, BUIS is almost as accurate as MCMC, while drastically reducing the computational times from 30 seconds to less than a second (Table 3.3). Such a major speedup is possible because IS simultaneously generates samples and computes the weights. MCMC, on the contrary, generates the samples sequentially. This could be a major advantage in modern applications, which require reconciling a large number of time series. A more detailed comparison of the computational times is given in Appendix A.4.

	Error wrt analytical solution		
ϵ	IS	BUIS	MCMC
0.1	0.17 %	0.11 %	0.12 %
0.3	0.33 %	0.11 %	0.10 %
0.5	1.75 %	0.13 %	0.08 %

Table 3.1: MAPE on the reconciled mean (binary hierarchy, Gaussian distributions)

	W_2		
ϵ	IS	BUIS	MCMC
0.1	0.041	0.028	0.031
0.3	0.094	0.031	0.030
0.5	0.521	0.042	0.031

Table 3.2: Average Wasserstein distance between the empirical and actual reconciled distribution (binary hierarchy, Gaussian distributions)

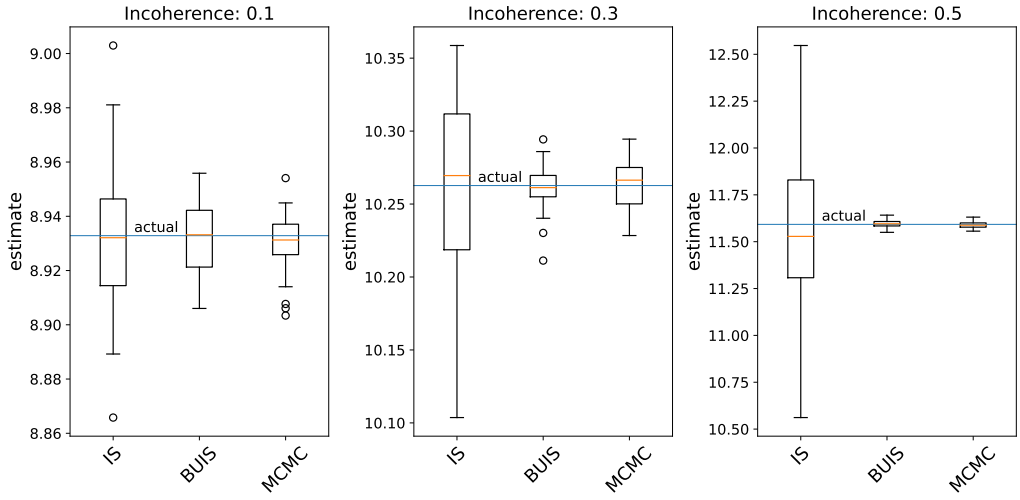


Figure 3.2: Boxplot of the reconciled mean of a bottom variable (binary hierarchy, Gaussian distributions)

	Average time		
hierarchy	IS	BUIS	MCMC
binary	0.06 s	0.17 s	33.9 s
weekly	-	1.47 s	1065.4 s

Table 3.3: Average computational times (Gaussian distributions)

Weekly hierarchy

We reconcile a weekly hierarchy with 52 bottom and 46 upper variables (Section 3.1.2). We run 30 experiments for the incoherence levels $\epsilon \in \{0.1, 0.3, 0.5\}$. We only compare BUIS and MCMC, since the dimension of the space is too large to use IS. Even with such a large hierarchy, BUIS achieves good results

(Tables 3.4 and 3.5), while its computational time is 3 orders of magnitude smaller than MCMC (second row of Table 3.3).

	Error	
ϵ	BUIS	MCMC
0.1	0.07 %	0.06 %
0.3	0.09 %	0.05 %
0.5	0.21 %	0.04 %

Table 3.4: MAPE on the reconciled mean (weekly hierarchy, Gaussian distributions)

	W_2	
ϵ	BUIS	MCMC
0.1	0.020	0.018
0.3	0.029	0.018
0.5	0.083	0.018

Table 3.5: Average Wasserstein distance between the empirical and actual reconciled distribution (weekly hierarchy, Gaussian distributions)

3.2.2 Reconciling Poisson forecasts

We now consider discrete base forecasts. We set a Poisson distribution on each bottom variable, with mean randomly chosen in the interval $[5, 10]$. We denote by $\widehat{\lambda}_b \in \mathbb{R}_+^8$ the vector of the base bottom means. As before, for each incoherence level $\epsilon \in \{0.1,\ 0.3,\ 0.5\}$, we set the mean of the upper variables as $\widehat{\lambda}_u = (1+\epsilon)\mathbf{A}\widehat{\lambda}_b$. In the Poisson case, the reconciled distribution cannot be analytically computed. We thus compare the results obtained using IS and BUIS with the results obtained using MCMC. Since probabilistic forecasts of count time series are typically given as samples (Liboschik et al. 2017), we also run sample-based BUIS (Section 3.1.1): we assume that the base distribution is unknown, and that only samples are available. The mean absolute percentage errors are computed with respect to the reconciled mean via MCMC (Table 3.6).

The boxplot of the reconciled mean of a bottom variable is shown in Figure 3.3. The results obtained using BUIS, in both cases, are similar to those

obtained using MCMC. Note that, for small incoherence levels, the standard deviation with MCMC is larger than with BUIS. Finally, the average computational times are reported in Table 3.7. Both BUIS and sample-based BUIS are two orders of magnitude faster than MCMC.

	Error wrt MCMC		
ϵ	IS	BUIS	BUIS w/samples
0.1	0.36 %	0.37 %	0.37 %
0.3	0.35 %	0.32 %	0.33 %
0.5	0.51 %	0.33 %	0.33 %

Table 3.6: MAPE on the reconciled mean (binary hierarchy, Poisson distributions)

	Average time			
hierarchy	IS	BUIS	BUIS w/samples	MCMC
binary	0.12 s	0.22 s	0.31 s	35.5 s
weekly	-	2.10 s	2.69 s	2417.8 s

Table 3.7: Average computational times (Poisson distributions)

Weekly hierarchy

The Mean absolute percentage errors over 30 experiments, using a weekly hierarchy with Poisson base distributions, are reported in Table 3.8; our reference method is MCMC. Note that the dimension of the space is too large to use IS. Even in the case of such a large hierarchy, using BUIS we are able to achieve a very small error. Finally, the average computational times are reported in Table 3.7. BUIS and sample-based BUIS are about 3 orders of magnitude faster than MCMC. Note that sample-based BUIS is almost as fast as BUIS, despite the density estimation step.

3.3 Experiments on real data

We now perform probabilistic reconciliation on temporal hierarchies, using time series extracted from two different data sets: *carparts*, available from

	Error wrt MCMC	
ϵ	BUIS	BUIS w/samples
0.1	0.34 %	0.33 %
0.3	0.36 %	0.36 %
0.5	1.09 %	1.07 %

Table 3.8: MAPE on the reconciled mean (weekly hierarchy, Poisson distributions)

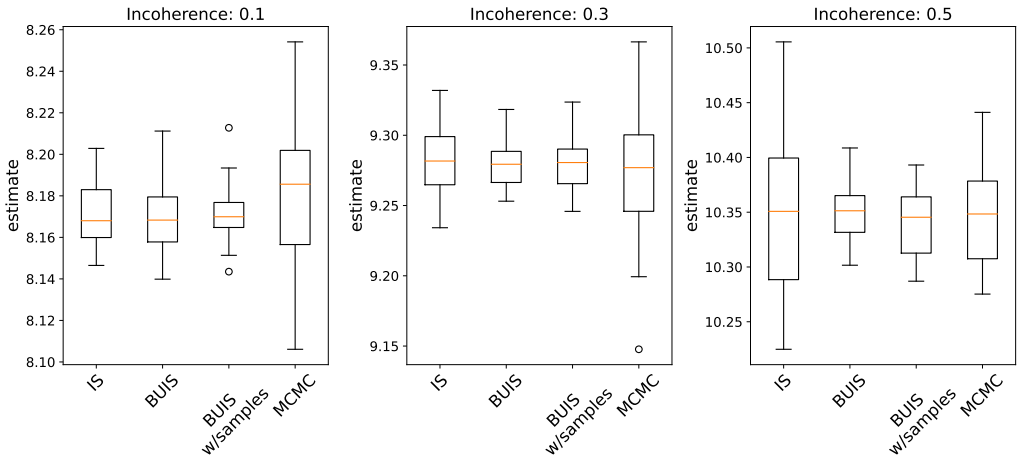


Figure 3.3: Boxplot of the reconciled mean of a bottom variable (binary hierarchy, Poisson distributions)

the R package *expsmooth* (Hyndman 2018), and *syph*, available from the R package *ZIM* (Yang et al. 2018).

The *carparts* data set is about monthly sales of car parts. As in Hyndman et al. 2008, Chapter 16, we remove time series with missing values, with less then 10 positive monthly demands and with no positive demand in the first 15 and final 15 months. After this selection, there are 1046 time series left. Note that we use less restrictive criteria in the selection of the time series than Corani et al. 2022, where only 219 time series from *carparts* were considered. Monthly data are aggregated into 2-months, 3-months, 4-months, 6-months and 12-months levels.

The *syph* data set is about the weekly number of syphilis cases in the United States. We remove the time series with ADI greater than 20. The ADI is computed as $ADI = \frac{\sum_{i=1}^{P} p_i}{P}$, where p_i is the time period between two non-zeros values and P is the total number of periods (Syntetos and Boylan 2005). We also remove the time series corresponding to the total number of

cases in the US. After this selection, there are 50 time series left. Weekly data are aggregated into 2-weeks, 4-weeks, 13-weeks, 26-weeks and 52-weeks levels.

For both data sets, we fit a Generalized Linear Models using the *tscount* package (Liboschik et al. 2017). We use a negative binomial predictive distribution, with a first-order regression on past observations. The test set has length 1 year for both data sets. We thus compute up to 12 steps ahead at monthly level, and up to 52 steps ahead at weekly level. Probabilistic forecasts are returned in the form of samples.

Reconciliation is performed in three different ways. In the first case, we fit a Gaussian distribution on the returned samples. Then, we follow (Corani et al. 2020) to analytically compute the Gaussian reconciled distribution. In the second case, we fit a negative binomial distribution on the samples, and we reconcile using the BUIS algorithm. Since for both data sets Assumption 2 does not hold, we use the method of Section 3.1.2 for grouped time series. Finally, we use the sample-based BUIS directly on the samples, as explained in Section 3.1.1. Although the sample-based algorithm is slightly slower, this method yields a computational gain over BUIS, as fitting a negative binomial distribution on the samples requires about 1.2 s for the monthly hierarchy and 3.9 s for the weekly hierarchy. We refer to these methods, respectively, as *N*, *NB*, and *samples*. Furthermore, we denote by *base* the unreconciled forecasts.

We use different indicators to assess the performance of each method. The mean scaled absolute error (MASE) (Hyndman 2006) is defined as

$$\text{MASE} = \frac{\text{MAE}}{Q},$$

where $\text{MAE} = \frac{1}{h}\sum_{j=1}^{h}|y_{t+j} - \hat{y}_{t+j|t}|$ and $Q = \frac{1}{T-1}\sum_{t=2}^{T}|y_t - y_{t-1}|$. Here, y_t denotes the value of the time series at time t, while $\hat{y}_{t+j|t}$ denotes the point forecast computed at time t for time $t+j$. The median of the distribution is used as point forecast, since it minimizes MASE (Kolassa 2016).

The mean interval score (MIS) (Gneiting 2011) is defined, for any $\alpha \in (0,1)$, as

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

where l and u are the lower and upper bounds of the $(1-\alpha)$ forecast coverage interval and y is the actual value of the time series. In the following, we use $\alpha = 0.1$. MIS penalizes wide prediction intervals, as well as intervals that do not contain the true value.

metric	hier-level	*N* vs *base*	*NB* vs *base*	*samples* vs *base*
ES		0.07	0.52	**0.53**
MASE	Monthly	-1.02	**0.14**	0.13
	2-Monthly	-0.53	0.25	**0.27**
	Quarterly	-0.42	0.21	**0.26**
	4-Monthly	-0.40	0.16	**0.21**
	Semiannual	-0.33	0.14	**0.16**
	Annual	-0.26	**0.18**	0.17
	average	-0.49	0.18	**0.20**
MIS	Monthly	-0.08	0.45	**0.63**
	2-Monthly	0.28	0.45	**0.56**
	Quarterly	0.22	0.43	**0.46**
	4-Monthly	0.03	0.35	**0.36**
	Semiannual	-0.07	**0.37**	0.26
	Annual	-0.17	**0.40**	0.22
	average	0.03	0.41	**0.42**

Table 3.9: Skill scores on the time series extracted from carparts, detailed by each level of the hierarchy

Finally, the Energy score (Székely and Rizzo 2013) is defined as

$$ES(P, \mathbf{y}) = \mathbb{E}_P\left[\|\mathbf{y} - \mathbf{s}\|^\alpha\right] - \frac{1}{2}\mathbb{E}_P\left[\|\mathbf{s} - \mathbf{s}'\|^\alpha\right],$$

where P is the forecast distribution on the whole hierarchy, $\mathbf{s}, \mathbf{s}' \sim P$ are a pair of independent random variables and $\mathbf{y}$ is the vector of the actual values of all the time series. The energy score is a proper scoring rule for distributions defined on the entire hierarchy (Panagiotelis et al. 2022). We compute ES, with $\alpha = 2$, using samples, as explained in Wickramasuriya 2021.

We use the skill score to compare the performance of a method with respect to a baseline method, in terms of percentage improvement. We use *base* as baseline method. For example, the skill score of *NB* on MASE is given by

$$\text{Skill}(NB, base) = \frac{\text{MASE}(base) - \text{MASE}(NB)}{\left(\text{MASE}(base) + \text{MASE}(NB)\right)/2}. \tag{3.2}$$

In the literature, the skill score is often defined using MASE(*base*) as the denominator in (3.2) (Wheatcroft 2019). However, we believe that our defi-

metric	hierc-level	N vs base	NB vs base	samples vs base
ES		0.08	0.11	**0.15**
MASE	Weekly	-0.63	**0.14**	**0.14**
	2-Weekly	-0.40	**0.16**	0.14
	4-Weekly	-0.22	**0.13**	0.12
	Quarterly	-0.10	0.01	**0.04**
	Semiannual	0.01	0.07	**0.15**
	Annual	-0.05	-0.00	**0.04**
	average	-0.23	0.08	**0.10**
MIS	Weekly	-0.06	**0.46**	0.45
	2-Weekly	0.08	0.33	**0.34**
	4-Weekly	0.03	0.19	**0.25**
	Quarterly	-0.15	-0.11	**-0.08**
	Semiannual	-0.34	-0.27	**-0.21**
	Annual	-0.33	-0.23	**-0.22**
	average	-0.13	0.06	**0.09**

Table 3.10: Skill scores on the time series extracted from syph, detailed by each level of the hierarchy

nition has two main advantages. First, it is symmetric. Second, the skill score is well-defined even if the baseline error is zero, and moreover it always lies between -2 and 2. For each level, since the skill score is scale-independent, we compute it for each forecasting horizon, and take the average.

The skill scores for *carparts* are reported in Table 3.9. Both *NB* and *samples* methods yield a significant improvement for all the indicators, and for all the hierarchy levels. For both methods, the average improvement is about 20% for MASE, 40% for MIS and 50% for ES. The skill scores for *syph* are reported in Table 3.10. As before, the average improvement of *NB* and *samples* is significant for all indicators. For both datasets, the *N* method performs poorly, in many cases yielding negative skill scores. As observed in Corani et al. 2022, this method does not capture the asymmetry of the base forecasts. Finally, *samples* appears to perform better that *NB*. Indeed, the step of fitting a Negative Binomial distribution on the forecast samples may yield an additional source of error.

Chapter 4

Reconciliation effects and application

This chapter is organized as follows. In Section 4.1, we focus on the Gaussian case. In Section 4.2 and 4.3, we analyze the effect of the reconciliation on (respectively) the variance and the mean of the forecast distribution, and we present some examples with Bernoulli and Poisson distributions. In Section 4.4, we present the multivariate score-driven model introduced by Agosto 2022 and the data set. Finally, in Section 4.5, we reconcile all the 3508 daily forecasts, obtaining a large improvement in the performance, and we observe the effects discussed before.

4.1 Gaussian case

As shown in Section 2.3, when the base forecast distribution is a multivariate Gaussian, the reconciled distribution can be analytically computed. Let

$$\widehat{\mathbf{Y}} = \begin{bmatrix} \widehat{\mathbf{U}} \\ \widehat{\mathbf{B}} \end{bmatrix} \sim \mathcal{N}\left(\widehat{\mathbf{y}},\, \widehat{\boldsymbol{\Sigma}}_Y\right) \tag{4.1}$$

be the base forecast distribution for the entire hierarchy, where

$$\widehat{\mathbf{y}} = \begin{bmatrix} \widehat{\mathbf{u}} \\ \widehat{\mathbf{b}} \end{bmatrix}, \quad \widehat{\boldsymbol{\Sigma}}_Y = \begin{bmatrix} \widehat{\boldsymbol{\Sigma}}_U & \widehat{\boldsymbol{\Sigma}}_{UB} \\ \widehat{\boldsymbol{\Sigma}}_{UB}^T & \widehat{\boldsymbol{\Sigma}}_B \end{bmatrix}.$$

The reconciled bottom and upper distributions are then multivariate Gaussian:

$$\widetilde{\mathbf{B}} \sim \mathcal{N}\left(\widetilde{\mathbf{b}},\, \widetilde{\boldsymbol{\Sigma}}_B\right), \qquad \widetilde{\mathbf{U}} \sim \mathcal{N}\left(\widetilde{\mathbf{u}},\, \widetilde{\boldsymbol{\Sigma}}_U\right),$$

where

$$\widetilde{\mathbf{b}} = \widehat{\mathbf{b}} + \left(\widehat{\mathbf{\Sigma}}_{UB}^T - \widehat{\mathbf{\Sigma}}_B A^T\right) \mathbf{Q}^{-1}(\mathbf{A}\widehat{\mathbf{b}} - \widehat{\mathbf{u}}), \tag{4.2}$$

$$\widetilde{\mathbf{\Sigma}}_B = \widehat{\mathbf{\Sigma}}_B - \left(\widehat{\mathbf{\Sigma}}_{UB}^T - \widehat{\mathbf{\Sigma}}_B \mathbf{A}^T\right) \mathbf{Q}^{-1} \left(\widehat{\mathbf{\Sigma}}_{UB}^T - \widehat{\mathbf{\Sigma}}_B \mathbf{A}^T\right)^T, \tag{4.3}$$

$$\widetilde{\mathbf{u}} = \widehat{\mathbf{u}} + \left(\widehat{\mathbf{\Sigma}}_U - \widehat{\mathbf{\Sigma}}_{UB} \mathbf{A}^T\right) \mathbf{Q}^{-1}(\mathbf{A}\widehat{\mathbf{b}} - \widehat{\mathbf{u}}), \tag{4.4}$$

$$\widetilde{\mathbf{\Sigma}}_U = \widehat{\mathbf{\Sigma}}_U - \left(\widehat{\mathbf{\Sigma}}_U - \widehat{\mathbf{\Sigma}}_{UB} \mathbf{A}^T\right) \mathbf{Q}^{-1} \left(\widehat{\mathbf{\Sigma}}_U - \widehat{\mathbf{\Sigma}}_{UB} \mathbf{A}^T\right)^T, \tag{4.5}$$

and $\mathbf{Q} := \widehat{\mathbf{\Sigma}}_U - \widehat{\mathbf{\Sigma}}_{UB}\mathbf{A}^T - \mathbf{A}\widehat{\mathbf{\Sigma}}_{UB}^T + \mathbf{A}\widehat{\mathbf{\Sigma}}_B\mathbf{A}^T$.

Note that the reconciled variance does not depend on the point forecasts, but only on the base variances. Moreover, the following proposition shows that the variance of each variable decreases after reconciliation (the proof is in Appendix A.5).

Proposition 3. *For each $i = 1, \ldots, m$, and $j = 1, \ldots, n - m$, we have*

$$\begin{aligned} \mathrm{Var}(\widetilde{B}_i) &\leq \mathrm{Var}(\widehat{B}_i), \\ \mathrm{Var}(\widetilde{U}_j) &\leq \mathrm{Var}(\widehat{U}_j). \end{aligned} \tag{4.6}$$

Moreover, we observe that the shift applied to the base forecast mean is proportional to $\mathbf{A}\widehat{\mathbf{b}} - \widehat{\mathbf{u}}$, which is often called *incoherence*. Let us now consider a simple hierarchy with 1 upper and 2 bottom variables, as in Figure 4.1.

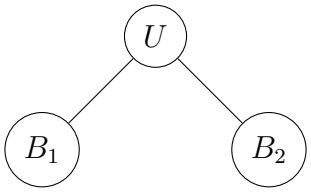


Figure 4.1: A simple hierarchy

Let us assume that $\widehat{B}_1$, $\widehat{B}_2$, and $\widehat{U}$ are independent and Gaussian-distributed:

$$\widehat{B}_1 \sim \mathcal{N}(\widehat{b}_1, \widehat{\sigma}_1^2), \quad \widehat{B}_2 \sim \mathcal{N}(\widehat{b}_2, \widehat{\sigma}_2^2), \quad \widehat{U} \sim \mathcal{N}(\widehat{u}, \widehat{\sigma}_U^2).$$

From (4.2) and (4.4), the reconciled means are given by

$$\begin{aligned} \widetilde{b}_1 &= (1 - g_1)\widehat{b}_1 + g_1(\widehat{u} - \widehat{b}_2), \\ \widetilde{b}_2 &= (1 - g_2)\widehat{b}_2 + g_2(\widehat{u} - \widehat{b}_1), \\ \widetilde{u} &= (1 - g_u)\widehat{u} + g_u(\widehat{b}_1 + \widehat{b}_2), \end{aligned} \tag{4.7}$$

where $g_1 = \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2+\sigma_u^2}$, $g_2 = \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2+\sigma_u^2}$, and $g_u = \frac{\sigma_u^2}{\sigma_1^2+\sigma_2^2+\sigma_u^2}$.

The reconciled mean of U is thus a convex combination of the base mean $\widehat{u}$ and the bottom-up mean $\widehat{b}_1 + \widehat{b}_2$. Indeed, the reconciliation merges the information coming from the base forecast distribution of the bottom and the upper variables. Note that, if $\sigma_U = 0$, we have $g_u = 0$ and thus $\widetilde{u} = \widehat{u}$. Indeed, there is no uncertainty in the forecast of U, hence only the bottom point forecast are adjusted in order to have $\widetilde{b}_1 + \widetilde{b}_2 = \widehat{u}$:

$$\widetilde{b}_1 = \widehat{b}_1 + \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}(\widehat{u} - \widehat{b}_1 - \widehat{b}_2),$$
$$\widetilde{b}_2 = \widehat{b}_2 + \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}(\widehat{u} - \widehat{b}_1 - \widehat{b}_2).$$

On the other hand, if σ_U is very large compared to σ_1 and σ_2, we have $g_1 \approx 0$, $g_2 \approx 0$, and $g_u \approx 1$, hence

$$\widetilde{b}_1 \approx \widehat{b}_1, \quad \widetilde{b}_2 \approx \widehat{b}_2, \quad \widetilde{u} \approx \widehat{b}_1 + \widehat{b}_2.$$

This corresponds to a bottom-up approach: if the uncertainty on the prediction of the upper variable is very large, only the information coming from the bottom variables is taken into account.

4.2 Reconciled variance

In the Gaussian case, the variance of the reconciled distribution is always smaller than the variance of the base distribution. This is somehow analogous to the Gaussian conjugate model in Bayesian statistics, where the variance of the posterior distribution is always guaranteed to decrease as we get more data. In the non-Gaussian case, however, the variance of the posterior may increase if the new observations are not coherent with prior beliefs (Gelman 2011). Analogously, the variance of the reconciled distribution may be larger than the variance of the base distribution. Intuitively, we expect this behaviour when there are conflicting information coming from the base forecast distributions.

We now show that in the case of count variables, the variance can in fact increase after reconciliation.

Proposition 4. *Let us assume that* $p := \mathbb{P}\big(\widehat{\mathbf{U}} = \mathbf{A}\widehat{\mathbf{B}}\big) > 0$. *Then, for any* $j = 1, \dots, m$, *we have*

$$\operatorname{Var}\big[\widetilde{B}_j\big] = \frac{\operatorname{Var}\big(\widehat{B}_j\big) - (1-p)\,\operatorname{Var}\big[\widehat{B}_j | \widehat{\mathbf{U}} \neq \mathbf{A}\widehat{\mathbf{B}}\big] - p(1-p)\,(a-b)^2}{p}, \tag{4.8}$$

where $a := \mathbb{E}\big[\widehat{B}_j|\widehat{\mathbf{U}} \neq \mathbf{A}\widehat{\mathbf{B}}\big]$ *and* $b := \mathbb{E}\big[\widehat{B}_j|\widehat{\mathbf{U}} = \mathbf{A}\widehat{\mathbf{B}}\big]$.

The proof is reported in Appendix A.6. The term $p = \mathbb{P}\big(\widehat{\mathbf{U}} = \mathbf{A}\widehat{\mathbf{B}}\big)$ represents the probability of coherence, according to the base forecasts. From (4.8), if p is small enough, the reconciled variance might be greater then the base variance. Indeed, in this case, there is conflict between the information coming from the bottom and the upper distributions. We present two examples using Bernoulli and Poisson distributions.

4.2.1 Bernoulli example

Let us consider the hierarchy in Figure 4.1, with 1 upper and 2 bottom variables. We now assume that the base bottom distributions are given by independent Bernoulli:

$$\widehat{B}_1 \sim \mathcal{B}(p_1), \quad \widehat{B}_2 \sim \mathcal{B}(p_2),$$

for some $p_1, p_2 \in [0,1]$. We denote by $\widehat{\pi}_1$ and $\widehat{\pi}_2$, respectively, the probability mass functions of $\widehat{B}_1$ and $\widehat{B}_2$, so that $\widehat{\pi}_1(0) = 1-p_1$, $\widehat{\pi}_1(1) = p_1$, and $\widehat{\pi}_1(k) = 0$ for any $k \neq 0, 1$.

The base distribution of the upper variable is given by

$$\widehat{U} = \begin{cases} 0 & \text{prob} = q_0 \\ 1 & \text{prob} = q_1 \\ 2 & \text{prob} = q_2, \end{cases}$$

hence the probability mass function π_U of $\widehat{U}$ is defined as $\widehat{\pi}_U(0) = q_0$, $\widehat{\pi}_U(1) = q_1$, $\widehat{\pi}_U(2) = q_2$, and $\widehat{\pi}_U(k) = 0$ for any $k \neq 0, 1, 2$.

Then, the probability mass function $\widetilde{\pi}$ of the reconciled distribution of the bottom variables is given by

$$\widetilde{\pi}(b_1, b_2) \propto \widehat{\pi}_1(b_1)\widehat{\pi}_2(b_2)\widehat{\pi}_U(b_1 + b_2),$$

so that the reconciled bottom distribution may be expressed as

$$(\widetilde{B}_1, \widetilde{B}_2) = \begin{cases} (0,0) & \text{prob} = (1-p_1)(1-p_2)q_0/S \\ (1,0) & \text{prob} = p_1(1-p_2)q_1/S \\ (0,1) & \text{prob} = (1-p_1)p_2q_1/S \\ (1,1) & \text{prob} = p_1p_2q_2/S, \end{cases}$$

where $S := (1-p_1)(1-p_2)q_0 + p_1(1-p_2)q_1 + (1-p_1)p_2q_1 + p_1p_2q_2$ is the normalizing constant. Hence

$$\widetilde{B}_1 \sim \mathcal{B}(\widetilde{p}_1), \quad \widetilde{B}_2 \sim \mathcal{B}(\widetilde{p}_2),$$

with

$$\begin{aligned}\widetilde{p}_1 &= \frac{[(1-p_2)q_1 + p_2 q_2]p_1}{S},\\ \widetilde{p}_2 &= \frac{[(1-p_1)q_1 + p_1 q_2]p_2}{S}.\end{aligned} \tag{4.9}$$

Moreover

$$\widetilde{U} = \begin{cases} 0 & \text{prob} = (1-p_1)(1-p_2)q_0/S \\ 1 & \text{prob} = (p_1+p_2-2p_1p_2)q_1/S \\ 2 & \text{prob} = p_1p_2q_2/S. \end{cases} \tag{4.10}$$

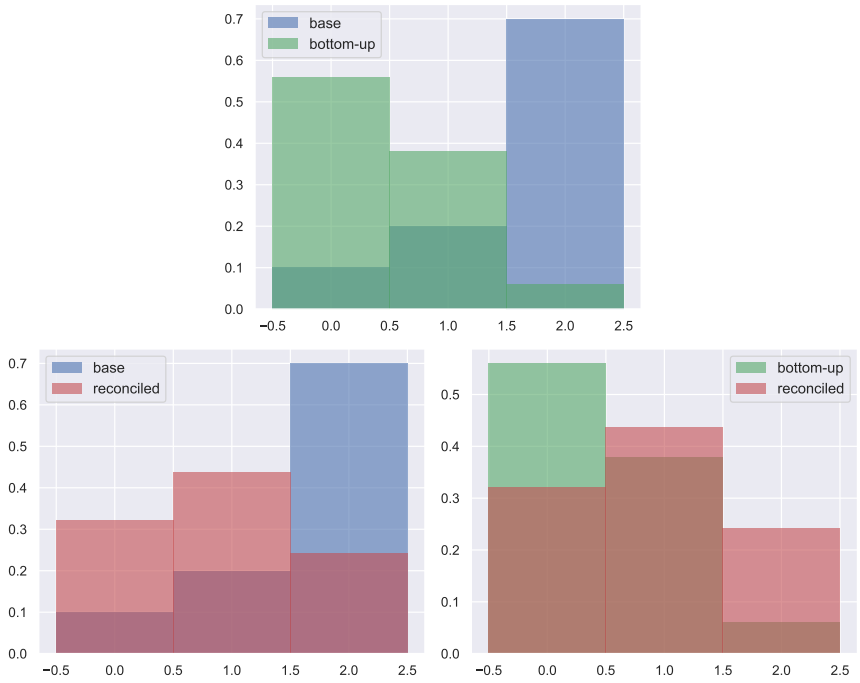


Figure 4.2: Probability mass function of U ($p_1 = 0.3$, $p_2 = 0.2$, $\boldsymbol{q} = [0.1, 0.2, 0.7]$)

Let us now set $p_1 = 0.3$, $p_2 = 0.2$, and $\boldsymbol{q} = [0.1, 0.2, 0.7]$. From (4.9) and (4.10), we have

$$\begin{aligned}\widetilde{p}_1 &\approx 0.52,\\ \widetilde{p}_2 &\approx 0.40,\\ \widetilde{\mathbf{q}} &\approx [0.32, 0.44, 0.24].\end{aligned} \tag{4.11}$$

In Figure 4.2, we compare the base, bottom-up, and reconciled distribution of the upper variable. Since the information provided by the bottom and the upper variables are in conflict, the resulting reconciled distribution is more spread across the domain. Indeed, the variance of all the variables increases after reconciliation:

$$\begin{aligned} \mathrm{Var}[\widehat{B}_1] = 0.21, \quad \mathrm{Var}[\widetilde{B}_1] \approx 0.25, \\ \mathrm{Var}[\widehat{B}_2] = 0.16, \quad \mathrm{Var}[\widetilde{B}_2] \approx 0.24, \\ \mathrm{Var}[\widehat{U}] = 0.44, \quad \mathrm{Var}[\widetilde{U}] \approx 0.56. \end{aligned}$$

4.2.2 Poisson example

Let us consider the same hierarchy, but we now assume to deal with independent Poisson base distributions:

$$\widehat{B}_1 \sim Poi(\lambda_1), \quad \widehat{B}_2 \sim Poi(\lambda_2), \quad \widehat{U} \sim Poi(\lambda_u),$$

for some $\lambda_1, \lambda_2, \lambda_u > 0$. In this case, reconciliation cannot be performed analytically. We thus use importance sampling to sample from the reconciled distribution.

We set $\lambda_1 = 0.5$, $\lambda_2 = 0.8$, and $\lambda_u = 6.0$. Then, we have

$$\begin{aligned} \mathrm{Var}[\widehat{B}_1] = 0.5, \quad \mathrm{Var}[\widetilde{B}_1] \approx 0.81, \\ \mathrm{Var}[\widehat{B}_2] = 0.8, \quad \mathrm{Var}[\widetilde{B}_2] \approx 1.13, \\ \mathrm{Var}[\widehat{U}] = 6.0, \quad \mathrm{Var}[\widetilde{U}] \approx 1.40. \end{aligned}$$

Since we have a large incoherence, the variance of the bottom variables increases. In Figure 4.3 we show the probability mass function of all the variables before and after reconciliation. In Figure 4.4, we compare the base, bottom-up, and reconciled distribution of the upper variable.

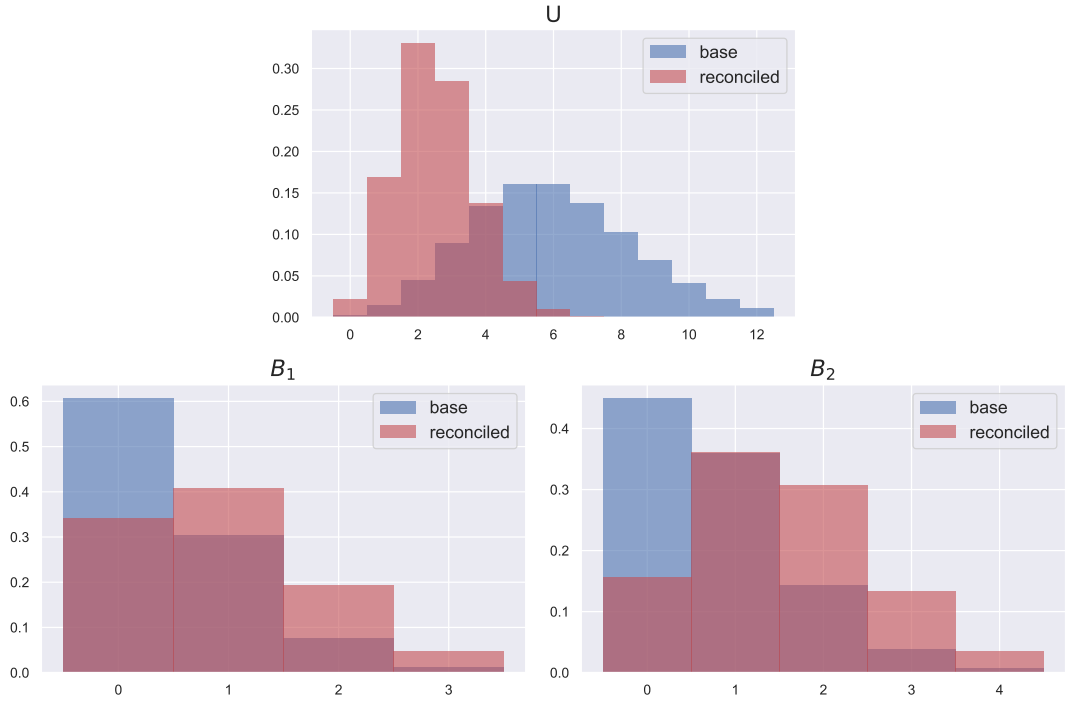


Figure 4.3: Base and reconciled probability mass functions of $B1$, B_2, and U ($\lambda_1 = 0.5$, $\lambda_2 = 0.8$, $\lambda_u = 6.0$)

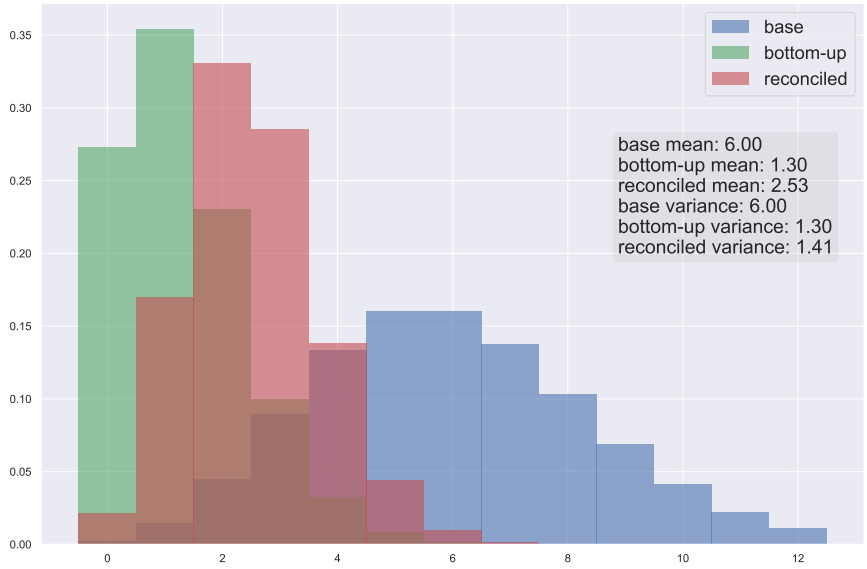


Figure 4.4: Base, bottom-up, and reconciled probability mass functions of U ($\lambda_1 = 0.5$, $\lambda_2 = 0.8$, $\lambda_u = 6.0$)

4.3 Reconciled mean

In the Gaussian case, the reconciled mean is given by a compromise between the base means of all the variables. In particular, from (4.7), we see that the reconciled mean of U is a convex combination of the base mean of U and the bottom-up mean. As before, this is analogous to the Gaussian conjugate model, where the posterior expectation is a convex combination of the prior expectation and the sample mean (Diaconis and Ylvisaker 1979).

This "compromise" effect has also been observed by Corani et al. 2022, using Poisson distributions. However, we show that this is not the only possible behavior. As explained before, if the incoherence is not large we typically observe a reduction of the variance of the forecast distribution. Indeed, in this case, the information provided by the bottom and the upper base distributions are consistent with each other: hence, the uncertainty decreases, and the mass gets concentrated on the values that are more likely. We refer to this effect as "strengthening" effect. The tail of the distribution is typically shortened as the variance decreases: if we deal with asymmetric distributions, this leads to a shift of the mean in the direction opposite to the tail. The reconciled mean of the upper variable may thus be lower than both the base upper mean and the bottom-up mean. We show these two different behaviors, first using a Poisson example, then through an application to financial count data time series (Section 4.5).

4.3.1 Poisson example

Let us consider the same example as in Section 4.2.2. Depending on the parameters of the base distributions, we observe the two different effects of the reconciliation described above.

Example 1: "strengthening" effect
We set $\lambda_1 = 0.5$, $\lambda_2 = 0.8$, and $\lambda_u = 1.5$. Then, we have

$$\begin{aligned} \mathbb{E}[\widehat{B}_1] = 0.5, \quad \mathbb{E}[\widetilde{B}_1] \approx 0.43, \\ \mathbb{E}[\widehat{B}_2] = 0.8, \quad \mathbb{E}[\widetilde{B}_2] \approx 0.68, \\ \mathbb{E}[\widehat{U}] = 1.50, \quad \mathbb{E}[\widetilde{U}] \approx 1.11. \end{aligned}$$

Note that the means of all the variables decrease after reconciliation. In Figure 4.5, we show the probability mass functions of all the variables before and after reconciliation. We observe a shift to the left as the tails of the distributions get thinner. In Figure 4.6, we compare the base distribution, the bottom-up distribution, and the reconciled distribution of the upper variable.

The effect of the reconciliation is to strengthen the information provided by the base and the bottom-up distributions, reducing the uncertainty.

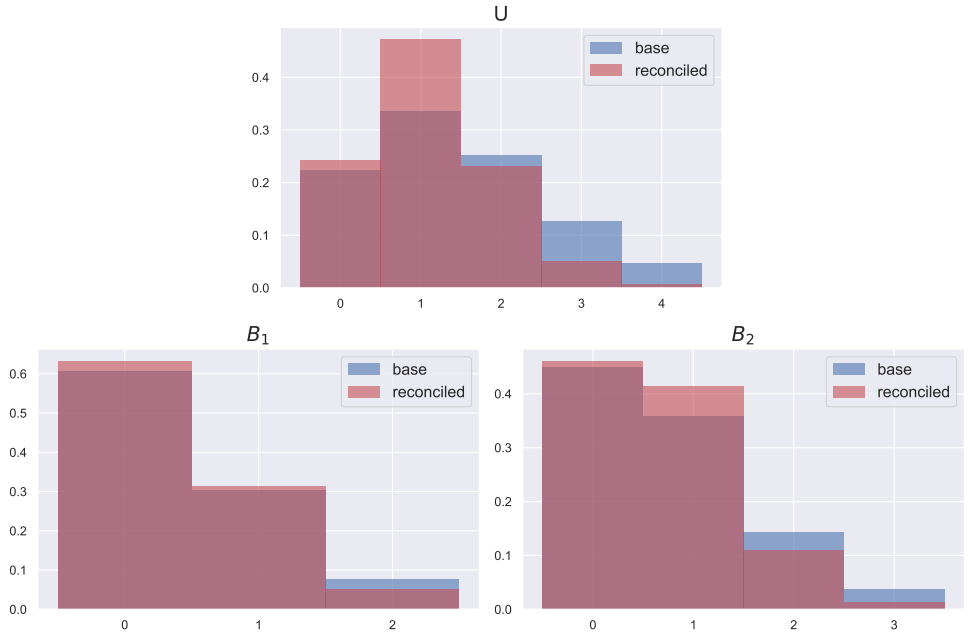


Figure 4.5: Effect of the reconciliation on the probability mass function of $B1$, B_2, and U ($\lambda_1 = 0.5$, $\lambda_2 = 0.8$, $\lambda_u = 1.5$)

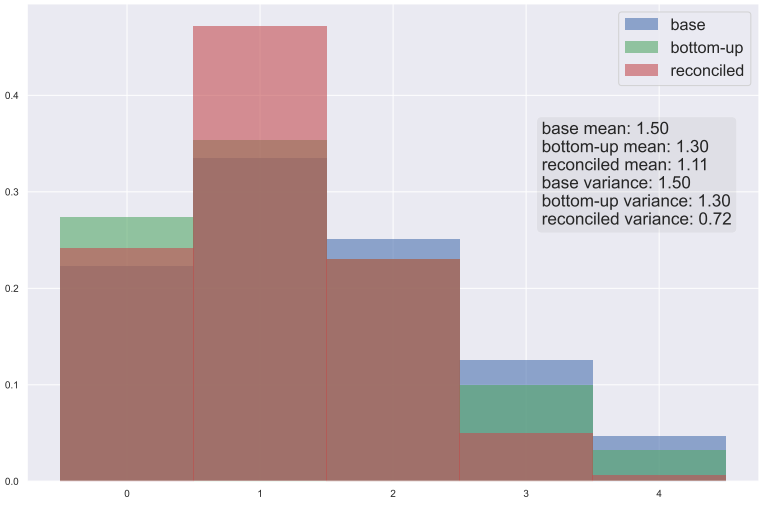


Figure 4.6: Effect of the reconciliation on the probability mass function of U ($\lambda_1 = 0.5$, $\lambda_2 = 0.8$, $\lambda_u = 1.5$)

Example 2: "compromise" effect

We now set $\lambda_1 = 5$, $\lambda_2 = 7$, and $\lambda_u = 18$. Then, we have

$$\begin{aligned} \mathbb{E}[\widehat{B}_1] &= 5, \quad \mathbb{E}[\widetilde{B}_1] \approx 6.02, \\ \mathbb{E}[\widehat{B}_2] &= 7, \quad \mathbb{E}[\widetilde{B}_2] \approx 8.43, \\ \mathbb{E}[\widehat{U}] &= 18, \quad \mathbb{E}[\widetilde{U}] \approx 14.44. \end{aligned}$$

In this case, the bottom means increase after reconciliation, while the upper mean decreases. In Figure 4.7, we show the probability mass functions of all the variables before and after reconciliation. We observe a shift to the left for the upper distribution, and to the right for the bottom distributions. In Figure 4.8, we compare the base, bottom-up, and reconciled distribution of the upper variable. The behavior is analogous to the Gaussian case: the reconciled distribution merges the information coming from the bottom and from the upper distributions.

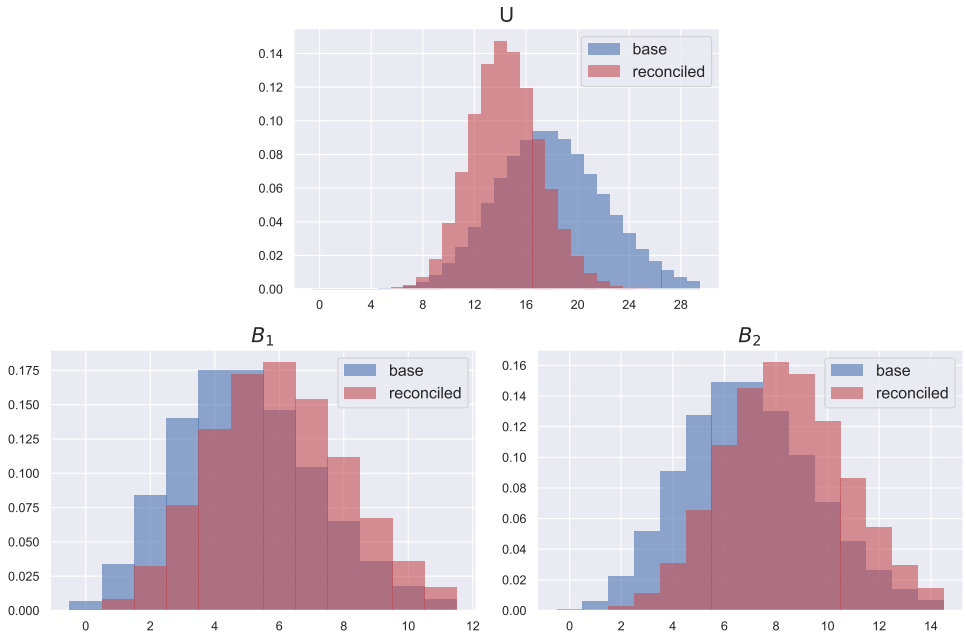


Figure 4.7: Effect of the reconciliation on the probability mass function of $B1$, B_2, and U ($\lambda_1 = 5$, $\lambda_2 = 7$, $\lambda_u = 18$)

Figure 4.8: Effect of the reconciliation on the probability mass function of U ($\lambda_1 = 5$, $\lambda_2 = 7$, $\lambda_u = 18$)

4.4 Model and data set

4.4.1 Multivariate score-driven models for count time series

Agosto 2022 proposed a multivariate negative binomial score-driven specification, assuming that the observations in each time series i follow a negative binomial distribution with a time-varying location parameter $\mu_{it} > 0$ and a static dispersion parameter $\alpha_i \geq 0$:

$$X_{it} \sim NB(\mu_{it}, \alpha_i) \tag{4.12}$$

The probability mass function is the following:

$$P[X_{it} = x_{it}|\mu_{it}, \alpha_i] = \frac{\Gamma(x_{it} + \alpha_i^{-1})}{\Gamma(x_{it} + 1)\Gamma(\alpha_i^{-1})} \left(\frac{\alpha_i^{-1}}{\alpha_i^{-1} + \mu_{it}}\right)^{\alpha_i^{-1}} \left(\frac{\mu_{it}}{\alpha_i^{-1} + \mu_{it}}\right)^{x_{it}} \tag{4.13}$$

for $i = 1, ..., k$ and $t = 1, ...T$.

The time-varying location parameters f_t follows a Generalized Autoregressive Score (GAS) specification (Creal, 2013, Harvey, 2013). In the general GAS specification, the dynamics of filtered parameters $\mathbf{f}_{t+1} = (f_1, ..., f_t)$ are captured by an autoregressive term and by the scaled score (gradient) of the conditional observation density through the recursions

$$\mathbf{f}_{t+1} = \mathbf{G} + \mathbf{H}\mathbf{f}_t + \mathbf{L}\ S(\mathbf{f}_t)\nabla(\mathbf{x}_t, \mathbf{f}_t) \tag{4.14}$$

where $\mathbf{f}_t = (f_{1t}, ..., f_{kt})$ is the vector of time-varying parameters, $\mathbf{G} = (g_1, ..., g_k)$ are the constant parameters, $\mathbf{H} = \text{diag}(h_1, ..., h_k)$ is the $k \times k$ diagonal matrix of autoregressive parameters, $\mathbf{L}$ is the $k \times k$ matrix of coefficients associated to the scaled score and $S(\mathbf{f}_t)$ is a scaling function for the score $\nabla(\mathbf{x}_t, \mathbf{f}_t)$.
Moreover, following Heinen and Rengifo 2007 and Escribano and Maggi 2019, Agosto 2022 assumes:

$$\mathbf{L} = \text{diag}(\mathbf{e}) + \boldsymbol{\gamma}\boldsymbol{\delta}' \tag{4.15}$$

where $\mathbf{e}$, $\boldsymbol{\gamma}$, $\boldsymbol{\delta} \in \mathbb{R}^k$ are column vectors. In addition, to be able to estimate the values of $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$, we impose $\delta_k = 1 - \sum_{i=1}^{k-1} \delta_i$.

The score $\nabla(\mathbf{x}_t, \mathbf{f}_t)$ corresponds to the first derivative of the negative binomial log-likelihood function:

$$\nabla(\mathbf{x}_t, \mathbf{f}_t) = \frac{\mathbf{x}_t - \exp(\mathbf{f}_t)}{\boldsymbol{\alpha}\exp(\mathbf{f}_t) + 1} \tag{4.16}$$

Agosto 2022 applied the model to the analysis of dependence between time series of extreme market event counts in different economic sectors. In such a context,the parameters entering the score filter dynamics can be interpreted as follows:

- The $\mathbf{G}$ constant parameters determine the unconditional and long-term mean of the number of events $\mathbf{f} = (\mathbf{I} - \mathbf{H})^{-1}\mathbf{G}$.
- The $\mathbf{H}$ coefficients express dependence of the expected number of extreme market events on the past expectations, allowing to capture long-memory effects in the count processes.
- The $\mathbf{L}$ matrix expresses in-sector and cross-sector dependence through the score. Being the latter calculated as the scaled difference between the observed and expected number of events at the previous time, the $\mathbf{L}$ coefficients determine the impact of shocks in the extreme event counts occurred in $t-1$ on the expected number of extreme events in t in the same sector (diagonal effects) and in other sectors (off-diagonal effects). Formulation (4.15) gives a further insight into the interpretation of parameters: $\mathbf{e}$ measures the own effect of shock events in sector i. The $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ vectors act instead as multipliers of the off-diagonal elements of $\mathbf{L}$.

The $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \mathbf{G}, \mathbf{H}, \mathbf{e}, \boldsymbol{\gamma}, \boldsymbol{\delta})'$ parameter vector describing the score-driven dynamics (4.14) can be estimated by maximum likelihood maximization (see Agosto, 2022 for details).

4.4.2 Empirical analysis

Data

This model has been applied to the count time series of extreme CDS returns of companies belonging to the Euro Stoxx 50 index in the period from 31 December 2004 to 19 December 2018, calculated from Bloomberg data and used in Agosto (2022). An extreme market event is considered to occur at time (day) t in a given asset if the corresponding observation exceeds the 90-th percentile of the CDS spread distribution of the same asset in the last trading year. Each count time series corresponds to one of the following industries: Financial (FIN), Information and Communication Technology (ICT), Manufacturing (MFG), Energy (ENG), Trade (TRD). As it can be seen from the descriptive statistics provided in Table 4.1, all the considered series show a high frequency of zeros and are overdispersed, i.e. their variance is higher than the mean. These features motivate the use of count data models for rare events allowing for possible zero-inflation and overdispersion.

Sector	Number of companies	Mean	Standard deviation	Frequency of zeros
FIN	10	0.96	2.12	0.74
ICT	4	0.38	0.86	0.79
MFG	7	0.67	1.40	0.73
ENG	5	0.48	1.13	0.80
TRD	3	0.29	0.66	0.81

Table 4.1: Number of companies, mean and standard deviation of extreme event counts and frequency of zero extreme event counts for each analyzed sector

Hierachical structure

As it can be easily noticed, the considered time series have a hierarchical structure: the bottom level is represented by the five count time series referred to the different economic sectors, while at the top level there is the aggregate time series obtained as the sum of the extreme events counts at each trading day in the sample period. While the dynamics of bottom time series is conditional on the other series' shocks, the aggregate series follows a purely autoregressive - GARCH-type - process.

Score-driven model estimation

The count predictions for the bottom time series model are obtained by fitting the negative binomial score-driven model introduced in Section 4.4.1. The $\boldsymbol{\alpha}$ coefficients are estimated using the univariate score-driven model by Blasques et al. 2018, and their values are shown in Table 4.2. As in Agosto 2022, we use a unit-scaling for the filtered dynamics, that is we set $s(\mathbf{f}_t) = \mathbf{I}_k$ in (4.14). The predictions for the top time series, which aggregates the counts of the individual sectors, are obtained by applying the univariate score-driven model by Blasques et al. 2018.

Parameter	FIN	ICT	MFG	ENG	TRD
$\hat{\alpha}$	0.40	0.15	0.02	0.14	0.10

Table 4.2: Estimated dispersion parameters (α)

4.5 Results

For each of the 3508 days, we reconcile the forecast distributions of the 6 time series. Since we deal with a small hierarchy, we use importance sampling (IS) to sample from the reconciled distribution, as explained in Chapter 2. We draw $100,000$ samples.

We use different indicators to assess the performance of the forecasts. The absolute error (AE) is defined as $\text{AE} := |y_t - \widehat{y}_{t|t-1}|$, while the squared error (SE) is defined as $\text{SE} := \left(y_t - \widehat{y}_{t|t-1}\right)^2$. Here, y_t denotes the value of the time series at time t, while $\widehat{y}_{t|t-1}$ denotes the point forecast computed at time $t-1$ for time t. We use the median of the distribution as point forecast for the AE, and the mean as point forecast for the SE (Kolassa 2016).

The mean interval score (MIS) (Gneiting 2011) and the Energy score (Székely and Rizzo 2013) have already been defined in Section 3.3. For the MIS, we set $\alpha = 0.1$, which corresponds to 90% coverage intervals. The ES, with $\alpha = 2$, is computed using samples, as explained in Wickramasuriya 2021.

We use the skill score (see Section 3.3), which is symmetric and scale-independent, to compare the performance of the reconciled forecast distribution with respect to the base forecast distribution, in terms of percentage improvement. We compute the skill score for each day, and for each time series. The average skill scores are reported in Table 4.3. In Figure 4.10, we show the boxplot of the skill score on ES, while in Figure 4.9, we show the

boxplot of the skill scores on AE, SE and MIS for all the time series. Both AE and SE measure the accuracy of the point forecasts. In most cases, there is no improvement in AE as the skill score is very close to 0. On the contrary, the improvement in SE is significant for all the time series. Indeed, SE is less robust with respect to extreme values, which are cut down using reconciliation. When we deal with intermittent time series, however, it is usually more important to compare the prediction intervals, rather then the point forecasts. We observe a very large improvement in MIS, which is a measure of the quality of the 90% coverage interval. Indeed, the average width of the coverage intervals decreases after reconciliation (Table 4.4), as in most cases the variance of the forecast distribution decreases. Note, however, that the reconciled forecast distribution is still calibrated, as for more than 90% of the days the actual value is contained within the interval (Table 4.5). Finally, we observe a very significant improvement also for ES.

metric	ALL	FIN	ICT	MFG	ENG	TRD
AE	-0.02	0.02	-0.02	-0.01	-0.03	-0.02
SE	0.82	1.10	1.11	1.07	1.12	1.11
MIS	0.87	1.11	0.2	1.07	0.22	0.18
ES			1.00			

Table 4.3: Average skill scores, for all the time series

	ALL	FIN	ICT	MFG	ENG	TRD
base	6.91	3.20	1.13	1.89	1.26	0.92
reconc.	3.33	2.10	0.95	1.25	1.04	0.77

Table 4.4: Average width of the 90% *coverage interval*

	ALL	FIN	ICT	MFG	ENG	TRD
base	96.2%	97.7%	97.9%	98.0%	97.9%	98.6%
reconc.	91.1%	95.5%	97.2%	96.7%	97.4%	98.0%

Table 4.5: Percentage of days for which the actual value in contained in the 90% *coverage interval.*

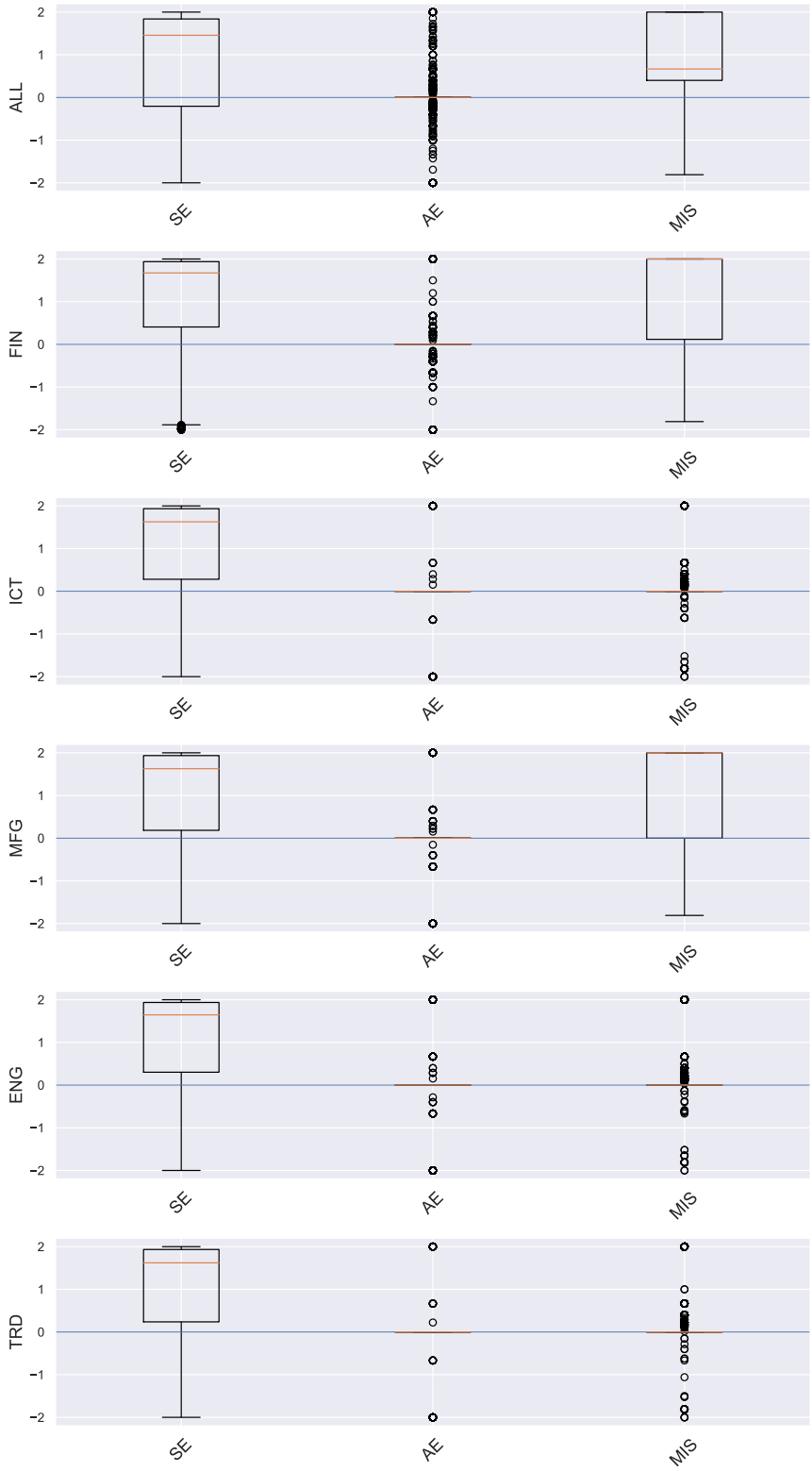


Figure 4.9: Boxplot of the skill scores on AE, SE, and MIS

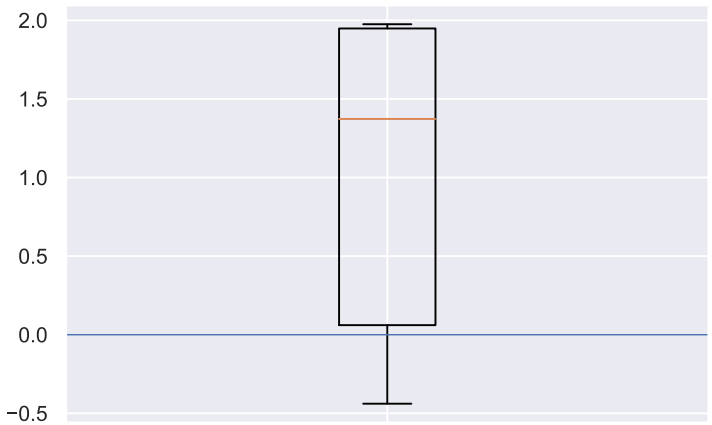


Figure 4.10: Boxplot of the skill scores on ES

4.5.1 Reconciled mean and variance

For each day, and for each time series, we compute the shift as the difference between the reconciled mean and the base mean. Hence, a positive shift means that the mass of the distribution has moved to the right, and vice versa.

We then divide all the 3508 days into two groups, depending on whether the sum of the bottom shifts has the same sign of the upper shift or not. Most of the days (3360, i.e. the 95.8%) fall within the first group: in this case, we thus observe the "strengthening" effect discussed in Section 4.3. For the remaining 148 days, the upper shift and the sum of the bottom shifts have a different sign. In this case, we observe the "compromise" effect: through the reconciliation, we merge the information coming from the bottom and the upper time series.

We visually represents the two groups using a scatter plot, with the bottom-up mean on the x axis and the base upper mean on the y axis; each point is a different day. As expected, the points from the first group are concentrated around the line $y = x$, which corresponds to coherence, while the points from the second group are more dispersed. We recall that the incoherence of the point forecasts is given by the difference between the bottom-up mean and the base upper mean, i.e. $\mathbf{A}\widehat{\mathbf{b}} - \widehat{\mathbf{u}}$. In Figure 4.12, we show the boxplot of the incoherence for all the days, divided in the two groups. As expected, the days with a small incoherence are those in which

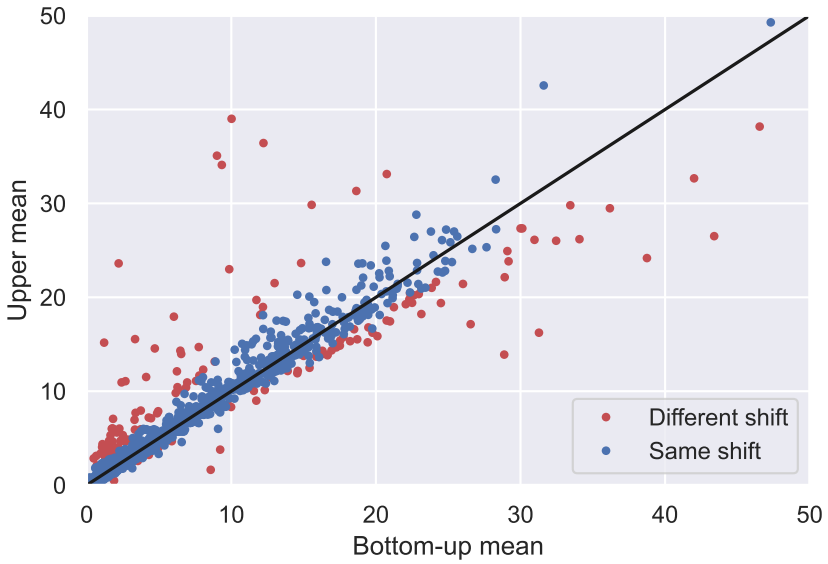


Figure 4.11: Bottom-up mean vs base upper mean

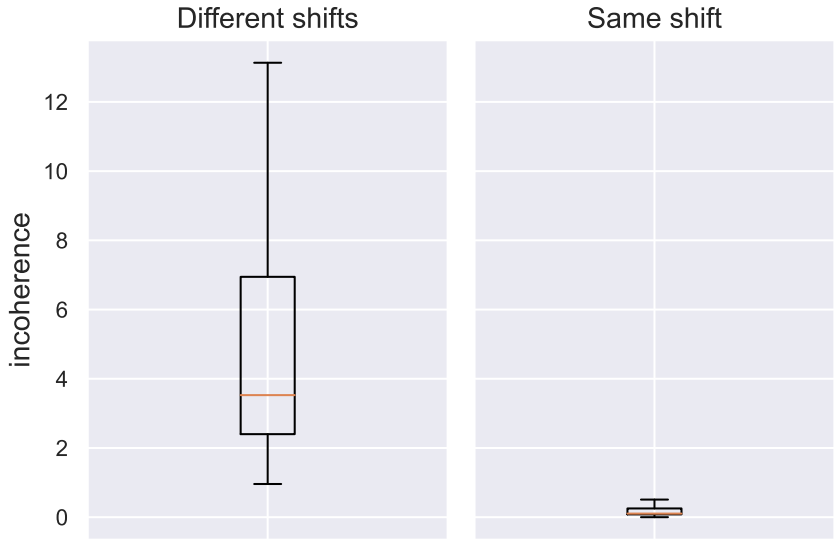


Figure 4.12: Boxplot of the incoherence in the two different cases

the shifts on the bottom and on the upper variables have the same sign.

In Figure 4.13, we show the probability mass functions of the bottom and upper time series before and after reconciliation, for one of the days in the first group. We also compare the base, bottom-up, and reconciled

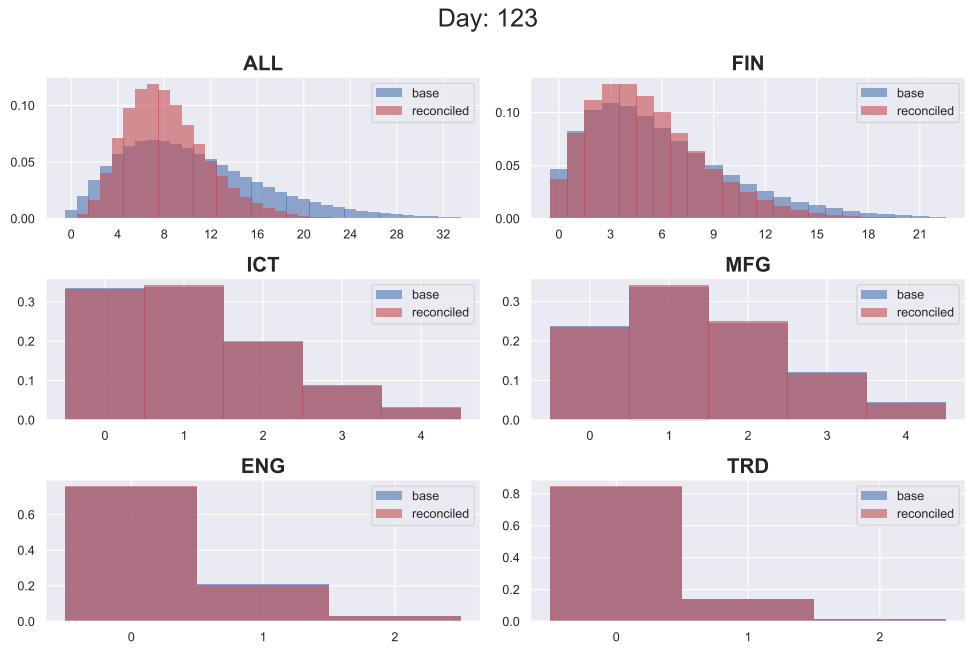


Figure 4.13: Probability mass function before and after reconciliation, day 123

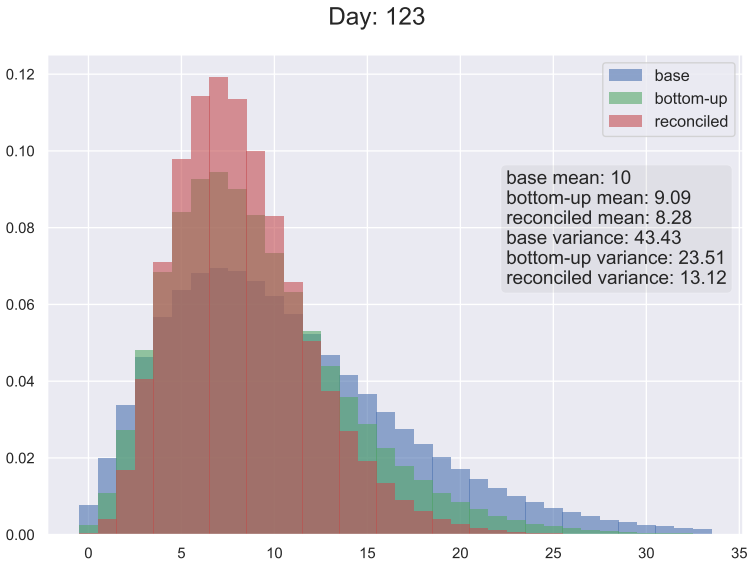


Figure 4.14: Base, bottom-up, and reconciled probability mass function of the upper time series, day 123

distribution of the upper time series (Figure 4.14). The main effect is a reduction of the variance, which leads to a flattening of the tail, and thus to

a negative shift since the distribution has a positive skewness. The effect of the reconciliation for one of the days of the second group is shown in Figures 4.15 and 4.16. In this case, we observe the "compromise" effect: through the reconciliation, we merge the information coming from the bottom and the upper time series.

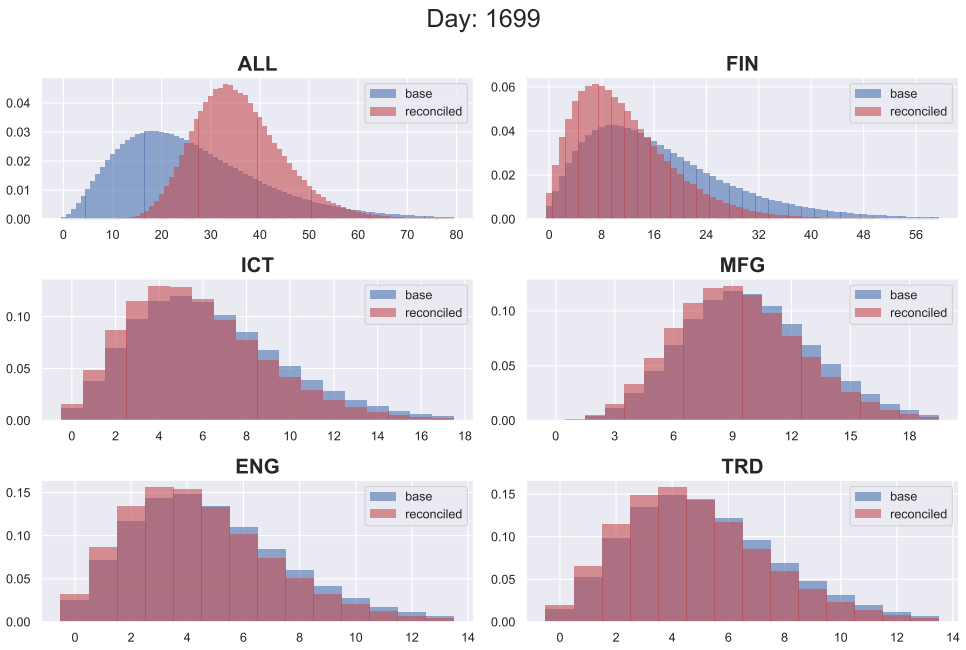


Figure 4.15: Probability mass function before and after reconciliation, day 1699

Finally, we compute the variance of the distributions before and after reconciliation. In most cases (3394 days, i.e. the 96.8%) the variance of all the variables decreases. There are some cases, however, in which the variance of one or more bottom variables increases. An example is shown in Figures 4.17 and 4.18: the information provided by the base upper distribution is in conflict with the information provided by the bottom, hence the variance of the bottom distributions increase.

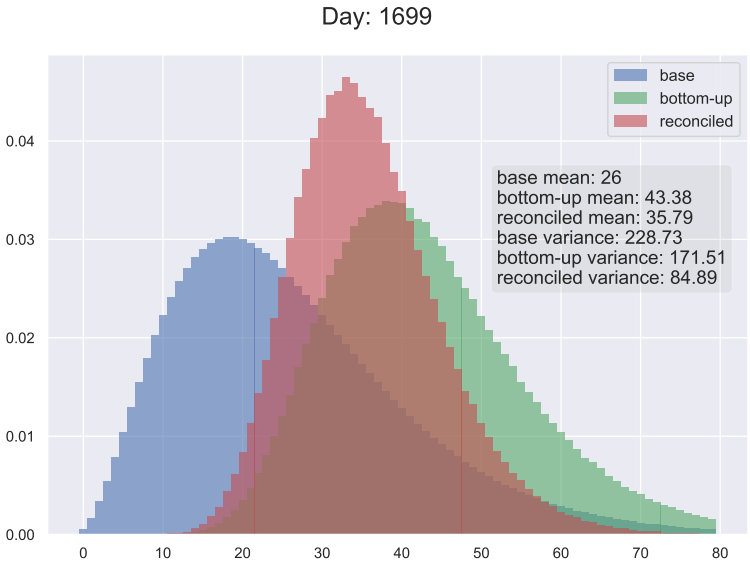


Figure 4.16: Base, bottom-up, and reconciled probability mass function of the upper time series, day 1699

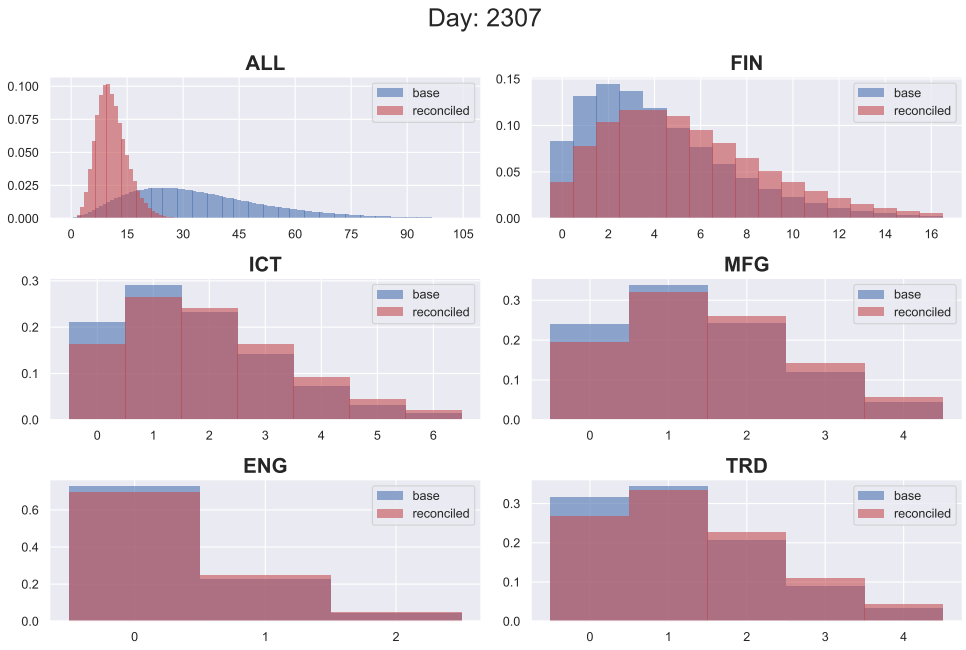


Figure 4.17: Probability mass function before and after reconciliation, day 2307

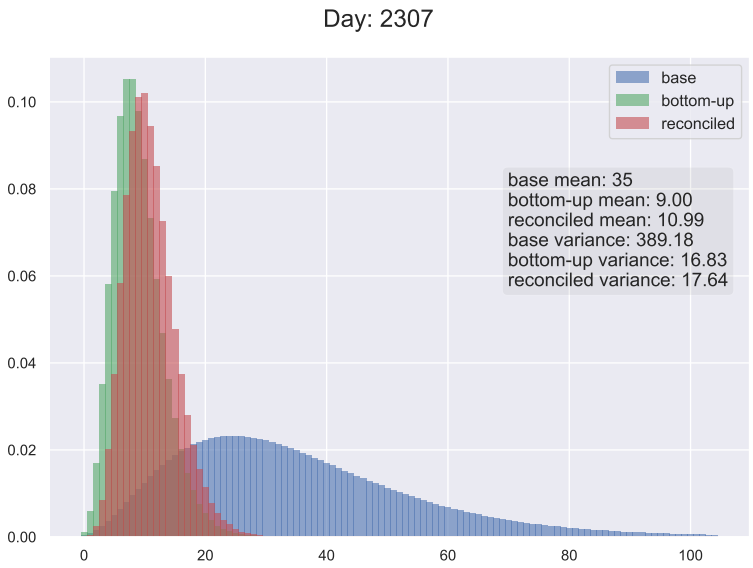


Figure 4.18: Base, bottom-up, and reconciled probability mass function of the upper time series, day 2307

Chapter 5

The Fourier Discrepancy Function

5.1 Discrepancies between probability measures

Discrepancies are becoming omnipresent tools in every applied fields that require the comparison of probability measures. Examples include computer vision (Angenent et al. 2004; Bassetti et al. 2020; Auricchio et al. 2019; Cuturi and Doucet 2014; Papadakis 2015; Vogel and Oman 1996; Fang et al. 2021; Ojha et al. 2021), supervised learning (Janocha and Czarnecki 2016; Bengio et al. 2017; Bishop 2006; Schmidhuber 2015; Frogner et al. 2015; Yu et al. 2012), and generative models (Arjovsky et al. 2017; Ansari et al. 2020; Li et al. 2017; Wang et al. 2019; Yu et al. 2018; Pan et al. 2020). Often the usage of these tools is bounded by their numerical complexity (Peyré and Cuturi 2019; Dvinskikh and Tiapkin 2021; Tarjan 1997; Nesterov 2007). To mitigate these issues, in recent years, several studies have been devoted to introduce new discrepancies (Lin 1991; Bonneel and Coeurjolly 2019) or to study the properties of the existing ones (Ling and Okada 2007; Auricchio et al. 2018). A special role is played by the study of the relationships between different discrepancies, usually through bounds. In particular, the problem of finding the tight bounds (Gilardoni 2006) in terms of the Total Variation has been particularly interesting for source coding (Csiszár 1967a; Csiszár 1967b; Sason 2014).

A well-known family of distances between probability measures is given by the χ_r-metrics. They are defined as the L^p distance between the characteristic functions of two given measures weighted by the function $\|k\|^{-rp}$. Despite the appealing properties they enjoy, the use of these metrics is bounded by

the fact that they are not well-defined unless the two measures we are comparing have equal moments up to the $\lceil r \rceil$-th one (Rachev 1991; Rachev et al. 2013). This is a standard assumption in some applied fields, such as kinetic theory (Carrillo and Toscani 2007; Baringhaus and Grübel 1997). In general, however, requiring two measures to have the same expectation is too restricting. In Auricchio et al. 2020, the authors studied the χ_r-metrics in the specific framework of discrete measures supported over a regular grid. In this framework, they proved that some requirements about the measures can be dropped while still preserving the appealing properties of their continuous counterparts. However, these distances are defined through an integral, and for $r \geq 2$ some conditions on the moments are still required to ensure the finiteness of the integral. In this chapter, we overcome this issue by introducing a discretized version of the χ_r-metrics, called Fourier Discrepancies.

The rest of the chapter is organized as follows. In Section 5.2, we recall the main notions about discrete probability measures and the Discrete Fourier Transform (DFT) (Rao and Yip 2018). In Section 5.3, we introduce a new family of distances between discrete probability measures, the p-Fourier Discrepancies. We show that they can be expressed as the square root of a bilinear form induced by a positive definite matrix, hence they are 1-homogeneous and convex. Moreover, we prove that the squared Fourier Discrepancy is twice differentiable and that both its gradient and Hessian have an explicit formula. In Section 5.4, we study the lower and upper tight bounds of the Fourier Discrepancy in terms of the Total Variation distance. In particular, we prove that the upper tight bound between any q-homogeneous and convex function and the Total Variation is attained in a finite set. We then present an open conjecture about the value of the upper tight bound of the Fourier Discrepancy. We conclude this chapter with a discussion about the appealing properties of the p-Fourier Discrepancies, and their possible applications to several applied fields (Section 5.5).

5.2 Preliminaries

In this section, we state the framework of our work and fix our notation. Throughout the chapter, we only consider one-dimensional discrete measures, but all the results may be extended to a multidimensional setting. Let us define the set $I_N \subset [0,1]$ as $I_N := \left\{0, \frac{1}{N}, \dots, \frac{N-1}{N}\right\}$. For the sake of simplicity, we will assume that N is an even number. A discrete measure μ on I_N is defined as

$$\mu := \sum_{j=0}^{N-1} \mu_j \delta_{\frac{j}{N}}, \tag{5.1}$$

where all the μ_j's are real values and, for any $k \in I_N$, δ_k is the Dirac's delta centered in k. We denote by $\mathcal{M}(I_N)$ the set of discrete measures over I_N and by $\mathcal{P}(I_N) := \{\mu \in \mathcal{M}(I_N) : \mu_j \geq 0,\ \sum_{j=0}^{N-1} \mu_j = 1\}$ the space of discrete probability measures.

Remark 1. *Since any discrete measure supported on I_N is fully characterised by the $N-$uple of positive values $(\mu_0, \ldots, \mu_{N-1})$, we refer to discrete measures and vectors interchangeably. Although this might lead to a slight abuse of notations, it allows us to express the Fourier Transform of a discrete measure through a linear operator.*

Definition 3. *The Discrete Fourier Transform (DFT) of $\mu \in \mathcal{P}(I_N)$ is the $N-$dimensional vector $\hat{\mu} := (\hat{\mu}_0, \ldots, \hat{\mu}_{N-1})$ defined as*

$$\hat{\mu}_k := \sum_{j=0}^{N-1} \mu_j e^{-2\pi i \frac{j}{N} k}, \qquad k \in \{0, \ldots, N-1\}. \tag{5.2}$$

Remark 2. *Since the complex exponential function $k \to e^{-2\pi i \frac{j}{N} k}$ is a $N-$periodic function for any integer j, we set $\hat{\mu}_k := \hat{\mu}_{mod_N(k)}$ for any $k \in \mathbb{Z}$, where $mod_N(k)$ is the $N-$modulo operation. In particular, $\hat{\mu}_{-k} = \hat{\mu}_{N-k}$ for any $k \in \{0, \ldots, N-1\}$.*

Remark 3. *The DFT of a discrete measure can be expressed as a linear map:*

$$(\hat{\mu}_0, \ldots, \hat{\mu}_{N-1}) = \Omega \cdot (\mu_0, \ldots, \mu_{N-1}), \tag{5.3}$$

where Ω is the $N \times N$ matrix defined as

$$\Omega := \begin{bmatrix} \omega_{0,0} & \omega_{0,1} & \ldots & \omega_{0,N-1} \\ \omega_{1,0} & \omega_{1,1} & \ldots & \omega_{1,N-1} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{N-1,0} & \omega_{N-1,1} & \ldots & \omega_{N-1,N-1} \end{bmatrix}, \tag{5.4}$$

and $\omega_{k,j} := e^{-2\pi i \frac{j}{N} k}$. Since the matrix Ω is invertible, the DFT is a bijective function.

For a complete discussion about the Discrete Fourier Transform (DFT), we refer to Rao and Yip 2018.

5.3 The Fourier Discrepancy Function

In this section we introduce the p-Fourier Discrepancy Functions, a family of discrete versions of the metrics introduced by Auricchio et al. 2020. The p-Fourier Discrepancies inherit from their continuous counterparts the property of being bounded by the Wasserstein distance. We show that the Fourier Discrepancies are convex and have an explicit derivative.

Definition 4. *For any $p \geq 1$, the p-Fourier Discrepancy Function is defined as $\mathbb{F}_p : \mathcal{P}(I_N) \times \mathcal{P}(I_N) \to [0, +\infty)$, where*

$$\mathbb{F}_p^2(\mu, \nu) := \sum_{k=1}^{\frac{N}{2}-1} \frac{|\hat{\mu}_k - \hat{\nu}_k|^2}{|k|^{2p}} + \frac{|\hat{\mu}_{\frac{N}{2}} - \hat{\nu}_{\frac{N}{2}}|^2}{|N|^{2p}}. \tag{5.5}$$

Remark 4. *It is easy to show that every $\mathbb{F}_p$ is a distance on $\mathcal{P}(I_N)$. In particular, unlike its continuous counterparts, $\mathbb{F}_p$ is finite even without requiring the two measures to have any equal moment.*

Remark 5. *Following Auricchio et al. 2020, it is possible to prove that*

$$\mathbb{F}_p \leq C_p W_1 \tag{5.6}$$

for any $p > \frac{3}{2}$, where W_1 is the 1-Wasserstein distance (Villani 2008) and C_p is a constant that only depends on p.

For any $p \geq 1$, let us introduce the matrix $\mathbb{K}_p := diag(b_p)$, where the vector b_p is defined as

$$b_p := \frac{1}{2}\left(1, 1^{-2p}, \ldots, \left(\frac{N}{2} - 1\right)^{-2p}, \frac{2}{N^{2p}}, \left(\frac{N}{2} - 1\right)^{-2p}, \ldots, 1^{-2p}\right). \tag{5.7}$$

Since $\hat{\mu}_k = \overline{\hat{\mu}_{N-k}}$, we can express the Fourier Discrepancy function as a quadratic form:

$$\mathbb{F}_p^2(\mu, \nu) = (\hat{\mu} - \hat{\nu})^T \mathbb{K}_p (\hat{\mu} - \hat{\nu}) = (\mu - \nu)^T \mathbb{H}_p (\mu - \nu), \tag{5.8}$$

where $\mathbb{H}_p := \Omega^T \mathbb{K}_p \Omega$ and Ω is the DFT matrix. Notice that we only consider the first $\frac{N}{2}$ frequencies as the last $\frac{N}{2}$ have the same magnitude, hence no information is lost by omitting them. Moreover, $\mathbb{H}_p$ is a symmetric and circulant matrix, since $(\mathbb{H}_p)_{i,j} = \mathrm{Re}((\hat{b}_p)_{i-j})$. Therefore, its eigenvalues can be explicitly computed (Davis 1979), leading us to the following result.

Lemma 1. *For any $p \geq 1$, the matrix $\mathbb{H}_p$ is positive definite and its eigenvalues are given by*

$$\lambda_i = N \cdot (b_p)_i, \qquad i = 0, \dots, N-1.$$

Since $\mathbb{H}_p$ is positive definite, there exists a matrix $\mathbb{L}_p$ such that $\mathbb{L}_p^T \mathbb{L}_p = \mathbb{H}_p$. We can then write $\mathbb{F}_p(\mu - \nu) = \|\mathbb{L}_p(\mu - \nu)\|_2$, where $\|\cdot\|_2$ is the l^2 norm. Hence, we have the following.

Proposition 5. *For any $p \geq 1$, the Fourier Discrepancy $\mathbb{F}_p$ is convex and $1-$homogeneous with respect to $\mu - \nu$.*

To conclude, we observe that we are able to explicitly compute the gradient and Hessian matrix of $\mathbb{F}_p^2$.

Proposition 6. *For any $p \geq 1$ and for any probability measure ν, the function $L_{p,\nu} : \mathcal{P}(I_n) \to \mathbb{R}$, defined as $L_{p,\nu}(\mu) := \mathbb{F}_p^2(\mu, \nu)$, is twice differentiable. Moreover, its gradient and Hessian matrix are expressed through the explicit formulae:*

$$(\nabla L_{p,\nu})_l(\mu) = \frac{\partial L_{p,\nu}}{\partial \mu_l}(\mu) = 2 \sum_{j=0}^{N-1} (\mu_j - \nu_j) \cdot \mathrm{Re}\left((\hat{b}_p)_{j-l}\right) \tag{5.9}$$

and

$$(HL_{p,\nu})_{h,l}(\mu) = \frac{\partial^2 L_{p,\nu}}{\partial \mu_h \partial \mu_l}(\mu) = 2\,\mathrm{Re}\left((\hat{b}_p)_{h-l}\right), \tag{5.10}$$

where $\hat{b}_p$ is the Fourier Transform of the vector b_p.

5.4 Tight Bounds

In this section, we study the tight bounds for the p-Fourier Discrepancy in terms of the Total Variation distance. We recall that, for any pair of discrete measures supported on I_N, the Total Variation is defined as

$$TV(\mu, \nu) := \frac{1}{2} \sum_{j=0}^{N-1} |\mu_j - \nu_j|.$$

Following Sason 2014, for any given $\theta \in (0, 1]$, we define the lower and the upper tight bounds, respectively $C_L(\theta)$ and $C_U(\theta)$, as

$$C_L(\theta) := \inf_{\mu,\nu:\, TV(\mu,\nu)=\theta} \mathbb{F}_p(\mu, \nu), \tag{5.11}$$

$$C_U(\theta) := \sup_{\mu,\nu:\, TV(\mu,\nu)=\theta} \mathbb{F}_p(\mu, \nu). \tag{5.12}$$

Due to the linearity of the DFT, we have that

$$\mathbb{F}_p^2(\mu,\nu) = \sum_{k=1}^{\frac{N}{2}-1} \frac{|\widehat{(\mu-\nu)}_k|^2}{|k|^{2p}} + \frac{|\widehat{(\mu-\nu)}_{\frac{N}{2}}|^2}{|N|^{2p}}, \tag{5.13}$$

we then set $\Delta := \mu - \nu$ and express $\mathbb{F}_p$ as a function of Δ, rather than μ and ν. Analogously, we will often write $TV(\Delta)$ instead of $TV(\mu,\nu)$, as long as $\Delta = \mu - \nu$. We now introduce the set of null-sum measures over I_N, $\mathcal{O}(I_N)$, defined as $\mathcal{O}(I_N) := \{\Delta \in \mathcal{M}(I_N) \quad \text{s.t.} \quad \sum_i \Delta_i = 0\}$. Given any pair of probability measures μ and ν, it is easy to see that $\mu - \nu \in \mathcal{O}(I_N)$. Up to a multiplicative constant, the converse is also true.

Proposition 7. *Given any non-zero $\Delta \in \mathcal{O}(I_N)$ and $\theta \in (0,1]$, there exists $C > 0$ and a pair of probability measures (μ,ν) such that*

$$\mu - \nu = C \cdot \Delta \quad \textit{and} \quad TV(\mu,\nu) = \theta.$$

The proof is reported in Appendix A.7.

Remark 6. *Thanks to Proposition 7, and for the 1-homogeneity of $\mathbb{F}_p$, we have that, for any $\theta \in [0,1)$*

$$C_L(\theta) = \inf_{\substack{\Delta \in \mathcal{O}(I_N): \\ \Delta \neq 0}} \mathbb{F}_p\left(\frac{\theta}{TV(\Delta)}\Delta\right) = \theta \cdot \inf_{\substack{\Delta \in \mathcal{O}(I_N): \\ \Delta \neq 0}} \frac{\mathbb{F}_p(\Delta)}{TV(\Delta)}, \tag{5.14}$$

and, analogously,

$$C_U(\theta) = \theta \cdot \sup_{\substack{\Delta \in \mathcal{O}(I_N): \\ \Delta \neq 0}} \frac{\mathbb{F}_p(\Delta)}{TV(\Delta)}. \tag{5.15}$$

5.4.1 Lower tight bound

Let us define $\omega_k \in \mathbb{C}^N$ as the $k-$th column of the DFT matrix Ω. Since $\{\omega_k\}_{k=0,\dots,N-1}$ is an orthogonal basis of $\mathbb{C}^n$ (Rao and Yip 2018), for any $\Delta \in \mathcal{O}(I_N)$ there exists a unique N-tuple of complex coefficients $\left(\lambda^{(k)}\right)_{k=0,\dots,N-1}$ such that

$$\Delta = \sum_{k=0}^{N-1} \lambda^{(k)} \omega_k.$$

We then define the set

$$\Xi := \left\{\Delta \in \mathcal{O}(I_N) : \sum_{k=0}^{N-1} |\lambda^{(k)}| = 1\right\}, \tag{5.16}$$

and notice that Ξ is not empty, as we have that $\omega_{\frac{N}{2}} = (-1, +1, -1, +1, -1, \dots, +1) \in \Xi$. Finally, since both TV and $\mathbb{F}_p$ are 1-homogeneous functions, we rewrite (5.14) as

$$C_L(\theta) = \theta \cdot \inf_{\substack{\Delta \in \mathcal{O}(I_N): \\ \Delta \neq 0}} \frac{\mathbb{F}_p\left(\frac{\Delta}{\sum |\lambda^{(k)}|}\right)}{TV\left(\frac{\Delta}{\sum |\lambda^{(k)}|}\right)} \frac{\sum |\lambda^{(k)}|}{\sum |\lambda^{(k)}|} = \theta \cdot \inf_{\Delta \in \Xi} \frac{\mathbb{F}_p(\Delta)}{TV(\Delta)}. \tag{5.17}$$

We now state the main result of the section.

Theorem 1. *The lower tight bound $C_L(\theta)$ is given by*

$$C_L(\theta) = 2\theta N^{-p}, \tag{5.18}$$

and is attained at $\omega_{\frac{N}{2}}$.

Proof. To prove the theorem, we show that $\omega_{\frac{N}{2}}$ both minimizes the Fourier Discrepancy and maximizes the Total Variation over the set Ξ. This is enough to conclude $C_L(\theta) = \theta \frac{\mathbb{F}_p(\omega_{\frac{N}{2}})}{TV(\omega_{\frac{N}{2}})}$ which, through a simple computation, proves (5.18). For the sake of clarity, we divide the proof into two steps.

First step ($\omega_{\frac{N}{2}}$ maximizes TV over Ξ).
For any $\Delta \in \Xi$, we have

$$\begin{aligned} TV(\Delta) &= TV\Big(\sum_{k=0}^{N-1} \lambda^{(k)} \omega_k\Big) = \frac{1}{2} \sum_{j=0}^{N-1} \Big| \sum_{k=0}^{N-1} \lambda^{(k)} (\omega_k)_j \Big| \\ &\leq \frac{1}{2} \sum_{j=0}^{N-1} \sum_{k=0}^{N-1} \Big| \lambda^{(k)} (\omega_k)_j \Big| = \frac{1}{2} \sum_{k=0}^{N-1} |\lambda^{(k)}| \sum_{j=0}^{N-1} |(\omega_k)_j| \\ &= \frac{N}{2} \sum_{k=0}^{N-1} |\lambda^{(k)}| = \frac{N}{2}. \end{aligned}$$

We then conclude the first step of the proof by noticing that $TV(\omega_{\frac{N}{2}}) = \frac{N}{2}$.

Second Step ($\omega_{\frac{N}{2}}$ minimizes $\mathbb{F}_p$ over Ξ). For any $j = 0, \dots, N-1$, the DFT of ω_j is given by

$$\widehat{(\omega_j)}_k = \sum_{l=0}^{N-1} e^{-i\frac{2\pi}{N}lk} (\omega_j)_l = \sum_{l=0}^{N-1} e^{-i\frac{2\pi}{N}l(k-j)} = N\delta_{k-j}.$$

From the linearity of the DFT, we infer

$$\widehat{\Delta}_k = \sum_{j=0}^{N-1} \lambda^{(j)} \widehat{(\omega_j)}_k = N \sum_{j=0}^{N-1} \lambda^{(j)} \delta_{k-j} = N\lambda^{(k)}, \tag{5.19}$$

therefore, for any $\Delta \in \mathcal{O}(I_N)$, we have

$$\mathbb{F}_p^2(\Delta) = N^2 \Bigg(\sum_{k=1}^{\frac{N}{2}-1} \frac{|\lambda^{(k)}|^2}{k^{2p}} + \frac{|\lambda^{(\frac{N}{2})}|^2}{|N|^{2p}} \Bigg). \tag{5.20}$$

Finally, we conclude the proof by showing

$$\inf_{\Delta \in \Xi} \mathbb{F}_p(\Delta) = \mathbb{F}_p(\omega_{\frac{N}{2}}) = N^{1-p}.$$

Let $\Delta \in \Xi$. From (5.19), we have that $\lambda^{(0)} = \frac{1}{N}\widehat{\Delta}_0 = \frac{1}{N}\sum_j \Delta_j = 0$. Moreover, since Δ is real, we have that $\widehat{\Delta}_k = \overline{\widehat{\Delta}_{N-k}}$ for any $k = 1, \dots, N-1$, hence $|\lambda^{(k)}| = |\lambda^{(N-k)}|$. Then, if we define

$$\gamma_j := \begin{cases} 2|\lambda^{(j)}| & j = 1, \dots, \frac{N}{2} - 1, \\ |\lambda^{(\frac{N}{2})}| & j = \frac{N}{2}, \end{cases}$$

the constraint (5.16) is written as

$$\sum_{j=1}^{\frac{N}{2}} \gamma_j = 1,$$

while from (5.20) we obtain $\mathbb{F}_p^2(\Delta) = \sum_{k=1}^{\frac{N}{2}} \alpha_k \gamma_k^2$, with

$$\alpha_k := \begin{cases} \left(\frac{N}{2}\right)^2 k^{-2p} & k = 1, \dots, \frac{N}{2} - 1, \\ N^{2-2p} & k = \frac{N}{2}. \end{cases}$$

Since the coefficient $\alpha_{\frac{N}{2}}$ is the lowest one, as long as $p \geq 1$, the minimum of $\mathbb{F}_p$ is achieved when $\gamma_{\frac{N}{2}} = 1$ and $\gamma_j = 0$ for $j = 1, \dots, \frac{N}{2} - 1$, and the proof is complete. □

5.4.2 Upper tight bound

We now show that it is possible to restrict the search space of the maximizer of (5.15) to a finite set with cardinality N. In particular, we prove that a

similar restriction may be applied whenever we search for the upper tight bound between the Total Variation and any convex and $p-$homogeneous function of $\Delta \in \mathcal{O}(I_N)$. To accomplish that, we show that every $\Delta \in \mathcal{O}(I_N)$ can be written as a linear combination of simpler null-sum measures, namely $\eta_{i,j}$, defined as

$$\eta_{i,j} := \delta_i - \delta_j,$$

for any $i, j \in \{0, \dots, N-1\}$ such that $i \neq j$. In particular, we have the following (the proof is in Appendix A.8).

Lemma 2. *Let Δ be a null-sum measure on I_N. Then, we can express Δ as $\Delta = TV(\Delta) \cdot \Delta'$, where Δ' is a convex combination of $\{\eta_{i_k,j_k}\}_k$ such that, for any $k \neq k'$, we have $i_k \neq j_{k'}$.*

This characterization allows us to restrict the set of possible maximizers of any convex and $p-$homogeneous function over the finite set $\{\eta_{i,j}\}_{i,j}$.

Theorem 2. *Let $\mathbb{G} : \mathcal{O}(I_N) \to [0, +\infty)$ be a convex and $p-$homogeneous function. Then, there exist $i^\star$, $j^\star \in \{0, \dots, N-1\}$ such that, for any $\theta \in (0, 1]$:*

$$\theta \cdot \eta_{i^\star,j^\star} = \underset{TV(\Delta)=\theta}{\operatorname{argmax}} \mathbb{G}(\Delta). \tag{5.21}$$

Proof. First, we notice that

$$(i^\star, j^\star) := \underset{i,j \in \{0,\dots,N-1\}}{\operatorname{argmax}} \mathbb{G}(\eta_{i,j}), \tag{5.22}$$

is well-defined as the maximum is taken over a finite set. Given any $\theta \in (0, 1]$, let Δ be a null-sum measure such that $TV(\Delta) = \theta$. Lemma 2 allows us to write $\Delta = \theta \cdot \sum_k \lambda_k \eta_{i_k,j_k}$, with $\lambda_k \geq 0$ for any k and $\sum_k \lambda_k = 1$. Finally, from the p-homogeneity and the convexity of $\mathbb{G}$, we obtain:

$$\begin{aligned}
\mathbb{G}(\Delta) &= \mathbb{G}\left(\theta \cdot \sum_k \lambda_k \eta_{i_k,j_k}\right) = \theta^p \cdot \mathbb{G}\left(\sum_k \lambda_k \eta_{i_k,j_k}\right) \\
&\leq \theta^p \cdot \sum_k \lambda_k \mathbb{G}\left(\eta_{i_k,j_k}\right) \leq \theta^p \cdot \sum_k \lambda_k \mathbb{G}\left(\eta_{i^\star,j^\star}\right) \\
&= \theta^p \cdot \mathbb{G}\left(\eta_{i^\star,j^\star}\right) = \mathbb{G}\left(\theta \cdot \eta_{i^\star,j^\star}\right),
\end{aligned}$$

which concludes the proof. □

Using the previous result we may recover the well-known upper tight bound between the l^p norm and the Total Variation. Indeed, since $||\eta_{i,j}||_p =$

$2^{\frac{1}{p}}$ for any p, we find that the inequality $||\mu - \nu||_p \leq 2^{\frac{1}{p}} TV(\mu, \nu)$ is tight.

Since $\mathbb{F}_p : \mathcal{O}(I_N) \to [0, +\infty)$ is convex and $1-$homogeneous, we infer $C_U(\theta) = \theta \cdot \mathbb{F}_p(\eta_{i^\star, j^\star})$, for some $i^\star, j^\star \in \{0, \dots, N-1\}$. Therefore, to find the upper tight bound of $\mathbb{F}_p$ we only need to search over a finite set of points, which correspond to the differences between two Dirac's deltas. Since the DFT is linear, we have that $\widehat{\eta_{l,j}} = \Theta_l - \Theta_j$, where $\Theta_k = \left(e^{i\frac{2\pi k}{N}0}, e^{i\frac{2\pi k}{N}1}, \dots, e^{i\frac{2\pi k}{N}(N-1)}\right)$ is the $k-$th column of the matrix Ω. Hence:

$$\mathbb{F}_p^2(\delta_l, \delta_j) = \mathbb{F}_p^2(\eta_{l,j}) = \sum_{k=1}^{\frac{N}{2}-1} \frac{|(\Theta_l - \Theta_j)_k|^2}{|k|^{2p}} + \frac{|(\Theta_l - \Theta_j)_{\frac{N}{2}}|^2}{|N|^{2p}},$$

which boils down to (see Appendix A.9)

$$\mathbb{F}_p^2(\eta_{j,l}) = \sum_{k=1}^{\frac{N}{2}-1} \frac{2 - 2\cos\left(\frac{2\pi|j-l|}{N}k\right)}{|k|^{2p}} + \frac{2 - 2\cos(\pi|j-l|)}{|N|^{2p}}, \tag{5.23}$$

for any $j,\ l \in \{0, \dots, N-1\}$. Finally, notice that $\mathbb{F}_p^2(\eta_{j,l})$ depends on j and l only through $d := |j - l|$. Hence, we can further restrict to measures of the form $\eta_{0,d}$, with $d \in \{1, \dots, N-1\}$.

Corollary 1. *For every $p \geq 1$, there exists $d \in \{0, 1, \dots, N-1\}$ such that*

$$C_U(\theta) = \theta \cdot \mathbb{F}_p(\eta_{0,d}).$$

Notice that, for any $d \in \{0, 1, \dots, N-1\}$, we have $\mathbb{F}_p^2(\eta_{0,d}) = C - 2g_p(d)$, where C is a constant and $g_p : [0, N] \to \mathbb{R}$ is defined as:

$$g_p(d) := \sum_{k=1}^{\frac{N}{2}-1} \frac{\cos\left(\frac{2\pi d}{N}k\right)}{|k|^{2p}} + \frac{\cos(\pi d)}{|N|^{2p}}. \tag{5.24}$$

By studying the derivatives with respect to d, it is possible to show that $d^* = \frac{N}{2}$ is a local minimum for g_p. This leads us to the following open conjecture.

Conjecture 1. *For every $p \geq 1$ and $d \in \{0, 1, \dots, N-1\}$, we have*

$$\mathbb{F}_p(\eta_{0,\frac{N}{2}}) \geq \mathbb{F}_p(\eta_{0,d}).$$

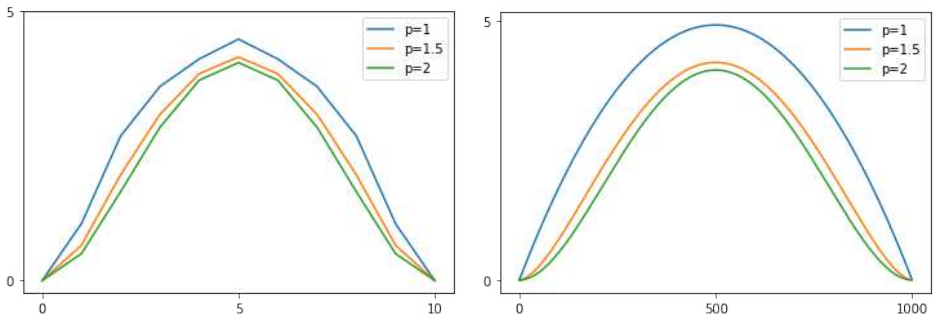


Figure 5.1: Plots of $\mathbb{F}_p(\eta_{0,d})$ for $p \in \{1, 1.5, 2\}$ and for $N = 10$ (left), $N = 1000$ (right). As conjectured, the maximum is attained at $d = \frac{N}{2}$.

If our conjecture was true, we would have

$$C_U(\theta) = \theta \cdot \sqrt{\sum_{k=1}^{\frac{N}{2}-1} \frac{2 - 2(-1)^k}{|k|^{2p}} + \frac{2 - 2(-1)^{\frac{N}{2}}}{|N|^{2p}}}. \tag{5.25}$$

Notice that, for $p = 1$, the value (5.25) converges to $\sqrt{\sum_{k=1}^{\infty} \frac{2-2(-1)^k}{k^2}} = \frac{\pi}{2}$ as $N \to \infty$.

We numerically verify that the conjecture is true for $p \in \{1, 1.5, 2\}$ and for any even N that ranges from 2 to 1000. In Figure 5.1, we report the graph of the function $d \to \mathbb{F}_p(\eta_{0,d})$ for $p \in \{1, 1.5, 2\}$ and $N \in \{10, 1000\}$.

5.5 Discussion

In this chapter, we have introduced a new class of metrics between discrete probability measures, the p-Fourier Discrepancy Functions. For any $p \geq 1$, $\mathbb{F}_p$ is a well-defined distance induced by a bilinear form. It is convex, and its square is twice differentiable with explicit formulae for both the gradient and Hessian. Moreover, as Figure 5.1 shows, the Fourier Discrepancy between two Dirac's deltas depends on the distance between their supports. Most common discrepancies, such as the Total Variation or the Kullback-Leibler, do not enjoy this property, which is instead a feature of the Wasserstein distance. This is consistent with the bound (5.6) and with the equivalence between Fourier-based and Wasserstein distances (Auricchio et al. 2020). In the last years, the Wasserstein distance has been widely used in several applied fields because of its topological weakness and its ability to deal with

the geometry of the underlying space (Arjovsky et al. 2017). However, its applicability, especially in higher dimensions, is bounded by the computational cost for both the distance and its gradient. On the other hand, the Fourier Discrepancy and its gradient are cheap to compute using the Fast Fourier Transform algorithm. We believe that the appealing properties of the Fourier Discrepancy make it a compelling alternative to the Wasserstein distance in several applied fields, such as machine learning (Frogner et al. 2015; Han et al. 2020; Hou et al. 2017), time series comparison (Zhang et al. 2020), or barycenters computation (Anderes et al. 2016; Bassetti et al. 2020; Cuturi and Doucet 2014). Finally, the Fourier Discrepancy may be easily generalized to a multidimensional setting.

Chapter 6

Conclusions

We have proposed a new approach for probabilistic reconciliation based on conditioning, rather than on projecting (Panagiotelis et al. 2022). We have also proposed the BUIS algorithm, which samples efficiently from the reconciled distribution. As a results, our approach is currently the only one which is both general (it reconciles both continuous and discrete distributions) and computationally fast. This algorithm can be used even if the base distributions are only available in the form of samples, which is often the case when forecasting count time series. The extensive numerical experiments, conducted on both standard data sets (*carparts* and *syph*) and on count time series of extreme events in the CDS market, show a clear improvement of the reconciled forecasts over the base forecasts. Finally, we have studied the effects of the reconciliation on the mean and variance of the forecast distribution.

Future research directions include:

- extending the algorithm to deal with correlations between the base forecasts
- writing an R package for probabilistic reconciliation
- studying a diagnostic, analogous to the ESS used for IS, to assess the performance of BUIS
- using the Fourier Discrepancy for time series comparison, as an alternative to dynamic time warping or time adaptive optimal transport.

Appendix A

A.1 Propositions for Section 2.2

Proposition 8. *Let $s : X \to Y$ be a measurable bijection between two measure spaces $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$. Then, the pushforward $s_{\#} : \mathcal{P}(X) \to \mathcal{P}(Y)$ is a bijection, with inverse given by $(s^{-1})_{\#}$.*

Proof. First, we recall that the pushforward $s_{\#}$ is defined, for any $\nu \in \mathcal{P}(X)$ and $F \in \mathcal{Y}$, as

$$s_{\#}\nu(F) = \nu(s^{-1}(F)).$$

Hence, for any $\nu \in \mathcal{P}(X)$ and $G \in \mathcal{X}$, we have

$$\begin{aligned}
\big((s^{-1})_{\#} \circ s_{\#}\big)\nu\,(G) &= (s^{-1})_{\#}\big(s_{\#}\nu\big)\,(G) \\
&= s_{\#}(\nu)\big((s^{-1})^{-1}(G)\big) \\
&= s_{\#}(\nu)\big(s(G)\big) \\
&= \nu\big(s^{-1}(s(G))\big) \\
&= \nu(G),
\end{aligned}$$

and therefore $(s^{-1})_{\#} \circ s_{\#}$ is the identity map. Analogously, for any $\mu \in \mathcal{P}(Y)$ and $F \in \mathcal{X}$, we have

$$\begin{aligned}
\big(s_{\#} \circ (s^{-1})_{\#}\big)\mu\,(F) &= s_{\#}\big((s^{-1})_{\#}\mu\big)\,(F) \\
&= (s^{-1})_{\#}(\mu)\big(s^{-1}(F)\big) \\
&= \mu\big((s^{-1})^{-1}\big(s^{-1}(F)\big)\big) \\
&= \mu(s(s^{-1}(F))) \\
&= \mu(F).
\end{aligned}$$

□

Proposition 9. *Let $\hat{\pi}$ be the joint density of the random vector $(\mathbf{U}, \mathbf{B})$. Then, the density of $(\mathbf{Z}, \mathbf{B})$, where $\mathbf{Z} := \mathbf{U} - \mathbf{AB}$, is given by*

$$\pi_{(\mathbf{Z},\mathbf{B})}(\mathbf{z}, \mathbf{b}) = \hat{\pi}(\mathbf{z} + \mathbf{Ab}, \mathbf{b}).$$

Proof. The joint density of $(\mathbf{Z}, \mathbf{B})$ can be computed using the rule of change of variables Billingsley 2008, Chapter 17. Let $\mathbf{H} : \mathbb{R}^n \to \mathbb{R}^n$ be defined as

$$\mathbf{H} : \begin{bmatrix} \mathbf{u} \\ \mathbf{b} \end{bmatrix} \to \begin{bmatrix} \mathbf{u} - \mathbf{Ab} \\ \mathbf{b} \end{bmatrix}.$$

$\mathbf{H}$ is invertible, with inverse given by

$$\mathbf{H}^{-1} : \begin{bmatrix} \mathbf{z} \\ \mathbf{b} \end{bmatrix} \to \begin{bmatrix} \mathbf{z} + \mathbf{Ab} \\ \mathbf{b} \end{bmatrix},$$

and we have that

$$\left| J\mathbf{H}^{-1}(\mathbf{b}, \mathbf{z}) \right| = \begin{vmatrix} \mathbf{I} & \mathbf{A}^T \\ \mathbf{0} & \mathbf{I} \end{vmatrix} = 1.$$

Then, the joint density of $(\mathbf{Z}, \mathbf{B})$ is given by

$$\begin{aligned} \pi_{(\mathbf{Z},\mathbf{B})}(\mathbf{z}, \mathbf{b}) &= \hat{\pi}\big(\mathbf{H}^{-1}(\mathbf{z}, \mathbf{b})\big) \cdot \left| J\mathbf{H}^{-1}(\mathbf{z}, \mathbf{b}) \right| \\ &= \hat{\pi}(\mathbf{z} + \mathbf{Ab}, \mathbf{b}). \end{aligned}$$

□

A.2 Proof of Proposition 1

First, we recall that, given a pair of absolutely continuous probability distributions μ and ν, the Kullback-Leibler (KL) divergence is defined as

$$KL(\mu \,\|\, \nu) = \int \log\left(\frac{p(x)}{q(x)}\right) p(x)\, dx,$$

where p and q are the densities of, respectively, μ and ν. The discrete case is completely analogous.

Now, let $\hat{\nu}_b$ be the base bottom forecast distribution, and $\widetilde{\nu}$ the reconciled distribution. We recall that the density of $\widetilde{\nu}$ is given by

$$\widetilde{\pi}(\mathbf{b}) = \frac{1}{c}\hat{\pi}(\mathbf{Ab}, \mathbf{b}),$$

where

$$c := \int \hat{\pi}(\mathbf{Ab}, \mathbf{b})\, d\mathbf{b} = \int \frac{\hat{\pi}(\mathbf{Ab}, \mathbf{b})}{\hat{\pi}_b(\mathbf{b})} \hat{\pi}_b(\mathbf{b})\, d\mathbf{b} = \mathbb{E}\left[\frac{\hat{\pi}(\mathbf{AB}, \mathbf{B})}{\hat{\pi}_b(\mathbf{B})}\right]$$

is the normalizing constant, and $\mathbf{B} \sim \hat{\nu}_b$. Then, we have

$$\begin{aligned} KL(\hat{\nu}_b \,\|\, \widetilde{\nu}) &= \int \log\left(c\, \frac{\hat{\pi}_b(\mathbf{b})}{\hat{\pi}(\mathbf{Ab}, \mathbf{b})}\right) \hat{\pi}_b(\mathbf{b})\, d\mathbf{b} \\ &= \log(c) - \int \log\left(\frac{\hat{\pi}(\mathbf{Ab}, \mathbf{b})}{\hat{\pi}_b(\mathbf{b})}\right) \hat{\pi}_b(\mathbf{b})\, d\mathbf{b} \\ &= \log\left(\mathbb{E}\left[\frac{\hat{\pi}(\mathbf{AB}, \mathbf{B})}{\hat{\pi}_b(\mathbf{B})}\right]\right) - \mathbb{E}\left[\log\left(\frac{\hat{\pi}(\mathbf{AB}, \mathbf{B})}{\hat{\pi}_b(\mathbf{B})}\right)\right] \\ &= \log\left(\mathbb{E}[W]\right) - \mathbb{E}[\log(W)]. \end{aligned} \tag{A.1}$$

A.3 Proof of Proposition 2

We show that the output $\left(\mathbf{b}^{(i)}\right)_i$ of the BUIS algorithm is approximately a sample drawn from the target distribution $\widetilde{\nu}$.

From (2.7), and from Assumption 1, we have that

$$\begin{aligned} \tilde{\pi}(\mathbf{b}) &\propto \hat{\pi}_b(\mathbf{b}) \cdot \hat{\pi}_u(\mathbf{Ab}) \\ &= \prod_{t=1}^{m} \pi_{b_t}(b_t) \cdot \prod_{l=1}^{L} \prod_{j=1}^{k_l} \pi_{u_{j,l}}\left(\sum_{k=1}^{q_{j,l}} b_{k,(j,l)}\right), \end{aligned}$$

where we are using the notation of Sect. 3.1. The initial distribution of the sample $\left(\mathbf{b}^{(i)}\right)_{i=1,\dots,N}$ is given by $\hat{\pi}_b = \prod_{t=1}^{m} \pi_{b_t}(b_t)$. We show that each iteration of the algorithm corresponds to multiplying by a $\pi_{u_{j,l}}\left(\sum_{k=1}^{q_{j,l}} b_{k,(j,l)}\right)$ term.

Let π_X be a density over $\mathbb{R}^d$, and $w : \mathbb{R}^d \to \mathbb{R}$ a continuous function. Let $X_1, \dots, X_N$ be independent samples from π_X, and compute the unnormalized weights $(\hat{w}^{(i)})_{i=1,\dots,N}$ as $\hat{w}^{(i)} = w(X_i)$. Then, if we draw $Y_1, \dots, Y_n$ from the discrete distribution given by

$$\mathbb{P}\left(Y = X_i\right) = w^{(i)}, \qquad i = 1, \dots, N,$$

where $w^{(i)} = \frac{\hat{w}^{(i)}}{\sum_{j=1}^{N} \hat{w}^{(j)}}$, then $(Y_i)_{i=1,\dots,n}$ is approximately an IID sample from the density $\pi_Y(x) \propto \pi_X(x) \cdot w(x)$. This technique is known as importance

resampling or weighted bootstrap (Smith and Gelfand 1992). The same holds also for discrete distributions, using the pmf instead of the density.

Hence, if we compute the weights $w^{(i)}$'s as in the algorithm and sample $\left(\tilde{\mathbf{b}}_j^{(i)}\right)_i$ from (3.1), it is approximately equivalent to sampling from $\hat{\pi}_b(\mathbf{b}) \cdot \pi_{u_{j,l}}\left(\sum_{t=1}^{q_{j,l}} b_t\right)$, where $\hat{\pi}_b$ is the original density of $\left(b_{1,(j,l)}, \ldots, b_{q_{j,l},(j,l)}\right)$. In other words, the weighting-resampling step corresponds to multiplying the density of the sample by a $\pi_{u_{j,l}}\left(\sum_{t=1}^{q_{j,l}} b_t\right)$ term.

Finally, note that in this way we are conditioning with respect to $u_{j,l}$. After the weighting-resampling step, $\left(b_{1,(j,l)}, \ldots, b_{q_{j,l},(j,l)}\right)$ are correlated. Since, from Assumption 2, the hierarchy is given by a tree, we are guaranteed that for any level l and for all $j = 1, \ldots, k_l$, $\tilde{\mathbf{b}}_j$ only depends on $b_{1,(j,l)}, \ldots, b_{q_{j,l},(j,l)}$, $u_{j,l}$ and each upper variable that is under $u_{j,l}$. From Assumption 1, we have that $\tilde{\mathbf{b}}_1, \ldots, \tilde{\mathbf{b}}_{k_l}$ are independent. Hence, the density of $\left[\tilde{\mathbf{b}}_1, \ldots, \tilde{\mathbf{b}}_{k_l}\right]$ is given by the product of the densities of all $\tilde{\mathbf{b}}_j$'s, and the proof is concluded.

A.4 MCMC-IS comparison

In order to fully understand the reasons for the significant difference in computational time between the MCMC and the IS approach, we compare the two methods on a minimal example. Le us consider a hierarchy given by two bottom variables, b_1 and b_2, and just one upper variable u, which is the sum of b_1 and b_2. We set a Gaussian distribution for each variables.

We implement a simple Metropolis-Hastings algorithm with a Gaussian proposal distribution with fixed variance τI to sample from the reconciled distribution $\tilde{\pi}(\mathbf{b}) = \pi_{b_1}(b_1) \cdot \pi_{b_2}(b_2) \cdot \pi_u(b_1 + b_2)$. The algorithm reads as follows:

Initialize $\mathbf{b}^{(0)}$
for $j = 1, \ldots, N$ **do**
  Sample $\mathbf{y}^{(j)} \sim \mathcal{N}(\mathbf{b}^{(j-1)}, \tau I)$
  $\alpha \leftarrow \min\left(1, \frac{\tilde{\pi}(\mathbf{y}^{(j)})}{\tilde{\pi}\left(\mathbf{b}^{(j-1)}\right)}\right)$
  $u \leftarrow Unif(0, 1)$
  if $u < \alpha$ **then**
    $\mathbf{b}^{(j)} \leftarrow \mathbf{y}^{(j)}$
  else
    $\mathbf{b}^{(j)} \leftarrow \mathbf{b}^{(j-1)}$
  end if
end for
return $\left(\mathbf{b}^{(i)}\right)_i$

On a standard laptop, it takes about 4 seconds to get $10,000$ samples from $\tilde{\pi}$. In particular, most of the time is employed by the computation of the acceptance probability α, which requires about $3.7 \cdot 10^{-4}$ seconds per loop. Sampling from the proposal distribution only requires about $3 \cdot 10^{-5}$ seconds.

We then implement an IS algorithm on the same hierarchy, using Python:

Sample $\mathbf{b}^{(1)}, \dots, \mathbf{b}^{(N)} \overset{\text{IID}}{\sim} \hat{\pi}_b$
$w_i \leftarrow \hat{\pi}_u \left(b_1^{(i)} + b_2^{(i)} \right)$
return $\left(\mathbf{b}^{(i)}, w_i \right)_i$

It takes about $7 \cdot 10^{-3}$ seconds to draw $100,000$ IID samples from $\hat{\pi}_b$, and about the same time to compute all the weights. The significant improvement in computational time using IS instead of MCMC is due to the fact that both sampling and computation of the weights are done simultaneously for all the samples, rather than sequentially as in MCMC.

A.5 Proof of Proposition 3

Since, from (2.3), the matrix $\mathbf{Q}$ is positive definite, $\mathbf{Q}^{-1}$ is also positive definite. Therefore, the matrices

$$\mathbf{G} := \left(\widehat{\mathbf{\Sigma}}_{UB}^T - \widehat{\mathbf{\Sigma}}_B \mathbf{A}^T \right) \mathbf{Q}^{-1} \left(\widehat{\mathbf{\Sigma}}_{UB}^T - \widehat{\mathbf{\Sigma}}_B \mathbf{A}^T \right)^T,$$
$$\mathbf{H} := \left(\widehat{\mathbf{\Sigma}}_U - \widehat{\mathbf{\Sigma}}_{UB} \mathbf{A}^T \right) \mathbf{Q}^{-1} \left(\widehat{\mathbf{\Sigma}}_U - \widehat{\mathbf{\Sigma}}_{UB} \mathbf{A}^T \right)^T,$$

are positive semi-definite.

From (4.3), we have that, for each $i = 1, \dots, m$

$$\text{Var}(\widetilde{B}_i) = \text{Var}(\widehat{B}_i) - G_{ii} \leq \text{Var}(\widehat{B}_i),$$

as $G_{ii} \geq 0$ since the matrix $\mathbf{G}$ is positive semi-definite. Analogously, we have

$$\text{Var}(\widetilde{U}_j) = \text{Var}(\widehat{U}_j) - H_{jj} \leq \text{Var}(\widehat{U}_j),$$

for all $j = 1, \dots, n - m$.

A.6 Proof of Proposition 4

Let us denote $Z := \mathbb{1}_{\{\mathbf{U}=\mathbf{AB}\}}$, so that $Z = 1$ when the constraint is satisfied, and 0 otherwise. By the law of total variance (Weiss 2005), for any

$j = 1, \ldots, m$, we have

$$\mathrm{Var}(B_j) = \mathbb{E}[\mathrm{Var}(B_j|Z)] + \mathrm{Var}(\mathbb{E}[B_j|Z]). \tag{A.2}$$

Since

$$\mathbb{E}[B_j|Z] = \begin{cases} \mathbb{E}[B_j|\mathbf{U} = \mathbf{AB}] & \text{if } Z = 1 \\ \mathbb{E}[B_j|\mathbf{U} \neq \mathbf{AB}] & \text{if } Z = 0, \end{cases}$$

we have that $\mathbb{E}[B_j|Z] = a + (b-a)\mathrm{Ber}$, where $\mathrm{Ber} \sim \mathrm{Bernoulli}(p)$, hence

$$\mathrm{Var}(\mathbb{E}[B_j|Z]) = (b-a)^2 p(1-p). \tag{A.3}$$

Moreover, since

$$\mathrm{Var}[B_j|Z] = \begin{cases} \mathrm{Var}[B_j|\mathbf{U} = \mathbf{AB}] & \text{if } Z = 1 \\ \mathrm{Var}[B_j|\mathbf{U} \neq \mathbf{AB}] & \text{if } Z = 0, \end{cases}$$

we have

$$\mathbb{E}[\mathrm{Var}(B_j|Z)] = p\,\mathrm{Var}[B_j|\mathbf{U} = \mathbf{AB}] + (1-p)\,\mathrm{Var}[B_j|\mathbf{U} \neq \mathbf{AB}]. \tag{A.4}$$

From (A.2), (A.3), and (A.4), we have

$$\begin{aligned} \mathrm{Var}(B_j) = {} & p\ \mathrm{Var}[B_j|\mathbf{U} = \mathbf{AB}] + (1-p)\ \mathrm{Var}[B_j|\mathbf{U} \neq \mathbf{AB}] \\ & + p(1-p)\,(a-b)^2, \end{aligned}$$

from which

$$\mathrm{Var}[B_j|\mathbf{U} = \mathbf{AB}] = \frac{\mathrm{Var}(B_j) - (1-p)\ \mathrm{Var}[B_j|\mathbf{U} \neq \mathbf{AB}] - p(1-p)\,(a-b)^2}{p}.$$

A.7 Proof of Proposition 7

Let $C := \frac{\theta}{TV(\Delta)}$ and $\widetilde{\Delta} := C \cdot \Delta$, which are well-defined since $TV(\Delta) \neq 0$ for any non-zero Δ. Then, for the $1-$homogeneity of TV, we have that $TV(\widetilde{\Delta}) = \frac{\theta}{TV(\Delta)} \cdot TV(\Delta) = \theta$.
Let $\widetilde{\mu}$ and $\widetilde{\nu}$ be, respectively, the positive and negative part of $\widetilde{\Delta}$. Therefore, $\widetilde{\Delta} = \widetilde{\mu} - \widetilde{\nu}$ and $\widetilde{\mu}_i, \widetilde{\nu}_i \geq 0$ for any i. We have that

$$2\theta = \sum_i |\widetilde{\Delta}_i| = \sum_i \widetilde{\mu}_i + \sum_i \widetilde{\nu}_i, \tag{A.5}$$

and moreover, since $\widetilde{\Delta}$ is a null-sum measure:

$$0 = \sum_i \widetilde{\Delta}_i = \sum_i \widetilde{\mu}_i - \sum_i \widetilde{\nu}_i. \tag{A.6}$$

From (A.5) and (A.6) follows easily that $\sum_i \widetilde{\mu}_i = \sum_i \widetilde{\nu}_i = \theta$.

We now define

$$\mu := \widetilde{\mu} + (1-\theta)\delta_0, \quad \nu := \widetilde{\nu} + (1-\theta)\delta_0.$$

We have that μ is a probability measure since $\mu_i \geq 0$ for any i and $\sum_i \mu_i = \sum_i \widetilde{\mu}_i + (1-\theta) = 1$. The same holds for ν. Moreover, $\mu - \nu = \widetilde{\Delta}$, hence $TV(\mu, \nu) = TV(\widetilde{\Delta}) = \theta$.

A.8 Proof of Lemma 2

Let Δ be a null-sum measure. Without loss of generality, we can reorder the values of Δ as follows:

$$\Delta = (\alpha_1, \ldots, \alpha_r, -\beta_1, \ldots, -\beta_l, 0, \ldots, 0),$$

where $r + l \leq N$, α_i, $\beta_j > 0$, $\alpha_i \leq \alpha_{i+1}$, $\beta_j \leq \beta_{j+1}$, for any i and j, and $\sum \alpha_i = \sum \beta_j$.

Without loss of generality, we assume that

$$\alpha_1 \leq \beta_1.$$

Hence, we can write

$$\Delta = \alpha_1 \eta_{0,r} + \Delta^{(1)},$$

where

$$\begin{aligned} \Delta^{(1)} &= (0, \alpha_2^{(1)}, \ldots, \alpha_r^{(1)}, -\beta_1^{(1)}, \ldots, -\beta_l^{(1)}, 0, \ldots, 0) \\ &:= (0, \alpha_2, \ldots, \alpha_r, -(\beta_1 - \alpha_1), -\beta_2, \ldots, -\beta_l, 0, \ldots, 0). \end{aligned}$$

Next, we compare $\alpha_2^{(1)}$ and $\beta_1^{(1)}$ and repeat the process until every entry vanishes. At the end, we find

$$\Delta = \lambda_1 \eta_{0,r} + \cdots + \lambda_k \eta_{r-1,N-1} =: \sum_k \lambda_k \eta_{i_k, j_k}. \tag{A.7}$$

Notice that each $\eta_{i,j}$ in (A.7) is such that $i < r$ and $j \geq r$ by construction, which implies $i \neq j$.

Since by hypothesis, for any $l = 0, \ldots, N-1$, all the l-th entries $(\eta_{i_k,j_k})_i$ have the same sign, we can write

$$|\Delta_l| = \Big| \sum_k \lambda_k (\eta_{i_k,j_k})_l \Big| = \sum_k \lambda_k |(\eta_{i_k,j_k})_l|.$$

Therefore:

$$\begin{aligned} TV(\Delta) &= \frac{1}{2} \sum_l |\Delta_l| = \frac{1}{2} \sum_l \sum_k \lambda_k |(\eta_{i_k,j_k})_l| \\ &= \frac{1}{2} \sum_k \sum_l \lambda_k |(\eta_{i_k,j_k})_l| \\ &= \frac{1}{2} \sum_k \lambda_k \sum_l |(\eta_{i_k,j_k})_l| = \sum_k \lambda_k, \end{aligned}$$

since $\sum_l |(\eta_{i,j})_l| = 2$ for any i, j. To conclude, it suffices to set

$$\Delta' := \frac{1}{TV(\Delta)} \Delta = \sum_k \widetilde{\lambda}_k \eta_{i_k,j_k},$$

where $\widetilde{\lambda}_k := \frac{\lambda_k}{\sum_l \lambda_l} > 0$, and $\sum_k \widetilde{\lambda}_k = 1$.

A.9 Computing $\mathbb{F}_p(\eta_{j,l})$

Let us consider null-sum measures of the form $\eta_{l,j}$. We recall that $\eta_{l,j} := \delta_l - \delta_j$. Since

$$\widehat{\eta_{l,j}} = \Omega \cdot \eta_{l,j},$$

we have

$$\widehat{\eta_{l,j}} = \Theta_l - \Theta_j, \tag{A.8}$$

where Θ_k is the $k-$th column of the matrix Ω. By the definition of Ω we have

$$\Theta_l = \Big(e^{i\frac{2\pi l}{N}0}, e^{i\frac{2\pi l}{N}1}, \ldots, e^{i\frac{2\pi l}{N}(N-1)} \Big),$$

therefore, the value $\mathbb{F}_p^2(\eta_{l,j})$ is then given by

$$\mathbb{F}_p^2(\eta_{l,j}) = \sum_{k=1}^{\frac{N}{2}-1} \frac{|(\Theta_l - \Theta_j)_k|^2}{k^{2p}} + \frac{|(\Theta_l - \Theta_j)_{\frac{N}{2}}|^2}{|N|^{2p}}. \tag{A.9}$$

Let us now compute explicitly $|(\Theta_l - \Theta_j)_k|^2$ for a given k. We have

$$(\Theta_l - \Theta_j)_k = \cos\left(\frac{2\pi l}{N} k\right) - \cos\left(\frac{2\pi j}{N} k\right) + i \sin\left(\frac{2\pi l}{N} k\right) - i \sin\left(\frac{2\pi j}{N} k\right),$$

therefore,

$$\begin{aligned}|(\Theta_l - \Theta_j)_k|^2 &= \left(\cos\left(\frac{2\pi l}{N}k\right) - \cos\left(\frac{2\pi j}{N}k\right)\right)^2 + \left(\sin\left(\frac{2\pi l}{N}k\right) - \sin\left(\frac{2\pi j}{N}k\right)\right)^2 \\ &= 2 - 2\left(\cos\left(\frac{2\pi l}{N}k\right)\cos\left(\frac{2\pi j}{N}k\right) + \sin\left(\frac{2\pi l}{N}k\right)\sin\left(\frac{2\pi j}{N}k\right)\right) \\ &= 2 - 2\cos\left(\frac{2\pi(j-l)}{N}k\right), \qquad \text{(A.10)}\end{aligned}$$

where the equality in (A.10) comes from the following trigonometric identity:

$$\cos(\alpha - \beta) = \cos(\alpha)\cos(\beta) + \sin(\alpha)\sin(\beta).$$

Therefore,

$$\mathbb{F}_p^2(\eta_{j,l}) = \sum_{k=1}^{\frac{N}{2}-1} \frac{2 - 2\cos\left(\frac{2\pi|j-l|}{N}k\right)}{k^{2p}} + \frac{2 - 2\cos(\pi|j-l|)}{N^{2p}}. \qquad \text{(A.11)}$$

Bibliography

[1] S. Agapiou, O. Papaspiliopoulos, D. Sanz-Alonso, and A. M. Stuart. "Importance sampling: Intrinsic dimension and computational cost". In: *Statistical Science* (2017), pp. 405–431.

[2] A. Agosto. "Multivariate Score-Driven Models for Count Time Series To Assess Financial Contagion". In: *SSRN: https://ssrn.com/abstract=4119895* (2022).

[3] E. Anderes, S. Borgwardt, and J. Miller. "Discrete Wasserstein barycenters: Optimal transport for discrete data". In: *Mathematical Methods of Operations Research* 84.2 (2016), pp. 389–409.

[4] S. Angenent, S. Haker, A. Tannenbaum, and L. Zhu. "Optimal mass transport for registration and warping". In: *International Journal of computer vision* 60.3 (2004), pp. 225–240.

[5] A. F. Ansari, J. Scarlett, and H. Soh. "A characteristic function approach to deep implicit generative modeling". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 7478–7487.

[6] M. Arjovsky, S. Chintala, and L. Bottou. "Wasserstein generative adversarial networks". In: *Proceedings of the 34th International Conference on Machine Learning*. 2017, pp. 214–223.

[7] G. Athanasopoulos, P. Gamakumara, A. Panagiotelis, R. J. Hyndman, and M. Affan. "Hierarchical forecasting". In: *Macroeconomic forecasting in the era of big data: Theory and practice* (2020), pp. 689–719.

[8] G. Athanasopoulos, R. J. Hyndman, N. Kourentzes, and F. Petropoulos. "Forecasting with temporal hierarchies". In: *European Journal of Operational Research* 262.1 (2017), pp. 60–74.

[9] G. Auricchio, F. Bassetti, S. Gualandi, and M. Veneroni. "Computing Wasserstein Barycenters via Linear Programming". In: *International Conference on Integration of Constraint Programming, Artificial Intelligence, and Operations Research*. 2019, pp. 355–363.

[10] G. Auricchio, A. Codegoni, S. Gualandi, G. Toscani, and M. Veneroni. "The equivalence of Fourier-based and Wasserstein metrics on imaging problems". In: *Rendiconti Lincei - Matematica e Applicazioni* 31 (2020), pp. 627–649.

[11] G. Auricchio, S. Gualandi, M. Veneroni, and F. Bassetti. "Computing Kantorovich-Wasserstein distances on d-dimensional histograms using $(d+1)$-partite graphs." In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 2018, pp. 5798–5808.

[12] L. Baringhaus and R. Grübel. "On a class of characterization problems for random convex combinations". In: *Annals of the Institute of Statistical Mathematics* 49.3 (1997), pp. 555–567.

[13] F. Bassetti, S. Gualandi, and M. Veneroni. "On the Computation of Kantorovich-Wasserstein Distances Between Two-Dimensional Histograms by Uncapacitated Minimum Cost Flows". In: *SIAM Journal on Optimization* 30.3 (2020), pp. 2441–2469.

[14] Y. Bengio, I. Goodfellow, and A. Courville. *Deep learning*. MIT press, 2017.

[15] P. Billingsley. *Probability and measure*. John Wiley & Sons, 2008.

[16] C. M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[17] F. Blasques, V. Holỳ, and P. Tomanová. "Zero-inflated autoregressive conditional duration model for discrete trade durations with excessive zeros". In: *arXiv preprint arXiv:1812.07318* (2018).

[18] N. Bonneel and D. Coeurjolly. "SPOT: Sliced Partial Optimal Transport". In: *ACM Transactions on Graphics* 38.4 (2019), pp. 1–13.

[19] J. Carrillo and G. Toscani. "Contractive probability metrics and asymptotic behavior of dissipative kinetic equations". In: *Rivista Matematica Università di Parma* 7.6 (2007), pp. 75–198.

[20] Y.-C. Chen. "A tutorial on kernel density estimation and recent advances". In: *Biostatistics & Epidemiology* 1.1 (2017), pp. 161–187.

[21] E. Çinlar. *Probability and stochastics*. Vol. 261. Springer, 2011.

[22] G. Corani, D. Azzimonti, J. P. Augusto, and M. Zaffalon. "Probabilistic Reconciliation of Hierarchical Forecast via Bayes' Rule." In: *Proc. European Conf. On Machine Learning and Knowledge Discovery in Database ECML/PKDD*. Vol. 3. 2020, pp. 211–226.

[23] G. Corani, N. Rubattu, D. Azzimonti, and A. Antonucci. "Probabilistic Reconciliation of Count Time Series". In: *arXiv preprint arXiv:2207.09322* (2022).

[24] I. Csiszár. "Information-type measures of difference of probability distributions and indirect observation". In: *studia scientiarum Mathematicarum Hungarica* 2 (1967), pp. 229–318.

[25] I. Csiszár. "Two remarks to noiseless coding". In: *Information and Control* 11.3 (1967), pp. 317–322.

[26] M. Cuturi and A. Doucet. "Fast computation of Wasserstein barycenters". In: *International Conference on Machine Learning*. 2014, pp. 685–693.

[27] A. Darwiche. *Modeling and reasoning with Bayesian networks*. Cambridge university press, 2009.

[28] P. J. Davis. *Circulant matrices*. Wiley, 1979.

[29] T. Di Fonzo and D. Girolimetto. "Forecast combination-based forecast reconciliation: Insights and extensions". In: *International Journal of Forecasting* (2022).

[30] T. Di Fonzo and D. Girolimetto. "Cross-temporal forecast reconciliation: Optimal combination method and heuristic alternatives". In: *International Journal of Forecasting* 39.1 (2023), pp. 39–57.

[31] P. Diaconis and D. Ylvisaker. "Conjugate priors for exponential families". In: *The Annals of statistics* (1979), pp. 269–281.

[32] D. Dvinskikh and D. Tiapkin. "Improved complexity bounds in wasserstein barycenter problem". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 1738–1746.

[33] V. Elvira and L. Martino. "Advances in Importance Sampling". In: *Wiley StatsRef-Statistics Reference Online* (2021).

[34] V. Elvira, L. Martino, and C. P. Robert. "Rethinking the Effective Sample Size". In: *International Statistical Review* (2022).

[35] A. Escribano and M. Maggi. "Intersectoral default contagion: A multivariate Poisson autoregression analysis". In: *Economic Modelling* 82 (2019), pp. 376–400.

[36] B. Fang, A. Guntuboyina, and B. Sen. "Multivariate extensions of isotonic regression and total variation denoising via entire monotonicity and Hardy–Krause variation". In: *The Annals of Statistics* 49.2 (2021), pp. 769–792.

[37] C. Frogner, C. Zhang, H. Mobahi, M. Araya-Polo, and T. Poggio. "Learning with a Wasserstein loss". In: *Proceedings of the 28th International Conference on Neural Information Processing Systems*. Vol. 2. 2015, pp. 2053–2061.

[38] A. Gelman. Adding more information can make the variance go up (depending on your model). `https://statmodeling.stat.columbia.edu/2011/07/20/adding_more_inf/`. 2011.

[39] G. Gilardoni. "On the minimum f-divergence for given total variation". In: *Comptes Rendus Mathematique* 343.11-12 (2006), pp. 763–766.

[40] T. Gneiting. "Quantiles as optimal point forecasts". In: *International Journal of forecasting* 27.2 (2011), pp. 197–207.

[41] T. Gneiting and M. Katzfuss. "Probabilistic forecasting". In: *Annual Review of Statistics and Its Application* 1 (2014), pp. 125–151.

[42] T. Gneiting, L. I. Stanberry, E. P. Grimit, L. Held, and N. A. Johnson. "Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds". In: *Test* 17.2 (2008), p. 211.

[43] H. Haario, E. Saksman, and J. Tamminen. "An adaptive Metropolis algorithm". In: *Bernoulli* (2001), pp. 223–242.

[44] J. M. Hammersley and K. W. Morton. "Poor man's monte carlo". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 16.1 (1954), pp. 23–38.

[45] Y. Han, X. Liu, Z. Sheng, Y. Ren, X. Han, J. You, R. Liu, and Z. Luo. "Wasserstein loss-based deep object detection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2020, pp. 998–999.

[46] J. Haughton and S. R. Khandker. *Handbook on poverty+ inequality*. World Bank Publications, 2009.

[47] A. Heinen and E. Rengifo. "Multivariate autoregressive modeling of time series count data using copulas". In: *Journal of Empirical Finance* 14.4 (2007), pp. 564–583.

[48] M. D. Hoffman, A. Gelman, et al. "The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo." In: *J. Mach. Learn. Res.* 15.1 (2014), pp. 1593–1623.

[49] R. Hollyman, F. Petropoulos, and M. E. Tipping. "Understanding forecast reconciliation". In: *European Journal of Operational Research* 294.1 (2021), pp. 149–160.

[50] L. Hou, C.-P. Yu, and D. Samaras. "Squared earth movers distance loss for training deep neural networks on ordered-classes". In: *NIPS Workshop*. 2017.

[51] R. Hyndman and G. Athanasopoulos. *Forecasting: principles and practice, 3rd edition,* OTexts: Melbourne, Australia, 2021. URL: `OTexts.com/fpp3`.

[52] R. Hyndman. "Another look at forecast-accuracy metrics for intermittent demand". In: *Foresight: The International Journal of Applied Forecasting* 4.4 (2006), pp. 43–46.

[53] R. Hyndman, A. B. Koehler, J. K. Ord, and R. D. Snyder. *Forecasting with exponential smoothing: the state space approach*. Springer Science & Business Media, 2008.

[54] R. J. Hyndman. *expsmooth: Data sets from "Exponential smoothing: a state space approach" by Hyndman, Koehler, Ord and Snyder (Springer, 2008)*. R package version 2.4. 2018. URL: `http://pkg.robjhyndman.com/expsmooth`.

[55] R. J. Hyndman, R. A. Ahmed, G. Athanasopoulos, and H. L. Shang. "Optimal combination forecasts for hierarchical time series". In: *Computational statistics & data analysis* 55.9 (2011), pp. 2579–2589. ISSN: 0167–9473.

[56] K. Janocha and W. M. Czarnecki. "On Loss Functions for Deep Neural Networks in Classification". In: *Schedae Informaticae* 25 (2016), pp. 49–59.

[57] J. Jeon, A. Panagiotelis, and F. Petropoulos. "Probabilistic forecast reconciliation with applications to wind power and electric load". In: *European Journal of Operational Research* 279.2 (2019), pp. 364–379.

[58] H. Kahn and A. W. Marshall. "Methods of reducing sample size in Monte Carlo computations". In: *Journal of the Operations Research Society of America* 1.5 (1953), pp. 263–278.

[59] S. Kolassa. "Evaluating predictive count data distributions in retail sales forecasting". In: *International Journal of Forecasting* 32.3 (2016), pp. 788–803.

[60] N. Kourentzes and G. Athanasopoulos. "Elucidate structure in intermittent demand series". In: *European Journal of Operational Research* 288.1 (2021), pp. 141–152.

[61] S. Kullback and R. A. Leibler. "On information and sufficiency". In: *The annals of mathematical statistics* 22.1 (1951), pp. 79–86.

[62] C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Póczos. "MMD GAN: Towards deeper understanding of moment matching network". In: *arXiv preprint arXiv:1705.08584* (2017).

[63] T. Liboschik, K. Fokianos, and R. Fried. "tscount: An R Package for Analysis of Count Time Series Following Generalized Linear Models". In: *Journal of Statistical Software* 82.5 (2017), pp. 1–51.

[64] J. Lin. "Divergence measures based on the Shannon entropy". In: *IEEE Transactions on Information theory* 37.1 (1991), pp. 145–151.

[65] H. Ling and K. Okada. "An efficient earth mover's distance algorithm for robust histogram comparison". In: *IEEE transactions on pattern analysis and machine intelligence* 29.5 (2007), pp. 840–853.

[66] S. Makridakis, E. Spiliotis, and V. Assimakopoulos. "The M5 competition: Background, organization, and implementation". In: *International Journal of Forecasting* (2021).

[67] Y. Nesterov. "Dual extrapolation and its applications to solving variational inequalities and related problems". In: *Mathematical Programming* 109.2 (2007), pp. 319–344.

[68] U. Ojha, Y. Li, J. Lu, A. A. Efros, Y. J. Lee, E. Shechtman, and R. Zhang. "Few-shot image generation via cross-domain correspondence". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 10743–10752.

[69] Z. Pan, W. Yu, B. Wang, H. Xie, V. S. Sheng, J. Lei, and S. Kwong. "Loss functions of generative adversarial networks (GANs): opportunities and challenges". In: *IEEE Transactions on Emerging Topics in Computational Intelligence* 4.4 (2020), pp. 500–522.

[70] A. Panagiotelis, G. Athanasopoulos, P. Gamakumara, and R. J. Hyndman. "Forecast reconciliation: A geometric view with new insights on bias correction". In: *International Journal of Forecasting* 37.1 (2021), pp. 343–359.

[71] A. Panagiotelis, P. Gamakumara, G. Athanasopoulos, and R. J. Hyndman. "Probabilistic forecast reconciliation: Properties, evaluation and score optimisation". In: *European Journal of Operational Research* (2022).

[72] V. M. Panaretos and Y. Zemel. "Statistical aspects of Wasserstein distances". In: *Annual review of statistics and its application* 6 (2019), pp. 405–431.

[73] N. Papadakis. "Optimal transport for image processing". PhD thesis. Université de Bordeaux; Habilitation thesis, 2015.

[74] G. Peyré and M. Cuturi. "Computational optimal transport: With applications to data science". In: *Foundations and Trends® in Machine Learning* 11.5-6 (2019), pp. 355–607.

[75] S. T. Rachev. *Probability Metrics and the Stability of Stochastic Models.* Vol. 269. Wiley, 1991.

[76] S. T. Rachev, L. B. Klebanov, S. V. Stoyanov, and F. Fabozzi. *The methods of distances in the theory of probability and statistics.* Vol. 10. Springer, 2013.

[77] K. R. Rao and P. C. Yip. *The transform and data compression handbook.* CRC press, 2018.

[78] J. Salvatier, T. V. Wiecki, and C. Fonnesbeck. "Probabilistic programming in Python using PyMC3". In: *PeerJ Computer Science* 2 (2016), e55.

[79] I. Sason. "Tight bounds for symmetric divergence measures and a refined bound for lossless source coding". In: *IEEE Transactions on Information Theory* 61.2 (2014), pp. 701–707.

[80] J. Schmidhuber. "Deep learning in neural networks: An overview". In: *Neural networks* 61 (2015), pp. 85–117.

[81] A. F. Smith and A. E. Gelfand. "Bayesian statistics without tears: a sampling–resampling perspective". In: *The American Statistician* 46.2 (1992), pp. 84–88.

[82] A. A. Syntetos and J. E. Boylan. "The accuracy of intermittent demand estimates". In: *International Journal of Forecasting* 21.2 (2005), pp. 303–314. ISSN: 0169-2070.

[83] G. J. Székely and M. L. Rizzo. "Energy statistics: A class of statistics based on distances". In: *Journal of statistical planning and inference* 143.8 (2013), pp. 1249–1272.

[84] S. B. Taieb, J. W. Taylor, and R. J. Hyndman. "Hierarchical probabilistic forecasting of electricity demand with smart meter data". In: *Journal of the American Statistical Association* 116.533 (2021), pp. 27–43.

[85] R. E. Tarjan. "Dynamic trees as search trees via euler tours, applied to the network simplex algorithm". In: *Mathematical Programming* 78.2 (1997), pp. 169–177.

[86] C. Villani. *Optimal transport: old and new.* Vol. 338. Springer Science & Business Media, 2008.

[87] C. R. Vogel and M. E. Oman. "Iterative methods for total variation denoising". In: *SIAM Journal on Scientific Computing* 17.1 (1996), pp. 227–238.

[88] Z. Wang, J. Li, and M. Enoh. "Removing ring artifacts in CBCT images via generative adversarial networks with unidirectional relative total variation loss". In: *Neural Computing and Applications* 31.9 (2019), pp. 5147–5158.

[89] N. A. Weiss. *A Course in Probability*. Pearson, 2005.

[90] E. Wheatcroft. "Interpreting the skill score form of forecast performance metrics". In: *International Journal of Forecasting* 35.2 (2019), pp. 573–579.

[91] S. L. Wickramasuriya. "Probabilistic forecast reconciliation under the Gaussian framework". In: *arXiv preprint arXiv:2103.11128* (2021).

[92] S. L. Wickramasuriya, G. Athanasopoulos, and R. J. Hyndman. "Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization". In: *Journal of the American Statistical Association* 114.526 (2019), pp. 804–819.

[93] S. L. Wickramasuriya, B. A. Turlach, and R. J. Hyndman. "Optimal non-negative forecast reconciliation". In: *Statistics and Computing* 30.5 (2020), pp. 1167–1182.

[94] D. Yang. "Reconciling solar forecasts: Probabilistic forecast reconciliation in a nonparametric framework". In: *Solar Energy* 210 (2020), pp. 49–58.

[95] M. Yang, G. Zamba, and J. Cavanaugh. *ZIM: Zero-Inflated Models (ZIM) for Count Time Series with Excess Zeros*. R package version 1.1.0. 2018. URL: `https://CRAN.R-project.org/package=ZIM`.

[96] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. "Generative image inpainting with contextual attention". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 5505–5514.

[97] J. Yu, M. Wang, and D. Tao. "Semisupervised multiview distance metric learning for cartoon synthesis". In: *IEEE Transactions on Image Processing* 21.11 (2012), pp. 4636–4648.

[98] Z. Zhang, P. Tang, and T. Corpetti. "Time adaptive optimal transport: A framework of time series similarity measure". In: *IEEE Access* 8 (2020), pp. 149764–149774.