ProCAST: A Projection Framework for Coupled Aggregation Constrained Multivariate Time Series Forecasting

Jiaqi Xue[1], **Hongji Dong**[1], **Yucen Gao**[1], **Xiaofeng Gao**[*1], **Guihai Chen**[1]

[1]Shanghai Key Laboratory of Scalable Computing and Systems, School of Computer Science,
Shanghai Jiao Tong University, China
{xuejiaqi, harlin671}@sjtu.edu.cn, gao-yc@outlook.com, {gao-xf, gchen}@cs.sjtu.edu.cn

Abstract

Aggregated time series are widely used in business and economics, where top-level sequences (e.g., category sales) aggregated from underlying sequences (e.g., individual items) often exhibit clearer trends and are therefore typically the primary focus of forecasting tasks. However, treating top-level sequences as ordinary multivariate time series is inappropriate in the presence of coupled aggregation constraints. The core challenge arises in coupled aggregation structures, where a single underlying sequence contributes to multiple top-level sequences, as simple nonnegativity constraints of underlying sequences induce highly complex constraints among top-level sequences. Existing methods fail to achieve high accuracy while satisfying these constraints. To address this, we propose ProCAST, a projection-based framework that adjusts forecasts from any multivariate base model to satisfy coupled aggregation constraints. By introducing virtual underlying sequences and leveraging orthogonal and oblique projection, our method ensures that the top-level forecasts are feasible without explicitly deriving complex constraints. Theoretically, we prove that the proposed method guarantees improved accuracy under distance-based loss functions. Experiments on real-world datasets show that our method completely eliminates constraint violations while achieving higher accuracy than current state-of-the-art approaches.

Code — https://github.com/HellOwhatAs/ProCAST

Introduction

Aggregation is a common modeling paradigm in business and economic sales forecasting (Hyndman et al. 2011). When forecasting product sale trends, aggregated statistics often provide more meaningful insights than the large, sparse sales figures of individual items (Teixeira, Oliveira, and Ramos 2024). We refer to the sequences produced through aggregation as the *top-level sequences*, and those used to form them as the *underlying sequences*. Although forecasting may focus solely on top-level sequences, we demonstrate that the long-overlooked *coupled aggregate constraints* among them can substantially compromise both forecasting accuracy and structural consistency when these sequences are treated as ordinary multivariate time series.

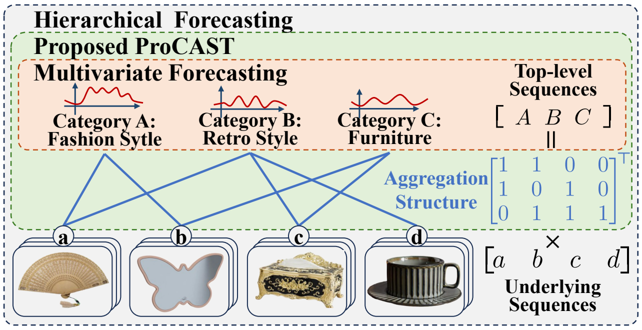


Figure 1: Comparison of how different forecasting methods utilize information within the aggregation structure.

If any underlying sequence does not aggregate to more than one top-level sequence, then we refer to the aggregation structure as not-coupled. Otherwise, it is referred to as a *coupled aggregate structure*. In such structures, simple range constraints on the underlying sequences (e.g., nonnegativity, which is common in statistical measures) can lead to extremely complex constraints among top-level sequences. We refer to these constraints as *coupled aggregate constraints*. For example, four non-negative underlying sequences, denoted as a, b, c, d, are aggregated to three top-level sequences labeled as A, B, C according to the aggregate structure in Figure 1.

This structure leads to the constraints $A + C \geq B$ and $B + C \geq A$, which can be derived through variable elimination. As the number of underlying sequences increases, these constraints become even more complex.

The challenge of this problem lies in balancing the tradeoff between prediction accuracy and meeting constraints.

First, multivariate time series forecasting methods will violate the constraints if we ignore them and directly forecast top-level sequences (Liu et al. 2024; Lu et al. 2024). This is because the feasible region that satisfies the constraints is extremely small. For example, in the coupled aggregate structure of one real-world dataset, we sampled 500,000 points uniformly in the $[0, 1000]$ range for each top-level dimension and only 931 (0.18%) satisfied the constraints. Without explicit enforcement, forecast values are extremely unlikely to lie in the feasible region.

[*] Corresponding author.

Second, the constraints among top-level sequences are hard to be explicitly written. A natural approach is to explicitly model the constraints and integrate them into the multivariate time series forecasting process for top-level sequences. However, in practical scenarios, the number of underlying sequences may be quite large, making it hard to explicitly derive equality and inequality constraints among top-level sequences. The time complexity of elimination is double-exponential (Fourier 1827) and proven no polynomial-time algorithm unless $P = NP$ (Tiwary 2012).

Third, underlying observations are often noisy or completely unavailable. Therefore, the Bottom-Up forecasting method, a basic approach in hierarchical forecasting that aggregates individual underlying forecasts into top-level predictions, suffers from low accuracy because the limited scale of the underlying data impedes pattern learning (Das et al. 2023b; Xu et al. 2024). Advanced hierarchical forecasting methods also heavily depend on underlying observations and thus fail to achieve sufficient accuracy when the goal is to forecast the top-level sequences.

To tackle these challenges, we propose ProCAST (a **Pro**jection framework for **C**oupled **A**ggregation con**S**trained multivariate **T**ime series forecasting). ProCAST introduces two projection approaches, orthogonal and oblique projection, that correct constraint-violating forecast points by projecting them into the feasible region. Unlike multivariate methods that only model top-level sequences, and hierarchical methods requiring both top-level and underlying data, our approach relies solely on top-level forecasts and the aggregation structure, without needing observed underlying values. During the projection process, we introduce virtual underlying sequences as optimization variables and impose simple constraints on them. Using the fixed aggregation procedure, we generate top-level forecasts that are inherently feasible, eliminating the need to explicitly derive constraints among top-level sequences. This approach additionally prevents the introduction of noise and remains effective even when underlying data is not available.

- We propose ProCAST built upon a multivariate time series forecasting base model, incorporating projection methods to enforce constraints. This approach ensures both accuracy and feasibility of the top-level sequences.
- We introduce two projection approaches, orthogonal and oblique projection, which come with rigorous theoretical guarantees and achieve competitive results.
- By leveraging the inherent aggregate structure rather than observed underlying values, the proposed method remains effective even when the underlying sequence is completely unavailable.

Related Work

Multivariate Time Series Forecasting

Multivariate time series forecasting aims to learn both temporal and cross-variable dependencies. Early models relied on mixed-channel inputs (Zhou et al. 2021; Zhang and Yan 2023), but subsequent work showed that modeling each variable separately can also yield strong performance. For example, (Nie et al. 2023) applies a Transformer to each variable independently, and (Zeng et al. 2023) demonstrates the power of simple linear decomposition. Later MLP-based methods explored alternative mixing strategies to combine information across series (Chen et al. 2023; Wang et al. 2024a,b). Other approaches employ specialized network architectures to capture complex cross-variable dependencies explicitly. (Cao et al. 2020) uses a graph-based method to learn dependencies among variables, while (Liu et al. 2024) applies attention across features instead of over time. KAN-based models revisit variable-specific modeling with expert gating networks (Han et al. 2024), and (Lu et al. 2024) captures cross-variable dependencies using a centralized strategy with linear complexity. Despite their advances, these automatic-learning methods still struggle with highly complex dependencies, making them unsuitable for our problem.

Constrained Optimization Learning

This type of method involves machine learning and constrained optimization techniques to form integrated models that incorporate constraints into the learning process (Kotary et al. 2021; Tanneau and Hentenryck 2024).

Implicit methods incorporate optimization layers for end-to-end training. (Agrawal et al. 2019) enables differentiation through convex programs at high computational cost, and (Amos and Kolter 2017) integrates quadratic programs via implicit differentiation but requires strict convexity.

Explicit methods require known equality and inequality constraints. (Donti, Rolnick, and Kolter 2021) applies equality completion and inequality correction, (Konstantinov and Utkin 2023) uses line-search projections for convex constraints, and (Qiu, Tanneau, and Van Hentenryck 2024) enforces feasibility constraints via dual proxies.

However, our problem cannot be formulated with explicit constraints, and existing implicit methods are either too costly or too restrictive.

Hierarchical Forecasting

In hierarchical time series forecasting, coherence is as important as accuracy. It requires that each aggregated forecast equals the sum of its components, but this cannot be guaranteed when forecasting each series independently.

Traditional hierarchical methods ensure coherence by forecasting a single level of the hierarchy and then reconcile forecasts using Top-Down (Athanasopoulos, Ahmed, and Hyndman 2009; Das et al. 2023b), Bottom-Up (Jain 1995; Kahn 1998), or Middle-Out (Hollyman, Petropoulos, and Tipping 2021) approaches.

On this basis, (Hyndman et al. 2011) introduced generalized least squares reconciliation, (Wickramasuriya, Athanasopoulos, and Hyndman 2019) uses the full error covariance, and (Ben Taieb and Koo 2019) relaxes the unbiasedness assumption. More recent works integrate reconciliation methods into end-to-end learning frameworks (Rangapuram et al. 2021; Tsiourvas et al. 2024).

However, these methods depend on noisy underlying observations, making it unlikely to improve accuracy and potentially even worsen it when the forecast target focus solely on top-level sequences.

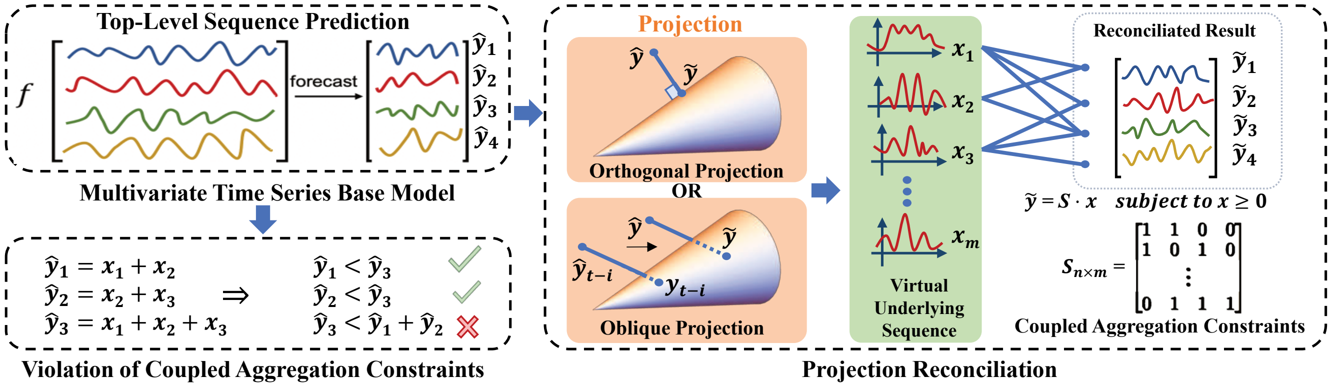

Top-Level Sequence Prediction
f
forecast
$\hat{y}_1$
$\hat{y}_2$
$\hat{y}_3$
$\hat{y}_4$
Multivariate Time Series Base Model
$\hat{y}_1 = x_1 + x_2$
$\hat{y}_2 = x_2 + x_3$
$\hat{y}_3 = x_1 + x_2 + x_3$
$\Rightarrow$
$\hat{y}_1 < \hat{y}_3$ ✓
$\hat{y}_2 < \hat{y}_3$ ✓
$\hat{y}_3 < \hat{y}_1 + \hat{y}_2$ ✗
Violation of Coupled Aggregation Constraints
Projection
$\hat{y}$
$\tilde{y}$
Orthogonal Projection
OR
$\hat{y}$
$\tilde{y}$
$\hat{y}_{t-i}$
y_{t-i}
Oblique Projection
x_1
x_2
x_3
x_m
Virtual Underlying Sequence
Reconciliated Result
$\tilde{y}_1$
$\tilde{y}_2$
$\tilde{y}_3$
$\tilde{y}_4$
$\tilde{y} = S \cdot x$ subject to $x \geq 0$
$S_{n\times m} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ & \vdots & & \\ 0 & 1 & 1 & 1 \end{bmatrix}$
Coupled Aggregation Constraints
Projection Reconciliation

Figure 2: Overview of ProCAST, including two optional projection methods: orthogonal and oblique.

Problem Formulation

In this section, we formally define the problem of multivariate time series forecasting with coupled aggregation constraints. We first provide a formal definition of the aggregation structure, which combines the underlying sequences to form the top-level sequences.

Definition 1 (Coupled Aggregation Structure)**.** Assume we have n top-level sequences, denoted as $Y = \{y_1, y_2, \ldots, y_n\}$, and m underlying sequences, denoted as $X = \{x_1, x_2, \ldots, x_m\}$. Top-level sequences are aggregated from underlying sequences by a "summing" matrix $S \in \{0,1\}^{n\times m}$ that satisfies $Y = S \cdot X$. In addition, if S does not satisfy the property in Eqn. (1), the structure have no coupling and the problem is trivial.

$$\left(\max_{i=1,\ldots,m} \sum_{j=1}^{n} S_{ij}\right) > 1 \tag{1}$$

The property in Eqn. (1) reveals coupling in the aggregation structure, where a single underlying sequence may belong to multiple top-level sequences as an aggregation component. Focusing on our problem, we care only about the accuracy of the top-level sequences and whether they satisfy the constraints. In Definition 2, we formally define the constraints among top-level sequences.

Definition 2 (Coupled-Coherence)**.** Let $\hat{Y}$ be a forecast for top-level sequences. If there *exist* underlying sequences $\hat{X} \geq 0$ that satisfy the coupled aggregation structure $S \cdot \hat{X} = \hat{Y}$, then we say that $\hat{Y}$ satisfies coupled-coherence.

Then we define our problem as a forecast for the top-level sequences satisfying coupled-coherence.

Definition 3 (Coupled-Coherence Forecasting Problem)**.** Given a coupled aggregation structure S and historical top-level sequences $Y_{t-h:t-1}$, the goal is to forecast the top-level sequences $\hat{Y}_t$ that minimize the forecast error in Eqn. (2) while satisfying coupled-coherence constraints in Eqn. (3).

$$\min \mathcal{L}(\hat{Y}_t, Y_t) \tag{2}$$

$$\text{s.t. } \exists \hat{X} \geq 0,\ S \cdot \hat{X} = \hat{Y}_t \tag{3}$$

$\mathcal{L}$ is a loss function that measures the forecast error between the forecasted top-level sequences $\hat{Y}_t$ and the true values Y_t.

Methodology

Overview

Figure 2 illustrates the framework of ProCAST, including two projection approaches for time series forecasting with coupled aggregate constraints. ProCAST begins with a base multivariate model that generates initial top-level sequence predictions $\hat{\mathbf{y}}$, which often violate the inherent coupled aggregation constraints. To address this, we first project these unconstrained predictions onto the feasible space via either orthogonal or oblique projection, resulting in a virtual underlying sequence $\tilde{\mathbf{y}}$ and then produce the final coherent forecasts that strictly satisfy the coupled aggregation constraints using the summing matrix S.

Preliminaries of Projection

In this subsection, we conduct a geometric analysis of the problem and deriving properties essential for our projection methods. Theorem 2 proves that the set of points satisfying coupled-coherence forms a convex cone, which serves as the foundation for our projection methods.

Since we have n top-level sequences and m underlying sequences, we work in the full space $\mathbb{R}^{m+n}$, where n dimensions correspond to n top-level sequences and m to the underlying sequences. Thus, at any given timestamp, the values of all time series can be represented by a point $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{m+n}$ where $\mathbf{y} \in \mathbb{R}^n$ is the values of top-level sequences and $\mathbf{x} \in \mathbb{R}^m$ is the values of underlying sequences.

Next, we discuss the coupled aggregation structure between the top-level and underlying sequences. As described in Definition 1, this relationship is mediated by matrix S that $\mathbf{y}$ and $\mathbf{x}$ must satisfy $\mathbf{y} = S \cdot \mathbf{x}$. This equation geometrically defines a m-dimensional hyperplane in an $(m+n)$-dimensional space, where a point satisfies the coupled aggregation constraints if and only if it lies on this hyperplane. Then we further incorporate the non-negativity constraint on the values of underlying sequences, i.e., $x \geq 0$.

Definition 4 (Coupled-Coherence Region)**.** Let $\mathcal{C} \subseteq \mathbb{R}^{m+n}$ denote the region satisfying the coupled-coherence. The region $\mathcal{C}$ is defined by both coupled-aggregation constraints and non-negativity constraints, as specified in Eqn. (4).

$$\mathcal{C} = \left\{ (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{m+n} \,\middle|\, \mathbf{x} \geq \mathbf{0},\ \mathbf{y} = \mathbf{S} \cdot \mathbf{x} \right\} \tag{4}$$

Theorem 1. *Region $\mathcal{C}$ in full $(m+n)$-dimensional space forms a convex cone.*

Since our problem concerns only top-level sequence prediction, we further consider the observable Coupled-Coherence Region in n-dimensional space, which we term the Projected Coupled-Coherence Region.

Definition 5 (Projected Coupled-Coherence Region)**.** We denote by $\pi(\cdot)$ the projection operation from an $(m+n)$-dimensional space to an n-dimensional space. As defined in Eqn. (5), the Projected Coupled-Coherence Region $\pi(\mathcal{C})$ is the projection of $\mathcal{C}$ onto the n-dimensional space. A value of top-level sequence $\mathbf{y}$ satisfies Coupled-Coherence if and only if $\mathbf{y} \in \pi(\mathcal{C})$ by Definition 2.

$$\pi(\mathcal{C}) = \{\mathbf{y} \in \mathbb{R}^n \mid \exists \mathbf{x} \in \mathbb{R}^m,\ \mathbf{x} \geq \mathbf{0},\ \mathbf{y} = \mathbf{S} \cdot \mathbf{x}\} \tag{5}$$

Theorem 2. *The Projected Coupled-Coherence Region $\pi(\mathcal{C})$ defined in Definition 5 is also a convex cone.*

Orthogonal Projection

Prediction points that violate the coupled-coherence property lie outside the projected coupled-coherence region $\pi(\mathcal{C})$. To enforce coupled-coherence, a straightforward strategy is to move those points into $\pi(\mathcal{C})$. We propose an orthogonal projection reconciliation method that eliminates any violation of coupled-coherence, and we prove that this method is guaranteed to improve forecast accuracy with respect to a class of loss functions based on a distance metric.

If a forecast point is inside the projected coupled-coherence region, it is a trivial case and should remain unchanged during reconciliation. If it lies outside, we must map it back into the region. Here, we employ the orthogonal projection method as defined in Definition 6 to find the closest point in the projected coupled-coherence region $\pi(\mathcal{C})$ to the forecast point $\hat{\mathbf{y}}$, as shown in Figure 3.

Definition 6 (Orthogonal Projection)**.** Given a forecast point $\hat{\mathbf{y}} \in \mathbb{R}^n$, the orthogonal projection onto the projected coupled-coherence region $\pi(\mathcal{C})$ is defined in Eqn. (6).

$$\tilde{\mathbf{y}} = \underset{\tilde{\mathbf{y}} \in \pi(\mathcal{C})}{\arg\min} \|\tilde{\mathbf{y}} - \hat{\mathbf{y}}\|_2^2 \tag{6}$$

Where $\|\cdot\|_2$ denotes the Euclidean norm, $\tilde{\mathbf{y}} \in \mathbb{R}^n$ is the reconciled top-level sequence.

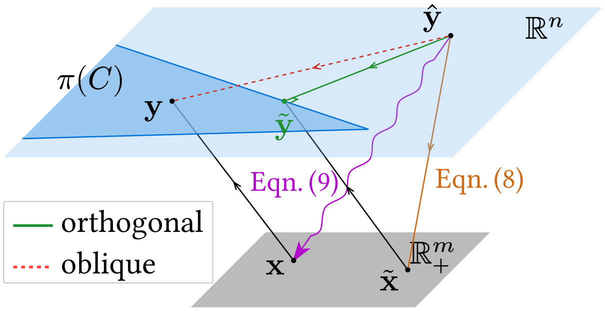


Figure 3: Orthogonal and Oblique projection of a forecast point onto the Projected Coupled-Coherence Region $\pi(\mathcal{C})$.

In this context, orthogonal projection does not refer to the usual perpendicular projection onto a subspace. It instead means finding the point in target set that is closest to the original point, which in fact projects onto a convex cone, as we discussed in Theorem 2. According to the Hilbert Projection Theorem (Rudin 1991), the orthogonal projection defined in Definition 6 guarantees to yield a unique solution $\tilde{\mathbf{y}}$.

Furthermore, we can prove in Theorem 3 that the orthogonal projection never increases the sum of squared errors.

Lemma 1. *Let $\hat{\mathbf{y}} \in \mathbb{R}^n$ be a forecast point and $\tilde{\mathbf{y}} \in \pi(\mathcal{C})$ be the orthogonal projection of $\hat{\mathbf{y}}$ onto $\pi(\mathcal{C})$. Then, $\forall \mathbf{y} \in \pi(\mathcal{C})$, we have $(\mathbf{y} - \tilde{\mathbf{y}})^\top (\tilde{\mathbf{y}} - \hat{\mathbf{y}}) \geq 0$ always holds.*

Proof. Let us define $f(\tilde{\mathbf{y}}) = \|\tilde{\mathbf{y}} - \hat{\mathbf{y}}\|_2^2$. According to the First-order optimality conditions, $\forall \mathbf{y} \in \pi(\mathcal{C})$, we have $\nabla f(\mathbf{y})^\top (\mathbf{y} - \tilde{\mathbf{y}}) \geq 0$. Since $\nabla f(\tilde{\mathbf{y}}) = 2(\tilde{\mathbf{y}} - \hat{\mathbf{y}})$, we have $2(\tilde{\mathbf{y}} - \hat{\mathbf{y}})^\top (\mathbf{y} - \tilde{\mathbf{y}}) \geq 0$ holds. □

Theorem 3. *Let $\hat{\mathbf{y}} \in \mathbb{R}^n$ be a forecast point and $\tilde{\mathbf{y}} \in \pi(\mathcal{C})$ be the orthogonal projection of $\hat{\mathbf{y}}$ onto $\pi(\mathcal{C})$. Then, we have Eqn. (7) always holds.*

$$\forall \mathbf{y} \in \pi(\mathcal{C}),\ \|\mathbf{y} - \tilde{\mathbf{y}}\|_2 \leq \|\mathbf{y} - \hat{\mathbf{y}}\|_2 \tag{7}$$

Proof. By Lemma 1, we have $(\mathbf{y} - \tilde{\mathbf{y}})^\top (\tilde{\mathbf{y}} - \hat{\mathbf{y}}) \geq 0$ for any $\mathbf{y} \in \pi(\mathcal{C})$.

$$\begin{aligned} 0 &\leq (\mathbf{y} - \tilde{\mathbf{y}})^\top (\tilde{\mathbf{y}} - \hat{\mathbf{y}}) \\ &\leq 2\,(\mathbf{y} - \tilde{\mathbf{y}})^\top (\tilde{\mathbf{y}} - \hat{\mathbf{y}}) + \|\tilde{\mathbf{y}} - \hat{\mathbf{y}}\|_2^2 \\ &= \|(\mathbf{y} - \tilde{\mathbf{y}}) + (\tilde{\mathbf{y}} - \hat{\mathbf{y}})\|_2^2 - \|\mathbf{y} - \tilde{\mathbf{y}}\|_2^2 \\ &= \|\mathbf{y} - \hat{\mathbf{y}}\|_2^2 - \|\mathbf{y} - \tilde{\mathbf{y}}\|_2^2 \end{aligned}$$

Thus, we have $\forall \mathbf{y} \in \pi(\mathcal{C}),\ \|\mathbf{y} - \tilde{\mathbf{y}}\|_2 \leq \|\mathbf{y} - \hat{\mathbf{y}}\|_2$. □

Since the ground truth of top-level sequences always lies in $\pi(\mathcal{C})$, Theorem 3 ensures that applying the orthogonal projection reconciliation method will never increase the prediction error based on a distance metric.

However, the problem posed in Eqn. (6) cannot be solved directly, since $\pi(\mathcal{C})$ cannot be explicitly characterized. To address this issue, we build on Eqn. (6) by introducing optimization variables $\tilde{\mathbf{x}}$ for underlying sequences, thus reformulating the problem as Eqn. (8).

$$\begin{aligned} &\min_{\tilde{\mathbf{x}} \in \mathbb{R}^m} \|\hat{\mathbf{y}} - S \cdot \tilde{\mathbf{x}}\|_2^2 \\ &\text{subject to} \quad \tilde{\mathbf{x}} \geq \mathbf{0} \end{aligned} \tag{8}$$

This reduces to a non-negative least squares (NNLS) problem, which can be solved efficiently using TNT-NN (Myre et al. 2018). Although the optimizer $\tilde{\mathbf{x}}$ in Eqn. (8) is not unique and undetermined, the result $S\tilde{\mathbf{x}}$ is well-defined and unique. Moreover, Theorem 4 shows that solving Eqn. (8) is equivalent to performing the orthogonal projection in Eqn. (6), which is also unique.

Theorem 4. *Let $\tilde{\mathbf{y}}$ be the optimal point of Eqn. (6) and $\tilde{\mathbf{x}}$ be the optimal point of the optimization problem in Eqn. (8). Then $\tilde{\mathbf{y}} = S \cdot \tilde{\mathbf{x}}$ always holds.*

Oblique Projection

Despite the fact that orthogonal projection admits a unique solution and guarantees an improvement over the original prediction, it is not necessarily the projection that will yield the maximum improvement in forecast accuracy as expected. To address this issue, we further enhance the orthogonal projection method by incorporating a data driven approach. We learn a "projection matrix" P from historical data, which enables the projected points to lie within the convex cone. To distinguish it from the orthogonal projection approach, we call this the *oblique projection* method.

Specifically, we modify the optimization problem in Eqn. (8) and embed it into our training loop so that the model learns patterns from the historical ground truth. We then rewrite the original decision variable $\tilde{\mathbf{x}}$ as $P \cdot \hat{\mathbf{y}}$ and shift the objective from "matching $\hat{\mathbf{y}}$" to "matching the historical ground truth $\mathbf{y}$." In this way, the learned matrix P serves as a mapping from the observed top-level sequence to a valid underlying sequence, as shown in Figure 3.

The deviation of a single forecast from its ground truth is essentially random, and it cannot define an oblique-projection direction by itself. To address this, we gather h consecutive historical points and share the projection matrix P to learn their overall deviation direction. As Definition 7 shows, we optimize P over the past h timestamps and then use the optimized matrix P as the direction of oblique projection the current timestamp to produce reconciled forecast.

Definition 7 (Oblique Projection)**.** Given a coupled aggregation structure S, a set of predictions for top-level sequences $\hat{\mathbf{y}}_{t-h:t}$ and historical ground truth of top-level sequences $\mathbf{y}_{t-h:t-1}$, the projection matrix P at timestamp t can be learned in Eqn. (9), where h is the number of historical timestamps used to learn the projection matrix. Then the reconciled forecast of $\hat{\mathbf{y}}_t$ via oblique projection is $\tilde{\mathbf{y}}_t = S \cdot \max(P \cdot \hat{\mathbf{y}}_t, 0)$.

$$\begin{aligned} \min_{P \in \mathbb{R}^{m \times n}} \quad & \sum_{i=1}^{h} \|S \cdot P \cdot \hat{\mathbf{y}}_{t-i} - \mathbf{y}_{t-i}\|_2^2 \\ \text{subject to} \quad & P \cdot \hat{\mathbf{y}}_{t-i} \geq \mathbf{0} \quad i = 1, \ldots, h \end{aligned} \tag{9}$$

The oblique projection minimizes the variance of the prediction error in reconciled forecast. Oblique projection is characterized by a statistical property, as it is optimal in expectation, as shown in Theorem 5.

Theorem 5. *Let $\hat{\mathbf{y}}_t$ be the forecast point at timestamp t and $\tilde{\mathbf{y}}_t$ be the oblique projection of $\hat{\mathbf{y}}_t$ onto $\pi(\mathcal{C})$. Assuming stationarity, i.e. the joint distribution of $(\hat{\mathbf{y}}_\tau, \mathbf{y}_\tau)$ is the same for all time $\tau \in [t-h, t]$, we have*

$$\mathbb{E}\left[\left\|\tilde{\mathbf{y}}_t - \mathbf{y}_t\right\|_2^2\right] \leq \mathbb{E}\left[\left\|\hat{\mathbf{y}}_t - \mathbf{y}_t\right\|_2^2\right]. \tag{10}$$

Proof. Define the population mean-squared error of any projection matrix $Q \in \mathbb{R}^{m \times n}$ by

$$M(Q) = \mathbb{E}\left[\left\|SQ\hat{\mathbf{y}}_t - \mathbf{y}_t\right\|_2^2\right].$$

By stationarity, the matrix

$$P = \underset{\substack{Q \in \mathbb{R}^{m \times n} \\ \text{s.t. } Q\hat{\mathbf{y}}_{t-i} \geq 0}}{\arg\min} \sum_{i=1}^{h} \left\|SQ\hat{\mathbf{y}}_{t-i} - \mathbf{y}_{t-i}\right\|_2^2$$

also satisfies $M(P) \leq M(Q)$ for every feasible Q. In particular, there exists some P_0 (satisfy $P_0 \cdot \hat{\mathbf{y}}_t = \hat{\mathbf{x}}_t$) such that $S \cdot \max(0, P_0\hat{\mathbf{y}}_t) = \hat{\mathbf{y}}_t$.

$$M(P_0) = \mathbb{E}\left[\left\|\hat{\mathbf{y}}_t - \mathbf{y}_t\right\|_2^2\right].$$

Therefore we have

$$\mathbb{E}\left[\left\|\tilde{\mathbf{y}}_t - \mathbf{y}_t\right\|_2^2\right] = M(P) \leq M(P_0) = \mathbb{E}\left[\left\|\hat{\mathbf{y}}_t - \mathbf{y}_t\right\|_2^2\right].$$

□

Theorem 5 assumes stationarity of the joint distribution of the forecast and ground truth. However, in practice, the distribution may shift over time due to factors such as seasonality or trend. For some relatively stable time series these shifts may occur only on an annual scale (Schwarzkopf, Tersine, and Morris 1988). To address this, we recommend learning the projection matrix P from the most recent data, as deviation patterns evolve smoothly between successive timestamps. If the historical window is too long, the learned P may no longer suit the current correction, degrading the accuracy of reconciled forecast.

Implementation Details

Orthogonal Projection The orthogonal projection method solves the non-negative least squares (NNLS) problem in Eqn. (8), which involves only m variables subject to m inequality constraints. We use the TNT-NN algorithm to solve this problem, which is 95 times faster than FNNLS algorithm (Myre et al. 2018). The algorithm builds and updates an active set, so a precise time complexity formula is not readily available. In most cases, its time complexity is roughly $O(m^2)$, and it performs well in practice.

Oblique Projection The oblique projection method learns the projection matrix P by solving the optimization problem in Eqn. (9). This problem involves $m \times n$ variables and $h \times m$ inequality constraints. Although oblique projection solves for more parameters than orthogonal projection, we exploit the smooth evolution of the projection matrix P between consecutive timestamps by using solution from each timestamp to initialize the next. This reduces computational cost by roughly two orders of magnitude.

Experiments

Datasets

To comprehensively evaluate the effectiveness of our proposed projection methods for multivariate time series with coupled aggregate constraints, we conducted experiments on two real-world multivariate time series datasets.

E-Commerce Dataset This dataset comprises all transactions recorded between December 1, 2010, and December 9, 2011, by a UK-based online retailer that operates without a physical store. The company primarily sells distinctive gifts for various occasions, with a significant portion of its customers being wholesalers (Chen 2012). The dataset contains the daily sales quantities of over 2,000 products as underlying sequences. These products are aggregated into 14 categories with coupling, which serve as the top-level sequences.

Dataset	E-Commerce Dataset										RH Dataset					
Metric	MAE					RMSE					MAE			RMSE		
Method	Raw	BU	MinT	Ortho.	Obliq.	Raw	BU	MinT	Ortho.	Obliq.	Raw	Ortho.	Obliq.	Raw	Ortho.	Obliq.
NBEATS	1092	1149	1083	1092	**1024**	1983	2052	1975	1983	**1877**	827.3	827.3	**752.7**	1262	1262	**1153**
NBEATSx	1092	1149	1083	1092	**1024**	1983	2052	1975	1983	**1877**	827.3	827.3	**752.7**	1262	1262	**1153**
NHITS	1103	1194	1100	1103	**1032**	2001	2102	2002	2001	**1901**	867.5	867.5	**807.0**	1331	1331	**1240**
TimesNet	1050	1175	1082	1050	**1038**	1949	2087	1989	1949	**1887**	584.1	584.6	**562.9**	880.4	879.9	**842.4**
TCN	1067	1258	1086	1067	**1054**	1959	2155	1985	1959	**1912**	674.7	674.7	**638.2**	1013	1013	**951.2**
BiTCN	1073	1174	1075	1073	**1069**	1978	2084	1979	1978	**1896**	3382	3379	**3046**	5674	5655	**5306**
DeepNPTS	1058	1209	1111	1058	**1039**	1944	2115	2015	1944	**1886**	664.0	664.0	**630.3**	1023	1022	**961.7**
TFT	1065	1236	1074	1065	**1048**	1957	2133	1969	1957	**1877**	608.7	610.2	**582.7**	913.7	909.3	**860.9**
TiDE	1080	1122	1120	1080	**1032**	1985	2033	2030	1985	**1886**	1258	1258	**1133**	1949	1949	**1762**
DLinear	1456	1549	1496	1414	**1319**	2411	2500	2406	2330	**2226**	3690	3311	**3283**	9510	7620	**7539**
Informer	3053	2890	3048	3053	**1406**	3702	3560	3698	3702	**2347**	2.2e4	2.2e4	**2.1e4**	2.7e4	2.7e4	**2.6e4**
Autoformer	**1050**	1114	1058	**1050**	1052	1950	2032	1963	1950	**1880**	1025	1025	**910.4**	1554	1554	**1388**
FEDformer	1053	1115	1059	1053	**1045**	1949	2029	1958	1949	**1874**	595.5	595.5	**571.9**	901.5	901.4	**854.7**
PatchTST	**1041**	1122	1054	**1041**	1057	1924	2030	1945	1924	**1881**	589.7	589.6	**561.8**	864.3	864.3	**817.9**
TimeXer	1074	1169	1082	1074	**1061**	1973	2075	1981	1972	**1897**	799.3	799.6	**741.3**	1214	1213	**1115**
TimeMixer	1063	1149	1067	1063	**1051**	1966	2062	1972	1966	**1887**	608.3	608.8	**585.9**	920.9	920.5	**878.1**
TSMixer	1057	1140	1065	1057	**1048**	1959	2054	1968	1959	**1877**	619.0	619.0	**589.0**	948.6	948.6	**888.6**
TSMixerx	1062	1134	1067	1062	**1057**	1963	2047	1966	1963	**1897**	731.6	733.5	**704.8**	1123	1118	**1061**
iTransformer	1097	1173	1091	1097	**1076**	1983	2075	1981	1982	**1928**	612.7	612.8	**595.2**	917.8	916.4	**883.7**
RMoK	1072	1162	1076	1072	**1047**	1962	2069	1969	1962	**1901**	682.1	682.6	**634.6**	1037	1037	**957.5**
SOFTS	1106	1174	1093	1106	**1091**	1984	2075	1981	1984	**1921**	669.7	670.5	**634.9**	986.8	986.3	**927.8**
StemGNN	1059	1236	1083	1059	**1055**	1950	2136	1981	1950	**1886**	890.7	894.0	**826.0**	1396	1367	**1258**

Table 1: Main results of the proposed method on the E-Commerce and RH datasets. Best results are **bolded**; second-best underlined; cells shaded in gray indicate predictions that violate Coupled-Coherence. "Raw" denotes uncorrected forecasts; "BU" (Bottom-Up) and "MinT" (MinTrace) are hierarchical methods; "Ortho." and "Obliq." indicate orthogonal and oblique projections of the proposed ProCAST framework, respectively.

RH Dataset The time series in this dataset are derived from the Exchange-Rate dataset (Lai et al. 2018) and are adjusted according to real-world coupled aggregation structures. The dataset contains 7 top-level sequences and simulates the scenario that values of the underlying sequences are not accessible. This highlights the practical significance of our proposed method: even when the underlying sequences are completely unavailable, our method remains effective.

Evaluation Metrics

We adopt a variety of evaluation metrics to assess prediction quality. Since our proposed method comes with theoretical guarantees for distance-based measures, we select Root Mean Squared Error (RMSE) as a representative distance-based metric. We also include Mean Absolute Error (MAE) to demonstrate the practical effectiveness of our approach.

Baselines

To validate the effectiveness of our proposed projection methods, we compare them against established forecasting techniques, including univariate models, multivariate models, and hierarchical forecasting approaches.

We begin by incorporating several state-of-the-art univariate forecasting models as base predictors in ProCAST. Although these models disregard cross-variable relationships, our projection technique guarantees that their forecasts lie within the Projected Coupled-Coherence Region $\pi(\mathcal{C})$. The univariate models include NBEATS, NBEATSx, NHITS, TimesNet, TiDE, DeepNPTS, TFT, TCN, BiTCN, and FEDformer (Oreshkin et al. 2020; Olivares et al. 2022; Challu et al. 2023; Wu et al. 2023; Das et al. 2023a; Rangapuram et al. 2023; Lim et al. 2021; Lea et al. 2016; Sprangers, Schelter, and de Rijke 2023; Zhou et al. 2022).

For multivariate models, we consider Informer, PatchTST, TimeXer, iTransformer, DLinear, TSMixer, TimeMixer, SOFTS, StemGNN, and RMoK (Zhou et al. 2021; Nie et al. 2023; Wang et al. 2024b; Liu et al. 2024; Zeng et al. 2023; Chen et al. 2023; Wang et al. 2024a; Lu et al. 2024; Cao et al. 2020; Han et al. 2024).

To further demonstrate the effectiveness of our approach, we also evaluate hierarchical reconciliation methods on the E-Commerce dataset, which provides underlying series observations. The tested methods include the Bottom-Up (Jain 1995) method and the MinTrace (Wickramasuriya, Athanasopoulos, and Hyndman 2019) method.

Main Results

We evaluated our method on two datasets using a historical window of $h = 72$ time steps for oblique projection. Table 1 presents the average results over 5 runs with different random seeds. On the E-Commerce dataset, which includes underlying observations, we compared our approach against the Bottom-Up and MinT hierarchical forecasting methods. The RH dataset, however, cannot support hierarchical techniques because it lacks any underlying sequences.

All base forecasting models on both datasets produce some predictions that violate Coupled-Coherence, as shown

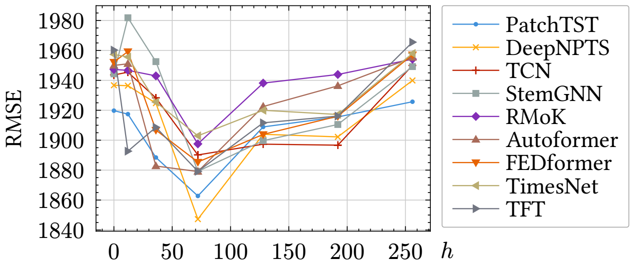


Figure 4: Hyperparameter sensitivity of h for the top-9 base forecasting models on the E-Commerce dataset.

in the gray cells in Table 1, with such violations occurring especially frequently in the RH dataset.

Across all datasets and base models, orthogonal projection consistently improves RMSE, as predicted by Theorem 3. Theorem 5 guarantees that oblique projection minimizes the expected loss and achieves the lowest RMSE. Empirically, our method achieves an average relative RMSE reduction of 5.15% on the E-Commerce dataset and 7.10% on the RH dataset, demonstrating its robustness across different settings. Even for MAE, where no theoretical guarantee is available, our projections still yield an average relative reduction of 4.33% and 6.19%, respectively, producing most of the best results and confirming practical effectiveness.

As for hierarchical forecasting methods, the Bottom-Up method performs worst, suffering from noise in the underlying observations. Although MinT provides moderate gains, it still typically underperforms the original forecasts.

Sensitivity Study

We evaluated the temporal continuity of the oblique projection matrix P by varying the historical window h and observing its effect on the reconciled forecasts. Figure 4 shows the 9 models with the lowest RMSE on the E-Commerce dataset (where $h = 0$ corresponds to the uncorrected forecasts). When $h < 40$, forecast error is highly volatile, reflecting insufficient historical information. As h grows from 40 to 80, error steadily decreases, demonstrating that P remains stable and can be learned reliably within this range. Beyond $h = 80$, error begins to rise again, since P gradually drifts and old observations no longer yield valid corrections.

Ablation Study

We use a matrix P in oblique projection method to map the top-level predictions ($\hat{\mathbf{y}}$) to virtual underlying sequence ($\tilde{\mathbf{x}}$). A natural question is why we adopt a linear mapping rather than a more complex nonlinear one. We conduct ablation study comparing different mapping structures.

Intuitively, aggregation $\mathbf{x} \to \mathbf{y}$ is linear via matrix S, so modeling $\hat{\mathbf{y}} \to \tilde{\mathbf{x}}$ via linear P is natural. We also tested replacing P with MLP, but observed accuracy degradation as model width/depth increased, as shown in Figure 5. Since infinitely many $\tilde{\mathbf{x}}$ satisfy the same $\mathbf{y}$, nonlinear models "overfit" easily, while linear P regularizes the underdetermined x, thus improving top-level accuracy.

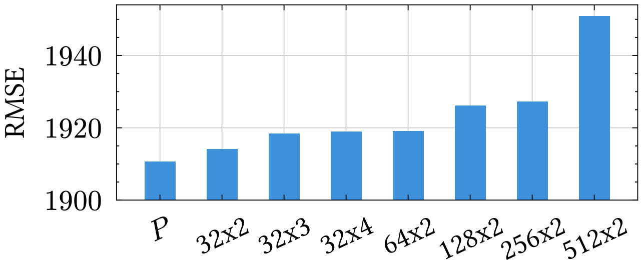
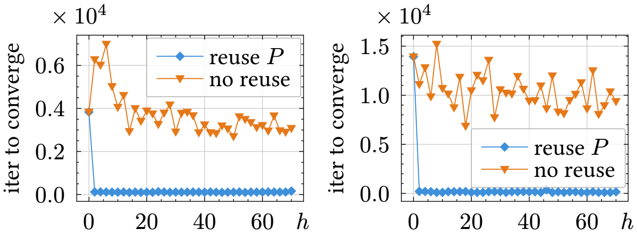


Figure 5: RMSE of Oblique Projection using different architecture of MLP as P on the E-Commerce dataset.



(a) E-Commerce Dataset

(b) RH Dataset

Figure 6: Iterations to coverage of oblique projection on predict days for reuse of P and no-reuse.

Time Consumption

As described in implementation details, we leverage the smooth temporal evolution of the projection matrix P by using solution of each timestamp to initialize the next. Comparative experiments on two datasets show that reusing P accelerates convergence, as shown in Figure 6, revealing strong temporal continuity. We observe up to 80× speedup on the E-Commerce dataset and 150× on the RH dataset.

On Platinum 8352V+RTX4090, 72-day inference of RH dataset: base model costs approximately 0.3s, Orthogonal Projection costs approximately 0.3s, Oblique Projection costs approximately 20s (while without reusing P costs 300s). For day-level forecast on our datasets and in practice, time is not a bottleneck, coherence and accuracy are critical.

Conclusion

ProCAST addresses the challenge of forecasting top-level series under coupled aggregation constraints by projecting unconstrained forecasts back into the valid aggregation region. By introducing virtual underlying sequences, it relies solely on top-level data and the known aggregation structure, ensuring accuracy and robustness even when underlying observations are noisy or missing. We prove that ProCAST guarantees error reduction under distance-based loss functions and show, on multiple real-world datasets, that it outperforms both standard multivariate and hierarchical methods while always satisfying aggregation constraints. By combining strong theoretical guarantees with practical resilience, ProCAST delivers a reliable solution for forecasting in complex business and economic environments.

Acknowledgments

This work was supported by the National Key R&D Program of China [2024YFF0617700], the National Natural Science Foundation of China [U23A20309, 62272302, 62372296]. We would like to express our sincere appreciation to the anonymous reviewers for their thoughtful feedback.

References

Agrawal, A.; Amos, B.; Barratt, S.; Boyd, S.; Diamond, S.; and Kolter, Z. 2019. Differentiable Convex Optimization Layers. In *Conference on Neural Information Processing Systems, NeurIPS*, 9562–9574.

Amos, B.; and Kolter, J. Z. 2017. OptNet: differentiable optimization as a layer in neural networks. In *International Conference on Machine Learning, ICML*, 136–145.

Athanasopoulos, G.; Ahmed, R. A.; and Hyndman, R. J. 2009. Hierarchical forecasts for Australian domestic tourism. *International Journal of Forecasting, IJF*, 25(1): 146–166.

Ben Taieb, S.; and Koo, B. 2019. Regularized Regression for Hierarchical Forecasting Without Unbiasedness Conditions. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD*, 1337–1347.

Cao, D.; Wang, Y.; Duan, J.; Zhang, C.; Zhu, X.; Huang, C.; Tong, Y.; Xu, B.; Bai, J.; Tong, J.; and Zhang, Q. 2020. Spectral Temporal Graph Neural Network for Multivariate Time-series Forecasting. In *Conference on Neural Information Processing Systems, NeurIPS*, 17766 – 17778.

Challu, C.; Olivares, K. G.; Oreshkin, B. N.; Garza Ramirez, F.; Mergenthaler Canseco, M.; and Dubrawski, A. 2023. NHITS: Neural Hierarchical Interpolation for Time Series Forecasting. *AAAI Conference on Artificial Intelligence, AAAI*, 37(6): 6989–6997.

Chen, D. 2012. Online Retail II. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5CG6D.

Chen, S.-A.; Li, C.-L.; Arik, S. O.; Yoder, N. C.; and Pfister, T. 2023. TSMixer: An All-MLP Architecture for Time Series Forecast-ing. *Transactions on Machine Learning Research, TMLR*.

Das, A.; Kong, W.; Leach, A.; Mathur, S.; Sen, R.; and Yu, R. 2023a. Long-term Forecasting with TiDE: Time-series Dense Encoder. *Trans. Mach. Learn. Res.*, 2023.

Das, A.; Kong, W.; Paria, B.; and Sen, R. 2023b. Dirichlet Proportions Model for Hierarchically Coherent Probabilistic Forecasting. In Evans, R. J.; and Shpitser, I., eds., *Conference on Uncertainty in Artificial Intelligence, UAI*, volume 216, 518–528.

Donti, P. L.; Rolnick, D.; and Kolter, J. Z. 2021. DC3: A learning method for optimization with hard constraints. In *International Conference on Learning Representations*.

Fourier, J. 1827. Histoire de l'Académie, partie mathématique (1824). In *Mémoires de l'Académie des sciences de l'Institut de France*, volume 7. Gauthier-Villars.

Han, X.; Zhang, X.; Wu, Y.; Zhang, Z.; and Wu, Z. 2024. KAN4TSF: Are KAN and KAN-based models Effective for Time Series Forecasting? *CoRR*.

Hollyman, R.; Petropoulos, F.; and Tipping, M. E. 2021. Understanding forecast reconciliation. *European Journal of Operational Research*, 294(1): 149–160.

Hyndman, R. J.; Ahmed, R. A.; Athanasopoulos, G.; and Shang, H. L. 2011. Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis*, 55(9): 2579–2589.

Jain, C. L. 1995. How to Determine the Approach to Forecasting. *The Journal of Business Forecasting*, 14: 2–3.

Kahn, K. B. 1998. Revisiting top-down versus bottom-up forecasting. *The Journal of Business Forecasting Methods & Systems*, 17(2): 14–19.

Konstantinov, A. V.; and Utkin, L. V. 2023. A New Computationally Simple Approach for Implementing Neural Networks with Output Hard Constraints. *Doklady Mathematics*, 108(2): S233–S241.

Kotary, J.; Fioretto, F.; Van Hentenryck, P.; and Wilder, B. 2021. End-to-End Constrained Optimization Learning: A Survey. In Zhou, Z.-H., ed., *International Joint Conference on Artificial Intelligence, IJCAI*, 4475–4482.

Lai, G.; Chang, W.-C.; Yang, Y.; and Liu, H. 2018. Modeling Long- and Short-Term Temporal Patterns with Deep Neural Networks. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 95–104.

Lea, C.; Vidal, R.; Reiter, A.; and Hager, G. D. 2016. Temporal Convolutional Networks: A Unified Approach to Action Segmentation. In *Computer Vision - ECCV 2016 Workshops - Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III*, volume 9915 of *Lecture Notes in Computer Science*, 47–54.

Lim, B.; Arik, S. O.; Loeff, N.; and Pfister, T. 2021. Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting, IJF*, 37(4): 1748–1764.

Liu, Y.; Hu, T.; Zhang, H.; Wu, H.; Wang, S.; Ma, L.; and Long, M. 2024. iTransformer: Inverted Transformers Are Effective for Time Series Forecasting. In *International Conference on Learning Representations, ICLR*.

Lu, H.; Chen, X.; Ye, H.; and Zhan, D. 2024. SOFTS: Efficient Multivariate Time Series Forecasting with Series-Core Fusion. In *Conference on Neural Information Processing Systems, NeurIPS*.

Myre, J. M.; Frahm, E.; Lilja, D. J.; and Saar, M. O. 2018. TNT: A Solver for Large Dense Least-Squares Problems that Takes Conjugate Gradient from Bad in Theory, to Good in Practice. In *International Parallel and Distributed Processing Symposium Workshops, IPDPSW*, 987–995.

Nie, Y.; Nguyen, N. H.; Sinthong, P.; and Kalagnanam, J. 2023. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In *International Conference on Learning Representations, ICLR*.

Olivares, K. G.; Challu, C.; Marcjasz, G.; Weron, R.; and Dubrawski, A. 2022. Neural basis expansion analysis with exogenous variables: Forecasting electricity prices with NBEATSx. *International Journal of Forecasting, IJF*.

Oreshkin, B. N.; Carpov, D.; Chapados, N.; and Bengio, Y. 2020. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations, ICLR*.

Qiu, G.; Tanneau, M.; and Van Hentenryck, P. 2024. Dual conic proxies for AC optimal power flow. *Electric Power Systems Research*, 236: 110661.

Rangapuram, S. S.; Gasthaus, J.; Stella, L.; Flunkert, V.; Salinas, D.; Wang, Y.; and Januschowski, T. 2023. Deep Non-Parametric Time Series Forecaster. *CoRR*.

Rangapuram, S. S.; Werner, L. D.; Benidis, K.; Mercado, P.; Gasthaus, J.; and Januschowski, T. 2021. End-to-End Learning of Coherent Probabilistic Forecasts for Hierarchical Time Series. In *International Conference on Machine Learning, ICML*, volume 139, 8832–8843.

Rudin, W. 1991. *Functional Analysis*. International Series in Pure and Applied Mathematics. Singapore: McGraw-Hill, 2nd edition. ISBN 0-07-054236-8. International Edition. When ordering, use ISBN 0-07-100944-2.

Schwarzkopf, A. B.; Tersine, R. J.; and Morris, J. S. 1988. Top-down versus bottom-up forecasting strategies. *International Journal of Production Research*, 26(11): 1833–1843.

Sprangers, O.; Schelter, S.; and de Rijke, M. 2023. Parameter-efficient deep probabilistic forecasting. *International Journal of Forecasting, IJF*, 39(1): 332–345.

Tanneau, M.; and Hentenryck, P. V. 2024. Dual Lagrangian Learning for Conic Optimization. arXiv:2402.03086.

Teixeira, M.; Oliveira, J. M.; and Ramos, P. 2024. Enhancing Hierarchical Sales Forecasting with Promotional Data: A Comparative Study Using ARIMA and Deep Neural Networks. *Machine Learning and Knowledge Extraction*, 6(4): 2659–2687.

Tiwary, H. R. 2012. On Computing the Shadows and Slices of Polytopes. arXiv:0804.4150.

Tsiourvas, A.; Sun, W.; Perakis, G.; Chen, P.-Y.; and Zhu, Y. 2024. Learning optimal projection for forecast reconciliation of hierarchical time series. In *International Conference on Machine Learning, ICML*.

Wang, S.; Wu, H.; Shi, X.; Hu, T.; Luo, H.; Ma, L.; Zhang, J. Y.; and Zhou, J. 2024a. TimeMixer: Decomposable Multiscale Mixing for Time Series Forecasting. In *International Conference on Learning Representations, ICLR*.

Wang, Y.; Wu, H.; Dong, J.; Qin, G.; Zhang, H.; Liu, Y.; Qiu, Y.; Wang, J.; and Long, M. 2024b. TimeXer: Empowering Transformers for Time Series Forecasting with Exogenous Variables. In *Conference on Neural Information Processing Systems, NeurIPS*.

Wickramasuriya, S. L.; Athanasopoulos, G.; and Hyndman, R. J. 2019. Optimal Forecast Reconciliation for Hierarchical and Grouped Time Series Through Trace Minimization. *Journal of the American Statistical Association*, 114(526): 804–819.

Wu, H.; Hu, T.; Liu, Y.; Zhou, H.; Wang, J.; and Long, M. 2023. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. In *International Conference on Learning Representations*.

Xu, K.; Yu, Z.; Gao, Y.; Zhang, S.; Fang, J.; Gao, X.; and Chen, G. 2024. MetaSTC: A Backbone Agnostic Spatio-Temporal Framework for Traffic Forecasting. In *IEEE International Conference on Data Mining, ICDM 2024, Abu Dhabi, United Arab Emirates, December 9-12, 2024*, 899–904. IEEE.

Zeng, A.; Chen, M.; Zhang, L.; and Xu, Q. 2023. Are Transformers Effective for Time Series Forecasting? In *AAAI Conference on Artificial Intelligence, AAAI*, 11121–11128.

Zhang, Y.; and Yan, J. 2023. Crossformer: Transformer Utilizing Cross-Dimension Dependency for Multivariate Time Series Forecasting. In *International Conference on Learning Representations*.

Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; and Zhang, W. 2021. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. In *AAAI Conference on Artificial Intelligence, AAAI*, 11106–11115.

Zhou, T.; Ma, Z.; Wen, Q.; Wang, X.; Sun, L.; and Jin, R. 2022. FEDformer: Frequency Enhanced Decomposed Transformer for Long-term Series Forecasting. In *International Conference on Machine Learning, ICML*, volume 162, 27268–27286.