Hierarchical Time Series Forecasting with Operator-Valued Kernels

Junjie Yang
LPSM, Sorbonne Université
junjie.yang@sorbonne-universite.fr

Adeline Fermanian
Califrais
adeline.fermanian@califrais.fr

Claire Boyer
LMO, Université Paris-Saclay
claire.boyer@universite-paris-saclay.fr

Gérard Biau
LPSM, Sorbonne Université
gerard.biau@sorbonne-universite.fr

Abstract

Hierarchical time series forecasting requires generating accurate predictions that are coherent across multiple levels of aggregation. Existing methods either rely on two-stage pipelines with post-hoc reconciliation or on end-to-end models that often lack statistical guarantees. We introduce a unified framework that formulates hierarchical forecasting as a multivariate regression problem in a vector-valued reproducing kernel Hilbert space. By designing structured operator-valued kernels, our approach enforces coherence directly within the hypothesis space, ensuring that all predictions satisfy hierarchical constraints by construction. We further propose tailored kernel constructions for hierarchical data and derive excess risk bounds that characterize generalization performance. Empirical results on synthetic and real-world datasets demonstrate the relevance of the proposed method.

1. Introduction

A *hierarchical time series* (HTS) is a collection of univariate time series organized across multiple levels of aggregation, where higher-level series are obtained by summing lower-level series according to a predefined structure (Hyndman and Athanasopoulos, 2018). A typical example arises in retail: daily demand can be recorded at the individual product level, aggregated into product categories, and further combined to form a total demand series. The problem of *hierarchical time series forecasting*, first introduced by Athanasopoulos et al. (2009) in the context of Australian domestic tourism, has recently attracted significant attention due to its relevance in a wide range of real-world applications. These include retail sales forecasting (Mancuso et al., 2021; Makridakis et al., 2022), energy consumption forecasting (Nespoli et al., 2020), and road traffic forecasting (Ben Taieb and Koo, 2019), among others.

A key challenge in hierarchical time series forecasting is that one must not only produce accurate forecasts for each individual series, but also ensure that these forecasts are *coherent*, i.e., they satisfy the aggregation constraints imposed by the hierarchy. Early approaches to this problem predominantly adopt a two-stage framework (e.g., Athanasopoulos et al., 2009; Hyndman et al., 2011; Ben Taieb and Koo, 2019; Wickramasuriya et al., 2019; Zhou et al., 2023). In the first stage, base forecasts are generated independently for each series using standard univariate time series models, such as ETS (Brown, 1959) or ARIMA (Box et al., 2015). In the second stage, these forecasts are adjusted to enforce coherence through a procedure known as reconciliation. More recently, several works (e.g., Rangapuram et al., 2021; Tsiourvas et al., 2024; Cini et al., 2024; Doumèche et al., 2025) have proposed to integrate base forecasting and reconciliation within a single end-to-end framework. In these approaches, model parameters are jointly learned across all series, enabling the direct generation of coherent forecasts without requiring a separate post-processing step. However, most of these methods rely on deep neural network architectures, e.g., DeepVAR (Salinas et al., 2019) or Autoformer (Wu et al., 2021), which can be difficult to train effectively in time series settings where data availability is limited. Furthermore, generalization guarantees remain limited for these end-to-end methods.

In this work, we recast the HTS forecasting problem as a (potentially non-linear) multivariate regression problem and introduce a unified kernel framework in which predictors are learned in a vector-valued reproducing ker-

. Correspondence to: Junjie Yang <junjie.yang@sorbonne-universite.fr>.

Preprint. October 8, 2026.

nel Hilbert space (vv-RKHS, Pedrick, 1957; Schwartz, 1964). This approach is motivated by several key advantages. First, by constructing operator-valued kernels (OVKs, Carmeli et al., 2006) whose associated vv-RKHSs consist only of functions taking values in the coherent subspace, we enforce coherence by construction. Second, the regularized least-squares problem has a unique global minimizer (see, e.g., Micchelli and Pontil, 2005). Finally, the well-established statistical learning theory for RKHSs (see, e.g., Smale and Zhou, 2007; Caponnetto and De Vito, 2007) allows us to derive generalization guarantees for our proposed predictor.

Our paper is organized as follows. In Section 2, we introduce the necessary background on hierarchical time series forecasting and operator-valued kernels. In Section 3, we introduce our predictor RVVKRR, based on vector-valued kernel ridge regression (Micchelli and Pontil, 2005), and provide efficient computational methods for both separable and finite-feature non-separable OVKs. In particular, we characterize the OVKs whose associated vv-RKHS, used as the hypothesis space, contains only functions that produce coherent forecasts. In Section 4, using integral operator theory in vv-RKHSs, we derive an excess risk bound for the proposed predictor under an i.i.d. regression model, explicitly highlighting the roles of both the hierarchical structure and the OVKs. In Section 5, we analyze the behavior of the proposed predictor on a synthetic dataset, and demonstrate its performance on two real-world datasets: TourismLarge and Traffic.

2. Preliminaries

2.1. Hierarchical Time Series Forecasting

A hierarchical time series is a multivariate time series $(y_t)_{t=1}^T \in (\mathbb{R}^d)^T$ that satisfies linear aggregation constraints, which can be typically encoded into a tree hierarchy (Hyndman et al., 2011). Here, $d \in \mathbb{N}^*$ represents the number of univariate time series. It can be divided into a *bottom-level series* $(y_t^b)_{t=1}^T$ with $y_t^b \in \mathbb{R}^{d_b}$ (i.e., the time series at the leaf nodes of the hierarchy), and an *aggregated series* $(y_t^a)_{t=1}^T$ with $y_t^a \in \mathbb{R}^{d_a}$ (i.e., those of the remaining nodes). Here, we have clearly $d = d_a + d_b$. Formally, at each timestamp t, the vector representing the whole time series y_t can be split into two such that $y_t = \begin{bmatrix}(y_t^a)^\mathsf{T} & (y_t^b)^\mathsf{T}\end{bmatrix}^\mathsf{T}$ and the whole time series can be obtained by the following transformation of the bottom-level series y_t^b:

$$y_t = S y_t^b \quad \text{with } S = \begin{bmatrix} S_{\text{sum}} \\ I_{d_b} \end{bmatrix} \in \{0,1\}^{d \times d_b}, \qquad (1)$$

where S is called the aggregation matrix and $S_{\text{sum}} \in \{0,1\}^{d_a \times d_b}$ is called the summation matrix. It should be noted that the summation matrix describes exactly the hierarchical structure of the time series. A time series that satisfies those constraints is said to be coherent, i.e., $y_t \in \text{Im}(S)$ for all $t \in \{1, \dots, T\}$. An example of such a hierarchical time series is provided in Appendix A.1.

Hierarchical Forecasting Methods. In the literature on hierarchical forecasting, to generate the coherent forecast $\hat{y}_t \in \text{Im}(S)$, a two-stage approach is typically followed. In the first stage, the multivariate and hierarchical structure of the data is completely ignored. A separate predictor is trained for each of the d univariate time series, with all predictors learned independently. Any classical time series forecasting method can be used at this stage, such as exponential smoothing (ETS, Brown, 1959), or kernel ridge regression (KRR, Saunders et al., 1998), to name a few. The resulting forecast of the multivariate time series at time step t, referred to as the *base forecast*, is denoted by $\tilde{y}_t \in \mathbb{R}^d$, which is not necessarily coherent. The second stage, called *reconciliation*, transforms the base forecast $\tilde{y}_t \in \mathbb{R}^d$ into a coherent forecast $\hat{y}_t \in \text{Im}(S) \subset \mathbb{R}^d$. Most reconciliation methods in the literature adopt a linear transformation $\hat{y}_t = SG\tilde{y}_t$ where $G \in \mathbb{R}^{d_b \times d}$ is a *bottom-extractor* matrix or a *reconciliation* matrix that forms only the bottom-level time series from the full set of forecasts across all hierarchy levels. Different choices of G correspond to different reconciliation strategies, detailed in Appendix A.2.

End-to-end approaches (e.g., Rangapuram et al., 2021; Tsiourvas et al., 2024; Doumèche et al., 2025) integrate reconciliation directly into the learning procedure. Typically, one chooses a hypothesis space $\mathcal{H} \subset (\mathbb{R}^d)^{\mathcal{X}}$ where $\mathcal{X}$ denotes the feature space, and parameterizes the predictor as $f(\cdot) = SGh(\cdot)$ with $h \in \mathcal{H}$. Learning is performed by solving an empirical risk minimization problem. For example, Rangapuram et al. (2021); Tsiourvas et al. (2024) parameterize $\mathcal{H}$ by neural network architectures, such as DeepVAR (Salinas et al., 2019) or Autoformer (Wu et al., 2021).

2.2. Operator-Valued Kernels

Let $\mathcal{X}$ be a nonempty set, $\mathcal{Y}$ be a real separable Hilbert space with inner space $\langle \cdot, \cdot \rangle_{\mathcal{Y}}$, and $\mathcal{L}(\mathcal{Y})$ be the space of bounded linear operators from $\mathcal{Y}$ to $\mathcal{Y}$. Operator-valued kernels (OVKs), defined as follows (see, e.g., Carmeli et al., 2006), can be seen as the generalization of scalar-valued kernels.

Definition 2.1 (Operator-Valued Kernels) *A function $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$ is said to be an operator-valued kernel if it is self-adjoint: for all $(x_i, x_j) \in \mathcal{X}^2$, $\mathcal{K}(x_i, x_j) = \mathcal{K}(x_j, x_i)^*$ where the superscript * denotes the adjoint of a bounded linear operator; and positive: for*

all $N \in \mathbb{N}^*$ *and* $y \in \mathcal{Y}$,

$$\sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j \langle \mathcal{K}(x_i, x_j)y, y\rangle_{\mathcal{Y}} \geq 0$$

for all $(x_1, \ldots, x_N) \in \mathcal{X}^N$ *and* $(a_1, \ldots, a_N) \in \mathbb{R}^N$.

The relevance of operator-valued kernels arises from their connection to function spaces known as vector-valued reproducing kernel Hilbert spaces (vv-RKHS, Pedrick, 1957; Schwartz, 1964), defined as follows.

Definition 2.2 (Vector-valued RKHSs) *A Hilbert space of functions* $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ *is called a* $\mathcal{Y}$*-valued reproducing kernel Hilbert space on* $\mathcal{X}$ *if for all* $x \in \mathcal{X}$ *there exists a positive constant* M_x *such that* $\|f(x)\|_{\mathcal{Y}} \leq M_x \|f\|_{\mathcal{H}}$ *for all* $f \in \mathcal{H}$.

Given an operator-valued kernel $\mathcal{K}$, one can show (see, e.g., Carmeli et al., 2006, Proposition 2.1) that there is a unique $\mathcal{Y}$-valued reproducing kernel Hilbert space (RKHS) $\mathcal{H}_{\mathcal{K}} \subset \mathcal{Y}^{\mathcal{X}}$ such that for all $x \in \mathcal{X}$ and for all $f \in \mathcal{H}_{\mathcal{K}}$,

$$f(x) = \mathcal{K}_x^* f \quad (\textit{reproducing property}), \tag{2}$$

where the linear operator $\mathcal{K}_x : \mathcal{Y} \to \mathcal{H}_{\mathcal{K}}$ is defined by $\mathcal{K}_x y(x') = \mathcal{K}(x', x)y$, for all $x' \in \mathcal{X}$ and for all $y \in \mathcal{Y}$. The converse is true by using the Riesz representation theorem (Carmeli et al., 2006).

A special class of OVKs, called the separable OVKs (see e.g., Caponnetto et al., 2008, Example 1), are often employed in real-world applications:

$$\mathcal{K}(x, x') = k(x, x')A, \tag{3}$$

where $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a scalar-valued kernel defined on the input space $\mathcal{X}$ and the operator $A \in \mathcal{L}(\mathcal{Y})$, referred to in some literature as the task kernel, is self-adjoint and positive, i.e., $A = A^*$ and $\langle Ay, y\rangle_{\mathcal{Y}} \geq 0$ for all $y \in \mathcal{Y}$. More details on the properties of separable OVKs are provided in Appendix A.3.

OVKs and vv-RKHSs have been employed to address a variety of supervised learning problems in which the output space can be modeled as a real separable Hilbert space. See Appendix A.4 for a detailed discussion. In this paper, the predictor we introduce for hierarchical time series forecasting is a function of a $\mathrm{Im}(S)$-valued RKHS.

3. Reconciled Forecasting with VVKRR

In this section, we first propose to construct a reconciled hierarchical time series predictor, $\hat{f} : \mathcal{X} \to \mathrm{Im}(S)$, using vector-valued kernel ridge regression (VVKRR). This approach, originally introduced by Micchelli and Pontil (2005), leverages OVKs to address regression problems with multidimensional outputs. We then present several strategies for designing appropriate OVKs tailored to the hierarchical time series setting.

3.1. Reconciled Vector-Valued Kernel Ridge Regression

Assume that we have observed a hierarchical time series along with its features, denoted as $((x_1, y_1), \ldots, (x_T, y_T)) \in (\mathcal{X} \times \mathrm{Im}(S))^T$. For instance, the feature vector $x_t \in \mathbb{R}^{d_b p+q}$ with $(p, q) \in (\mathbb{N}^*)^2$ could include p lagged observations of each of the d_b bottom-level series, along with q additional covariates capturing temporal effects such as seasonality or calendar-based features. $\mathrm{Im}(S)$ is a finite-dimensional subspace of $\mathbb{R}^d$, thus a real Hilbert space with the inner product inherited from $\mathbb{R}^d$. Moreover, one can show that $\dim(\mathrm{Im}(S)) = d_b$ where d_b is the number of bottom-level series (see Lemma B.1).

As a result, it is natural to design an OVK of the type $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathrm{Im}(S))$ and consider the vv-RKHS $\mathcal{H}_{\mathcal{K}} \subset \mathrm{Im}(S)^{\mathcal{X}}$ associated to it as the hypothesis space of our predictor. We choose to find such a function $\hat{f}$ by solving the following optimization problem:

$$\hat{f} = \operatorname*{arg\,min}_{f \in \mathcal{H}_{\mathcal{K}}} \frac{1}{T}\sum_{t=1}^{T} \|f(x_t) - y_t\|_{\mathbb{R}^d}^2 + \lambda \|f\|_{\mathcal{H}_{\mathcal{K}}}^2, \tag{4}$$

where $\lambda > 0$ is a regularization parameter. In the rest of the paper, we refer to such a predictor as the *reconciled vector-valued kernel ridge regression* (RVVKRR).

By the representer theorem (Micchelli and Pontil, 2005, Theorem 4.1), we know that $\hat{f}$ takes the form $\hat{f} = \sum_{t=1}^{T} \mathcal{K}_{x_t} c_t$ with $(c_t)_{t=1}^T \in (\mathrm{Im}(S))^T \subset (\mathbb{R}^d)^T$ satisfying the following linear system:

$$T\lambda c_t + \sum_{j=1}^{T} \mathcal{K}(x_t, x_j)c_j = y_t. \tag{5}$$

For separable OVKs, Equation (5) can be reduced to a discrete-time Sylvester equation that can be solved by the Hessenberg-Schur method (Golub et al., 1979; Sima, 1996). When a non-separable OVK admitting an explicit finite-dimensional feature map is employed, we solve the equivalent primal problem using matrix-free preconditioned conjugate gradient method (Hestenes and Stiefel, 1952; Cutajar et al., 2016). More details on the computational methods are provided in Appendix B.9. It can be shown that $\hat{f}$ can also be written as $\hat{f} = (\hat{\Sigma}_{\mathcal{K}} + \lambda I)^{-1} \frac{1}{T}\sum_{t=1}^{T} \mathcal{K}_{x_t} y_t$ with $\hat{\Sigma}_{\mathcal{K}} := \frac{1}{T}\sum_{t=1}^{T} \mathcal{K}_{x_t}\mathcal{K}_{x_t}^*$ (see e.g., Caponnetto and De Vito, 2007, Proposition 1). While the first expression of the predictor is more suitable from a computational point of view, the latter is more commonly

used in the statistical analysis of the predictor, as will be discussed further in Section 4.

3.2. Coherent Operator-Valued Kernels for Hierarchical Time Series.

In this section, we introduce two complementary constructions of operator-valued kernels tailored to hierarchical time series. First, we characterize separable coherent OVKs and show how to derive their task kernels from similarity matrices that need not be positive semi-definite. Second, we start from an arbitrary, possibly non-separable, $\mathcal{L}(\mathbb{R}^d)$-valued kernel and map it into the coherent output space through a reconciliation operator.

3.2.1. Intrinsic Construction of Separable Coherent OVKs

As we have shown in Section 2.2, the task kernel A of a separable OVK $\mathcal{K}(x, x') = k(x, x')A$ must be a self-adjoint and positive operator in $\mathcal{L}(\mathrm{Im}(S))$. In this section, we propose a method to construct such a task kernel A from a symmetric similarity matrix $M \in \mathbb{S}^d$ that encodes the pairwise similarities between the time series. First, we will characterize all coherent task kernels A by the following lemma, whose proof can be found in Appendix B.2.

Lemma 3.1 *Let $(u_1, \ldots, u_{d_b})$ be an orthonormal basis for $\mathrm{Im}(S)$ with $u_i \in \mathbb{R}^d$ for all $i \in \{1, \ldots, d_b\}$ (for example, obtained by the singular value decomposition of the matrix S). Let $U_S = \begin{bmatrix} u_1 & u_2 & \cdots & u_{d_b} \end{bmatrix} \in \mathbb{R}^{d \times d_b}$. Then any self-adjoint and positive operator $A \in \mathcal{L}(\mathrm{Im}(S))$ admits a matrix representation $U_S A' U_S^\mathsf{T}$ where $A' \in \mathbb{R}^{d_b \times d_b}$ is a symmetric and positive semi-definite matrix.*

As a result, designing a coherent separable OVK is equivalent to designing a positive semi-definite task kernel $A' \in \mathbb{S}_+^{d_b}$ in the d_b-dimensional coherent coordinate system. One can simply choose $A' = I_{d_b}$ to obtain a coherent task kernel $A = U_S U_S^\mathsf{T}$, which is the identity operator on $\mathrm{Im}(S)$. Another simple choice for A' is a diagonal matrix $A' = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_{d_b})$ where where $\lambda_i > 0$ for all $i \in \{1, \ldots, d_b\}$. However, those choices do not explicitly encode any prior knowledge about the relationships between time series, beyond the aggregation structure. In fact, it can be shown that this choice corresponds to a learning algorithm based on d_b independent kernel ridge regression predictors (KRR, Saunders et al., 1998; Evgeniou et al., 2000). See Appendix B.3 for more details.

Instead, we propose to construct the task kernel $A \in \mathcal{L}(\mathrm{Im}(S))$ from a similarity matrix $M \in \mathbb{S}^d$, assumed to be available a priori, that encodes the pairwise similarities between the time series. Since such a matrix may not lie in $\mathcal{L}(\mathrm{Im}(S))$ or may not be positive semi-definite, we seek a valid task kernel A by finding $A' \in \mathbb{S}_+^{d_b}$ that minimizes $\left\|M - U_S A' U_S^\mathsf{T}\right\|_\mathrm{F}^2$. The solution is given by the following lemma, whose proof can be found in Appendix B.4.

Lemma 3.2 *The solution of the optimization problem $\min_{A' \in \mathbb{S}_+^{d_b}} \left\|M - U_S A' U_S^\mathsf{T}\right\|_F^2$ is given by $A'_\mathrm{opt} = Q\mathrm{diag}(\max(\sigma_i, 0))Q^\mathsf{T}$, where $Q \in \mathbb{R}^{d_b \times d_b}$ and $(\sigma_i)_{i=1}^{d_b}$ are from the eigendecomposition of $U_S^\mathsf{T} M U_S$, i.e., $U_S^\mathsf{T} M U_S = Q\mathrm{diag}(\sigma_i)Q^\mathsf{T}$.*

In the case where $M \in \mathbb{S}_+^d$, the task kernel obtained via the above construction admits a simpler form. This result is summarized in the following corollary, whose proof is provided in Appendix B.5.

Corollary 3.3 *When $M \in \mathbb{S}_+^d$, the minimizer of the optimization problem $\min_A \|M - A\|_F^2$ with $A \in \left\{U_S A' U_S^\mathsf{T} \mid A' \in \mathbb{S}_+^{d_b}\right\}$ is written as $A_\mathrm{opt} = PMP$ where $P \in \mathbb{R}^{d \times d}$ is the orthogonal projection matrix onto $\mathrm{Im}(S)$.*

Following Corollary 3.3, we give an example of a coherent separable OVK constructed from the graph relationship between the time series, which will be used in the synthetic experiment of Section 5.1.

Example 1 (Graph-based separable OVK) *Let L be the Laplacian matrix of a graph encoding known relationships between the time series. Since L is positive semi-definite, we may take $M = L^\dagger$ in Corollary 3.3, where $\dagger$ denotes the Moore–Penrose inverse. This gives the coherent task kernel $A_\mathrm{Lap} = PL^\dagger P$ and the corresponding separable OVK $\mathcal{K}_\mathrm{Lap}(x, x') = k(x, x')PL^\dagger P$. For comparison, taking $M = I_d$ gives $\mathcal{K}_\mathrm{Id}(x, x') = k(x, x')P$.*

Remark 3.4 *Constructing a task kernel A based on the relationships between tasks modeled by a given structure is common in the multi-task learning literature. In the context where the structure is a general graph, Micchelli and Pontil (2004) propose considering the Moore–Penrose inverse of the graph Laplacian L. A non-singular variant, $A = (\mu L + \lambda I)^{-1}$, also appeared in Sheldon (2008). When the tasks are organized in a hierarchical fashion, which is closely related to our setting, Widmer et al. (2010) propose defining the similarity matrix between the organisms (leaf nodes) based on their distance in a taxonomic tree.*

3.2.2. General Coherent OVKs by Linear Composition

We now consider the case where a $\mathcal{L}(\mathbb{R}^d)$-valued kernel is available and propose a method to build a $\mathcal{L}(\mathrm{Im}(S))$-valued kernel on top of it. In this case, we can leverage the reconciliation matrix G. Specifically, given a $\mathcal{L}(\mathbb{R}^d)$-valued kernel $\widetilde{\mathcal{K}}$, define

$$\mathcal{K}_{SG}(x, x') \coloneqq SG\widetilde{\mathcal{K}}(x, x')G^\mathsf{T}S^\mathsf{T} \tag{6}$$

for all $(x, x') \in \mathcal{X}^2$, where $S \in \{0,1\}^{d \times d_b}$ is the aggregation matrix defined in Equation (1) and $G \in \mathbb{R}^{d_b \times d}$ is a fixed reconciliation matrix (see Section 2.1). One can verify that $\mathcal{K}_{SG}$ is a valid $\mathcal{L}(\mathrm{Im}(S))$-valued kernel. The base kernel $\widetilde{\mathcal{K}}$ determines how inputs and tasks interact before imposing coherence, whereas SG determines the geometry through which these relationships are transferred into the coherent output space. It should be noted that the base kernel $\widetilde{\mathcal{K}}$ will not be required to be separable. We give first an example of a non-separable base kernel for multivariate time series.

Example 2 (Non-separable Base OVK) *Suppose that the input space can be decomposed as $\mathcal{X} = (\mathcal{X}_0)^d$. Let $x = (x_1, \ldots, x_d) \in \mathcal{X}$, where $x_r \in \mathcal{X}_0$ contains the covariates of series $r \in \{1, \ldots, d\}$. Let $B \in \mathbb{R}^{d \times q}$ be an arbitrary task embedding with nonzero rows normalized to unit Euclidean norm. For example, $B_{r,:}$ could be the one-hot encoding that represents the hierarchical level of the series r or other non-hierarchical labels on the series (see, e.g., Section 5.2). Then we define*

$$A_\rho = (1-\rho) I_d + \rho B B^\mathsf{T}, \qquad \text{with } \rho \in [0,1]. \tag{7}$$

Let $\lambda_1, \ldots, \lambda_d > 0$. Given two inputs $x = (x_1, \ldots, x_d)$ and $x' = (x'_1, \ldots, x'_d)$, we define a non-separable base kernel $\widetilde{\mathcal{K}}_\rho : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathbb{R}^d)$ as follows: for any $r, s \in \{1, \ldots, d\}$,

$$\left[\widetilde{\mathcal{K}}_\rho(x, x')\right]_{rs} = \frac{(A_\rho)_{rs}}{\sqrt{\lambda_r \lambda_s}} k(x_r, x'_s), \tag{8}$$

where $k : \mathcal{X}_0 \times \mathcal{X}_0 \to \mathbb{R}$ is a scalar-valued kernel admitting an explicit finite-dimensional feature map $\phi : \mathcal{X}_0 \to \mathbb{R}^{p_\phi}$, i.e., $k(x_r, x'_s) = \langle \phi(x_r), \phi(x'_s) \rangle_{\mathbb{R}^{p_\phi}}$. Such a kernel allows for task-specific covariates to influence the cross-task relationships. Moreover, $\widetilde{\mathcal{K}}_\rho$ is a valid OVK, as it admits a feature operator Γ_ρ that can be explicitly characterized. More details can be found in Appendix B.6.

Now we give one example of such general coherent OVKs with the MinT reconciliation matrices.

Example 3 (MinT) *Let $W \in \mathbb{S}^d_{++}$ be an estimate of the covariance matrix of the base-forecast errors used in the MinT reconciliation (Wickramasuriya et al., 2019). The common MinT variants, e.g., OLS, WLS-variance and shrinkage, correspond to*

$$W_{\mathrm{OLS}} = I_d, \quad W_{\mathrm{WLS}} = \mathrm{diag}(\widehat{W}), \tag{9}$$

$$W_{\mathrm{shr}} = (1-\alpha)\widehat{W} + \alpha \mathrm{diag}(\widehat{W}), \tag{10}$$

where $\widehat{W}$ is the empirical covariance matrix of the base-forecast residuals and $\alpha \in (0,1]$ is the shrinkage intensity (see Appendix A.2 for more details). Then the MinT reconciliation matrix (Wickramasuriya et al., 2019) and the corresponding coherent projection are $G_W = (S^\mathsf{T} W^{-1} S)^{-1} S^\mathsf{T} W^{-1}$ and $P_W = S G_W$.

For any base OVK $\widetilde{\mathcal{K}}$, we define the associated MinT-composed OVK as

$$\mathcal{K}_W(x, x') = P_W \widetilde{\mathcal{K}}(x, x') P_W^\mathsf{T}. \tag{11}$$

For $W = I_d$, P_W is the orthogonal projector $P = S(S^\mathsf{T} S)^{-1} S^\mathsf{T}$. Thus, if $\widetilde{\mathcal{K}}(x, x') = k(x, x')\widetilde{A}$ is separable, Equation (11) reduces to $k(x, x') P \widetilde{A} P$ and coincides with Corollary 3.3.

Interestingly, the predictor obtained by choosing an OVK of this form coincides with that produced by an alternative learning algorithm previously studied in the kernel literature for other types of outputs (see Appendix A.4 for further details). Specifically, for any $\lambda > 0$, if $\hat{h}$ is the minimizer of the optimization problem:

$$\min_{h \in \mathcal{H}_{\widetilde{\mathcal{K}}}} \frac{1}{T} \sum_{t=1}^{T} \|SGh(x_t) - y_t\|^2_{\mathbb{R}^d} + \lambda \|h\|^2_{\mathcal{H}_{\widetilde{\mathcal{K}}}}, \tag{12}$$

where $\mathcal{H}_{\widetilde{\mathcal{K}}}$ is the RKHS associated to $\widetilde{\mathcal{K}}$, then $\hat{f} = SG \circ \hat{h}$ is the minimizer of the following one:

$$\min_{f \in \mathcal{H}_{\widetilde{\mathcal{K}}, SG}} \frac{1}{T} \sum_{t=1}^{T} \|f(x_t) - y_t\|^2_{\mathbb{R}^d} + \lambda \|f\|^2_{\mathcal{H}_{\widetilde{\mathcal{K}}, SG}}, \tag{13}$$

where $\mathcal{H}_{\widetilde{\mathcal{K}}, SG}$ is the RKHS associated to $\mathcal{K}_{SG}$ as defined in Equation (6). The proof of the observation is provided in Appendix B.7. It also allows us to reinterpret the WeaKL estimators introduced by Doumèche et al. (2025) into our framework, by viewing them as instances where the OVKs are special cases of the construction in Equation (6). Additional discussion on this connection is provided in Appendix B.8.

4. Statistical Analysis

4.1. Statistical Setting

In this section, from now on, we consider hierarchical time series forecasting as a possibly nonlinear regression problem. Let $(\mathsf{X}, \mathsf{Y}_b, \mathsf{Y})$ be three random variables, valued in $\mathcal{X}$, $\mathcal{Y}_b$ and $\mathcal{Y}$ respectively. In the hierarchical forecasting setting, Y_b denotes the d_b bottom-level time-series observations, Y denotes the d time-series observations across all levels, and X denotes the associated covariates. We assume that $\mathcal{X}$ is a Polish space, $\mathcal{Y}_b$ and $\mathcal{Y}$ are two real separable Hilbert spaces. We consider that the following relationships exist between the variables:

$$\mathsf{Y}_b = g^\star(\mathsf{X}) + \xi, \quad \mathsf{Y} = S\mathsf{Y}_b, \tag{14}$$

where $g^\star : \mathcal{X} \to \mathcal{Y}_b$ is an unknown target function and $S \in \mathcal{L}(\mathcal{Y}_b, \mathcal{Y})$ is a known bounded linear operator. Here, ξ is a $\mathcal{Y}_b$-valued random variable modeling the noise. We assume that $\mathrm{Im}(S)$ is closed, so that it is a valid Hilbert space with the inner product inherited from $\mathcal{Y}$.

Let $\mathbb{P}$ be the joint probability distribution of (X, Y), and ρ be its marginal distribution on X. Given T observed data points $((x_1, y_1), \ldots, (x_T, y_T)) \in (\mathcal{X} \times \mathcal{Y})^T$ independently distributed by $\mathbb{P}$, the goal is to obtain an estimator $\hat{f} : \mathcal{X} \to \mathrm{Im}(S)$ of $S \circ g^\star$ whose excess risk can be controlled, defined as follows:

$$\mathcal{R}(\hat{f}) - \mathcal{R}(S \circ g^\star) = \mathbb{E}\Big[\big\|\hat{f}(\mathsf{X}) - \mathsf{Y}\big\|_{\mathcal{Y}}^2\Big] - \mathbb{E}\Big[\big\|S \circ g^\star(\mathsf{X}) - \mathsf{Y}\big\|_{\mathcal{Y}}^2\Big]. \quad (15)$$

It should be noted that the statistical analysis in this section assumes i.i.d. input–output pairs. Extending the excess-risk bound to temporally dependent observations is left for future work.

Let $L^2(\mathcal{X}, \rho, \mathcal{Y})$ be the Hilbert space of measurable function $f : \mathcal{X} \to \mathcal{Y}$ that are square-integrable with respect to the probability measure ρ, i.e., $\int_{\mathcal{X}} \|f(x)\|_{\mathcal{Y}}^2 \, d\rho(x) < \infty$. Such a space is endowed with the inner product defined by $\langle f_1, f_2 \rangle_\rho = \int_{\mathcal{X}} \langle f_1(x), f_2(x) \rangle_{\mathcal{Y}} \, d\rho(x)$ and the associated norm $\|\cdot\|_\rho$ induced by the inner product.

We make the following assumptions.

Assumption 1 $\|g^\star(\mathsf{X})\|_{\mathcal{Y}_b}^2 \le C_1 < \infty$ *almost surely and* $\|\xi\|_{\mathcal{Y}_b}^2 \le C_2 < \infty$ *almost surely.*

Assumption 2 $\|\mathcal{K}_x\|_2^2 = \mathrm{tr}\,(\mathcal{K}_x^* \mathcal{K}_x) \le \kappa < \infty$ *for all* $x \in \mathcal{X}$.

Assumption 3 *The random variables* X *and* ξ *are independent and* $\mathbb{E}[\xi] = 0$.

Assumption 4 *There exists an* $h^\star \in L^2(\mathcal{X}, \rho, \mathcal{Y})$ *such that, for some* $0 < r \le 1$, $S \circ g^\star = L_{\mathcal{K}}^r h^\star$ *where the integral operator* $L_{\mathcal{K}} : L^2(\mathcal{X}, \rho, \mathcal{Y}) \to L^2(\mathcal{X}, \rho, \mathcal{Y})$ *is defined by* $L_{\mathcal{K}} h := \int_{\mathcal{X}} \mathcal{K}_x h(x) d\rho(x)$ *for any* $h \in L^2(\mathcal{X}, \rho, \mathcal{Y})$.

The source condition in the last assumption is standard in the kernel literature and controls the smoothness of the target function $S \circ g^\star$ relative to the kernel integral operator $L_{\mathcal{K}}$.

4.2. Main Results

We now present a compact excess-risk bound for the RVVKRR estimator, whose proof is deferred to Appendix C. The proof is obtained using the technique of integral operators in RKHS, particularly introduced by Smale and Zhou (2007); Caponnetto and De Vito (2007). To state the result, define the covariance operator and its effective dimension by $\Sigma_{\mathcal{K}} := \int_{\mathcal{X}} \mathcal{K}_x \mathcal{K}_x^* \, d\rho(x)$ and $\mathcal{N}(\lambda) := \mathrm{tr}\left((\Sigma_{\mathcal{K}} + \lambda I)^{-1} \Sigma_{\mathcal{K}}\right)$.

Proposition 4.1 (Excess-risk Bound) *Under Assumptions 1 to 4, let* $\delta \in (0, 1)$ *and suppose that* $\lambda \ge 8 \log\left(\frac{4}{\delta}\right) \frac{\kappa}{\sqrt{T}}$. *Then, with probability at least* $1 - \delta$,

$$\mathcal{R}(\hat{f}) - \mathcal{R}(S \circ g^\star) \le \frac{128}{T} \log^2\left(\frac{4}{\delta}\right) \left(\mathcal{N}(\lambda)\sigma_S^2 + \frac{\kappa}{\lambda}(3B_S^2 + M_S^2)\right) + 2\lambda^{2r} R_r^2,$$

with $\sigma_S^2 := \mathbb{E}\left[\|S\xi\|_{\mathcal{Y}}^2\right]$, $B_S^2 := \|S\|_{\mathrm{op}}^2 C_1$, $M_S^2 := \|S\|_{\mathrm{op}}^2 C_2$ *and* $R_r := \left\|L_{\mathcal{K}}^{-r}(S \circ g^\star)\right\|_\rho$.

The first term in Proposition 4.1 is the estimation error, controlled by the effective dimension and the transformed noise level. The second term is the approximation error induced by regularization and the source condition.

Remark 4.2 *In the HTS setting, we have* $\mathcal{Y} = \mathrm{Im}(S) \subset \mathbb{R}^d$. *Denote* $\Sigma_\xi = \mathbb{E}\left[\xi\xi^\top\right] \in \mathbb{R}^{d_b \times d_b}$. *Then by the definition of* σ_S^2, *we have*

$$\sigma_S^2 \le \|S\|_{\mathrm{op}}^2 \, \mathbb{E}\left[\|\xi\|_2^2\right] = \|S\|_{\mathrm{op}}^2 \, \mathrm{tr}\,(\Sigma_\xi) \le \|S\|_{\mathrm{F}}^2 \, \mathrm{tr}\,(\Sigma_\xi).$$

Thus, the estimation error in Proposition 4.1 depends on both the total bottom-level noise level $\mathrm{tr}\,(\Sigma_\xi)$ *and the aggregation structure. For a binary summing matrix,* $\|S\|_{\mathrm{F}}^2$ *counts the total number of bottom-series memberships across the hierarchy. Indeed,* $\|S\|_{\mathrm{F}}^2 = \sum_{i=1}^{d} \sum_{j=1}^{d_b} S_{ij}^2 = \sum_{i=1}^{d} \sum_{j=1}^{d_b} S_{ij}$, *and* $S_{ij} = 1$ *exactly when the bottom-level series* j *contributes to series* i. *Consequently, a deeper or more densely aggregated hierarchy can increase this upper bound on the propagated noise.*

5. Numerical Experiments

In this section, we present an empirical study of the proposed predictors on both synthetic and real-world datasets. For the two-stage baseline models, the reconciliation methods are implemented using the *HierarchicalForecast* library[1] (Olivares et al., 2022). For evaluation, we report the mean squared error (MSE) computed at each hierarchical level as well as across all levels.

5.1. Synthetic Dataset

Generation Process. We generate a three-level hierarchy with $d = 8$ series: one top-level, two middle-level, and

[1] https://pypi.org/project/hierarchicalforecast/

$d_b = 5$ bottom-level series. Each bottom-level series follows an additive trend–seasonality model with correlated Gaussian noise; series sharing a parent have the same trend and seasonal components. The remaining series are obtained by aggregation with S, ensuring coherence by construction. A scalar $\alpha \geq 0$ controls the noise level. Exact parameter values, the covariance construction, and an example realization are provided in Appendix D.1.

Experimental Setting. We investigate how model performance depends on the training length T, with the covariance scaling parameter fixed at $\alpha = 0.6$. We take $T \in \{40, 80, 160, 320\}$ and run 32 Monte Carlo simulations for each value. In each simulation, we generate hierarchical time series of length $T + 200$: the first T timestamps are used for training, the next 100 for validation and hyperparameter selection, and the last 100 for testing. For the input covariates, we use $x_t = (t, \sin(2\pi t/7), \sin(2\pi t/30)) \in \mathbb{R}^3$ for all series in the hierarchy. We compare KRR followed by reconciliation with our RVVKRR predictor. For KRR, the ridge parameter λ is cross-validated separately for each series, and we consider some reconciliation strategies described in Appendix A.2. For RVVKRR, we use the two separable OVKs introduced in Example 1: the identity kernel (RVVKRR-I) and the graph-based kernel (RVVKRR-Lap). The graph connects bottom-level series sharing the same parent, and the corresponding edge weights are selected by validation. The node-wise diagonal weights are set to the KRR ridge parameters.

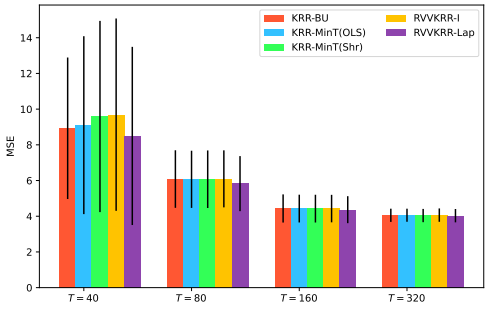


Figure 1. All-level MSE on the synthetic dataset as a function of the training length T, with $\alpha = 0.6$.

Experimental Results. Figure 1 illustrates the impact of the training length T on the MSE across all levels. Employing a Laplacian OVK yields better results than using an identity kernel. Moreover, RVVKRR consistently outperforms the two-stage methods, particularly in low-data regimes ($T = 40$ or $T = 80$). Additional results on the effect of the covariance scaling parameter α are reported in Appendix D.1.

5.2. TourismLarge Dataset

Dataset. We next consider the Australian domestic TourismLarge dataset (Wickramasuriya et al., 2019), which contains 228 monthly observations from January 1998 to December 2016. The hierarchy comprises $d = 415$ series: one country, 7 states, 27 zones, 76 regions, and 304 bottom-level region–purpose series. The four bottom-level purposes are holiday, visiting friends and relatives (VFR), business, and other.

Experimental Setting. We use the preceding 24 observations of each series as its covariates. After constructing these lags, 204 observations remain; the last 24 form the test set and the first 180 form the training set. Hyperparameters are selected by three-fold expanding-window cross-validation. Independent OLS and KRR each use only the 24 lags of the series being predicted. KRR uses a linear input kernel. For both independent estimators, we report unreconciled forecasts and forecasts reconciled using the reconciliation methods described in Appendix A.2. RVVKRR receives the concatenation of all series-specific lag blocks, hence $x_t \in \mathbb{R}^{24d}$. The OVK used by RVVKRR combines the non-separable base kernel of Example 2 with the MinT (shrinkage) projection of Example 3. The scalar-valued kernel in Equation (8) is the linear kernel. We take $B = B_{\text{pur}} \in \mathbb{R}^{d\times 5}$, whose r-th row is the one-hot encoding of the membership of series r in either the aggregate category or one of the four travel-purpose categories. These relationships are not explicitly encoded by the aggregation matrix. The parameters λ_r are inherited from the independent KRR. Cross-validation selects $\rho = 0.1$.

Table 1. Test MSE for KRR and RVVKRR at each aggregation level of TourismLarge, reported in millions (10^6). All sums the level-wise MSEs. † denotes unreconciled forecasts. Lower is better.

Method	Ctry.	State	Zone	Reg.	Purp.	All
KRR base forecasts						
Unreconciled †	2.71	1.69	1.30	1.19	1.13	8.02
Bottom-Up	8.06	2.53	1.53	1.21	1.13	14.46
Top-Down (AHP)	2.71	3.94	2.38	2.05	1.92	13.00
Top-Down (PHA)	2.71	3.89	2.37	2.04	1.92	12.94
MinT (OLS)	2.72	1.58	1.25	1.12	1.11	7.78
MinT (shrinkage)	3.05	1.60	**1.21**	**1.08**	**1.06**	7.99
MinT (WLS-var)	4.56	1.83	1.29	1.11	1.09	9.89
ERM (closed form)	5.36	2.50	2.61	2.57	2.52	15.57
RVVKRR	**2.02**	**1.52**	1.30	1.24	1.57	**7.64**

Experimental Results. Table 1 reports the test MSE at each aggregation level; the full comparison, including OLS base forecasts and alternative RVVKRR projections, is given in Appendix D.2. Relative to the matched two-stage KRR with MinT (shrinkage) baseline, the RVVKRR predictor with the CV-selected purpose prior ($\rho = 0.1$) reduces

the all-level MSE from 7.99×10^6 to 7.64×10^6. The gains are concentrated at the country and state levels, with larger errors at the lower levels.

5.3. Traffic Dataset

Dataset. We finally consider the SF Bay Area daily Traffic dataset, originally preprocessed by Ben Taieb and Koo (2019) and distributed with the *HierarchicalForecast* benchmark (Olivares et al., 2022). It contains 366 daily observations from January 1 to December 31, 2008, for $d = 207$ series arranged in four levels: one total series, two Level 2 aggregates, four Level 3 aggregates, and $d_b = 200$ bottom-level series.

Experimental Setting. For each series r, the covariate vector $x_{t,r} \in \mathbb{R}^{23}$ contains 16 raw lag values at lags $1, \ldots, 14$, 21, and 28, six Fourier features corresponding to harmonics 1, 2, and 3 of the weekly period $P = 7$, and a linear trend. As for TourismLarge, independent KRR uses only $x_{t,r}$ to predict series r, whereas RVVKRR receives the concatenation $x_t = (x_{t,1}, \ldots, x_{t,d}) \in \mathbb{R}^{23d}$. No feature standardization is applied. Independent KRR uses a linear kernel, with a separate ridge parameter λ_r selected for each series by time-series cross-validation. For RVVKRR, we set $\rho = 0$ in the base OVK of Example 2 and compose this kernel with a MinT projection as in Example 3. The parameters λ_r are inherited from the fitted independent KRR and set the relative regularization scales. We consider three projections inside the kernel: MinT (OLS), MinT (shrinkage), and MinT (WLS-var). Within each training window, the covariance proxies for the latter two are estimated from in-sample residuals of auxiliary independent KRR predictors. For each RVVKRR variant, the global regularization parameter λ is selected by cross-validation. The corresponding KRR+MinT baselines apply the same OLS, shrinkage, or variance-weighted projection to independently fitted KRR forecasts post hoc. The full comparison with additional models and further experimental details are provided in Appendix D.3.

Experimental Results. Table 2 shows that post-hoc MinT reconciliation does not necessarily improve independent KRR: OLS and shrinkage yield only marginal changes, while WLS-var degrades performance. In contrast, placing the shrinkage or WLS-var projection inside the joint RVVKRR fit substantially reduces the all-level MSE. This benefit is not due to coherence alone, since RVVKRR-MinT (OLS) remains close to independent KRR. Unlike post-hoc reconciliation, which only transforms already fitted forecasts, an internal projection changes the RKHS geometry and allows the series-specific feature blocks to be estimated jointly. The resulting gains occur mainly at the aggregate levels, at the cost of some bottom-level accuracy under the stated all-level MSE objective.

Table 2. Test MSE at each aggregation level of Traffic, reported in thousands (10^3). All sums the level-wise MSEs. † denotes incoherent forecasts. Lower is better.

Method	L1	L2	L3	Bottom	All
KRR base forecasts					
Unreconciled †	8.95	5.05	2.84	0.63	17.47
MinT (OLS)	9.01	4.95	2.85	0.61	17.42
MinT (shrinkage)	9.16	4.83	2.69	**0.60**	17.28
MinT (WLS-var)	11.68	6.11	3.34	0.62	21.76
RVVKRR forecasts					
MinT (OLS)	8.99	4.93	2.79	0.61	17.32
MinT (shrinkage)	3.84	2.14	2.20	3.39	11.57
MinT (WLS-var)	**0.32**	**0.24**	**0.44**	3.02	**4.03**

6. Conclusion

We introduced RVVKRR, an end-to-end kernel framework for hierarchical time series forecasting that enforces coherence directly in the hypothesis space. Experiments on synthetic and real datasets show that embedding an appropriate hierarchical geometry within joint estimation can outperform post-hoc reconciliation, while coherence alone does not guarantee an improvement. Future work should learn the coherent kernel geometry from data, accommodate level-specific forecasting priorities, scale to larger hierarchies, and extend the analysis to dependent time-series observations.

References

Athanasopoulos, G., Ahmed, R. A., and Hyndman, R. J. (2009). Hierarchical Forecasts for Australian Domestic Tourism. *International Journal of Forecasting*, 25(1):146–166.

Ben Taieb, S. and Koo, B. (2019). Regularized Regression for Hierarchical Forecasting Without Unbiasedness Conditions. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pages 1337–1347, New York, NY, USA. Association for Computing Machinery.

Bonilla, E. V., Chai, K., and Williams, C. (2007). Multi-task Gaussian Process Prediction. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems*, volume 20, pages 153–160. Curran Associates, Inc.

Bouche, D. (2023). *Function-valued regression with kernels : Improving speed, flexibility and robustness*. Theses, Institut Polytechnique de Paris.

Bouche, D., Clausel, M., Roueff, F., and d'Alché Buc, F. (2021). Nonlinear Functional Output Regression: A Dictionary Approach. In Banerjee, A. and Fukumizu, K., editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 235–243. PMLR.

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control*. John Wiley & Sons.

Brouard, C., d'Alché Buc, F., and Szafranski, M. (2011). Semi-supervised Penalized Output Kernel Regression for Link Prediction. In Getoor, L. and Scheffer, T., editors, *Proceedings of the 28th International Conference on Machine Learning*, pages 593–600, Bellevue, Washington, USA. Omnipress.

Brouard, C., Shen, H., Dührkop, K., d'Alché Buc, F., Böcker, S., and Rousu, J. (2016). Fast Metabolite Identification with Input Output Kernel Regression. *Bioinformatics (Oxford, England)*, 32(12):i28–i36.

Brown, R. G. (1959). *Statistical Forecasting for Inventory Control*. McGraw-Hill.

Caponnetto, A. and De Vito, E. (2007). Optimal Rates for the Regularized Least-Squares Algorithm. *Foundations of Computational Mathematics*, 7(3):331–368.

Caponnetto, A., Micchelli, C. A., Pontil, M., and Ying, Y. (2008). Universal Multi-Task Kernels. *Journal of Machine Learning Research*, 9(52):1615–1646.

Carmeli, C., De Vito, E., and Toigo, A. (2006). Vector Valued Reproducing Kernel Hilbert Spaces of Integrable Functions and Mercer Theorem. *Analysis and Applications*, 4(04):377–408.

Carmeli, C., De Vito, E., Toigo, A., and Umanitá, V. (2010). Vector Valued Reproducing Kernel Hilbert Spaces and Universality. *Analysis and Applications*, 8(01):19–61.

Cini, A., Mandic, D., and Alippi, C. (2024). Graph-based Time Series Clustering for End-to-End Hierarchical Forecasting. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F., editors, *Proceedings of the 41st International Conference on Machine Learning*, pages 8985–8999. PMLR.

Cutajar, K., Osborne, M., Cunningham, J., and Filippone, M. (2016). Preconditioning Kernel Matrices. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 2529–2538, New York, New York, USA. PMLR.

Doumèche, N., Bach, F. R., Bedek, E., Biau, G., Boyer, C., and Goude, Y. (2025). Forecasting Time Series with Constraints. *CoRR*, abs/2502.10485.

Evgeniou, T., Micchelli, C. A., and Pontil, M. (2005). Learning Multiple Tasks with Kernel Methods. *Journal of Machine Learning Research*, 6(21):615–637.

Evgeniou, T., Pontil, M., and Poggio, T. (2000). Regularization Networks and Support Vector Machines. *Advances in Computational Mathematics*, 13(1):1–50.

Golub, G., Nash, S., and Van Loan, C. (1979). A Hessenberg-Schur method for the problem AX + XB= C. *IEEE Transactions on Automatic Control*, 24(6):909–913.

Gross, C. W. and Sohl, J. E. (1990). Disaggregation Methods to Expedite Product Line Forecasting. *Journal of Forecasting*, 9(3):233–254.

Hestenes, M. R. and Stiefel, E. (1952). Methods of Conjugate Gradients for Solving Linear Systems. *Journal of Research of the National Bureau of Standards*, 49(6):409–436.

Higham, N. J. (1988). Computing a Nearest Symmetric Positive Semidefinite Matrix. *Linear Algebra and its Applications*, 103:103–118.

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.

Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*. OTexts.

Kadri, H., Duflos, E., Preux, P., Canu, S., Rakotomamonjy, A., and Audiffren, J. (2016). Operator-valued Kernels for Learning from Functional Response Data. *Journal of Machine Learning Research*, 17(20):1–54.

Kadri, H., Ghavamzadeh, M., and Preux, P. (2013). A Generalized Kernel Approach to Structured Output Learning. In Dasgupta, S. and McAllester, D., editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 471–479, Atlanta, Georgia, USA. PMLR. Issue: 1.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2022). The M5 Competition: Background, Organization, and Implementation. *International Journal of Forecasting*, 38(4):1325–1336.

Mancuso, P., Piccialli, V., and Sudoso, A. M. (2021). A Machine Learning Approach for Forecasting Hierarchical Time Series. *Expert Systems with Applications*, 182:115102.

Micchelli, C. and Pontil, M. (2004). Kernels for Multi–task Learning. In Saul, L., Weiss, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems*, volume 17. MIT Press.

Micchelli, C. A. and Pontil, M. (2005). On Learning Vector-Valued Functions. *Neural Computation*, 17(1):177–204.

Minh, H. Q., Bazzani, L., and Murino, V. (2016). A Unifying Framework in Vector-valued Reproducing Kernel Hilbert Spaces for Manifold Regularization and Co-Regularized Multi-view Learning. *Journal of Machine Learning Research*, 17(25):1–72.

Nespoli, L., Medici, V., Lopatichki, K., and Sossan, F. (2020). Hierarchical Demand Forecasting Benchmark for the Distribution Grid. *Electric Power Systems Research*, 189:106755.

Olivares, K. G., Garza, F., Luo, D., Challu, C., Mergenthaler, M., and Dubrawski, A. (2022). HierarchicalForecast: A Reference Framework for Hierarchical Forecasting in Python. *CoRR*, abs/2207.03517.

Pedrick, G. (1957). *Theory of Reproducing Kernels for Hilbert Spaces of Vector Valued Functions*. PhD Thesis, University of Kansas.

Peypouquet, J. (2015). *Convex Optimization in Normed Spaces: Theory, Methods and Examples*. Springer Cham.

Pinelis, I. (1994). Optimum Bounds for the Distributions of Martingales in Banach Spaces. *The Annals of Probability*, 22(4).

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 Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8832–8843. PMLR.

Rudi, A., Carratino, L., and Rosasco, L. (2017). FALKON: An Optimal Large Scale Kernel Method. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, pages 3888–3898. Curran Associates, Inc.

Salinas, D., Bohlke-Schneider, M., Callot, L., Medico, R., and Gasthaus, J. (2019). High-Dimensional Multivariate Forecasting with Low-Rank Gaussian Copula Processes. In Wallach, H., Larochelle, H., Beygelzimer, A., dAlché Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32, pages 6827–6837. Curran Associates, Inc.

Saunders, C., Gammerman, A., and Vovk, V. (1998). Ridge Regression Learning Algorithm in Dual Variables. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 515–521, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Schwartz, L. (1964). Sous-espaces hilbertiens d'espaces vectoriels topologiques et noyaux associés (Noyaux reproduisants). *Journal d'Analyse Mathématique*, 13(1):115–256.

Sheldon, D. (2008). Graphical Multi-Task Learning. In *NeurIPS Workshop on Structured Input and Structured Output*.

Sima, V. (1996). *Algorithms for Linear-Quadratic Optimization*. CRC Press LLC, Milton.

Simoncini, V. (2016). Computational Methods for Linear Matrix Equations. *SIAM Review*, 58(3):377–441.

Sindhwani, V., Minh, H. Q., and Lozano, A. C. (2013). Scalable Matrix-Valued Kernel Learning for High-Dimensional Nonlinear Multivariate Regression and Granger Causality. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence*, pages 586–595, Bellevue, WA. AUAI Press.

Smale, S. and Zhou, D.-X. (2005). Shannon Sampling II: Connections to Learning Theory. *Computational Harmonic Analysis - Part 1*, 19(3):285–302.

Smale, S. and Zhou, D.-X. (2007). Learning Theory Estimates via Integral Operators and Their Approximations. *Constructive Approximation*, 26(2):153–172.

Tsiourvas, A., Sun, W., Perakis, G., Chen, P.-Y., and Zhu, Y. (2024). Learning Optimal Projection for Forecast Reconciliation of Hierarchical Time Series. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F., editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 48713–48727. PMLR.

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.

Widmer, C., Leiva, J., Altun, Y., and Rätsch, G. (2010). Leveraging Sequence Classification by Taxonomy-Based Multitask Learning. In *Research in Computational Molecular Biology*, pages 522–534, Berlin, Heidelberg.

Wu, H., Xu, J., Wang, J., and Long, M. (2021). Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 22419–22430. Curran Associates, Inc.

Zhou, F., Pan, C., Ma, L., Liu, Y., Wang, S., Zhang, J., Zhu, X., Hu, X., Hu, Y., Zheng, Y., Lei, L., and Yun, H. (2023). SLOTH: Structured Learning and Task-Based Optimization for Time Series Forecasting on Hierarchies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11417–11425.

Álvarez, M. A., Rosasco, L., and Lawrence, N. D. (2012). Kernels for Vector-Valued Functions: A Review. *Foundations and Trends® in Machine Learning*, 4(3):195–266.

A. Additional Preliminaries

A.1. Hierarchical Time Series Example

Example 4 gives an illustration of such a hierarchical time series.

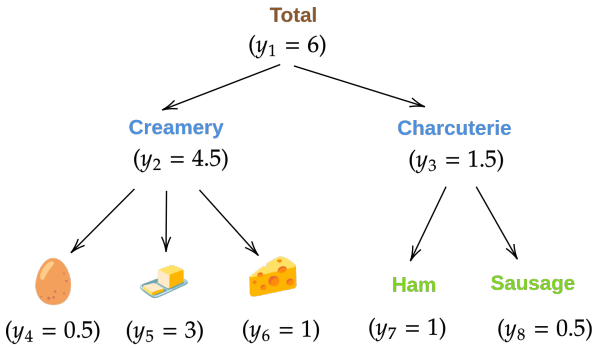


Figure 2. A three level hierarchical time series example for the daily raw material demand of a crêperie.

Example 4 *Imagine that we have 5 univariate time series representing the daily raw material demand of a crêperie, for example: eggs, butter, cheese, sausage, and ham. The demand for these products can be aggregated to form two new univariate time series representing, respectively, the charcuterie and creamery categories. These two time series can be further aggregated to obtain the total demand. The hierarchical structure is shown in Figure 2. In this example, the summation matrix is defined as follows:*

$$S_{\text{sum}} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}. \tag{16}$$

A.2. Reconciliation Strategies

As explained in Section 2.1, a linear reconciliation method transforms a base forecast $\tilde{y}_t \in \mathbb{R}^d$ into the coherent forecast

$$\hat{y}_t = SG\tilde{y}_t \in \text{Im}(S),$$

where $G \in \mathbb{R}^{d_b \times d}$ maps forecasts from all levels of the hierarchy to the bottom level. The main classical strategies differ in how they choose G.

Bottom-up reconciliation. The bottom-up approach retains the bottom-level base forecasts and discards the forecasts at the aggregate levels. Its reconciliation matrix is

$$G_{\text{BU}} = \begin{bmatrix} 0_{d_b \times d_a} & I_{d_b} \end{bmatrix}.$$

Thus, $G_{\text{BU}}\tilde{y}_t$ selects the bottom-level forecasts, which are then aggregated by S to recover forecasts at every level of the hierarchy.

Top-down reconciliation. The top-down approach starts from the base forecast of the top-level series and disaggregates it across the bottom-level series (Gross and Sohl, 1990). It uses

$$G_{\text{TD}} = \begin{bmatrix} p & 0_{d_b \times (d-1)} \end{bmatrix}, \qquad p \in \mathbb{R}^{d_b},$$

where p contains the disaggregation proportions. Let $y_t^{\text{top}} \in \mathbb{R}$ denote the observed top-level series and let $y_{t,i}^b$ denote the i-th bottom-level series, for $t \in \{1, \dots, T\}$ and $i \in \{1, \dots, d_b\}$. The *average historical proportions* (AHP) method averages the share of each bottom-level series at every training time:

$$p_i^{\text{AHP}} = \frac{1}{T} \sum_{t=1}^{T} \frac{y_{t,i}^b}{y_t^{\text{top}}}, \qquad i \in \{1, \dots, d_b\}.$$

By contrast, the *proportions of the historical averages* (PHA) method computes the share from the historical mean of each series:

$$p_i^{\mathrm{PHA}} = \frac{T^{-1}\sum_{t=1}^{T} y_{t,i}^{b}}{T^{-1}\sum_{t=1}^{T} y_t^{\mathrm{top}}} = \frac{\sum_{t=1}^{T} y_{t,i}^{b}}{\sum_{t=1}^{T} y_t^{\mathrm{top}}}, \qquad i \in \{1, \dots, d_b\}.$$

For coherent training data, both choices satisfy $\sum_{i=1}^{d_b} p_i = 1$. The vector $p = (p_1, \dots, p_{d_b})^{\mathsf{T}}$ obtained with either method is then used in G_{TD} above.

MinT reconciliation. Let $W \in \mathbb{S}_{++}^{d}$ be a positive-definite covariance proxy for the base-forecast errors. The minimum trace (MinT) reconciliation matrix and its associated coherent projection are (Wickramasuriya et al., 2019)

$$G_W = (S^{\mathsf{T}} W^{-1} S)^{-1} S^{\mathsf{T}} W^{-1}, \qquad P_W = S G_W.$$

Ideally, W is the covariance matrix $\mathrm{cov}(y_t - \tilde{y}_t)$, but this quantity is unknown and must be estimated. Let

$$\widehat{W} = \frac{1}{T}\sum_{t=1}^{T} (y_t - \tilde{y}_t)(y_t - \tilde{y}_t)^{\mathsf{T}}$$

denote the empirical covariance of the base-forecast errors over the T training observations. The variants used in our experiments correspond to the following choices of the covariance proxy.

OLS. The OLS reconciliation of Hyndman et al. (2011) is the special case of MinT obtained with

$$W_{\mathrm{OLS}} = I_d.$$

Indeed,

$$G_{W_{\mathrm{OLS}}} = (S^{\mathsf{T}} S)^{-1} S^{\mathsf{T}}, \qquad P_{W_{\mathrm{OLS}}} = S(S^{\mathsf{T}} S)^{-1} S^{\mathsf{T}},$$

so $P_{W_{\mathrm{OLS}}}$ is the orthogonal projector onto $\mathrm{Im}(S)$.

WLS-variance. The weighted least squares variance variant retains only the marginal base-forecast error variances:

$$W_{\mathrm{WLS}} = \mathrm{diag}(\widehat{W}) = \mathrm{diag}(\widehat{\sigma}_1^2, \dots, \widehat{\sigma}_d^2),$$

where $\widehat{\sigma}_r^2$ is the empirical error variance of series r. This diagonal proxy ignores cross-series error correlations.

Shrinkage. The shrinkage variant combines the full empirical covariance matrix with its diagonal counterpart:

$$W_{\mathrm{shr}} = (1-\alpha)\widehat{W} + \alpha \mathrm{diag}(\widehat{W}), \qquad \alpha \in (0, 1].$$

It therefore preserves estimated cross-series error correlations while shrinking them toward zero. For a comprehensive review of classical reconciliation strategies, see Hyndman and Athanasopoulos (2018) and the references therein.

A.3. Separable Operator-Valued Kernels

Separable OVKs provides a clear decoupling between the contribution of input and output, where k measures input similarity and A encodes the relationships between output components, which makes them highly interpretable compared to general OVKs. Moreover, they simplify massively the computation burden for some kernel-based learning algorithms such as multi-task kernel regression (Micchelli and Pontil, 2004) or multi-task Gaussian processes (Bonilla et al., 2007). See, e.g., Álvarez et al. (2012, Section 6) for the discussion on the computation consideration.

A.4. Operator-Valued Kernels in Supervised Learning

In multi-task learning, for instance, it is natural to take the output space to be $\mathbb{R}^d$ where d denotes the number of related tasks whose joint learning may yield mutual benefits (Micchelli and Pontil, 2004; Evgeniou et al., 2005). More recently, several works have focused on predicting functional data by setting $\mathcal{Y} = L^2(\Omega)$, the set of square integrable functions on a given compact set $\Omega \subset \mathbb{R}^d$ (Kadri et al., 2016; Bouche et al., 2021; Bouche, 2023). In the context of structured prediction, complex output objects can be embedded into an RKHS $\mathcal{H}_k$ using a scalar-valued kernel $k : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, so that the

learning problem can be performed by finding a function in an $\mathcal{H}_k$-valued RKHS (Brouard et al., 2011; Kadri et al., 2013; Brouard et al., 2016).

Moreover, learning functions of the form $f = P \circ h$ where $P : \mathcal{W} \to \mathcal{Y}$ is a bounded linear operator and $h : \mathcal{X} \to \mathcal{W}$ is a function from a $\mathcal{W}$-valued RKHS, is not totally new in the kernel literature. In vector-valued multi-view learning, Minh et al. (2016) consider P as the combination operator that transforms the output from the m view in $\mathcal{W} = \mathcal{Y}^m$ into a single output in $\mathcal{Y}$. In Bouche et al. (2021), the operator P is a projection operator that transforms the expansion coefficient $u \in \mathcal{W} = \mathbb{R}^d$ to a function $\sum_{l=1}^{d} u_l \phi_l \in \mathcal{Y} = L^2(\Omega)$ where $(\phi_l)_{l=1}^{d} \in L^2(\Omega)^d$ is a dictionary of functions.

B. Additional Details for Section 3

B.1. Dimension of the Coherent Forecast Space

Lemma B.1 *Let S be the aggregation matrix defined in Equation 1. Then $\dim(\mathrm{Im}(S)) = d_b$, where d_b is the number of bottom-level series.*

Proof Suppose that we have $Sv = 0$ for some $v \in \mathbb{R}^{d_b}$. By definition of the matrix S, we have

$$\begin{bmatrix} S_{\mathrm{sum}} \\ I_{d_b} \end{bmatrix} v = \begin{bmatrix} S_{\mathrm{sum}} v \\ v \end{bmatrix} = 0.$$

The lower block gives $v = 0$. So the only solution of $Sv = 0$ is $v = 0$. Thus, we have $\mathrm{Ker}(S) = \{0\}$ and $\dim(\mathrm{Ker}(S)) = 0$. By the rank–nullity theorem, we can conclude that $\dim(\mathrm{Im}(S)) = \dim(\mathbb{R}^{d_b}) - \dim(\mathrm{Ker}(S)) = d_b$. ■

B.2. Proof of Lemma 3.1

Proof Let $A \in \mathcal{L}(\mathrm{Im}(S))$ be self-adjoint and positive, and set $P \coloneqq U_S U_S^{\mathsf{T}}$, the orthogonal projection onto $\mathrm{Im}(S)$. Define

$$A' \coloneqq U_S^{\mathsf{T}} A U_S \in \mathbb{R}^{d_b \times d_b}.$$

This composition is well-defined since U_S maps $\mathbb{R}^{d_b}$ into $\mathrm{Im}(S)$. For every $y \in \mathrm{Im}(S)$, Ay belongs to $\mathrm{Im}(S)$, so $Py = y$ and $PAy = Ay$. Consequently,

$$U_S A' U_S^{\mathsf{T}} y = PAPy = Ay, \qquad \forall y \in \mathrm{Im}(S).$$

Thus, $U_S A' U_S^{\mathsf{T}}$ is a matrix representation of A.

For any $v, w \in \mathbb{R}^{d_b}$, self-adjointness of A gives

$$\langle A'v, w \rangle_{\mathbb{R}^{d_b}} = \langle AU_S v, U_S w \rangle_{\mathbb{R}^d} = \langle U_S v, AU_S w \rangle_{\mathbb{R}^d} = \langle v, A'w \rangle_{\mathbb{R}^{d_b}}.$$

Hence A' is symmetric. Positivity of A also yields

$$\langle A'v, v \rangle_{\mathbb{R}^{d_b}} = \langle AU_S v, U_S v \rangle_{\mathbb{R}^d} \geq 0, \qquad \forall v \in \mathbb{R}^{d_b},$$

so A' is positive semi-definite. ■

B.3. Diagonal-based Separable OVKs

A specific choice of the separable OVK is to consider a diagonal matrix as A', i.e., $\mathcal{K}(x, x') = k(x, x') U_S A' U_S^{\mathsf{T}}$ with

$$A' = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_{d_b}) \tag{17}$$

where $\lambda_i > 0$ for all $i \in \{1, \ldots, d_b\}$. We will show here actually this predictor corresponds to the one of a learning algorithm based on d_b independent predictors of kernel ridge regression (KRR, Saunders et al., 1998; Evgeniou et al., 2000).

When A' is defined as in Equation (17), solving Equation (52) equals to solving the following linear system

$$\frac{T\lambda}{\lambda_i} e_i + K e_i = \left(Y^\mathsf{T} U_S\right)_{:,i} \quad \text{with } e_i \in \mathbb{R}^T \tag{18}$$

for all $i \in \{1, \ldots, d_b\}$, and then defining E such that $E_{i,:} = \frac{e_i^\mathsf{T}}{\lambda_i}$. We observe that Equation (18) actually corresponds to the linear system arising in KRR (see, e.g., Evgeniou et al., 2000, Equation (4.3)), where the regularization parameter is $\frac{\lambda}{\lambda_i}$ and the training label corresponds to the i-th component of the transformed output $y' = U_S^\mathsf{T} y \in \mathbb{R}^{d_b}$.

For a new input $x \in \mathcal{X}$, let $k_x = (k(x, x_1), \ldots, k(x, x_T))^\mathsf{T}$ and denote the corresponding scalar KRR predictions by $\hat{g}_i(x) = e_i^\mathsf{T} k_x$. Since the i-th row of $A'E$ is e_i^T, the RVVKRR prediction is

$$\hat{f}(x) = U_S A' E k_x = U_S \begin{bmatrix} \hat{g}_1(x) \\ \vdots \\ \hat{g}_{d_b}(x) \end{bmatrix}. \tag{19}$$

Thus, the predictor is obtained by fitting d_b independent KRR predictors to the transformed output coordinates and mapping their predictions back to $\mathrm{Im}(S)$ through U_S.

B.4. Proof of Lemma 3.2

Proof We denote the objective function by $J(A') = \left\|M - U_S A' U_S^\mathsf{T}\right\|_\mathrm{F}^2$. Since the objective function is continuous and coercive, and the PSD cone $\mathbb{S}_+^{d_b}$ is closed, there exists a minimizer. Moreover, since J is strictly convex, the minimizer is unique. Actually, the minimization problem is a so-called convex quadratic SDP problem. Due to the fact that

$$J(A') = \left\|U_S A' U_S^\mathsf{T}\right\|_\mathrm{F}^2 - 2\left\langle M, U_S A' U_S^\mathsf{T}\right\rangle_\mathrm{F} + \|M\|_\mathrm{F}^2 \tag{20}$$

$$= \|A'\|_\mathrm{F}^2 - 2\left\langle U_S^\mathsf{T} M U_S, A'\right\rangle_\mathrm{F} + \|M\|_\mathrm{F}^2, \tag{21}$$

we have

$$\operatorname*{arg\,min}_{A' \in \mathbb{S}_+^{d_b}} J(A') = \operatorname*{arg\,min}_{A' \in \mathbb{S}_+^{d_b}} \|A'\|_\mathrm{F}^2 - 2\left\langle U_S^\mathsf{T} M U_S, A'\right\rangle_\mathrm{F} + \left\|U_S^\mathsf{T} M U_S\right\|_\mathrm{F}^2 \tag{22}$$

$$= \operatorname*{arg\,min}_{A' \in \mathbb{S}_+^{d_b}} \left\|A' - U_S^\mathsf{T} M U_S\right\|_\mathrm{F}^2 \tag{23}$$

whose solution is the nearest PSD matrix of $U_S^\mathsf{T} M U_S$ in Frobenius norm. Higham (1988) shows that the solution is obtained by eigenvalue decomposition of $U_S^\mathsf{T} M U_S$ and zeroing out negative eigenvalues, thus concluding the proof of the lemma. ■

B.5. Proof of Corollary 3.3

Proof By Lemma 3.2, we have $\operatorname{arg\,min}_{A' \in \mathbb{S}_+^{d_b}} \left\|M - U_S A' U_S^\mathsf{T}\right\|_\mathrm{F}^2 = \operatorname{arg\,min}_{A' \in \mathbb{S}_+^{d_b}} \left\|A' - U_S^\mathsf{T} M U_S\right\|_\mathrm{F}^2$. Since $M \in \mathbb{S}_+^d$, it follows that $U_S^\mathsf{T} M U_S \in \mathbb{S}_+^{d_b}$. Consequently, $A'_\mathrm{opt} = U_S^\mathsf{T} M U_S \in \mathbb{S}_+^{d_b}$. It then follows that

$$A_\mathrm{opt} = U_S A'_\mathrm{opt} U_S^\mathsf{T} = U_S U_S^\mathsf{T} M U_S U_S^\mathsf{T} = PMP \tag{24}$$

where $P \coloneqq U_S U_S^\mathsf{T}$ is actually the orthogonal projection onto $\mathrm{Im}(S)$. ■

B.6. Existence of Feature Map for the Non-separable OVK Defined in Example 2

Proposition B.2 *$\widetilde{\mathcal{K}}_\rho$ defined in Example 2 is a valid OVK since it admits a feature operator Γ_ρ such that for all $x, x' \in \mathcal{X}$,*

$$\widetilde{\mathcal{K}}_\rho(x, x') = \Gamma_\rho(x) \Gamma_\rho(x')^*, \tag{25}$$

where the feature operator $\Gamma_\rho(x) : \mathbb{R}^{q_\rho \times p_\phi} \to \mathbb{R}^d$ *is defined, for all* $r \in \{1, \dots, d\}$*, by*

$$[\Gamma_\rho(x)\Theta]_r = \frac{1}{\sqrt{\lambda_r}} \left\langle \phi(x_r), \Theta^\mathsf{T}(L_\rho)_{r,:}^\mathsf{T} \right\rangle, \tag{26}$$

and its adjoint is

$$\Gamma_\rho(x)^* v = \sum_{r=1}^{d} \frac{v_r}{\sqrt{\lambda_r}} (L_\rho)_{r,:}^\mathsf{T} \phi(x_r)^\mathsf{T} \in \mathbb{R}^{q_\rho \times p_\phi}, \tag{27}$$

where $L_\rho \in \mathbb{R}^{d \times q_\rho}$ *is any factor satisfying* $A_\rho = L_\rho L_\rho^\mathsf{T}$*.*

Proof First, A_ρ is positive semidefinite. Indeed, for every $v \in \mathbb{R}^d$,

$$v^\mathsf{T} A_\rho v = (1-\rho) \|v\|_{\mathbb{R}^d}^2 + \rho \left\| B^\mathsf{T} v \right\|_{\mathbb{R}^q}^2 \geq 0. \tag{28}$$

It therefore admits a factorization $A_\rho = L_\rho L_\rho^\mathsf{T}$, where we may take $q_\rho = \operatorname{rank}(A_\rho)$ and $L_\rho \in \mathbb{R}^{d \times q_\rho}$.

Equip the feature space $\mathcal{F}_\rho = \mathbb{R}^{q_\rho \times p_\phi}$ with the Frobenius inner product. For $x \in \mathcal{X}$, define the linear operator $\Gamma_\rho(x) : \mathcal{F}_\rho \to \mathbb{R}^d$ as

$$[\Gamma_\rho(x)\Theta]_r = \frac{1}{\sqrt{\lambda_r}} \left\langle \phi(x_r), \Theta^\mathsf{T}(L_\rho)_{r,:}^\mathsf{T} \right\rangle. \tag{29}$$

For every $\Theta \in \mathcal{F}_\rho$ and $v \in \mathbb{R}^d$, we have

$$\langle \Gamma_\rho(x)\Theta, v \rangle_{\mathbb{R}^d} = \sum_{r=1}^{d} \frac{v_r}{\sqrt{\lambda_r}} \left\langle \phi(x_r), \Theta^\mathsf{T}(L_\rho)_{r,:}^\mathsf{T} \right\rangle_{\mathbb{R}^{p_\phi}} \tag{30}$$

$$= \left\langle \Theta, \sum_{r=1}^{d} \frac{v_r}{\sqrt{\lambda_r}} (L_\rho)_{r,:}^\mathsf{T} \phi(x_r)^\mathsf{T} \right\rangle_{\mathcal{F}_\rho}. \tag{31}$$

Hence its adjoint is precisely the following operator:

$$\Gamma_\rho(x)^* v = \sum_{r=1}^{d} \frac{v_r}{\sqrt{\lambda_r}} (L_\rho)_{r,:}^\mathsf{T} \phi(x_r)^\mathsf{T} \in \mathbb{R}^{q_\rho \times p_\phi}. \tag{32}$$

It remains to verify the factorization of the kernel. For any $x, x' \in \mathcal{X}$, $v \in \mathbb{R}^d$, and $r \in \{1, \dots, d\}$,

$$[\Gamma_\rho(x)\Gamma_\rho(x')^* v]_r = \frac{1}{\sqrt{\lambda_r}} \left\langle \phi(x_r), \left(\sum_{s=1}^{d} \frac{v_s}{\sqrt{\lambda_s}} (L_\rho)_{s,:}^\mathsf{T} \phi(x'_s)^\mathsf{T} \right)^\mathsf{T} (L_\rho)_{r,:}^\mathsf{T} \right\rangle_{\mathbb{R}^{p_\phi}} \tag{33}$$

$$= \sum_{s=1}^{d} \frac{(L_\rho)_{r,:}(L_\rho)_{s,:}^\mathsf{T}}{\sqrt{\lambda_r \lambda_s}} \langle \phi(x_r), \phi(x'_s) \rangle_{\mathbb{R}^{p_\phi}} v_s \tag{34}$$

$$= \sum_{s=1}^{d} \frac{(A_\rho)_{rs}}{\sqrt{\lambda_r \lambda_s}} k(x_r, x'_s) v_s = [\widetilde{\mathcal{K}}_\rho(x, x')v]_r. \tag{35}$$

Since this identity holds for every $v \in \mathbb{R}^d$, it follows that

$$\widetilde{\mathcal{K}}_\rho(x, x') = \Gamma_\rho(x)\Gamma_\rho(x')^*. \tag{36}$$

Thus $\widetilde{\mathcal{K}}_\rho$ is a valid operator-valued kernel, and Γ_ρ is one of its feature operators. ■

B.7. Proof of the Equivalence of the Two Learning Formulations

The following lemma summarizes the relationship between the RKHS associated to $\mathcal{K}_{SG}$ and the one associated to $\widetilde{\mathcal{K}}$.

Lemma B.3 *Let $\widetilde{\mathcal{K}}$ be a $\mathcal{L}(\mathbb{R}^d)$-valued kernel, and let $\mathcal{H}_{\widetilde{\mathcal{K}}}$ be its associated RKHS. Let $SG : \mathbb{R}^d \to \mathbb{R}^d$ be a bounded operator. Define*

$$\mathcal{H}_{\widetilde{\mathcal{K}},SG} \coloneqq \left\{SG \circ h \mid h \in \mathcal{H}_{\widetilde{\mathcal{K}}}\right\} \subset (\mathrm{Im}(S))^{\mathcal{X}}, \tag{37}$$

then $\mathcal{H}_{\widetilde{\mathcal{K}},SG}$ is the RKHS associated to the operator-valued kernel $\mathcal{K}_{SG}$ defined in Equation (6). Moreover,

$$\|f\|_{\mathcal{H}_{\widetilde{\mathcal{K}},SG}} = \min\left\{\|h\|_{\mathcal{H}_{\widetilde{\mathcal{K}}}} \mid h \in \mathcal{H}_{\widetilde{\mathcal{K}}}, SG \circ h = f\right\}. \tag{38}$$

Proof Define the operator $W : \mathcal{H}_{\widetilde{\mathcal{K}}} \to \mathcal{H}_{\widetilde{\mathcal{K}},SG}$ such that $W(h) = SG \circ h$. Let $w = SG$, then the first part of the lemma can be proved by directly applying Proposition 6 of Carmeli et al. (2010). Moreover, by Proposition 1 of Carmeli et al. (2010), we know that W is a partial isometry from $\mathcal{H}_{\widetilde{\mathcal{K}}}$ onto $\mathcal{H}_{\widetilde{\mathcal{K}},SG}$, which proves the second part of the lemma. ■

The following lemma establishes the existence and uniqueness of the minimizer $\hat{h}$ of Problem (12) and shows that it admits a finite kernel expansion.

Lemma B.4 (Representation Theorem) *Problem (12) admits a unique minimizer $\hat{h}$, which is of the form $\hat{h} = \sum_{t=1}^{T} \widetilde{\mathcal{K}}_{x_t} c_t$ where $c_t \in \mathbb{R}^d$ for all $t \in \{1, \ldots, T\}$.*

Proof Let $\mathcal{U} = \left\{h \mid h = \sum_{t=1}^{T} \widetilde{\mathcal{K}}_{x_t} c_t, (c_t)_{t=1}^{T} \in (\mathbb{R}^d)^T\right\}$ be a subspace of $\mathcal{H}_{\widetilde{\mathcal{K}}}$. Since $\mathcal{U}$ is finite-dimensional and hence closed, we can write $\mathcal{H}_{\widetilde{\mathcal{K}}} = \mathcal{U} \oplus \mathcal{U}^{\perp}$ where $\mathcal{U}^{\perp}$ denotes the orthogonal complement of $\mathcal{U}$. Any $h \in \mathcal{H}_{\widetilde{\mathcal{K}}}$ can be decomposed as $h = h_{\mathcal{U}} + h_{\mathcal{U}^{\perp}}$ with $h_{\mathcal{U}} \in \mathcal{U}$ and $h_{\mathcal{U}^{\perp}} \in \mathcal{U}^{\perp}$. Then by the reproducing property, we have for all $t \in \{1, \ldots, T\}$,

$$\langle h_{\mathcal{U}^{\perp}}(x_t), y\rangle_{\mathbb{R}^d} = \left\langle \widetilde{\mathcal{K}}^*_{x_t} h_{\mathcal{U}^{\perp}}, y\right\rangle_{\mathbb{R}^d} = \left\langle h_{\mathcal{U}^{\perp}}, \widetilde{\mathcal{K}}_{x_t} y\right\rangle_{\mathcal{H}_{\widetilde{\mathcal{K}}}} = 0, \quad \forall y \in \mathbb{R}^d. \tag{39}$$

As a result, $h_{\mathcal{U}^{\perp}}(x_t) = 0$ for all $t \in \{1, \ldots, T\}$. For the first part of the objective function, we have

$$\|SGh(x_t) - y_t\|^2_{\mathbb{R}^d} = \|SG\left(h_{\mathcal{U}^{\perp}}(x_t) + h_{\mathcal{U}}(x_t)\right) - y_t\|^2_{\mathbb{R}^d} = \|SGh_{\mathcal{U}}(x_t) - y_t\|^2_{\mathbb{R}^d}. \tag{40}$$

For the regularization term, we have

$$\|h\|^2_{\mathcal{H}_{\widetilde{\mathcal{K}}}} = \|h_{\mathcal{U}} + h_{\mathcal{U}^{\perp}}\|^2_{\mathcal{H}_{\widetilde{\mathcal{K}}}} = \|h_{\mathcal{U}}\|^2_{\mathcal{H}_{\widetilde{\mathcal{K}}}} + \|h_{\mathcal{U}^{\perp}}\|^2_{\mathcal{H}_{\widetilde{\mathcal{K}}}} \geq \|h_{\mathcal{U}}\|^2_{\mathcal{H}_{\widetilde{\mathcal{K}}}}. \tag{41}$$

Hence, if there exists a minimizer $\hat{h} \in \mathcal{H}_{\widetilde{\mathcal{K}}}$, we know that $\hat{h} \in \mathcal{U}$.

Recall that our objective function is $L(h) = \frac{1}{T}\sum_{t=1}^{T}\left\|SG\widetilde{\mathcal{K}}^*_{x_t}h - y_t\right\|^2_{\mathbb{R}^d} + \lambda\|h\|^2_{\mathcal{H}_{\widetilde{\mathcal{K}}}}$. Since $\mathcal{H}_{\widetilde{\mathcal{K}}}$ is a reflexive Banach space, and $L : \mathcal{H}_{\widetilde{\mathcal{K}}} \to \mathbb{R}$ is continuous, coercive, and strictly convex, we conclude that the minimum can be reached, and it admits a unique minimizer (see, e.g., Theorem 2.19 of Peypouquet, 2015). ■

Proof Denoting $L(f) \coloneqq \frac{1}{T}\sum_{t=1}^{T}\|f(x_t) - y_t\|^2_{\mathbb{R}^d} + \lambda\|f\|^2_{\mathcal{H}_{\widetilde{\mathcal{K}},SG}}$, we want to prove that for all $f \in \mathcal{H}_{\widetilde{\mathcal{K}},SG}$, $L(\hat{f}) \leq L(f)$. By Lemma B.3 and the fact that $\hat{f} = SG \circ \hat{h}$, we have

$$\begin{aligned} L(\hat{f}) &= \frac{1}{T}\sum_{t=1}^{T}\left\|SG\hat{h}(x_t) - y_t\right\|^2_{\mathbb{R}^d} + \lambda \min\left\{\|h\|^2_{\mathcal{H}_{\widetilde{\mathcal{K}}}} \mid h \in \mathcal{H}_{\widetilde{\mathcal{K}}}, SG \circ h = \hat{f}\right\} \\ &\leq \frac{1}{T}\sum_{t=1}^{T}\left\|SG\hat{h}(x_t) - y_t\right\|^2_{\mathbb{R}^d} + \lambda\left\|\hat{h}\right\|^2_{\mathcal{H}_{\widetilde{\mathcal{K}}}}. \end{aligned}$$

Since $\hat{h} = \arg\min_{h \in \mathcal{H}_{\widetilde{\mathcal{K}}}} \frac{1}{T} \sum_{t=1}^{T} \|SGh(x_t) - y_t\|_{\mathbb{R}^d}^2 + \lambda \|h\|_{\mathcal{H}_{\widetilde{\mathcal{K}}}}^2$, it follows that

$$L(\hat{f}) \leq \frac{1}{T} \sum_{t=1}^{T} \|SGh(x_t) - y_t\|_{\mathbb{R}^d}^2 + \lambda \|h\|_{\mathcal{H}_{\widetilde{\mathcal{K}}}}^2, \quad \forall h \in \mathcal{H}_{\widetilde{\mathcal{K}}}. \tag{42}$$

Let $f \in \mathcal{H}_{\widetilde{\mathcal{K}},SG}$, and let

$$\tilde{h} \coloneqq \operatorname*{arg\,min}_{h \in \mathcal{H}_{\widetilde{\mathcal{K}}}, SG \circ h = f} \|h\|_{\mathcal{H}_{\widetilde{\mathcal{K}}}}^2 .$$

Since the inequality described by Equation (42) is valid for all $h \in \mathcal{H}_{\widetilde{\mathcal{K}}}$, we have then

$$\begin{aligned} L(\hat{f}) &\leq \frac{1}{T} \sum_{t=1}^{T} \left\| SG\tilde{h}(x_t) - y_t \right\|_{\mathbb{R}^d}^2 + \lambda \left\| \tilde{h} \right\|_{\mathcal{H}_{\widetilde{\mathcal{K}}}}^2 \\ &= \frac{1}{T} \sum_{t=1}^{T} \|f(x_t) - y_t\|_{\mathbb{R}^d}^2 + \lambda \min \left\{ \|h\|_{\mathcal{H}_{\widetilde{\mathcal{K}}}}^2 \mid h \in \mathcal{H}_{\widetilde{\mathcal{K}}}, SG \circ h = f \right\} \\ &= L(f). \end{aligned}$$

The proof is thus concluded. ■

B.8. Link with WeaKL Estimators

The general estimator proposed in Doumèche et al. (2025, Section 2) tries to solve the following optimization problem:

$$\min_{\theta \in \mathbb{R}^{d_1}} \frac{1}{T} \sum_{t=1}^{T} \ell(\Phi(x_t)\theta, y_t) + \|M\theta\|_{\mathbb{R}^{d_2}}^2 \tag{43}$$

where, for each $i \in \{1, \dots, d\}$, $\phi_i : \mathcal{X} \to \mathbb{R}^{p_i}$ is a feature map with $p_i \in \mathbb{N}^*$, and $d_1 = \sum_{i=1}^{d} p_i$. Since in Doumèche et al. (2025, Proposition 2.1), M is assumed to be injective, hence the inverse of $M^\mathsf{T} M$ is well-defined. For any $x \in \mathcal{X}$, the feature matrix is defined as

$$\Phi(x) \coloneqq \begin{bmatrix} \phi_1(x)^\mathsf{T} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \phi_d(x)^\mathsf{T} \end{bmatrix} \in \mathbb{R}^{d \times d_1} \tag{44}$$

Here, $M \in \mathbb{R}^{d_2 \times d_1}$ is a regularization penalty matrix, and $\ell : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ is a loss function defined as $\ell(y, y') = \|\Lambda(y - y')\|_{\mathbb{R}^{d_3}}^2$ with $\Lambda \in \mathbb{R}^{d_3 \times d}$. Since $\|M\theta\|_{\mathbb{R}^{d_2}}^2 = \langle M^\mathsf{T} M\theta, \theta \rangle_{\mathbb{R}^{d_1}} = \left\|(M^\mathsf{T} M)^{1/2}\theta\right\|_{\mathbb{R}^{d_1}}^2$, denoting $w \coloneqq (M^\mathsf{T} M)^{1/2}\theta \in \mathbb{R}^{d_1}$, then the above optimization problem can be rewritten as

$$\min_{w \in \mathbb{R}^{d_1}} \frac{1}{T} \sum_{t=1}^{T} \ell\left(\Phi(x_t)(M^\mathsf{T} M)^{-1/2} w, y_t\right) + \|w\|_{\mathbb{R}^{d_1}}^2 . \tag{45}$$

This problem is equivalent to the following vector-valued regression problem:

$$\min_{f \in \mathcal{H}_\mathcal{K}} \frac{1}{T} \sum_{t=1}^{T} \ell\left(f(x_t), y_t\right) + \|f\|_{\mathcal{H}_\mathcal{K}}^2 . \tag{46}$$

with the operator-valued kernel defined as, for all $(x, x') \in \mathcal{X}^2$,

$$\mathcal{K}(x, x') = \Phi(x)(M^\mathsf{T} M)^{-1} \Phi(x')^\mathsf{T} \in \mathbb{R}^{d \times d}. \tag{47}$$

To justify this equivalence, set $\Psi(x) = \Phi(x)(M^\mathsf{T} M)^{-1/2}$, so that $\mathcal{K}(x, x') = \Psi(x)\Psi(x')^\mathsf{T}$. By Proposition 1 of Carmeli et al. (2010), the associated RKHS consists of the functions $f_w = \Psi(\cdot)w$, with norm

$$\|f\|_{\mathcal{H}_\mathcal{K}}^2 = \min \left\{ \|w\|_{\mathbb{R}^{d_1}}^2 \mid w \in \mathbb{R}^{d_1},\ \Psi(x)w = f(x) \text{ for all } x \in \mathcal{X} \right\}. \tag{48}$$

Since the empirical loss depends on w only through f_w, minimizing the parameter objective first over all representations of each function yields exactly Equation (46).

The WeaKL-BU estimator proposed in Doumèche et al. (2025, Section 4.1) learns parameters only for the bottom-level series. Let $\theta_b \in \mathbb{R}^{p_b}$ collect these parameters, let $\Phi_b(x) \in \mathbb{R}^{d_b \times p_b}$ be their block-diagonal feature matrix, and let $M_b \in \mathbb{R}^{q_b \times p_b}$ be an injective penalty matrix. The predictor for all hierarchy levels is $f_{\theta_b}(x) = S\Phi_b(x)\theta_b$, with penalty $\|M_b\theta_b\|^2_{\mathbb{R}^{q_b}}$. Applying the preceding construction to this parameterization shows that the resulting predictor is the minimizer of Equation (46) with the OVK

$$\mathcal{K}_{\mathrm{BU}}(x, x') = S\Phi_b(x)(M_b^\mathsf{T} M_b)^{-1}\Phi_b(x')^\mathsf{T} S^\mathsf{T} \in \mathbb{R}^{d\times d}. \tag{49}$$

Similarly, the WeaKL-T estimator adds transfer learning constraints to the same bottom-level parameterization. We use the notation J, D, Π_J, and P_J from Doumèche et al. (2025, Section 4.1), to which we refer for their definitions. In particular, P_J is the orthogonal projector onto the parameter-sharing subspace, so $(I_{D|J|} - P_J)^\mathsf{T}(I_{D|J|} - P_J) = I_{D|J|} - P_J$. The estimator corresponds to the minimizer with the OVK

$$\mathcal{K}_{\mathrm{T}}(x, x') = S\Phi_b(x)\left(\mu\Pi_J^\mathsf{T}(I_{D|J|} - P_J)\Pi_J + M_b^\mathsf{T} M_b\right)^{-1}\Phi_b(x')^\mathsf{T} S^\mathsf{T} \in \mathbb{R}^{d\times d}, \tag{50}$$

with $\mu \geq 0$ since $\mu\left\|(I_{D|J|} - P_J)\Pi_J\theta_b\right\|^2 + \|M_b\theta_b\|^2 = \left\|(\mu\Pi_J^\mathsf{T}(I_{D|J|} - P_J)\Pi_J + M_b^\mathsf{T} M_b)^{1/2}\theta_b\right\|^2_{\mathbb{R}^{p_b}}$.

To express both kernels in the form of Equation (6), let $G_{\mathrm{BU}} = [0_{d_b\times d_a} \;\; I_{d_b}]$ and embed the bottom-level feature matrix into the full output space as $\overline{\Phi}(x) = G_{\mathrm{BU}}^\mathsf{T}\Phi_b(x) \in \mathbb{R}^{d\times p_b}$. For either penalty matrix $Q = M_b^\mathsf{T} M_b$ or $Q = \mu\Pi_J^\mathsf{T}(I_{D|J|} - P_J)\Pi_J + M_b^\mathsf{T} M_b$, define the base OVK $\widetilde{\mathcal{K}}(x, x') = \overline{\Phi}(x)Q^{-1}\overline{\Phi}(x')^\mathsf{T}$. Since $G_{\mathrm{BU}}G_{\mathrm{BU}}^\mathsf{T} = I_{d_b}$, composing this base kernel with SG_{BU} yields the corresponding kernel above. Thus, both are special cases of Equation (6).

However, it should be noted that the WeaKL estimators use a weighted loss function and allow complex-valued features, which can be easily extended into our RVVKRR framework.

B.9. Computation Methods for Solving the Linear System in Equation (5)

Now we describe the computational methods for solving the linear system in Equation (5). We will first consider the case of separable OVKs, which is more common in practice, and then discuss the non-separable case.

B.9.1. Separable OVK Case

In the case of a separable operator-valued kernel, i.e., $\mathcal{K}(x, x') = k(x, x')A$ where $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a scalar-valued kernel defined on the input space and the task kernel $A \in \mathcal{L}(\mathrm{Im}(S)) \subset \mathbb{R}^{d\times d}$ is self-adjoint and positive, Equation (5) can be expressed in the following matrix form:

$$T\lambda C + ACK = Y, \tag{51}$$

with $Y = \begin{bmatrix} y_1 & y_2 & \cdots & y_T \end{bmatrix} \in \mathbb{R}^{d\times T}$, $C = \begin{bmatrix} c_1 & c_2 & \cdots & c_T \end{bmatrix} \in \mathrm{Im}(S)^T$ and $K = (k(x_i, x_j)) \in \mathbb{R}^{T\times T}$. As a result of Lemma 3.1, Equation (51) becomes

$$T\lambda E + A'EK = U_S^\mathsf{T} Y. \tag{52}$$

with $E = U_S^\mathsf{T} C \in \mathbb{R}^{d_b\times T}$, which is a discrete-time Sylvester equation and can therefore be solved with the Hessenberg-Schur method (Golub et al., 1979; Sima, 1996), with a time complexity $\mathcal{O}(T^3 + d_b^3)$ and a space complexity $\mathcal{O}(T^2 + d_b^2)$[2]. For a discrete-time Sylvester equation, the solution is unique if and only if $\sigma_i\mu_j \neq -1$ for all the eigenvalues σ_i of A' and all the eigenvalues μ_j of K (Simoncini, 2016, Section 6), which is satisfied here since A' and K are both positive semi-definite. Then, the prediction for a given input $x \in \mathcal{X}$ is given by: $\hat{f}(x) = \sum_{t=1}^T U_S A' U_S^\mathsf{T} k(x, x_t)c_t = U_S A' E k_x$ with $k_x = \begin{bmatrix} k(x, x_1) & \cdots & k(x, x_T) \end{bmatrix}^\mathsf{T} \in \mathbb{R}^T$.

Remark B.5 *Equation (52) provides another interpretation of our method. We first represent the training outputs as $y' = U_S y \in \mathbb{R}^{d_b}$ in the new orthonormal basis $(u_1, \ldots, u_{d_b})$ of $\mathrm{Im}(S)$ .Using this transformed dataset, we then perform the VVKRR with a separable OVK, where $A \in \mathbb{S}_+^{d_b}$.*

[2]For example, we can use the implementation SB04QD provided by the SLICOT library (https://www.slicot.org/) and the corresponding python wrapper Slycot (https://github.com/python-control/Slycot).

B.9.2. NON-SEPARABLE OVK CASE

In the case of a non-separable OVK, the Sylvester formulation in Equation (5) is no longer applicable. A different computational route is available whenever the OVK admits an explicit finite-dimensional feature factorization. Suppose that, for some $m \in \mathbb{N}^*$, there exists a feature operator $\Gamma(x) : \mathbb{R}^m \longrightarrow \mathrm{Im}(S)$ such that $\mathcal{K}(x, x') = \Gamma(x)\Gamma(x')^*$. By Proposition 2.2 of Carmeli et al. (2006), the RKHS $\mathcal{H}_\mathcal{K}$ is then the set of all functions $f_\theta(x) = \Gamma(x)\theta$ with $\theta \in \mathbb{R}^m$, and the RKHS norm is defined as $\|f_\theta\|_{\mathcal{H}_\mathcal{K}} = \min\{\|\theta'\|_{\mathbb{R}^m} \mid f_{\theta'} = f_\theta\}$. It follows that the RVVKRR defined in Equation (4) can be obtained by solving the following finite-dimensional optimization problem:

$$\widehat{\theta} = \underset{\theta \in \mathbb{R}^m}{\arg\min} \frac{1}{T} \sum_{t=1}^{T} \|\Gamma(x_t)\theta - y_t\|_{\mathbb{R}^d}^2 + \lambda \|\theta\|_{\mathbb{R}^m}^2 . \tag{53}$$

The first-order optimality condition gives the symmetric positive-definite linear system

$$\left(\sum_{t=1}^{T} \Gamma(x_t)^* \Gamma(x_t) + T\lambda I_m \right) \widehat{\theta} = \sum_{t=1}^{T} \Gamma(x_t)^* y_t. \tag{54}$$

Although this system could be solved by explicitly forming and factorizing its $m \times m$ coefficient matrix, doing so can be prohibitive. We instead apply the preconditioned conjugate gradient method (Hestenes and Stiefel, 1952; Cutajar et al., 2016) in matrix-free form. Each iteration only requires evaluations of $\Gamma(x_t)$ and its adjoint; the matrix $\sum_{t=1}^{T} \Gamma(x_t)^*\Gamma(x_t) + T\lambda I_m$ need not be materialized. In the OVK literature, this strategy has previously been used to solve large-scale discrete-time Sylvester equations induced by the separable matrix-valued kernel regression (Sindhwani et al., 2013). For example, consider the coherent OVK in Section 3.2.2 with the base kernel of Example 2. For a generic input $x = (x_1, \ldots, x_d)$, write $\phi_r = \phi(x_r) \in \mathbb{R}^{p_\phi}$, $\ell_r = (L_\rho)_{r,:}^\mathsf{T} \in \mathbb{R}^{q_\rho}$, $\theta = \mathrm{vec_r}(\Theta) = \mathrm{vec}(\Theta^\mathsf{T})$ with $\Theta \in \mathbb{R}^{q_\rho \times p_\phi}$, and $R = SG$. Since $\Gamma(x) = R\Gamma_\rho(x)$, its forward evaluation is

$$\Gamma(x)\theta = R \begin{bmatrix} \lambda_1^{-1/2} \ell_1^\mathsf{T} \Theta \phi_1 \\ \vdots \\ \lambda_d^{-1/2} \ell_d^\mathsf{T} \Theta \phi_d \end{bmatrix} \in \mathbb{R}^d. \tag{55}$$

and $\forall v \in \mathbb{R}^d$, its adjoint evaluation is

$$\Gamma(x)^* v = \mathrm{vec_r} \left(\sum_{r=1}^{d} \frac{(R^\mathsf{T} v)_r}{\sqrt{\lambda_r}} \ell_r \phi_r^\mathsf{T} \right) \in \mathbb{R}^{q_\rho p_\phi}. \tag{56}$$

We describe now the implementation in reduced coherent coordinates. Let $U_S \in \mathbb{R}^{d \times d_b}$ have orthonormal columns spanning $\mathrm{Im}(S)$, let $R = SG$, we have

$$\|R\Gamma_\rho(x_t)\Theta - y_t\|_{\mathbb{R}^d}^2 = \left\|U_S^\mathsf{T} R\Gamma_\rho(x_t)\Theta - U_S^\mathsf{T} y_t\right\|_{\mathbb{R}^{d_b}}^2 + \left\|(I_d - U_S U_S^\mathsf{T}) y_t\right\|_{\mathbb{R}^d}^2 . \tag{57}$$

Since the second term on the right-hand side is independent of Θ, we can equivalently minimize the reduced loss described by the first term.

For $x = (x_1, \ldots, x_d)$, write $\phi_r = \phi(x_r)$ and $\ell_r = (L_\rho)_{r,:}^\mathsf{T}$. Now we define the reduced forward and adjoint operators $\overline{\Gamma}_{R,\rho}(x)$ and $\overline{\Gamma}_{R,\rho}^*(x)$ by

$$\overline{\Gamma}_{R,\rho}(x)\Theta = U_S^\mathsf{T} R\Gamma_\rho(x)\Theta = U_S^\mathsf{T} R \begin{bmatrix} \lambda_1^{-1/2} \ell_1^\mathsf{T} \Theta \phi_1 \\ \vdots \\ \lambda_d^{-1/2} \ell_d^\mathsf{T} \Theta \phi_d \end{bmatrix} \in \mathbb{R}^{d_b}, \qquad \forall \Theta \in \mathbb{R}^{q_\rho \times p_\phi}, \tag{58}$$

$$\overline{\Gamma}_{R,\rho}(x)^* c = \Gamma_\rho(x)^* R^\mathsf{T} U_S c = \sum_{r=1}^{d} \frac{(R^\mathsf{T} U_S c)_r}{\sqrt{\lambda_r}} \ell_r \phi_r^\mathsf{T} \in \mathbb{R}^{q_\rho \times p_\phi}, \qquad \forall c \in \mathbb{R}^{d_b}. \tag{59}$$

With this notation, the linear system to solve can be rewritten as

$$\left(\sum_{t=1}^{T} \overline{\Gamma}_{R,\rho}(x_t)^* \overline{\Gamma}_{R,\rho}(x_t) + T\lambda I_{\mathbb{R}^{q_\rho \times p_\phi}}\right) \Theta = \sum_{t=1}^{T} \overline{\Gamma}_{R,\rho}(x_t)^* U_S^\mathsf{T} y_t. \tag{60}$$

Let $p = p_\phi$, $q = q_\rho$. In the real implementation, it is more practical to define the batched forward operator $\mathcal{F} : \mathbb{R}^{q \times p} \to \mathbb{R}^{d_b \times T}$ and its adjoint $\mathcal{F}^* : \mathbb{R}^{d_b \times T} \to \mathbb{R}^{q \times p}$ as follows:

$$\mathcal{F}(\Theta) = \begin{bmatrix} \overline{\Gamma}_{R,\rho}(x_1)\Theta & \dots & \overline{\Gamma}_{R,\rho}(x_T)\Theta \end{bmatrix} \in \mathbb{R}^{d_b \times T}, \tag{61}$$

and

$$\mathcal{F}^*(C) = \sum_{t=1}^{T} \overline{\Gamma}_{R,\rho}(x_t)^* c_t, \quad \forall C = \begin{bmatrix} c_1 & \cdots & c_T \end{bmatrix} \in \mathbb{R}^{d_b \times T}. \tag{62}$$

One can check that $\mathcal{F}^*$ is indeed the adjoint of $\mathcal{F}$ with respect to the Frobenius inner product, i.e., for all $\Theta \in \mathbb{R}^{q \times p}$ and $C \in \mathbb{R}^{d_b \times T}$, we have $\langle \mathcal{F}(\Theta), C \rangle_{\mathbb{R}^{d_b \times T}} = \langle \Theta, \mathcal{F}^*(C) \rangle_{\mathbb{R}^{q \times p}}$. As a result, Equation (60) can be rewritten as

$$\left(\mathcal{F}^*\mathcal{F} + T\lambda I_{\mathbb{R}^{q \times p}}\right) \Theta = \mathcal{F}^*(\overline{Y}), \tag{63}$$

where $Y = \begin{bmatrix} y_1 & \cdots & y_T \end{bmatrix} \in \mathbb{R}^{d \times T}$ and $\overline{Y} = U_S^\mathsf{T} Y \in \mathbb{R}^{d_b \times T}$.

We propose solving Equation (63) using the matrix-free preconditioned conjugate gradient method (Hestenes and Stiefel, 1952; Cutajar et al., 2016), with the preconditioner and custom initialization[3], where only the forward operator $\mathcal{F}$ and its adjoint $\mathcal{F}^*$ are needed. The following paragraphs describe how to compute $\mathcal{F}$ and $\mathcal{F}^*$ efficiently using tensor contractions.

Tensor Computation of $\mathcal{F}$ and $\mathcal{F}^*$ Let $\mathsf{X} \in \mathbb{R}^{T \times d \times p}$ be the feature tensor defined by $\mathsf{X}_{t,r,:} = \phi(x_{t,r})^\mathsf{T}$. Define the intermediate matrix $A_\Theta \in \mathbb{R}^{d \times T}$ by

$$A_\Theta = \begin{bmatrix} \Gamma_\rho(x_1)\Theta & \cdots & \Gamma_\rho(x_T)\Theta \end{bmatrix} \in \mathbb{R}^{d \times T}. \tag{64}$$

Then

$$\mathcal{F}(\Theta) = U_S^\mathsf{T} R A_\Theta \in \mathbb{R}^{d_b \times T}. \tag{65}$$

The entries of A_Θ can be computed by the tensor contraction

$$(A_\Theta)_{rt} = \lambda_r^{-1/2} \ell_r^\mathsf{T} \Theta \phi(x_{t,r}) = \frac{1}{\sqrt{\lambda_r}} \sum_{j=1}^{p} \mathsf{X}_{t,r,j} (L_\rho \Theta)_{rj}. \tag{66}$$

For $C \in \mathbb{R}^{d_b \times T}$, define the intermediate matrix $G_C \in \mathbb{R}^{d \times p}$ by

$$(G_C)_{rj} = \sum_{t=1}^{T} \frac{(R^\mathsf{T} U_S C)_{rt}}{\sqrt{\lambda_r}} \phi(x_{t,r})_j = \frac{1}{\sqrt{\lambda_r}} \sum_{t=1}^{T} \mathsf{X}_{t,r,j} (R^\mathsf{T} U_S C)_{rt}. \tag{67}$$

Then $\mathcal{F}^*(C) = L_\rho^\mathsf{T} G_C$. Moreover, to apply conjugate gradients, we identify $\Theta \in \mathbb{R}^{q \times p}$ with $\theta \in \mathbb{R}^{qp}$ using row-wise vectorization.

Preconditioner We denote by $\mathrm{vec_r}$ the row-wise vectorization of a matrix, i.e., for $\Theta \in \mathbb{R}^{q \times p}$, we have

$$\vartheta_k = \Theta_{k,:}^\mathsf{T} \in \mathbb{R}^p, \qquad \theta = \mathrm{vec_r}(\Theta) = \begin{bmatrix} \vartheta_1^\mathsf{T} & \cdots & \vartheta_q^\mathsf{T} \end{bmatrix}^\mathsf{T}. \tag{68}$$

Let $\mathrm{mat_r}$ denote the inverse of the row-wise vectorization introduced above. We denote by $H \in \mathbb{R}^{qp \times qp}$ the matrix representation of the operator $\mathcal{F}^*\mathcal{F} + T\lambda I_{\mathbb{R}^{q \times p}}$ and set

$$H\theta = \mathrm{vec_r}(\mathcal{F}^*\mathcal{F}(\mathrm{mat_r}(\theta)) + T\lambda\, \mathrm{mat_r}(\theta)), \tag{69}$$

$$b = \mathrm{vec_r}\left(\mathcal{F}^*(\overline{Y})\right). \tag{70}$$

[3] We use the implementation provided by SciPy library (https://docs.scipy.org/doc/scipy/reference/generated/scipy.sparse.linalg.cg.html).

Thus, Equation (63) is equivalent to $H\theta = b$. Let $H_{k\ell} \in \mathbb{R}^{p\times p}$ denote block (k,ℓ) of H, then it can be shown that

$$H_{k\ell} = \sum_{r=1}^{d}\sum_{s=1}^{d} \Omega_{rs} J_{rk} J_{s\ell} \Phi_r^\mathsf{T} \Phi_s + T\lambda\, \mathbf{1}_{\{k=\ell\}} I_p. \tag{71}$$

where

$$D_\lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d), \qquad J = D_\lambda^{-1/2} L_\rho, \qquad \Omega = R^\mathsf{T} U_S U_S^\mathsf{T} R = R^\mathsf{T} R, \tag{72}$$

and $\Phi_r \in \mathbb{R}^{T\times p}$ the feature matrix of series r, whose t-th row is $\phi(x_{t,r})^\mathsf{T}$.

To increase the convergence speed of conjugate gradients, we use a preconditioner $M_\rho \approx H^{-1}$, which maps the residual r_j to $z_j = M_\rho r_j$. Following the general principle of approximate-inverse preconditioning for kernel systems (Cutajar et al., 2016; Rudi et al., 2017), we use one of two approximations depending on whether the kernel contains task coupling. When $\rho > 0$, we use the following Kronecker approximation of H:

$$\widetilde{H}_{\text{task}} := A_{\text{task}} \otimes I_p \approx H, \qquad A_{\text{task}} \in \mathbb{R}^{q\times q}. \tag{73}$$

where

$$(A_{\text{task}})_{k\ell} = \frac{1}{p}\mathrm{tr}\,(H_{k\ell}) = \sum_{r=1}^{d}\sum_{s=1}^{d} \Omega_{rs} J_{rk} J_{s\ell} \frac{1}{p}\mathrm{tr}\left(\Phi_r^\mathsf{T}\Phi_s\right) + T\lambda\, \mathbf{1}_{\{k=\ell\}}. \tag{74}$$

It follows that

$$A_{\text{task}} = J^\mathsf{T}(\Omega \odot C_X) J + T\lambda I_q. \tag{75}$$

where $(C_X)_{rs} = \frac{1}{p}\mathrm{tr}\left(\Phi_r^\mathsf{T}\Phi_s\right)$. Thus the task-only preconditioner is

$$M_{\text{task}} = \widetilde{H}_{\text{task}}^{-1} = A_{\text{task}}^{-1} \otimes I_p. \tag{76}$$

It retains all interactions between the q task factors and treats the feature directions as isotropic. When $\rho = 0$, there is no task-factor mixing: $q = d$, $L_0 = I_d$, and $J = D_\lambda^{-1/2}$. We instead use a feature-only block-Jacobi approximation that discards the off-diagonal blocks of H, i.e.,

$$\widetilde{H}_{\text{feature}} := \mathrm{blockdiag}(H_{11}, \ldots, H_{qq}) \approx H. \tag{77}$$

In this case, Equation (71) shows that its diagonal blocks are exactly

$$A_k := H_{kk} = \frac{\Omega_{kk}}{\lambda_k}\Phi_k^\mathsf{T}\Phi_k + T\lambda I_p. \tag{78}$$

The feature-only preconditioner is therefore

$$M_{\text{feature}} = \mathrm{blockdiag}(A_1^{-1}, \ldots, A_q^{-1}). \tag{79}$$

Indeed, $\Omega \succeq 0$ and $C_X \succeq 0$, so the Schur product theorem and $T\lambda > 0$ imply $A_{\text{task}} \succ 0$ and $A_k \succ 0$ for every k; consequently, both M_{task} and M_{feature} are symmetric positive-definite and are valid PCG preconditioners.

Initialization Instead of starting the conjugate gradient method from the zero vector, we propose to use a warm start based on the independently fitted KRR coefficients, which are denoted by

$$\beta_r = \left(\Phi_r^\mathsf{T}\Phi_r + T\lambda\,\lambda_r I_p\right)^{-1} \Phi_r^\mathsf{T} y_r. \tag{80}$$

where $y_r = (y_{1,r}, \ldots, y_{T,r})^\mathsf{T} \in \mathbb{R}^T$ and $\Phi_r \in \mathbb{R}^{T\times p}$ has rows $\phi(x_{t,r})^\mathsf{T}$ as we defined before. Specifically, we begin from $\theta_0 = \mathrm{vec_r}(\Theta_0)$ where Θ_0 is constructed as follows:

$$\Theta_0 := L_\rho^\dagger \begin{bmatrix} \sqrt{\lambda_1}\beta_1^\mathsf{T} \\ \vdots \\ \sqrt{\lambda_d}\beta_d^\mathsf{T} \end{bmatrix}, \tag{81}$$

where $L_\rho^\dagger$ is the Moore–Penrose inverse. As a result, for $\rho = 0$, $L_0 = I_d$, it can be shown that Θ_0 reproduces the d independent-KRR predictors before the coherent transformation $U_S^\mathsf{T} R$ is applied since

$$(\Gamma_\rho(x)\Theta_0)_r = \lambda_r^{-1/2}(\Theta_0)_{r,:}\phi(x_r) = \beta_r^\mathsf{T}\phi(x_r). \tag{82}$$

For $\rho > 0$, Θ_0 minimizes $\|L_\rho\Theta - W\|^2_{\mathbb{R}^{d\times p}}$ where $W = \begin{bmatrix}\sqrt{\lambda_1}\beta_1 & \cdots & \sqrt{\lambda_d}\beta_d\end{bmatrix}^\mathsf{T}$.

Once PCG returns $\widehat{\theta}$, we reshape it as $\widehat{\Theta} = \mathrm{mat}_\mathrm{r}(\widehat{\theta}) \in \mathbb{R}^{q_\rho \times p_\phi}$ and predict, for any $x \in \mathcal{X}$,

$$\widehat{f}(x) = R\Gamma_\rho(x)\widehat{\Theta} = U_S\overline{\Gamma}_{R,\rho}(x)\widehat{\Theta} \in \mathrm{Im}(S), \tag{83}$$

so coherence holds by construction.

C. Excess-Risk Bounds and Proofs

Notation for linear operators. Let $\mathcal{H}$ and $\mathcal{H}'$ be two Hilbert spaces. Let $(e_k)_{k\in I}$ be any orthonormal basis of $\mathcal{H}$.

Let $\mathcal{B}(\mathcal{H},\mathcal{H}')$ be the space of bounded linear operators from $\mathcal{H}$ to $\mathcal{H}'$. It's a normed space endowed with the operator norm $\|\cdot\|_{\mathrm{op}}$ defined by $\|A\|_{\mathrm{op}} \coloneqq \sup_{h\in\mathcal{H},\|h\|_\mathcal{H}\leq 1}\|Ah\|_{\mathcal{H}'}$. Moreover, for any bounded operators A, B, when the output space of B is the input space of A, we have

$$\|AB\|_{\mathrm{op}} \leq \|A\|_{\mathrm{op}}\|B\|_{\mathrm{op}}. \tag{84}$$

We define also $|A| \coloneqq (A^*A)^{\frac{1}{2}}$.

Let $\mathcal{B}_0(\mathcal{H},\mathcal{H}')$ be the space of compact bounded linear operators from $\mathcal{H}$ to $\mathcal{H}'$.

Let $\mathcal{B}_1(\mathcal{H},\mathcal{H}')$ be the space of Schatten 1-class or trace class operators from $\mathcal{H}$ to $\mathcal{H}'$, i.e.,

$$\mathcal{B}_1(\mathcal{H},\mathcal{H}') = \left\{A \in \mathcal{B}_0(\mathcal{H},\mathcal{H}') \mid \sum_{k\in I}\langle|A|e_k, e_k\rangle_\mathcal{H} < \infty\right\}. \tag{85}$$

It is a Banach space endowed with the so-called trace class norm $\|\cdot\|_1$ defined by $\|A\|_1 \coloneqq \sum_{k\in I}\langle|A|e_k,e_k\rangle_\mathcal{H} = \sum_{k\in I}\sigma_k(A)$, where $(\sigma_k(A))_{k\in I}$ are the singular values of A. For a trace class operator $A \in \mathcal{B}_1(\mathcal{H})$, its trace is well-defined: $\mathrm{tr}\,(A) \coloneqq \sum_{l\in I}\langle Ae_l, e_l\rangle_\mathcal{H} \leq \|A\|_1 < \infty$.

Let $\mathcal{B}_2(\mathcal{H},\mathcal{H}')$ be the space of Schatten 2-class or Hilbert-Schmidt operators from $\mathcal{H}$ to $\mathcal{H}'$, i.e.,

$$\mathcal{B}_2(\mathcal{H},\mathcal{H}') = \left\{A \in \mathcal{B}_0(\mathcal{H},\mathcal{H}') \mid \sum_{k\in I}\left\langle|A|^2e_k, e_k\right\rangle_\mathcal{H} < \infty\right\} \tag{86}$$

Then we can define a so-called Hilbert-Schmidt norm on $\mathcal{B}_2(\mathcal{H},\mathcal{H}')$ as follows: $\|A\|_{\mathrm{HS}} \coloneqq \left(\sum_{k\in I}\left\langle|A|^2e_k,e_k\right\rangle_\mathcal{H}\right)^{1/2} = \left(\sum_{k\in I}\|Ae_k\|^2_{\mathcal{H}'}\right)^{1/2} = \left(\sum_{k\in I}\sigma_k^2(A)\right)^{1/2}$. For any $(A,B) \in \mathcal{B}_2(\mathcal{H},\mathcal{H}')$, A^*B is in $\mathcal{B}_1(\mathcal{H})$. As a result, $\mathrm{tr}\,(A^*B)$ is well-defined. We can therefore endow $\mathcal{B}_2(\mathcal{H},\mathcal{H}')$ with an inner product defined by $\langle A,B\rangle_2 \coloneqq \mathrm{tr}\,(A^*B)$, and show that it is a Hilbert space. The corresponding Hilbert-Schmidt norm is denoted by $\|\cdot\|_2$ and we have $\|A\|_2 = \sqrt{\mathrm{tr}\,(A^*A)} = \|A\|_{\mathrm{HS}}$.

One can show that a trace-class operator is automatically Hilbert-Schmidt, so we have

$$\mathcal{B}_1(\mathcal{H},\mathcal{H}') \subset \mathcal{B}_2(\mathcal{H},\mathcal{H}') \subset \mathcal{B}_0(\mathcal{H},\mathcal{H}'). \tag{87}$$

Moreover, for all $A \in \mathcal{B}_1(\mathcal{H},\mathcal{H}')$, one has

$$\|A\|_{\mathrm{op}} \leq \|A\|_2 \leq \|A\|_1. \tag{88}$$

And for all $A \in \mathcal{B}(\mathcal{H})$, $B \in \mathcal{B}_p(\mathcal{H})$ with $p \in \{1,2\}$, we have

$$\|AB\|_p \leq \|A\|_{\mathrm{op}}\|B\|_p. \tag{89}$$

Before presenting the main results of this section, we recall that it can be shown that the estimator of the RVVKRR can also be written in the following form:

$$\hat{f} = (\hat{\Sigma}_{\mathcal{K}} + \lambda I)^{-1} \frac{1}{T} \sum_{t=1}^{T} \mathcal{K}_{x_t} y_t \tag{90}$$

with

$$\hat{\Sigma}_{\mathcal{K}} \coloneqq \frac{1}{T} \sum_{t=1}^{T} \mathcal{K}_{x_t} \mathcal{K}_{x_t}^{*}. \tag{91}$$

We introduce f_λ which is a solution of the following optimization problem:

$$f_\lambda \in \operatorname*{arg\,min}_{f \in \mathcal{H}_{\mathcal{K}}} \left\{ \|f - S \circ g^\star\|_\rho^2 + \lambda \|f\|_{\mathcal{H}_{\mathcal{K}}}^2 \right\}. \tag{92}$$

Since $\lambda > 0$, a unique minimizer f_λ exists, and is given by (see Proposition 1 of Caponnetto and De Vito (2007) for example):

$$f_\lambda = (\Sigma_{\mathcal{K}} + \lambda I)^{-1} \int_{\mathcal{X}} \mathcal{K}_x S \circ g^\star(x) d\rho(x) \tag{93}$$

where

$$\Sigma_{\mathcal{K}} \coloneqq \int_{\mathcal{X}} \mathcal{K}_x \mathcal{K}_x^{*} d\rho(x). \tag{94}$$

To ensure that the Bochner integral in Equation (93) is well-defined, we assume additionally that $\mathcal{H}_{\mathcal{K}}$ is separable, and $(x, x') \mapsto \langle \mathcal{K}_x y, \mathcal{K}_{x'} y' \rangle_{\mathcal{H}_{\mathcal{K}}}$ is measurable for all $y, y' \in \mathcal{Y}$ (see, e.g., Caponnetto and De Vito, 2007, Hypothesis 1).

C.1. Auxiliary Results

First, we show some useful lemmas that are the direct consequences of the assumptions we made in Section 4.1.

Lemma C.1 *Under Assumption 1, we have*

$$\|S \circ g^\star\|_\rho \leq \|S\|_{\text{op}} \sqrt{C_1} \tag{95}$$

$$\|f_\lambda - S \circ g^\star\|_\rho \leq \|S\|_{\text{op}} \sqrt{C_1} \tag{96}$$

$$\|f_\lambda\|_{\mathcal{H}_{\mathcal{K}}} \leq \|S\|_{\text{op}} \sqrt{\frac{C_1}{\lambda}}. \tag{97}$$

Proof By definition of f_λ (Equation (92)), we have

$$\|f_\lambda - S \circ g^\star\|_\rho^2 + \lambda \|f_\lambda\|_{\mathcal{H}_{\mathcal{K}}}^2 \leq \|S \circ g^\star\|_\rho^2. \tag{98}$$

The fact that

$$\|S \circ g^\star\|_\rho^2 = \int_{\mathcal{X}} \|S \circ g^\star(x)\|_{\mathcal{Y}}^2 \, d\rho(x) \leq \|S\|_{\text{op}}^2 \, C_1 \tag{99}$$

concludes the proof. ■

Lemma C.2 *Under Assumption 2, $\mathcal{K}_x^* \mathcal{K}_x$ is a trace-class operator, and we have*

$$\|\mathcal{K}_x^* \mathcal{K}_x\|_{\text{op}} \leq \|\mathcal{K}_x^* \mathcal{K}_x\|_2 \leq \|\mathcal{K}_x^* \mathcal{K}_x\|_1 \leq \kappa \tag{100}$$

Proof Since $\mathcal{K}_x^* \mathcal{K}_x$ is a self-adjoint, positive operator, we have $|\mathcal{K}_x^* \mathcal{K}_x| = \mathcal{K}_x^* \mathcal{K}_x$. Then by definition of the trace norm and the trace, $\|\mathcal{K}_x^* \mathcal{K}_x\|_1 = \operatorname{tr}(\mathcal{K}_x^* \mathcal{K}_x) \leq \kappa$. The other inequalities are due to the properties of the trace-class operator. ■

Lemma C.3 *Under Assumption 3, it can be shown that*

$$\mathcal{R}(\hat{f}) - \mathcal{R}(S \circ g^\star) = \mathbb{E}\left[\left\|\hat{f}(\mathsf{X}) - S \circ g^\star(\mathsf{X})\right\|_{\mathcal{Y}}^2\right] = \left\|\hat{f} - S \circ g^\star\right\|_\rho^2. \tag{101}$$

The source condition in Assumption 4 is standard in the kernel literature and controls the smoothness of the target function. It is convenient for bounding the approximation error through the following lemma, which is similar to Proposition 3 in Caponnetto and De Vito (2007). The scalar-valued kernel version and its proof can be found in Theorem 4 of Smale and Zhou (2005); we omit the proof here.

Lemma C.4 *Under Assumption 4, we have*

$$\|f_\lambda - S \circ g^\star\|_\rho \leq \lambda^r \left\|L_{\mathcal{K}}^{-r}(S \circ g^\star)\right\|_\rho. \tag{102}$$

Before giving the explicit excess-risk bound and its proof, we present several useful auxiliary results. We first restate Bennett's inequality for Hilbert-space-valued random variables. This is a restatement of Lemma 2 in Smale and Zhou (2007), derived from Theorem 3.4 of Pinelis (1994).

Proposition C.5 (Bennett Inequality in Hilbert Space) *Let $\mathcal{H}$ be a separable Hilbert space and let Z be a random variable on a probability space $(\Omega, \mathcal{F}, \rho)$ with values in $\mathcal{H}$. Assume that $\|\mathsf{Z}\|_{\mathcal{H}} \leq M < \infty$ almost surely. Denote $\sigma^2 = \mathbb{E}\left[\|\mathsf{Z}\|_{\mathcal{H}}^2\right]$. Let $\{\omega_i\}_{i=1}^n$ be independent draws from ρ. For any $0 < \delta < 1$, with a probability greater than $1 - \delta$,*

$$\left\|\frac{1}{n}\sum_{i=1}^n (\mathsf{Z}(\omega_i) - \mathbb{E}[\mathsf{Z}])\right\|_{\mathcal{H}} \leq \frac{2M\log(2/\delta)}{n} + \sqrt{\frac{2\sigma^2\log(2/\delta)}{n}} \leq 2\log\left(\frac{2}{\delta}\right)\left(\frac{M}{n} + \sqrt{\frac{\sigma^2}{n}}\right) \tag{103}$$

The second inequality is due to the fact that $\sqrt{2\log(2/\delta)} < 2\log(2/\delta)$.

Using Proposition C.5, the following two lemmas give some concentration results that are useful to prove the main proposition.

Lemma C.6 *Let $0 < \delta < 1$. Under Assumption 2, with a probability greater than $1 - \delta$, we have*

$$\left\|\hat{\Sigma}_{\mathcal{K}} - \Sigma_{\mathcal{K}}\right\|_2 \leq 4\log\left(\frac{2}{\delta}\right)\frac{\kappa}{\sqrt{T}}. \tag{104}$$

Moreover, we have

$$\left\|(\Sigma_{\mathcal{K}} + \lambda I)^{-1}\left(\hat{\Sigma}_{\mathcal{K}} - \Sigma_{\mathcal{K}}\right)\right\|_2 \leq \frac{1}{2} \tag{105}$$

when $\lambda \geq 8\log\left(\frac{2}{\delta}\right)\frac{\kappa}{\sqrt{T}}$.

Remark C.7 *The above lemma can be seen as the generalization of Lemma 4 in Smale and Zhou (2007) in the case of OVK, and we give here a more concise proof.*

Proof We consider the random variable $\mathsf{Z} := \mathcal{K}_{\mathsf{X}}\mathcal{K}_{\mathsf{X}}^*$ valued in $\mathcal{B}_2(\mathcal{H}_{\mathcal{K}})$. Then by Lemma C.2, $\|\mathsf{Z}\|_2 = \|\mathcal{K}_{\mathsf{X}}\mathcal{K}_{\mathsf{X}}^*\|_2 = \|\mathcal{K}_{\mathsf{X}}^*\mathcal{K}_{\mathsf{X}}\|_2 \leq \kappa$ and $\mathbb{E}\left[\|\mathsf{Z}\|_2^2\right] \leq \kappa^2$. Since $\mathcal{B}_2(\mathcal{H}_{\mathcal{K}})$ is a Hilbert space, applying Proposition C.5 to Z, we obtain:

$$\left\|\hat{\Sigma}_{\mathcal{K}} - \Sigma_{\mathcal{K}}\right\|_2 \leq 2\log\left(\frac{2}{\delta}\right)\left(\frac{\kappa}{T} + \frac{\kappa}{\sqrt{T}}\right) \leq 4\log\left(\frac{2}{\delta}\right)\frac{\kappa}{\sqrt{T}}.$$

For the second bound, by Equation (89),

$$\begin{aligned}\left\|(\Sigma_{\mathcal{K}} + \lambda I)^{-1}\left(\hat{\Sigma}_{\mathcal{K}} - \Sigma_{\mathcal{K}}\right)\right\|_2 &\leq \left\|(\Sigma_{\mathcal{K}} + \lambda I)^{-1}\right\|_{\text{op}}\left\|\hat{\Sigma}_{\mathcal{K}} - \Sigma_{\mathcal{K}}\right\|_2 \\ &\leq 4\log\left(\frac{2}{\delta}\right)\frac{\kappa}{\lambda\sqrt{T}}\end{aligned}$$

since $\left\|(\Sigma_{\mathcal{K}} + \lambda I)^{-1}\right\|_{\text{op}} \leq \frac{1}{\lambda}$. And the proof is concluded by setting $4\log\left(\frac{2}{\delta}\right)\frac{\kappa}{\lambda\sqrt{T}} \leq \frac{1}{2}$. ■

Lemma C.8 *Let $0 < \delta < 1$ and let $\lambda > 0$. Under Assumptions 1 to 3, with a probability greater than $1 - \delta$, we have*

$$\left\| (\Sigma_{\mathcal{K}} + \lambda I)^{-\frac{1}{2}} \left(\frac{1}{T} \sum_{t=1}^{T} \mathcal{K}_{x_t} S(g^\star(x_t) + \xi_t) - \hat{\Sigma}_{\mathcal{K}} f_\lambda - \left(\int_{\mathcal{X}} \mathcal{K}_x S \circ g^\star(x) d\rho(x) - \Sigma_{\mathcal{K}} f_\lambda \right) \right) \right\|_{\mathcal{H}_{\mathcal{K}}}$$
$$\leq 2 \log\left(\frac{2}{\delta}\right) \left(\frac{1}{T} \sqrt{\frac{\kappa}{\lambda}} \|S\|_{\mathrm{op}} \left(\left(1 + \sqrt{\frac{\kappa}{\lambda}} \right) \sqrt{C_1} + \sqrt{C_2} \right) + \sqrt{\frac{2 \left(\mathcal{N}(\lambda) \mathbb{E}\left[\|S\xi\|_{\mathcal{Y}}^2 \right] + \frac{\kappa}{\lambda} \|S\|_{\mathrm{op}}^2 C_1 \right)}{T}} \right)$$

with $\mathcal{N}(\lambda) \coloneqq \operatorname{tr}\left((\Sigma_{\mathcal{K}} + \lambda I)^{-1} \Sigma_{\mathcal{K}} \right)$. $\mathcal{N}(\lambda)$ is often referred to as the effective dimension in the literature. Here, we write $y_t = S(g^\star(x_t) + \xi_t)$ where (x_t, ξ_t) are i.i.d. samples of (X, ξ).

Proof To apply Proposition C.5, we define $\mathsf{Z} \coloneqq (\Sigma_{\mathcal{K}} + \lambda I)^{-\frac{1}{2}} \mathcal{K}_{\mathsf{X}} \left(S\xi + S \circ g^\star(\mathsf{X}) - \mathcal{K}_{\mathsf{X}}^* f_\lambda \right)$ valued in $\mathcal{H}_{\mathcal{K}}$. First, we try to bound the $\mathcal{H}_{\mathcal{K}}$-norm of the random variable Z:

$$\begin{aligned}
\|\mathsf{Z}\|_{\mathcal{H}_{\mathcal{K}}} &= \left\| (\Sigma_{\mathcal{K}} + \lambda I)^{-\frac{1}{2}} \mathcal{K}_{\mathsf{X}} \left(S\xi + S \circ g^\star(\mathsf{X}) - \mathcal{K}_{\mathsf{X}}^* f_\lambda \right) \right\|_{\mathcal{H}_{\mathcal{K}}} \\
&\leq \left\| (\Sigma_{\mathcal{K}} + \lambda I)^{-\frac{1}{2}} \right\|_{\mathrm{op}} \|\mathcal{K}_{\mathsf{X}}\|_{\mathrm{op}} \|S\xi + S \circ g^\star(\mathsf{X}) - \mathcal{K}_{\mathsf{X}}^* f_\lambda\|_{\mathcal{Y}} \\
&\leq \left\| (\Sigma_{\mathcal{K}} + \lambda I)^{-\frac{1}{2}} \right\|_{\mathrm{op}} \sqrt{\kappa} \|S\xi + S \circ g^\star(\mathsf{X}) - \mathcal{K}_{\mathsf{X}}^* f_\lambda\|_{\mathcal{Y}} \qquad \text{(since } \|\mathcal{K}_{\mathsf{X}}\|_{\mathrm{op}} \leq \|\mathcal{K}_{\mathsf{X}}\|_2\text{)} \\
&\leq \sqrt{\frac{\kappa}{\lambda}} \|S\xi + S \circ g^\star(\mathsf{X}) - \mathcal{K}_{\mathsf{X}}^* f_\lambda\|_{\mathcal{Y}} \\
&\leq \sqrt{\frac{\kappa}{\lambda}} \left(\|S\xi\|_{\mathcal{Y}} + \|S \circ g^\star(\mathsf{X})\|_{\mathcal{Y}} + \|\mathcal{K}_{\mathsf{X}}^* f_\lambda\|_{\mathcal{Y}} \right) \\
&\leq \sqrt{\frac{\kappa}{\lambda}} \left(\|S\|_{\mathrm{op}} \|\xi\|_{\mathcal{Y}_b} + \|S\|_{\mathrm{op}} \|g^\star(\mathsf{X})\|_{\mathcal{Y}_b} + \|\mathcal{K}_{\mathsf{X}}^* f_\lambda\|_{\mathcal{Y}} \right) \\
&\leq \sqrt{\frac{\kappa}{\lambda}} \left(\|S\|_{\mathrm{op}} \sqrt{C_2} + \|S\|_{\mathrm{op}} \sqrt{C_1} + \|\mathcal{K}_{\mathsf{X}}^* f_\lambda\|_{\mathcal{Y}} \right) \qquad \text{(by Assumption 1).}
\end{aligned}$$

Lemma C.1 and Assumption 2 give $\|\mathcal{K}_{\mathsf{X}}^* f_\lambda\|_{\mathcal{Y}} \leq \|\mathcal{K}_{\mathsf{X}}^*\|_{\mathrm{op}} \|f_\lambda\|_{\mathcal{H}_{\mathcal{K}}} \leq \|S\|_{\mathrm{op}} \sqrt{\frac{\kappa C_1}{\lambda}}$. As a result, we have

$$\|\mathsf{Z}\|_{\mathcal{H}_{\mathcal{K}}} \leq \sqrt{\frac{\kappa}{\lambda}} \|S\|_{\mathrm{op}} \left(\left(1 + \sqrt{\frac{\kappa}{\lambda}} \right) \sqrt{C_1} + \sqrt{C_2} \right).$$

Then we focus on obtaining an upper bound of $\mathbb{E}\left[\|\mathsf{Z}\|_{\mathcal{H}_{\mathcal{K}}}^2 \right]$. Denoting $\mathsf{U} \coloneqq S\xi + S \circ g^\star(\mathsf{X}) - \mathcal{K}_{\mathsf{X}}^* f_\lambda$ valued in $\mathcal{Y}$, we can rewrite $\mathbb{E}\left[\|\mathsf{Z}\|_{\mathcal{H}_{\mathcal{K}}}^2 \right]$ as follows:

$$\mathbb{E}\left[\|\mathsf{Z}\|_{\mathcal{H}_{\mathcal{K}}}^2 \right] = \mathbb{E}\left[\left\langle \mathcal{K}_{\mathsf{X}}^* (\Sigma_{\mathcal{K}} + \lambda I)^{-1} \mathcal{K}_{\mathsf{X}} \mathsf{U}, \mathsf{U} \right\rangle_{\mathcal{Y}} \right].$$

It can be shown that $\mathcal{K}_{\mathsf{X}}^* (\Sigma_{\mathcal{K}} + \lambda I)^{-1} \mathcal{K}_{\mathsf{X}}$ is self-adjoint, which yields:

$$\mathbb{E}\left[\|\mathsf{Z}\|_{\mathcal{H}_{\mathcal{K}}}^2 \right] \leq \mathbb{E}\left[\left\| \mathcal{K}_{\mathsf{X}}^* (\Sigma_{\mathcal{K}} + \lambda I)^{-1} \mathcal{K}_{\mathsf{X}} \right\|_{\mathrm{op}} \|\mathsf{U}\|_{\mathcal{Y}}^2 \right].$$

The fact that $\|\mathsf{U}\|_{\mathcal{Y}}^2 \leq 2 \|S\xi\|_{\mathcal{Y}}^2 + 2 \|S \circ g^\star(\mathsf{X}) - \mathcal{K}_{\mathsf{X}}^* f_\lambda\|_{\mathcal{Y}}^2$ results in

$$\mathbb{E}\left[\|\mathsf{Z}\|_{\mathcal{H}_{\mathcal{K}}}^2 \right] \leq 2 \left(\mathbb{E}\left[\left\| \mathcal{K}_{\mathsf{X}}^* (\Sigma_{\mathcal{K}} + \lambda I)^{-1} \mathcal{K}_{\mathsf{X}} \right\|_{\mathrm{op}} \|S\xi\|_{\mathcal{Y}}^2 \right] + \mathbb{E}\left[\left\| \mathcal{K}_{\mathsf{X}}^* (\Sigma_{\mathcal{K}} + \lambda I)^{-1} \mathcal{K}_{\mathsf{X}} \right\|_{\mathrm{op}} \|S \circ g^\star(\mathsf{X}) - \mathcal{K}_{\mathsf{X}}^* f_\lambda\|_{\mathcal{Y}}^2 \right] \right).$$

It follows that

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathcal{K}_{\mathsf{X}}^*(\Sigma_{\mathcal{K}}+\lambda I)^{-1}\mathcal{K}_{\mathsf{X}}\right\|_{\mathrm{op}}\|S\xi\|_{\mathcal{Y}}^2\right]\\
=&\mathbb{E}\left[\left\|\mathcal{K}_{\mathsf{X}}^*(\Sigma_{\mathcal{K}}+\lambda I)^{-1}\mathcal{K}_{\mathsf{X}}\right\|_{\mathrm{op}}\right]\mathbb{E}\left[\|S\xi\|_{\mathcal{Y}}^2\right] && \text{(by Assumption 3)}\\
\leq&\mathbb{E}\left[\mathrm{tr}\left(\mathcal{K}_{\mathsf{X}}^*(\Sigma_{\mathcal{K}}+\lambda I)^{-1}\mathcal{K}_{\mathsf{X}}\right)\right]\mathbb{E}\left[\|S\xi\|_{\mathcal{Y}}^2\right] && \text{(since } \mathcal{K}_{\mathsf{X}}^*(\Sigma_{\mathcal{K}}+\lambda I)^{-1}\mathcal{K}_{\mathsf{X}} \text{ is positive)}\\
=&\mathbb{E}\left[\mathrm{tr}\left((\Sigma_{\mathcal{K}}+\lambda I)^{-1}\mathcal{K}_{\mathsf{X}}\mathcal{K}_{\mathsf{X}}^*\right)\right]\mathbb{E}\left[\|S\xi\|_{\mathcal{Y}}^2\right]\\
=&\mathcal{N}(\lambda)\mathbb{E}\left[\|S\xi\|_{\mathcal{Y}}^2\right] && \text{(by definition of } \mathcal{N}(\lambda))
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathcal{K}_{\mathsf{X}}^*(\Sigma_{\mathcal{K}}+\lambda I)^{-1}\mathcal{K}_{\mathsf{X}}\right\|_{\mathrm{op}}\|S\circ g^\star(\mathsf{X})-\mathcal{K}_{\mathsf{X}}^* f_\lambda\|_{\mathcal{Y}}^2\right]\\
\leq&\mathbb{E}\left[\|\mathcal{K}_{\mathsf{X}}^*\|_{\mathrm{op}}\left\|(\Sigma_{\mathcal{K}}+\lambda I)^{-1}\right\|_{\mathrm{op}}\|\mathcal{K}_{\mathsf{X}}\|_{\mathrm{op}}\|S\circ g^\star(\mathsf{X})-\mathcal{K}_{\mathsf{X}}^* f_\lambda\|_{\mathcal{Y}}^2\right] && \text{(Equation (84))}\\
\leq&\mathbb{E}\left[\frac{\kappa}{\lambda}\|S\circ g^\star(\mathsf{X})-\mathcal{K}_{\mathsf{X}}^* f_\lambda\|_{\mathcal{Y}}^2\right]\\
=&\frac{\kappa}{\lambda}\|S\circ g^\star-f_\lambda\|_\rho^2\\
\leq&\frac{\kappa\|S\|_{\mathrm{op}}^2 C_1}{\lambda} && \text{(Lemma C.1)}
\end{aligned}$$

As a result, we obtain

$$\mathbb{E}\left[\|\mathsf{Z}\|_{\mathcal{H}_\mathcal{K}}^2\right]\leq 2\left(\mathcal{N}(\lambda)\mathbb{E}\left[\|S\xi\|_{\mathcal{Y}}^2\right]+\frac{\kappa\|S\|_{\mathrm{op}}^2 C_1}{\lambda}\right).$$

We conclude by applying Proposition C.5 to the random variable Z, using the following expression for $\mathbb{E}[\mathsf{Z}]$:

$$\begin{aligned}
\mathbb{E}[\mathsf{Z}]&=(\Sigma_{\mathcal{K}}+\lambda I)^{-\frac{1}{2}}\left(\mathbb{E}[\mathcal{K}_{\mathsf{X}}S\xi]+\mathbb{E}[\mathcal{K}_{\mathsf{X}}S\circ g^\star(\mathsf{X})]-\mathbb{E}[\mathcal{K}_{\mathsf{X}}\mathcal{K}_{\mathsf{X}}^* f_\lambda]\right)\\
&=(\Sigma_{\mathcal{K}}+\lambda I)^{-\frac{1}{2}}\left(\mathbb{E}[\mathcal{K}_{\mathsf{X}}S\circ g^\star(\mathsf{X})]-\mathbb{E}[\mathcal{K}_{\mathsf{X}}\mathcal{K}_{\mathsf{X}}^* f_\lambda]\right) && \text{(Assumption 3)}\\
&=(\Sigma_{\mathcal{K}}+\lambda I)^{-\frac{1}{2}}\left(\int_{\mathcal{X}}\mathcal{K}_x S\circ g^\star(x)d\rho(x)-\Sigma_{\mathcal{K}} f_\lambda\right)
\end{aligned}$$

■

We also need the following lemma which links the norms $\|\cdot\|_\rho$ and $\|\cdot\|_{\mathcal{H}_\mathcal{K}}$.

Lemma C.9 *For any $(f,g)\in\mathcal{H}_\mathcal{K}^2$, we have*

$$\langle f,g\rangle_\rho=\langle f,\Sigma_{\mathcal{K}}g\rangle_{\mathcal{H}_\mathcal{K}} \tag{106}$$

with $\Sigma_{\mathcal{K}}\coloneqq\int_{\mathcal{X}}\mathcal{K}_x\mathcal{K}_x^* d\rho(x)\in\mathcal{L}(\mathcal{H}_\mathcal{K})$. As a result, $\|f\|_\rho=\left\|\Sigma_{\mathcal{K}}^{1/2}f\right\|_{\mathcal{H}_\mathcal{K}}$.

Proof By the definition of the $L^2(\mathcal{X},\rho,\mathcal{Y})$-inner product, we have, for any $(f,g)\in\mathcal{H}_\mathcal{K}^2$,

$$\begin{aligned}
\langle f,g\rangle_\rho&=\int_{\mathcal{X}}\langle f(x),g(x)\rangle_{\mathcal{Y}}\,d\rho(x)\\
&=\int_{\mathcal{X}}\langle\mathcal{K}_x^* f,\mathcal{K}_x^* g\rangle_{\mathcal{Y}}\,d\rho(x)\\
&=\int_{\mathcal{X}}\langle f,\mathcal{K}_x\mathcal{K}_x^* g\rangle_{\mathcal{H}_\mathcal{K}}\,d\rho(x)\\
&=\left\langle f,\left(\int_{\mathcal{X}}\mathcal{K}_x\mathcal{K}_x^* d\rho(x)\right)g\right\rangle_{\mathcal{H}_\mathcal{K}}
\end{aligned}$$

The proof is concluded with the definition of $\Sigma_{\mathcal{K}}$. ■

C.2. Explicit Excess-Risk Bound

Proposition C.10 (Explicit excess-risk bound) *Under Assumptions 1 to 4, let* $0 < \delta < 1$*, then with a probability greater than* $1 - \delta$*,*

$$\mathcal{R}(\hat{f}) - \mathcal{R}(S \circ g^\star) \leq 128 \log^2\left(\frac{4}{\delta}\right) \left[\frac{\kappa \|S\|_{\text{op}}^2}{\lambda T^2}\left(\left(1 + \sqrt{\frac{\kappa}{\lambda}}\right)^2 C_1 + C_2\right) + \frac{1}{T}\left(\mathcal{N}(\lambda)\mathbb{E}\left[\|S\xi\|_{\mathcal{Y}}^2\right] + \frac{\kappa}{\lambda}\|S\|_{\text{op}}^2 C_1\right)\right] + 2\lambda^{2r}\left\|L_{\mathcal{K}}^{-r}(S \circ g^\star)\right\|_\rho^2,$$

provided that

$$\lambda \geq 8 \log\left(\frac{4}{\delta}\right) \frac{\kappa}{\sqrt{T}}. \tag{107}$$

Proof First we decompose $\hat{f} - S \circ g^\star$ by

$$\hat{f} - S \circ g^\star = \hat{f} - f_\lambda + f_\lambda - S \circ g^\star.$$

From Equation (101) and due to the inequality $\|f_1 + f_2\|_\rho^2 \leq 2(\|f_1\|_\rho^2 + \|f_2\|_\rho^2)$ for any $(f_1, f_2) \in L^2(\mathcal{X}, \rho, \mathcal{Y})^2$, the excess risk can be upper bounded by

$$\mathcal{R}(\hat{f}) - \mathcal{R}(S \circ g^\star) \leq 2\left(\left\|\hat{f} - f_\lambda\right\|_\rho^2 + \|f_\lambda - S \circ g^\star\|_\rho^2\right)$$

The first step of the proof is dedicated to obtain the probabilistic bound on $\left\|\hat{f} - f_\lambda\right\|_\rho$. By the definition of $\hat{f}$, we have

$$\begin{aligned}
\hat{f} - f_\lambda &= (\hat{\Sigma}_{\mathcal{K}} + \lambda I)^{-1} \frac{1}{T}\sum_{t=1}^T \mathcal{K}_{x_t} S(g^\star(x_t) + \xi_t) - f_\lambda \\
&= (\hat{\Sigma}_{\mathcal{K}} + \lambda I)^{-1}\left(\frac{1}{T}\sum_{t=1}^T \mathcal{K}_{x_t} S(g^\star(x_t) + \xi_t) - (\hat{\Sigma}_{\mathcal{K}} + \lambda I) f_\lambda\right) \\
&= (\hat{\Sigma}_{\mathcal{K}} + \lambda I)^{-1}\left(\frac{1}{T}\sum_{t=1}^T \mathcal{K}_{x_t} S(g^\star(x_t) + \xi_t) - (\hat{\Sigma}_{\mathcal{K}} - \Sigma_{\mathcal{K}}) f_\lambda - (\Sigma_{\mathcal{K}} + \lambda I) f_\lambda\right)
\end{aligned}$$

It follows that

$$\begin{aligned}
&\hat{f} - f_\lambda \\
=&(\hat{\Sigma}_{\mathcal{K}} + \lambda I)^{-1}\left(\frac{1}{T}\sum_{t=1}^T \mathcal{K}_{x_t} S(g^\star(x_t) + \xi_t) - (\hat{\Sigma}_{\mathcal{K}} - \Sigma_{\mathcal{K}}) f_\lambda - \int_{\mathcal{X}} \mathcal{K}_x S \circ g^\star(x) d\rho(x)\right) \\
=&(\hat{\Sigma}_{\mathcal{K}} + \lambda I)^{-1}\left(\frac{1}{T}\sum_{t=1}^T \mathcal{K}_{x_t} S(g^\star(x_t) + \xi_t) - \hat{\Sigma}_{\mathcal{K}} f_\lambda - \left(\int_{\mathcal{X}} \mathcal{K}_x S \circ g^\star(x) d\rho(x) - \Sigma_{\mathcal{K}} f_\lambda\right)\right)
\end{aligned}$$

since $(\Sigma_{\mathcal{K}} + \lambda I) f_\lambda = \int_{\mathcal{X}} \mathcal{K}_x S \circ g^\star(x) d\rho(x)$ from Equation (93). Denoting $u = \frac{1}{T}\sum_{t=1}^T \mathcal{K}_{x_t} S(g^\star(x_t) + \xi_t) - \hat{\Sigma}_{\mathcal{K}} f_\lambda - \left(\int_{\mathcal{X}} \mathcal{K}_x S \circ g^\star(x) d\rho(x) - \Sigma_{\mathcal{K}} f_\lambda\right)$, $\left\|\hat{f} - f_\lambda\right\|_\rho$ can be bounded as follows

$$\begin{aligned}
\left\|\hat{f} - f_\lambda\right\|_\rho &= \left\|(\hat{\Sigma}_{\mathcal{K}} + \lambda I)^{-1} u\right\|_\rho \\
&= \left\|\Sigma_{\mathcal{K}}^{1/2}(\hat{\Sigma}_{\mathcal{K}} + \lambda I)^{-1} u\right\|_{\mathcal{H}_{\mathcal{K}}} && \text{(Lemma C.9)} \\
&\leq \left\|\Sigma_{\mathcal{K}}^{\frac{1}{2}}(\hat{\Sigma}_{\mathcal{K}} + \lambda I)^{-1}(\Sigma_{\mathcal{K}} + \lambda I)^{\frac{1}{2}}\right\|_{\text{op}} \left\|(\Sigma_{\mathcal{K}} + \lambda I)^{-\frac{1}{2}} u\right\|_{\mathcal{H}_{\mathcal{K}}}.
\end{aligned}$$

To bound $\left\|\Sigma_{\mathcal{K}}^{\frac{1}{2}}(\hat{\Sigma}_{\mathcal{K}}+\lambda I)^{-1}(\Sigma_{\mathcal{K}}+\lambda I)^{\frac{1}{2}}\right\|_{\text{op}}$, Step 3.1 in the proof of Theorem 4 of Caponnetto and De Vito (2007) shows

$$\left\|\Sigma_{\mathcal{K}}^{\frac{1}{2}}(\hat{\Sigma}_{\mathcal{K}}+\lambda I)^{-1}(\Sigma_{\mathcal{K}}+\lambda I)^{\frac{1}{2}}\right\|_{\text{op}} \leq 2$$

provided that $\left\|(\Sigma_{\mathcal{K}}+\lambda I)^{-1}\left(\hat{\Sigma}_{\mathcal{K}}-\Sigma_{\mathcal{K}}\right)\right\|_2 \leq \frac{1}{2}$. Lemma C.6 implies that it holds, with a probability greater than $1-\delta/2$, when $\lambda \geq 8\log\left(\frac{4}{\delta}\right)\frac{\kappa}{\sqrt{T}}$. To bound $\left\|(\Sigma_{\mathcal{K}}+\lambda I)^{-\frac{1}{2}}u\right\|_{\mathcal{H}_{\mathcal{K}}}$, we apply Lemma C.8, which indicates, with a probability greater than $1-\delta/2$,

$$\begin{aligned}
&\left\|(\Sigma_{\mathcal{K}}+\lambda I)^{-\frac{1}{2}}u\right\|_{\mathcal{H}_{\mathcal{K}}} \\
\leq& 2\log\left(\frac{4}{\delta}\right)\left(\frac{1}{T}\sqrt{\frac{\kappa}{\lambda}}\left\|S\right\|_{\text{op}}\left(\left(1+\sqrt{\frac{\kappa}{\lambda}}\right)\sqrt{C_1}+\sqrt{C_2}\right)+\sqrt{\frac{2\left(\mathcal{N}(\lambda)\mathbb{E}\left[\left\|S\xi\right\|_{\mathcal{Y}}^2\right]+\frac{\kappa}{\lambda}\left\|S\right\|_{\text{op}}^2 C_1\right)}{T}}\right).
\end{aligned}$$

It follows that with a probability greater than $1-\delta$, when $\lambda \geq 8\log\left(\frac{4}{\delta}\right)\frac{\kappa}{\sqrt{T}}$,

$$\left\|\hat{f}-f_\lambda\right\|_\rho^2 \leq 64\log^2\left(\frac{4}{\delta}\right)\left(\frac{\kappa\left\|S\right\|_{\text{op}}^2}{\lambda T^2}\left((1+\sqrt{\frac{\kappa}{\lambda}})^2 C_1+C_2\right)+\frac{1}{T}\left(\mathcal{N}(\lambda)\mathbb{E}\left[\left\|S\xi\right\|_{\mathcal{Y}}^2\right]+\frac{\kappa}{\lambda}\left\|S\right\|_{\text{op}}^2 C_1\right)\right)$$

since $(\frac{1}{T}\sqrt{\frac{\kappa}{\lambda}}\left\|S\right\|_{\text{op}}\left(\left(1+\sqrt{\frac{\kappa}{\lambda}}\right)\sqrt{C_1}+\sqrt{C_2}\right))^2 \leq \frac{2\kappa\left\|S\right\|_{\text{op}}^2}{\lambda T^2}\left(\left(1+\sqrt{\frac{\kappa}{\lambda}}\right)^2 C_1+C_2\right)$.

The second step to upper bound on the approximation error $\left\|f_\lambda - S\circ g^\star\right\|_\rho$, which can be achieved by applying directly Lemma C.4.

The proof is concluded by combining the two bounds.

■

C.3. Derivation of the Compact Bound

We now derive the compact bound stated in the main paper from Proposition C.10.

Proof [Proof of Proposition 4.1] Let $\ell_\delta \coloneqq \log(4/\delta)$. The condition on λ in Proposition 4.1 implies

$$\frac{\kappa}{\lambda} \leq \frac{\sqrt{T}}{8\ell_\delta} \leq \frac{\sqrt{T}}{8},$$

where the second inequality follows from $\ell_\delta > \log(4) > 1$. Since $T \geq 1$, we therefore have

$$\frac{1}{T}\left(1+\sqrt{\frac{\kappa}{\lambda}}\right)^2 \leq \frac{1}{T}\left(1+\frac{T^{1/4}}{\sqrt{8}}\right)^2 \leq \frac{1}{\sqrt{T}}\left(1+\frac{1}{\sqrt{8}}\right)^2 \leq 2.$$

Using the notation introduced in Proposition 4.1, this yields

$$\frac{\kappa\left\|S\right\|_{\text{op}}^2}{\lambda T^2}\left[\left(1+\sqrt{\frac{\kappa}{\lambda}}\right)^2 C_1+C_2\right] = \frac{\kappa}{\lambda T}\left[\frac{1}{T}\left(1+\sqrt{\frac{\kappa}{\lambda}}\right)^2 B_S^2+\frac{M_S^2}{T}\right] \leq \frac{\kappa}{\lambda T}(2B_S^2+M_S^2).$$

Furthermore,

$$\frac{1}{T}\left[\mathcal{N}(\lambda)\mathbb{E}\left[\left\|S\xi\right\|_{\mathcal{Y}}^2\right]+\frac{\kappa}{\lambda}\left\|S\right\|_{\text{op}}^2 C_1\right] = \frac{1}{T}\left[\mathcal{N}(\lambda)\sigma_S^2+\frac{\kappa}{\lambda}B_S^2\right].$$

It follows that the bound in Proposition C.10 can be simplified to, with probability at least $1-\delta$,

$$\mathcal{R}(\hat{f})-\mathcal{R}(S\circ g^\star) \leq \frac{128}{T}\log^2\left(\frac{4}{\delta}\right)\left[\mathcal{N}(\lambda)\sigma_S^2+\frac{\kappa}{\lambda}(3B_S^2+M_S^2)\right]+2\lambda^{2r}R_r^2,$$

which is precisely the claim of Proposition 4.1. ■

D. Additional Details on Numerical Experiments

All experiments were conducted on a Linux-based personal computer equipped with an Intel Core Ultra 7 165U processor (12 cores, 14 threads, up to 4.9 GHz). No GPU was used.

Evaluation Metrics. To evaluate methods in the numerical experiments, we report the mean squared error (MSE) computed at each hierarchical level as well as across all levels. Specifically, given an observed time series $(y_t)_{t=1}^T$ with $y_t \in \mathbb{R}^r$ where r denotes either the number of series at a given hierarchical level or the total number of series across all levels, and corresponding predictions $(\hat{y}_t)_{t=1}^T$, the MSE is defined as $\frac{1}{T}\sum_{t=1}^T \|y_t - \hat{y}_t\|_{\mathbb{R}^r}^2$.

D.1. Synthetic dataset

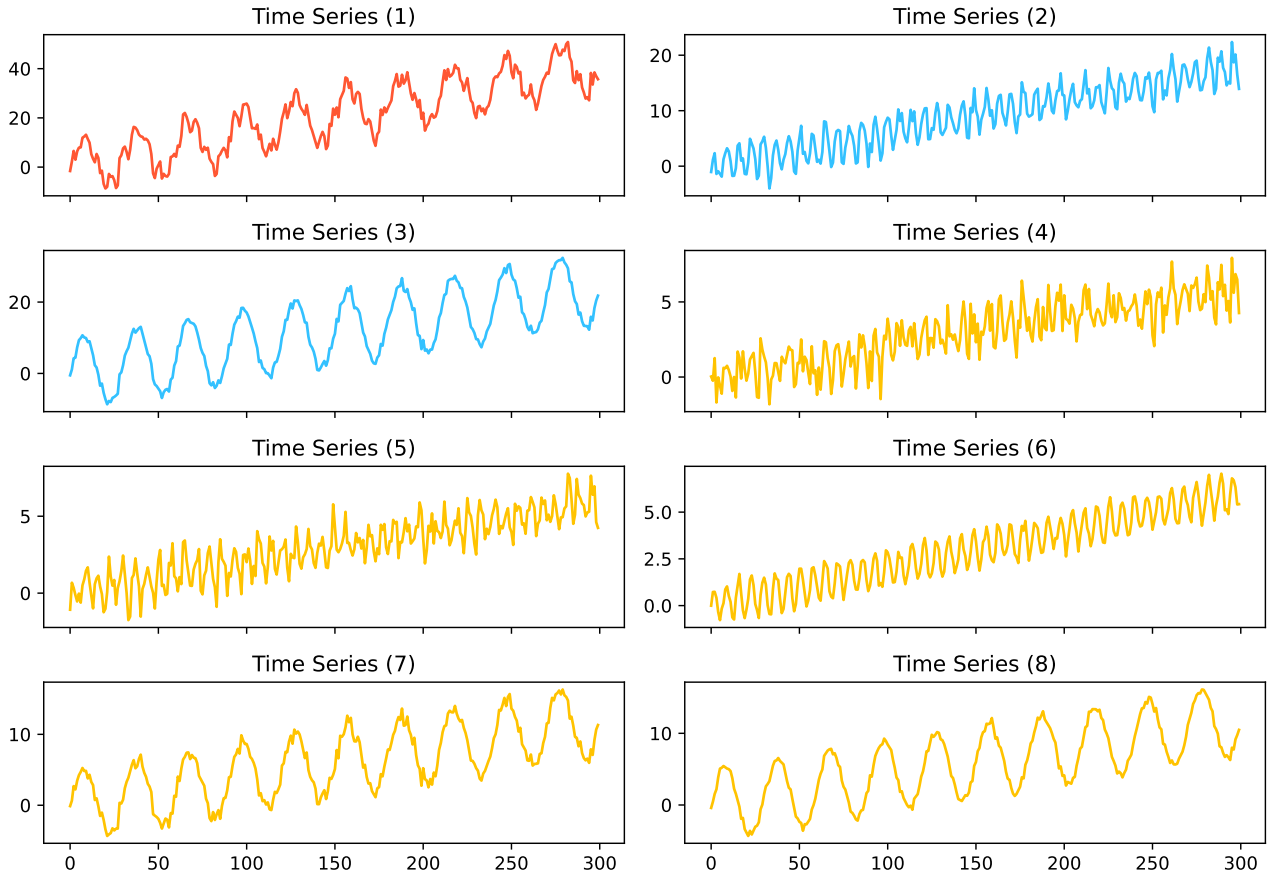


Figure 3. A simulation of the synthetic hierarchical time series with $\alpha = 1$ and a length of 300. Time series belonging to the same hierarchical level are represented using the same color.

Generation Process. The hierarchy contains $d = 8$ series distributed over three levels: one top-level series, two middle-level series, and $d_b = 5$ bottom-level series, with the structure shown in Figure 2. Each bottom-level series i is generated using the additive decomposition model (Hyndman and Athanasopoulos, 2018)

$$y_t^{(i)} = a^{(i)}t + b^{(i)} \sin\left(\frac{2\pi t}{p^{(i)}}\right) + \epsilon_t^{(i)}.$$

Series sharing a parent have identical trend and seasonal parameters. Specifically, $(a^{(i)}, b^{(i)}, p^{(i)}) = (0.02, 1, 7)$ for $i \in \{4, 5, 6\}$ and $(a^{(i)}, b^{(i)}, p^{(i)}) = (0.04, 5, 30)$ for $i \in \{7, 8\}$. Let $\xi_t = [\epsilon_t^{(4)}, \ldots, \epsilon_t^{(8)}]^\mathsf{T} \in \mathbb{R}^5$. We draw

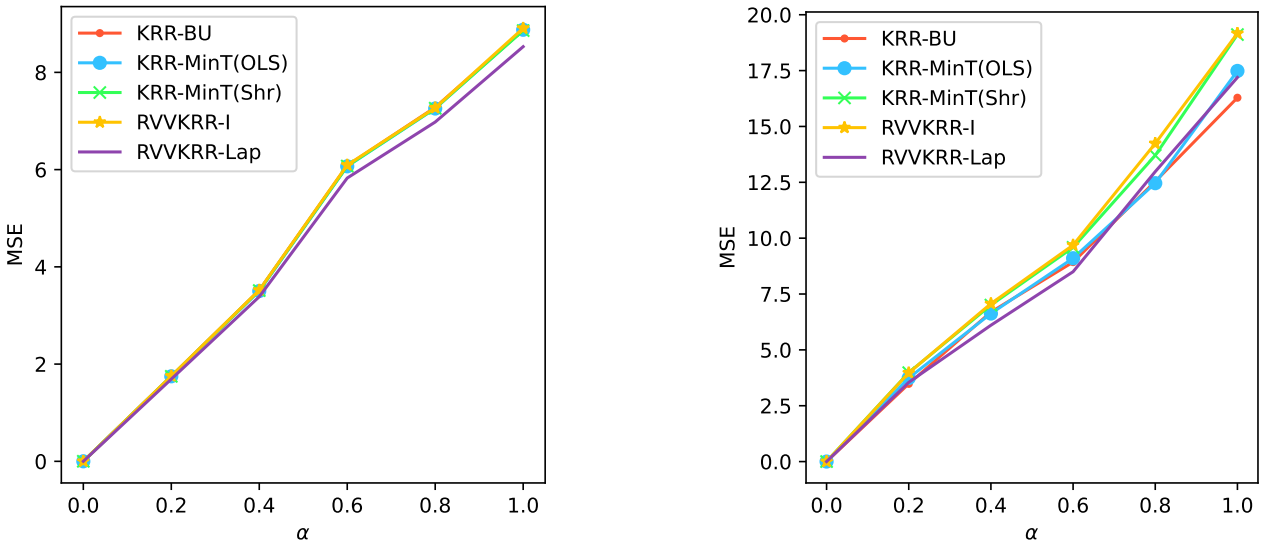
$$\xi_t \sim \mathcal{N}(0, \Sigma_\xi), \qquad \Sigma_\xi = Q(\alpha\Lambda)Q^\mathsf{T}, \qquad \Lambda = \mathrm{diag}(1, 0.5, 0.1, 0.05, 0.01),$$

where Q is orthogonal and $\alpha \geq 0$ controls the noise scale. In particular,

$$\mathbb{E}\left[\|S\xi_t\|_{\mathcal{Y}}^2\right] = \alpha \mathrm{tr}\left(SQ\Lambda Q^{\mathsf{T}} S^{\mathsf{T}}\right).$$

Writing $y_t^b = [y_t^{(4)}, \ldots, y_t^{(8)}]^{\mathsf{T}}$, the aggregate series are generated as $y_t^{(i)} = S_{i,:} y_t^b$ for $i \in \{1, 2, 3\}$.

Figure 3 shows one realization of the synthetic dataset.



(a) $T = 80$. *(b)* $T = 40$.

Figure 4. All-level MSE on the synthetic dataset as a function of the covariance scaling parameter α.

Sensitivity to the covariance scale. We additionally study the effect of the covariance scaling parameter α, varying it over $[0, 1]$ for training lengths $T = 80$ and $T = 40$. We run 32 Monte Carlo simulations for each configuration, using the same validation and test lengths as in the main experiment.

Figure 4 shows a quasi-linear relationship between α and the MSE. RVVKRR-Lap consistently outperforms the other methods for all values of α when $T = 80$; when $T = 40$, this holds only for $\alpha < 0.8$.

D.2. TourismLarge dataset

Experimental Setting. Hyperparameters are selected using three-fold expanding-window cross-validation, with training/validation window lengths of $108/24$, $132/24$, and $156/24$. Independent KRR selects a separate ridge parameter λ_r for each series from 64 logarithmically spaced values between 10^{-5} and 10^{8}. For each RVVKRR projection, ρ is selected from $\{10^{-3}, 3\times10^{-3}, 10^{-2}, 3\times10^{-2}, 10^{-1}\}$ by minimizing the mean Overall MSE across the three validation folds. This selects $\rho = 0.03$ for MinT (OLS) and $\rho = 0.1$ for both MinT (shrinkage) and MinT (WLS-var). The covariance proxies for the latter two are estimated from in-sample residuals of auxiliary independent KRR predictors within each training window. No feature standardization is applied.

D.3. Traffic Dataset

Experimental Setting. The largest lag used as the feature is 28, leaving 338 observations from the original 366-day sample. We reserve the last 28 observations for testing and use the preceding 310 for model selection and fitting. Hyperparameters are selected using three-fold expanding-window cross-validation by minimizing the MSE. Each validation window contains 28 observations, giving training/validation window lengths of $226/28$, $254/28$, and $282/28$. After selection, every model is refitted on all 310 pre-test observations.

Independent KRR selects a separate ridge parameter λ_r for each series r from 64 logarithmically spaced values between

Table 3. Full comparison of test MSE at each aggregation level of TourismLarge, reported in millions (10^6). All sums the level-wise MSEs. † denotes unreconciled forecasts. Lower is better; column-wise minima are bold.

Method	Ctry.	State	Zone	Reg.	Purp.	All
OLS base forecasts						
Unreconciled †	3.33	1.98	1.55	1.48	1.32	9.66
Bottom-Up	8.97	2.79	1.88	1.52	1.32	16.48
Top-Down (AHP)	3.33	4.00	2.39	2.06	1.91	13.69
Top-Down (PHA)	3.33	3.95	2.38	2.05	1.91	13.61
MinT (OLS)	3.48	1.77	1.47	1.32	1.28	9.32
MinT (shrinkage)	5.33	1.97	1.49	1.30	1.24	11.33
MinT (WLS-var)	6.59	2.19	1.56	1.33	1.25	12.92
ERM (closed form)	4.03	2.16	2.10	1.98	2.04	12.32
KRR base forecasts						
Unreconciled †	2.71	1.69	1.30	1.19	1.13	8.02
Bottom-Up	8.06	2.53	1.53	1.21	1.13	14.46
Top-Down (AHP)	2.71	3.94	2.38	2.05	1.92	13.00
Top-Down (PHA)	2.71	3.89	2.37	2.04	1.92	12.94
MinT (OLS)	2.72	1.58	1.25	1.12	1.11	7.78
MinT (shrinkage)	3.05	1.60	**1.21**	**1.08**	**1.06**	7.99
MinT (WLS-var)	4.56	1.83	1.29	1.11	1.09	9.89
ERM (closed form)	5.36	2.50	2.61	2.57	2.52	15.57
RVVKRR forecasts						
MinT (OLS), $\rho = 0$	2.75	1.55	1.23	1.10	1.10	7.73
MinT (OLS), $\rho = 0.03$	2.74	1.54	1.24	1.11	1.10	7.72
MinT (shrinkage), $\rho = 0$	**2.01**	**1.51**	1.28	1.23	1.58	**7.61**
MinT (shrinkage), $\rho = 0.1$	2.02	1.52	1.30	1.24	1.57	7.64
MinT (WLS-var), $\rho = 0$	2.52	1.81	1.45	1.40	1.75	8.93
MinT (WLS-var), $\rho = 0.1$	2.39	1.83	1.49	1.44	1.71	8.85

10^{-5} and 10^{8}. The same selected parameters are used by the auxiliary independent KRR fits that estimate the residual-based MinT projections. Independent OLS uses the same features and has no hyperparameter; the cross-validation results are used only for diagnostic purposes.

For each RVVKRR variant, the global regularization parameter λ is selected by the cross-validation over $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^{2}, 10^{3}\}$. All three variants select $\lambda = 1$. Covariance-based projections are estimated from in-sample residuals using only the current fold's training window. They are re-estimated using all 310 pre-test observations for the final fit.

Table 4. Full comparison of test MSE at each aggregation level of Traffic dataset, reported in thousands (10^3). All sums the level-wise MSEs. † denotes unreconciled forecasts. Lower is better; column-wise minima are bold.

Method	L1	L2	L3	Bottom	All
OLS base forecasts					
Unreconciled †	8.43	4.87	3.03	0.60	16.93
Bottom-Up	12.33	6.46	3.49	0.60	22.88
Top-Down (AHP)	8.43	4.56	3.04	1.33	17.36
Top-Down (PHA)	8.43	4.57	3.06	1.32	17.39
MinT (OLS)	8.41	4.90	2.85	0.58	16.74
MinT (shrinkage)	8.62	4.62	2.55	**0.57**	16.36
MinT (WLS-var)	10.73	5.70	3.12	0.60	20.15
ERM (closed form)	33.78	17.33	8.96	1.38	61.45
KRR base forecasts					
Unreconciled †	8.95	5.05	2.84	0.63	17.47
Bottom-Up	13.30	6.91	3.74	0.63	24.58
Top-Down (AHP)	8.95	4.81	3.17	1.33	18.27
Top-Down (PHA)	8.95	4.83	3.19	1.33	18.29
MinT (OLS)	9.01	4.95	2.85	0.61	17.42
MinT (shrinkage)	9.16	4.83	2.69	0.60	17.28
MinT (WLS-var)	11.68	6.11	3.34	0.62	21.76
ERM (closed form)	72.00	36.45	18.62	1.34	128.41
RVVKRR forecasts					
MinT (OLS)	8.99	4.93	2.79	0.61	17.32
MinT (shrinkage)	3.84	2.14	2.20	3.39	11.57
MinT (WLS-var)	**0.32**	**0.24**	**0.44**	3.02	**4.03**