Regularized Regression for Hierarchical Forecasting Without Unbiasedness Conditions

Souhaib Ben Taieb*
University of Mons
Mons, Belgium
souhaib.bentaieb@umons.ac.be
Monash University
Melbourne, Australia
souhaib.bentaieb@monash.edu

Bonsoo Koo
Monash University
Melbourne, Australia
bonsoo.koo@monash.edu

ABSTRACT

Forecasting a large set of time series with hierarchical aggregation constraints is a central problem for many organizations. However, it is particularly challenging to forecast these hierarchical structures. In fact, it requires not only good forecast accuracy at each level of the hierarchy, but also the coherency between different levels, i.e. the forecasts should satisfy the hierarchical aggregation constraints. Given some incoherent base forecasts, the state-of-the-art methods compute revised forecasts based on forecast combination which ensures that the aggregation constraints are satisfied. However, these methods assume the base forecasts are unbiased and constrain the revised forecasts to be also unbiased. We propose a new forecasting method which relaxes these unbiasedness conditions, and seeks the revised forecasts with the best tradeoff between bias and forecast variance. We also present a regularization method which allows us to deal with high-dimensional hierarchies, and provide its theoretical justification. Finally, we compare the proposed method with the state-of-the-art methods both theoretically and empirically. The results on both simulated and real-world data indicate that our methods provide competitive results compared to the state-of-the-art methods.

CCS CONCEPTS

• **Mathematics of computing** → **Time series analysis**; • **Computing methodologies** → *Regularization*.

KEYWORDS

time series, hierarchical forecasting, regularization, sparsity

ACM Reference Format:
Souhaib Ben Taieb and Bonsoo Koo. 2019. Regularized Regression for Hierarchical Forecasting Without Unbiasedness Conditions . In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4–8, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3292500.3330976

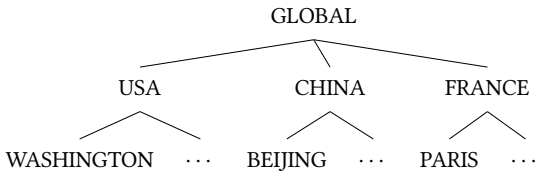
*Souhaib is an Adjunct Senior Research Fellow in the Department of Econometrics and Business Statistics at Monash University.

KDD '19, August 4–8, 2019, Anchorage, AK, USA

ACM ISBN 978-1-4503-6201-6/19/08…$15.00
https://doi.org/10.1145/3292500.3330976

1 INTRODUCTION

Forecasting a large set of time series with hierarchical aggregation constraints is a central problem for many organizations. For example, the store-level sales of a multinational company can be grouped in a hierarchical structure composed of countries, cities and regions [13]. Smart grid is another context where a hierarchy with different levels of aggregation naturally arises. The bottom level can include demand for single households or buildings and supply from single solar panels or wind turbines, while the intermediate levels could be the total demand and supply of entire regions, states and countries as a whole [1]. The next Figure gives an example of hierarchical aggregation.



It is particularly challenging to forecast these hierarchical structures (see Section 2.1). In fact, in addition to providing good forecast accuracy, it is often essential for decision-making purposes to produce "coherent" forecasts, i.e. the forecasts of aggregates should be equal to the sum of the corresponding disaggregated forecasts. Unfortunately, independently forecasting all the series in a hierarchy, also called base forecasts, is unlikely to deliver coherent forecasts. Another feature of these hierarchical structures is the highly noisy series in the most disaggregated bottom level. As a result, computing bottom-up forecasts, i.e. summing the individual bottom-level forecasts is unlikely to provide good accuracy at upper levels in the hierarchy [11].

The state-of-the art methods for hierarchical forecasting combine both base and bottom-up forecasts [10]. Specifically, starting from (probably) incoherent base forecasts, coherent revised forecasts are computed by applying forecast combination and using a bottom-up procedure. In other words, the base forecasts are adjusted so that they become coherent.

[17] proposed a new forecasting method called MinT which computes the optimal combination weights for the revised forecasts. More precisely, MinT assumes the base forecasts are unbiased and seeks the best combination weights under the constraint that the

revised forecasts will also be unbiased. In other words, MinT seeks the revised forecasts with minimum variance among all the unbiased revised forecasts (see Section 2.2). Although the MinT optimal combination weights are available in closed-form, the two conditions are hard to justify in practice. In fact, the base forecasts can be biased and revised forecasts with low mean squared forecast errors is often more important than the unbiasedness property.

We make the following contributions to hierarchical forecasting:

- First, we propose a new forecasting method which does not assume or constrain the base or the revised forecasts to be unbiased. To do so, we formulate the hierarchical forecasting problem as an empirical risk minimization problem which directly minimizes the mean squared forecast errors. We also provide the associated closed-form solution (see Section 3.1).
- Second, we address the problem of forecasting high dimensional hierarchies, i.e. hierarchies for which the total number of series is much larger than the number of historical observations. We adjust our objective function by adding a regularization term and provide its theoretical justification. We also derive its asymptotic properties including its limiting distribution (see Section 3.2).
- Third, we provide a comparison between our methods and the state-of-the-art MinT method, including similarities and differences, as well as the asymptotic properties of the methods. For example, we show that MinT is a particular case of our method, and under certain conditions, it is asymptotically equivalent to our method (see Section 4).
- Finally, we evaluate and compare the different hierarchical forecasting methods using both simulated and real-world data sets (see Section 5).

2 PRELIMINARIES

2.1 Hierarchical Forecasting

We let $\boldsymbol{b}_t \in \mathbb{R}^m$ represent the observations at time t for the m series in the most disaggregated (bottom) level. Then $\boldsymbol{a}_t = \boldsymbol{A}\boldsymbol{b}_t \in \mathbb{R}^k$ contains the observations at time t for the k aggregated series, where $\boldsymbol{A} \in \{0,1\}^{k\times m}$. Each entry $\boldsymbol{A}_{ij}$ is equal to 1 if the ith aggregate series contains the jth bottom-level series, where $i = 1,\dots,k$ and $j = 1,\dots,m$. Finally, $\boldsymbol{y}_t = \boldsymbol{S}\boldsymbol{b}_t = (\boldsymbol{a}_t' \ \boldsymbol{b}_t')' \in \mathbb{R}^n$ contains observations both at the aggregate and bottom levels, where $\boldsymbol{S}' = \begin{bmatrix}\boldsymbol{A}' & \boldsymbol{I}_m\end{bmatrix}$ and $\boldsymbol{I}_m$ is an $m \times m$ identity matrix. To avoid pathological hierarchies, we will assume that $m \geq 2$, $k \geq 1$ and $\sum_{j=1}^{m} \boldsymbol{A}_{i,j} > 1$ for $i = 1,\dots,k$.

Figure 1 gives the tree and the matrix representation of a hierarchical time series with $m = 4$ bottom level series with $\boldsymbol{b}_t = (y_{aa,t}, y_{ab,t}, y_{ba,t}, y_{bb,t})'$, and $k = 3$ aggregate series with $\boldsymbol{a}_t = (y_t, y_{a,t}, y_{b,t})'$, $y_t = y_{a,t} + y_{b,t}$, $y_{a,t} = y_{aa,t} + y_{ab,t}$ and $y_{b,t} = y_{ba,t} + y_{bb,t}$.

While $\boldsymbol{\mathcal{I}}_T = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_T\}$ denotes data observed up to time T, let $\hat{\boldsymbol{y}}_T(h)$ denote h-period ahead forecasts based on $\boldsymbol{\mathcal{I}}_T$ and $h = 1,\dots,H$ where h is the forecast horizon. The optimal h-period ahead forecasts which minimizes the conditional expectation:

$$\mathrm{E}\left[\|\boldsymbol{y}_{T+h} - \hat{\boldsymbol{y}}_T(h)\|_2^2 \,|\, \boldsymbol{\mathcal{I}}_T\right], \tag{1}$$

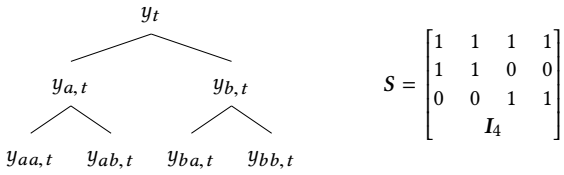


Figure 1: The tree and the matrix representation of a hierarchical time series with $m = 4$ bottom-level series and $k = 3$ aggregate series.

where $\|\cdot\|_p$ denotes an L_p norm over the relevant space, are given by

$$\boldsymbol{\mu}_T(h) = \mathrm{E}[\boldsymbol{y}_{T+h}|\boldsymbol{\mathcal{I}}_T] = \boldsymbol{S}\, E[\boldsymbol{b}_{T+h}|\boldsymbol{\mathcal{I}}_T]. \tag{2}$$

See [6] for more details on optimal forecasts.

In other words, if (1) is the error measure, any hierarchical forecasting method aims to provide the best estimate of $\boldsymbol{\mu}_T(h)$.

A natural plug-in estimator for expression (2), known as *bottom-up* (BU), is given by

$$\hat{\boldsymbol{y}}_T(h) = \boldsymbol{S}\,\hat{\boldsymbol{b}}_T(h), \tag{3}$$

where $\hat{\boldsymbol{b}}_T(h)$ is an estimate of $E[\boldsymbol{b}_{T+h}|\boldsymbol{\mathcal{I}}_T]$. The bottom-up method forecasts each of the bottom-level series, and then uses simple aggregation to obtain forecasts at higher levels of the hierarchy. Although this method is conceptually simple, it can provide poor forecasts on highly disaggregated and noisy data, especially at upper levels of the hierarchy.

Another approach, known as *base*, consists in forecasting all series in the hierarchy without considering the hierarchical constraints. In other words, we compute

$$\hat{\boldsymbol{y}}_T(h) = (\hat{\boldsymbol{a}}_T(h)' \ \hat{\boldsymbol{b}}_T(h)')', \tag{4}$$

where $\hat{\boldsymbol{a}}_T(h)$ and $\hat{\boldsymbol{b}}_T(h)$ are estimates of $E[\boldsymbol{a}_{T+h}|\boldsymbol{\mathcal{I}}_T]$ and $E[\boldsymbol{b}_{T+h}|\boldsymbol{\mathcal{I}}_T]$, respectively. This approach has the advantage that the typical poor forecasts for highly noisy series at the bottom level will not affect the forecasts for the series at upper levels of the hierarchy. However, it is unlikely that the resulting set of forecasts will be coherent, i.e. satisfy the aggregation constraints $\hat{\boldsymbol{a}}_T(h) = \boldsymbol{A}\hat{\boldsymbol{b}}_T(h)$. Imposing the aggregation constraints seems sensible since the optimal forecasts in expression (2) are coherent by definition. Furthermore, coherent forecasts will allow coherent decisions over the entire hierarchy.

2.2 The MinT approach to hierarchical forecasting

The state-of-the art method for hierarchical forecasting combines the best of both *bottom-up* and *base* forecasts, given in (3) and (4), respectively. Starting from a set of (probably incoherent) base forecasts $\hat{\boldsymbol{y}}_T(h)$, [10] proposed the computation of revised forecasts

$$\tilde{\boldsymbol{y}}_T(h) = \boldsymbol{S}\boldsymbol{P}\hat{\boldsymbol{y}}_T(h), \tag{5}$$

for some appropriately chosen matrix $\boldsymbol{P} \in \boldsymbol{\mathcal{P}} \subseteq \mathbb{R}^{m\times n}$ where $\boldsymbol{\mathcal{P}}$ is the domain of $\boldsymbol{P}$. This approach has multiple advantages: (1) the forecasts are coherent by construction; (2) a forecast combination from all levels is applied through the weight matrix $\boldsymbol{P}$; and (3) many

hierarchical forecasting methods are represented as particular cases, including the bottom-up forecasts, for which

$$\boldsymbol{P} = \left[\mathbf{0}_{m\times k} \mid \boldsymbol{I}_m\right], \tag{6}$$

where $\mathbf{0}_{m\times k}$ is an $m \times k$ zero matrix.

More recently, [17] proposed to optimally combine the base forecasts. Specifically, assuming the base forecasts $\hat{\boldsymbol{y}}_T(h)$ are unbiased, [17] computes the weight $\boldsymbol{P}$ giving the minimum variance unbiased revised forecasts. In other words, assuming $\mathrm{E}[\hat{\boldsymbol{y}}_T(h)|\mathcal{I}_T] = \mathrm{E}[\boldsymbol{y}_{T+h}|\mathcal{I}_T]$, denoted as **A1**, [17] considered the following optimization problem:

$$\begin{aligned}&\min_{\boldsymbol{P}\in\mathcal{P}} \mathrm{E}\left[\|\boldsymbol{y}_{T+h} - \tilde{\boldsymbol{y}}_T(h)\|_2^2 \,|\mathcal{I}_T\right]\\ &\text{subject to } \mathrm{E}[\tilde{\boldsymbol{y}}_T(h)|\mathcal{I}_T] = \mathrm{E}[\boldsymbol{y}_{T+h}|\mathcal{I}_T]\ (\mathbf{C1}).\end{aligned} \tag{7}$$

Under the unbiasedness conditions, **A1** and **C1**, [17] showed that minimizing (7) reduces to the following problem:

$$\min_{\boldsymbol{P}\in\mathcal{P}} \mathrm{Tr}(\mathrm{Var}[\boldsymbol{y}_{T+h} - \tilde{\boldsymbol{y}}_T(h)|\mathcal{I}_T]) \text{ subject to } \boldsymbol{SPS} = \boldsymbol{S}, \tag{8}$$

where $\mathrm{Tr}(\cdot)$ denotes the trace of a matrix.

The previous optimization problem has a closed-form solution given by

$$\boldsymbol{P}^* = (\boldsymbol{S}'\boldsymbol{W}_c^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}_c^{-1}, \tag{9}$$

where $\boldsymbol{W}_c^{-1}$ is the Moore-Penrose generalized inverse of $\boldsymbol{W}_c = \mathrm{E}[\hat{\boldsymbol{e}}_T(h)\hat{\boldsymbol{e}}_T(h)'|\mathcal{I}_T]$ and $\hat{\boldsymbol{e}}_T(h) = \boldsymbol{y}_{T+h} - \hat{\boldsymbol{y}}_T(h)$. We refer to this approach as the infeasible MinT (minimum trace) reconciliation given that $\boldsymbol{W}_c$ is unavailable in practice.

In [17], the authors proposed to compute

$$\hat{\boldsymbol{P}}_{\mathrm{MinT}} = (\boldsymbol{S}'\hat{\boldsymbol{W}}^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\hat{\boldsymbol{W}}^{-1}, \tag{10}$$

where $\hat{\boldsymbol{W}}$ is an estimate of $\boldsymbol{W}_c$. A natural estimator could be $\hat{\boldsymbol{W}} = T^{-1}\sum_{t=1}^{T}\hat{\boldsymbol{e}}_t(h)\hat{\boldsymbol{e}}_t(h)'$. However, since $\boldsymbol{W}_c$ is hard to estimate for $h > 1$, [17] assumed $\boldsymbol{W}_c = k_h\boldsymbol{W}_c^{(1)}$ where $k_h > 0$ and $\boldsymbol{W}_c^{(1)}$ is the covariance matrix of the one-step-ahead base forecast errors, i.e $h = 1$. Among the various covariance estimators for $\boldsymbol{W}_c^{(1)}$ considered in [17], the most effective one is the shrinkage estimator with diagonal target given by

$$\hat{\boldsymbol{W}} = (1-\alpha)\hat{\boldsymbol{W}}_{\mathrm{s}} + \alpha\hat{\boldsymbol{W}}_{\mathrm{d}}, \quad \hat{\boldsymbol{W}}_s = \frac{1}{T}\sum_{t=1}^{T}\hat{\boldsymbol{e}}_t(1)\hat{\boldsymbol{e}}_t(1)', \tag{11}$$

where $\hat{\boldsymbol{W}}_d = \mathrm{diag}(\hat{\boldsymbol{W}}_s)$ and $\alpha \in (0, 1]$. This covariance matrix is always invertible which is particularly useful for the computation of (9) when $n > T$ or $n = O(T)$ since then $\hat{\boldsymbol{W}}_s$ is singular, and hence non-invertible.

3 A NEW METHOD FOR HIERARCHICAL FORECASTING

3.1 Relaxation of the unbiasedness assumptions

We can decompose the objective function of MinT in (7) as

$$\mathrm{E}\left[\|\boldsymbol{y}_{T+h} - \tilde{\boldsymbol{y}}_T(h)\|_2^2 \,|\mathcal{I}_T\right] \tag{12}$$

$$= \|\boldsymbol{SP}(\mathrm{E}[\hat{\boldsymbol{y}}_T(h)|\mathcal{I}_T] - \mathrm{E}[\boldsymbol{y}_{T+h}|\mathcal{I}_T]) + (\boldsymbol{S} - \boldsymbol{SPS})\mathrm{E}[\boldsymbol{b}_{T+h}|\mathcal{I}_T]\|_2^2 \tag{13}$$

$$+ \mathrm{Tr}(\mathrm{Var}[\boldsymbol{y}_{T+h} - \tilde{\boldsymbol{y}}_T(h)|\mathcal{I}_T]), \tag{14}$$

where (13) and (14) are the bias and variance terms of the revised forecasts $\tilde{\boldsymbol{y}}_T(h)$, respectively.

Since MinT assumes the base forecasts are unbiased, the first term in (13) cancels. And since MinT requires the revised forecasts to be unbiased, equivalently $\boldsymbol{S} = \boldsymbol{SPS}$, the last term in (13) also cancels. As a result, minimizing (12) reduces to minimizing (14) with the constraint $\boldsymbol{S} = \boldsymbol{SPS}$. This is exactly the optimization problem given in (8).

We propose a new forecasting algorithm which do not impose the two unbiasedness conditions of MinT, **A1** and **C1**. In other words, we do not assume the base forecasts are unbiased, and we do not seek to produce the minimum variance unbiased revised forecasts. Instead, we aim to find the best tradeoff between bias and estimation variance of the revised forecasts by directly minimizing the mean squared forecast errors in (12).

Let $\hat{\boldsymbol{y}}_t(h)$ be the h-period ahead base forecasts based on $\mathcal{I}_t = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_t\}$ computed for $t = T_1, \ldots, T-h$ where $T_1 < T$ is the number of observations used for model fitting. We consider the following empirical risk minimization (ERM) problem:

$$\min_{\boldsymbol{P}\in\mathcal{P}} L_T(\boldsymbol{P}), \tag{15}$$

where

$$L_T(\boldsymbol{P}) = \frac{1}{(T-T_1-h+1)n}\sum_{t=T_1}^{T-h}\|\boldsymbol{y}_{t+h} - \boldsymbol{SP}\hat{\boldsymbol{y}}_t(h)\|_2^2, \tag{16}$$

$$= \left\|Y - \hat{Y}\boldsymbol{P}'\boldsymbol{S}'\right\|_F^2/Nn, \tag{17}$$

where $\|\cdot\|_F$ is the Frobenius norm defined as $\|\boldsymbol{X}\|_F = \sqrt{\mathrm{Tr}(\boldsymbol{X}'\boldsymbol{X})}$, $N = T - T_1 - h + 1$, and

$$\begin{aligned}Y &= [\boldsymbol{y}_{T_1+h}, \boldsymbol{y}_{T_1+h+1}, \ldots, \boldsymbol{y}_{T-h}]' \in \mathbb{R}^{N\times n},\\ \hat{Y} &= [\hat{\boldsymbol{y}}_{T_1}(h), \hat{\boldsymbol{y}}_{T_1+1}(h), \ldots, \hat{\boldsymbol{y}}_{T-h}(h)]' \in \mathbb{R}^{N\times n}.\end{aligned}$$

Proposition 1. *If the matrix $\hat{Y}'\hat{Y}$ is invertible, the solution to* (15) *is given by*

$$\hat{\boldsymbol{P}}_{ERM} = (\boldsymbol{S}'\boldsymbol{S})^{-1}\boldsymbol{S}'Y'\hat{Y}(\hat{Y}'\hat{Y})^{-1} \tag{18}$$

$$= \boldsymbol{B}'\hat{Y}(\hat{Y}'\hat{Y})^{-1}, \tag{19}$$

where $\boldsymbol{S}'\boldsymbol{S}$ is invertible by construction.

Proof. Let $\mathrm{vec}(\boldsymbol{X})$ denote the vectorization of the matrix $\boldsymbol{X}$, i.e. all the columns of $\boldsymbol{X}$ are sequentially stacked in a one-column vector. Using the fact that $\|\boldsymbol{X}\|_F^2 = \|\mathrm{vec}(\boldsymbol{X})\|_2^2$, we can write (17) multiplied by Nn as

$$\left\|Y - \hat{Y}\boldsymbol{P}'\boldsymbol{S}'\right\|_F^2 = \left\|\mathrm{vec}(Y) - \mathrm{vec}(\hat{Y}\boldsymbol{P}'\boldsymbol{S}')\right\|_2^2 \tag{20}$$

$$= \left\|\mathrm{vec}(Y) - (\boldsymbol{S}\otimes\hat{Y})\mathrm{vec}(\boldsymbol{P}')\right\|_2^2, \tag{21}$$

The vector $\mathrm{vec}(\boldsymbol{P}')$ which minimizes (21) is given by

$$\mathrm{vec}(\boldsymbol{P}') = \left[(\boldsymbol{S}\otimes\hat{Y})'(\boldsymbol{S}\otimes\hat{Y})\right]^{-1}(\boldsymbol{S}\otimes\hat{Y})'\mathrm{vec}(Y) \tag{22}$$

$$= \left[(\boldsymbol{S}'\boldsymbol{S})^{-1}\boldsymbol{S}'\otimes(\hat{Y}'\hat{Y})^{-1}\hat{Y}'\right]\mathrm{vec}(Y) \tag{23}$$

$$= \mathrm{vec}\left((\hat{Y}'\hat{Y})^{-1}\hat{Y}'Y\boldsymbol{S}(\boldsymbol{S}'\boldsymbol{S})^{-1}\right). \tag{24}$$

Unvectorizing and transposing (24) gives expression (18). Finally, by noting that $\boldsymbol{Y} = \boldsymbol{B}\boldsymbol{S}^{'}$, (18) can be simplified into (19).

□

Remark 2. *When $\hat{\boldsymbol{Y}}^{'}\hat{\boldsymbol{Y}}$ is not invertible, $\hat{\boldsymbol{P}}_{ERM}$ in* (19) *is not unique. This can notably happen with coherent base forecasts or in high-dimensional setting.*

If the base forecasts are coherent, we can write $\hat{\boldsymbol{Y}} = \boldsymbol{D}\boldsymbol{S}^{'}$. In that case, the columns of $\hat{\boldsymbol{Y}}$ are linearly dependent. Consequently, we will have

$$rank(\hat{\boldsymbol{Y}}^{'}\hat{\boldsymbol{Y}}) = rank(\hat{\boldsymbol{Y}}) = rank(\boldsymbol{S}\boldsymbol{D}^{'}) \leq min(rank(\boldsymbol{S}), rank(\boldsymbol{D})) \leq m.$$

Therefore, $\hat{\boldsymbol{Y}}^{'}\hat{\boldsymbol{Y}} \in \mathbb{R}^{n\times n}$ is not invertible. In other words, incoherent base forecasts is a necessary condition for $\hat{\boldsymbol{Y}}^{'}\hat{\boldsymbol{Y}}$ to be invertible.

Obtaining coherent base forecasts is unlikely to happen in practice due to estimation errors in $\hat{\boldsymbol{Y}}$. However, asymptotically, i.e. when $T_1 \to \infty$, and if the model is well-specified, the base forecasts $\hat{\boldsymbol{y}}_t(h)$ will converge to the conditional mean given in expression (2)*. In that case, the rows of matrix $\boldsymbol{D}$ will contain $E[\boldsymbol{b}_{t+h}|\mathcal{I}_t]$ at times $t = T_1 + h, \ldots, T - h$. This implies that the model is unidentified in the limit when the forecasts are coherent eventually as the number of observations T_1 tends to infinity.*

Nevertheless, we can compute a unique solution using the Moore-Penrose generalized inverse. If $\hat{\boldsymbol{Y}} = \boldsymbol{U}_{\hat{\boldsymbol{Y}}}\Sigma_{\hat{\boldsymbol{Y}}}\boldsymbol{V}_{\hat{\boldsymbol{Y}}}^{'}$ is the thin SVD of $\hat{\boldsymbol{Y}}$, i.e. without the zeroes in $\Sigma_{\hat{\boldsymbol{Y}}}$, then this is equivalent to computing

$$\hat{\boldsymbol{P}}_{ERM} = \boldsymbol{V}_{\hat{\boldsymbol{Y}}}\Sigma_{\hat{\boldsymbol{Y}}}^{-1}\boldsymbol{U}_{\hat{\boldsymbol{Y}}}^{'}\boldsymbol{B}. \tag{25}$$

Note that the same observation applies to the MinT solution in (9)*. In fact, if the model is well-specified, $\boldsymbol{W}_c$ is asymptotically singular since $\boldsymbol{W}_c = \boldsymbol{S}\boldsymbol{V}_b\boldsymbol{S}^{'}$ where $\boldsymbol{V}_b$ is the covariance of the random error terms in the bottom level.*

3.2 Regularization for high-dimensional hierarchies

The ERM method has the advantage of relaxing the unbiasedness assumptions of MinT. The associated optimization problem in (15) involves the estimation of $m \times n = m^2 + mk$ parameters where m is the number of bottom-level series. However, common hierarchies can have a large number of bottom-level series m compared to the number of historical observations. The larger number of parameters to estimate results in an increase in estimation variance due to the accumulation of estimation errors. This is in turn transmitted into higher forecast variance, leading to poor forecast accuracy.

To overcome this issue, we propose to apply regularization to our objective in (17). More precisely, we propose a new method, called ERMreg, which solves the following problem:

$$\hat{\boldsymbol{P}}_{\text{ERMreg}} = \underset{\boldsymbol{P}\in\mathcal{P}}{\text{argmin}} \left\{ \left\| \boldsymbol{Y} - \hat{\boldsymbol{Y}}\boldsymbol{P}^{'}\boldsymbol{S}^{'} \right\|_F^2 / Nn + \lambda \left\| \text{vec}(\boldsymbol{P}) \right\|_1 \right\}, \tag{26}$$

where $\lambda \geq 0$ is a hyperparameter which controls the amount of regularization to apply. When $\lambda = 0$, the problem reduces to (15) with a closed-form solution given by (19).

The major benefit of imposing a complexity penalty lies in the reduction in estimation variance by sacrificing some bias. In fact, parsimonious models often provide better forecasts than complex true models. Imposing an L_1 penalty is motivated by the fact that, asymptotically, the optimal forecasts are given by $\boldsymbol{P} = \boldsymbol{P}_{\text{BU}} = \left[\boldsymbol{0}_{m\times k} \mid \boldsymbol{I}_m\right]$, which is sparse.

By replacing the first term in (26) with (21), the previous problem can be reduced to a standard LASSO problem with vec($\boldsymbol{Y}$) as dependent variable and ($\boldsymbol{S} \otimes \hat{\boldsymbol{Y}}$) as design matrix. In other words, we need to solve a LASSO problem with $N \times n$ observations and $m^2 + mk$ variables.

We can use various efficient LASSO solvers for high-dimensional problems, for example cyclical coordinate descent methods (see Section 5.4 in [7] and [5]) and since ($\boldsymbol{S} \otimes \hat{\boldsymbol{Y}}$) is a sparse matrix, exploiting sparse matrix algebra can further reduce the computational load.

Adding an L_1 penalty term to the objective function will encourage sparsity in $\boldsymbol{P}$. This is notably motivated by the fact that the optimal forecasts have a sparse $\boldsymbol{P}$ matrix. However, this will also have the effect of shrinking the revised forecasts towards zero, which is not desirable when dealing with strictly positive observations.

In traditional forecast combination, the combination weights are often penalized towards the average combination, i.e. the same weight is given to all forecasts [16]. In the context of hierarchical forecasting, average combination is not possible since the forecasts are associated to different variables and their aggregates.

We propose to penalize the matrix $\boldsymbol{P}$ towards the $\boldsymbol{P}_{\text{BU}} = \left[\boldsymbol{0}_{m\times k} \mid \boldsymbol{I}_m\right]$, i.e. no combination is applied.

In other words, we propose to solve the following problem:

$$\hat{\boldsymbol{P}}_{\text{ERMregbu}} = \underset{\boldsymbol{P}\in\mathcal{P}}{\text{argmin}} \left\{ \left\| \boldsymbol{Y} - \hat{\boldsymbol{Y}}\boldsymbol{P}^{'}\boldsymbol{S}^{'} \right\|_F^2 / Nn + \lambda \left\| \text{vec}(\boldsymbol{P} - \boldsymbol{P}_{\text{BU}}) \right\|_1 \right\}. \tag{27}$$

Note that we do not use $\boldsymbol{P}_{\text{BU}}$ as weight matrix, but we rather penalize our weights towards $\boldsymbol{P}_{\text{BU}}$ where the amount of regularization is controlled by λ.

Using $\boldsymbol{P}_{\text{BU}}$ as reference matrix is motivated by the fact that, asymptotically, the optimal forecasts given in (2) can be computed as revised forecasts with $\boldsymbol{P} = \boldsymbol{P}_{\text{BU}}$. Furthermore, the matrix $\boldsymbol{P}_{\text{BU}}$ has the advantage of parsimony. If $\hat{\boldsymbol{y}}_T(h) = \boldsymbol{\mu}_T(h)$, any matrix $\boldsymbol{P}$ that satisfy $\boldsymbol{P}\boldsymbol{S} = \boldsymbol{I}$ will produce bottom-up revised forecasts. But, $\boldsymbol{P}_{\text{BU}}$ is the sparsest (in L_0 "norm") matrix $\boldsymbol{P}$ for which $\boldsymbol{P}\boldsymbol{S} = \boldsymbol{I}$, as summarized in the following theorem.

Theorem 3. *$\boldsymbol{P}_{BU} = \left[\boldsymbol{0}_{m\times k} \mid \boldsymbol{I}_m\right]$ is the unique solution of the following optimization problem:*

$$\min_{\boldsymbol{P}\in\mathcal{P}} \sum_{i=1}^{m}\sum_{j=1}^{n} |P_{ij}|^0 \text{ subject to } \boldsymbol{P}\boldsymbol{S} = \boldsymbol{I}. \tag{28}$$

The proof is given in the Appendix B.1.

Using the reparametrization $\boldsymbol{C} = \boldsymbol{P} - \boldsymbol{P}_{\text{BU}}$, Problem (27) can be rewritten in the same form as (26). Furthermore, Problem (27) can be written as a standard LASSO problem as stated by the next Proposition.

Proposition 4. *Let $\boldsymbol{\beta} = vec(\boldsymbol{P}^{'} - \boldsymbol{P}_{BU}^{'})$. Problem* (27) *can be formulated as*

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \|\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 / Nm + \lambda \|\boldsymbol{\beta}\|_1 \right. , \tag{29}$$

where $\boldsymbol{X} = (\boldsymbol{S} \otimes \hat{\boldsymbol{Y}})$ and $\boldsymbol{z} = vec(\boldsymbol{Y}) - \boldsymbol{X}\ vec(\boldsymbol{P}_{BU}^{'})$.

Proof. The result follows from plugging $\boldsymbol{\beta} = \boldsymbol{P}^{'} - \boldsymbol{P}^{'}_{\text{BU}}$ in (27) and applying vectorization as in (21). □

Remark 5. *As stated by Proposition 4, Problem* (27) *can be reduced to a standard LASSO problem and therefore, one can use readily available standard algorithms embedded in statistical software. Furthermore, thanks to the L_1 regularization, a unique solution to* (29) *may exist even though $(\boldsymbol{X}^{'}\boldsymbol{X})$ is singular, where $\boldsymbol{X}$ is defined in* (29).

However, standard inference tools based on well-established statistical properties of LASSO estimation are not directly applicable since it is subject to the issue of nearly singular design introduced in [12]. Using the method developed in [12], for the purpose of statistical inference, the following theorem provides the limiting distribution of appropriately standardized estimator of $\boldsymbol{\beta}$ in (29) *even in the nearly singular design.*

Theorem 6. *Under the assumptions **D1**-**D2** in Appendix A.2,*

$$r_q(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \operatorname{argmin} Q_0(\boldsymbol{g})$$

with

$$Q_0(\boldsymbol{g}) = -2\boldsymbol{g}^{'}\boldsymbol{V} + \boldsymbol{g}^{'}\boldsymbol{D}\boldsymbol{g} + \lambda_0 \sum_{j=1}^{m\times n}\{g_j sgn(\beta_j) + |g_j| I(\beta_j = 0)\},$$

where $\boldsymbol{M} = plim_{q\to\infty} 1/q \sum_{t=1}^{q} \boldsymbol{x}_t\boldsymbol{x}_t^{'}$, $q = Nm$, $\boldsymbol{V}$ is defined as a zero-mean multivariate normal random vector such that $Var(\boldsymbol{g}'\boldsymbol{V}) = \sigma^2\boldsymbol{g}^{'}\boldsymbol{D}\boldsymbol{g}$ for $\boldsymbol{g}$ satisfying $\boldsymbol{M}\boldsymbol{g} = \boldsymbol{0}$ and $sgn(\cdot)$ denotes the sign function.

The proof is given in Appendix B.2.

4 COMPARISON WITH MINT

In this section, we give more details about the differences between the MinT method and our new forecasting methods presented in Sections 2.2 and 3, respectively.

4.1 Unbiasedness assumptions

Under some regularity conditions (see Appendix A), the objective function in (16) is the sample version of the following population objective

$$\mathcal{L}(\boldsymbol{P}) = \mathrm{E}\left[\|\boldsymbol{y}_{T+h} - \tilde{\boldsymbol{y}}_T(h)\|_2^2\right] \tag{30}$$

$$= \mathrm{E}\left[\mathrm{E}\left[\|\boldsymbol{y}_{T+h} - \tilde{\boldsymbol{y}}_T(h)\|_2^2 \,|\boldsymbol{\mathcal{I}}_T\right]\right]. \tag{31}$$

Let us compare our population objective in (31) with the MinT population objective function in (7). First, we can see that we have relaxed the unbiasedness assumptions of MinT. Secondly, our objective involves *averaged* forecast errors while MinT deals with *path-dependent* forecast errors.

Although *path-dependent* forecast errors are more general than *averaged* forecast errors [14], it is hard to compute a reliable estimate of the *path-dependent* forecast errors. This is due to fact that we only observe one realization of the underlying data generating process (see Section 7.12 in [8] for a related discussion).

Furthermore, the MinT solution depends on the conditional covariance matrix $\boldsymbol{W}_c$ defined in (9). However, by the law of large numbers, $\hat{\boldsymbol{W}}$ in (11) asymptotically converges to the unconditional covariance matrix $\boldsymbol{W}_u = \mathrm{E}\left[\hat{\boldsymbol{e}}_T(h)\hat{\boldsymbol{e}}_T^{'}(h)\right]$. In other words, MinT asymptotically minimize (31) under the constraints **A1** and **C1**. This is summarized in the following lemma.

Lemma 7. *Suppose that the conditions **A1** and **C1** hold. Then, the weight matrix that minimizes the criterion* (31) *is given by*

$$\boldsymbol{P}_U = \left(\boldsymbol{S}^{'}\boldsymbol{W}_u^{-1}\boldsymbol{S}\right)^{-1}\boldsymbol{S}^{'}\boldsymbol{W}_u^{-1}, \tag{32}$$

where $\boldsymbol{W}_u = \mathrm{E}\left[\hat{\boldsymbol{e}}_T(h)\hat{\boldsymbol{e}}_T^{'}(h)\right]$ and $\boldsymbol{W}_u^{-1}$ is the Moore-Penrose generalized inverse of $\boldsymbol{W}_u$.

Proof. Using expression (14) and Lemma 1 in [17], we can write

$$\mathrm{E}\left[\mathrm{E}\left[\|\boldsymbol{y}_{T+h} - \tilde{\boldsymbol{y}}_T(h)\|_2^2 \,|\boldsymbol{\mathcal{I}}_T\right]\right] \tag{33}$$

$$= \mathrm{E}\left[\mathrm{Tr}(\mathrm{Var}[\boldsymbol{y}_{T+h} - \tilde{\boldsymbol{y}}_T(h)|\boldsymbol{\mathcal{I}}_T])\right] \tag{34}$$

$$= \mathrm{Tr}\left[\mathrm{E}((\mathrm{Var}[\boldsymbol{y}_{T+h} - \tilde{\boldsymbol{y}}_T(h)|\boldsymbol{\mathcal{I}}_t]))\right] \tag{35}$$

$$= \mathrm{Tr}\left[\boldsymbol{S}\boldsymbol{P}\boldsymbol{W}_u\boldsymbol{P}^{'}\boldsymbol{S}^{'}\right], \tag{36}$$

where $\boldsymbol{W}_u = \mathrm{E}\left[\hat{\boldsymbol{e}}_T(h)\hat{\boldsymbol{e}}_T^{'}(h)\right]$ with $\hat{\boldsymbol{e}}_T(h) = \boldsymbol{y}_{T+h} - \hat{\boldsymbol{y}}_T(h)$, i.e. $\boldsymbol{W}_u$ is the variance covariance matrix of the unconditional h-step-ahead base forecast errors. Then, the solution that minimizes (36) is given by

$$\boldsymbol{P}_U = \left(\boldsymbol{S}^{'}\boldsymbol{W}_u^{-1}\boldsymbol{S}\right)^{-1}\boldsymbol{S}^{'}\boldsymbol{W}_u^{-1}, \tag{37}$$

due to the Moore-Penrose generalized inverse. □

A direct consequence of Lemma (7) is that, under the conditions **A1** and **C1**, $\hat{\boldsymbol{P}}_{\text{MinT}}$ in (10) and $\hat{\boldsymbol{P}}_{\text{ERM}}$ in (15) are asymptotically equivalent to $\boldsymbol{P}_U$ in (32). This is summarized in the following Corollary.

Corollary 8. *Suppose that the regularity conditions B1-B3 in Appendix A hold. Then, under the assumptions **A1** and **C1**, $\hat{\boldsymbol{P}}_{MinT}$ in* (10) *and $\hat{\boldsymbol{P}}_{ERM}$ in* (15) *are asymptotically equivalent to $\boldsymbol{P}_U$ in* (32).

The proof is given in the Appendix B.3.

In summary, asymptotically, our ERM method includes MinT as a particular case, and reduces to MinT under the conditions **A1** and **C1**.

4.2 Regularization

Let $\tilde{\boldsymbol{b}}_T(h) = \boldsymbol{P}^*\hat{\boldsymbol{y}}_T(h)$ be the bottom-level infeasible MinT forecasts where $\boldsymbol{P}^*$ is defined in (9). Using the proof of Theorem 1 in [17], we can show that $\tilde{\boldsymbol{b}}_T(h) = \hat{\boldsymbol{\beta}}$ where $\hat{\boldsymbol{\beta}}$ is the solution of the following generalized least squares problem:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^m} (\hat{\boldsymbol{y}}_T(h) - \boldsymbol{S}\boldsymbol{\beta})'\boldsymbol{W}^{-1}(\hat{\boldsymbol{y}}_T(h) - \boldsymbol{S}\boldsymbol{\beta}). \tag{38}$$

Although, the previous optimization problem does not include any explicit regularization term, the specific structure of the matrix $\boldsymbol{S}$ has an implicit regularization effect. For simplicity of exposition, let us assume that $\boldsymbol{W}$ is a block diagonal matrix, i.e. $\boldsymbol{W} = \mathrm{diag}(\boldsymbol{W}_a\ \boldsymbol{W}_b)$. Then, since $\boldsymbol{S}' = \left[\boldsymbol{A}'\ \boldsymbol{I}_m\right]'$ and $(\hat{\boldsymbol{a}}_T(h)\ \hat{\boldsymbol{b}}_T(h))'$, we can rewrite problem (38) as

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^m} (\hat{\boldsymbol{a}}_T(h) - \boldsymbol{A}\boldsymbol{\beta})'\boldsymbol{W}_a^{-1}(\hat{\boldsymbol{a}}_T(h) - \boldsymbol{A}\boldsymbol{\beta}) \tag{39}$$

$$+(\boldsymbol{\beta} - \hat{\boldsymbol{b}}_T(h))'\boldsymbol{W}_b^{-1}(\boldsymbol{\beta} - \hat{\boldsymbol{b}}_T(h)). \tag{40}$$

The previous optimization problem reveals that the objective function of MinT is similar to a generalized ridge regression problem, where the first term is a generalized weighted least squares criterion, and the second term is the generalized ridge penalty. A

closer look at (40) shows that MinT seeks to find the bottom-level forecasts $\boldsymbol{\beta}$ for which the associated bottom-up forecasts $\boldsymbol{A\beta}$ are closest to the base forecasts at the aggregate levels where the importance of each difference is controlled by $\boldsymbol{W}_a^{-1}$. Furthermore, the bottom-level forecasts $\boldsymbol{\beta}$ are shrunk towards the bottom-level base forecasts where the shrinkage amount is controlled by $\boldsymbol{W}_b^{-1}$.

For MinT in high-dimensional setting, a shrinkage estimator becomes inevitable due to the accumulation of estimation errors. We can see this by noting that $\left|\mathrm{Tr}(\boldsymbol{SP}\hat{\boldsymbol{W}}\boldsymbol{P}'\boldsymbol{S}') - \mathrm{Tr}(\boldsymbol{SPW}_c\boldsymbol{P}'\boldsymbol{S}')\right| \leq \left\|\hat{\boldsymbol{W}} - \boldsymbol{W}_c\right\|_\infty \left\|\boldsymbol{P}'\boldsymbol{S}'\right\|_1^2$. Related discussion can be found in [2] and [3] among many others.

Let us consider $\hat{\boldsymbol{W}}$ as defined in (11) with $\hat{\boldsymbol{W}}_s = \mathrm{diag}(\hat{\boldsymbol{W}}_{s,a}\ \hat{\boldsymbol{W}}_{s,b})$ and $\hat{\boldsymbol{W}}_\mathrm{d} = \mathrm{diag}(\hat{\boldsymbol{W}}_{d,a}\ \hat{\boldsymbol{W}}_{d,b})$. Let us plug $\hat{\boldsymbol{W}}$ in (40), i.e. $\boldsymbol{W}_a = (1-\alpha)\hat{\boldsymbol{W}}_{s,\mathrm{a}} + \alpha\hat{\boldsymbol{W}}_{d,\mathrm{a}}$ and $\boldsymbol{W}_b = (1-\alpha)\hat{\boldsymbol{W}}_{s,\mathrm{b}} + \alpha\hat{\boldsymbol{W}}_{d,\mathrm{b}}$. Then, we can see that the value of α will notably affect the amount of shrinkage towards the bottom-level base forecasts. In particular, when $\alpha = 1$, which has been considered in [11], the penalty term reduces to $\sum_{j=1}^{m} \frac{1}{\hat{w}_j}(\boldsymbol{\beta}_j - \hat{b}_{j,T}(h))^2$ where $\hat{w}_j$ is the jth diagonal element of the matrix $\hat{\boldsymbol{W}}_{d,b}$. Assuming the series are in the same scale, this shows that more shrinkage is applied when the base forecasts $\hat{b}_{j,T}(h)$ are more accurate.

Although MinT implicitly regularizes the revised forecasts, the solution is still based on the assumption **A1** and constraint **C1**. In other words, if **A1** and **C1** are not appropriate for the problem at hand, shrinking the revised forecasts towards the base forecasts can lead to poor forecast accuracy. Our method ERMreg is more robust since we try to find the best tradeoff between bias and estimation variance by explicitly adding a regularization term.

5 EXPERIMENTS

We perform multiple experiments to compare our forecasting methods with the state-of-the art methods in different conditions. We consider both simulated and real-world datasets. For each dataset, we sample 100 different hierarchies, and the results are averaged over these 100 hierarchies.

5.1 Data

In the following, we describe the different data sets used in the experiments.

5.1.1 Simulated data (see Figure 3a). We use the data generating mechanism developed in [17] (see Sections 3.3 and 3.5) to reflect common characteristics of hierarchical time series such as correlations across series, and smoother data with more aggregation. The bottom-level series are sampled from correlated ARMA processes where the coefficients are sampled uniformly from a specific parameter space that guarantees stationarity. The random error terms have a multivariate Gaussian distribution with a covariance structure that induces a strongly positive correlation among series with the same parents, but a moderately positive correlation among series with different parents. For each series, we generate 600 observations split in training and validation sets, 2/3 and 1/3 respectively. An additional sample of 200 observations is used for testing. We consider both a small and a large two-level hierarchy where $m = 4$ and $k = 3$, and $m = 100$ and $k = 26$, respectively. More specifically, the bottom level series were aggregated in groups of two for the next level in the small hierarchy, and groups of four in the large hierarchy. These series were then aggregated to obtain the top level series.

5.1.2 Road traffic [1] *(see Figure 3b).* The dataset gives the occupancy rate (between 0 and 1) of 963 car lanes of San Francisco bay area freeways. The measurements are sampled every 10 minutes from Jan. 1st 2008 to Mar. 30th 2009. This dataset has been notably used in [4] for classification tasks. We aggregate the data to 366 daily observations split in 120, 120 and 126 observations for training, validation and testing, respectively. We consider hierarchies with $m = 200$ and $k = 7$, where each hierarchy is constructed as follows. We sample 200 bottom level series from the 963 series, and compute the upper level series by aggregation. More specifically, 200 series at the bottom level were aggregated in groups of 50 for the next level, resulting in 4 series. These 4 series were then aggregated in groups of two to obtain two aggregate series and the top level series.

5.1.3 Wikipedia webpage views [2] *(see Figure 3c).* The dataset gives the number of daily views of 145,000 different Wikipedia articles starting from July, 1st, 2015 up until December 31st, 2016. We use the 366 observations for 2016 split in 86, 160 and 120 observations for training, validation and testing, respectively. We consider hierarchies where $m = 150$ and $k = 49$, where each hierarchy is constructed as follows. We sample 150 bottom level series from the 145k series, and compute the upper level series by aggregation. More specifically, the bottom level series are hierarchically aggregated into type of agent ("all-agents" and "spider"), type of access ("all-access", "desktop" and "mobile-web") and country codes ("en", "fr", "de", "ja", "ru", "zh"). The hierarchy's architecture is summarized in Figure 2, where the top level series is in the center and the most disaggregated bottom level series are on the circumference of the circle.

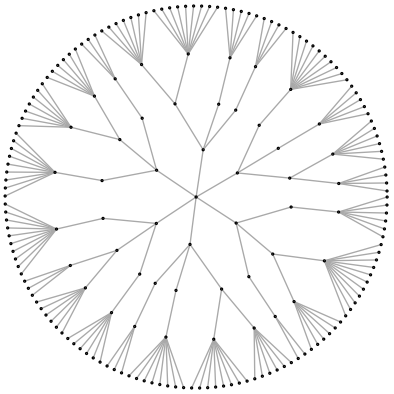
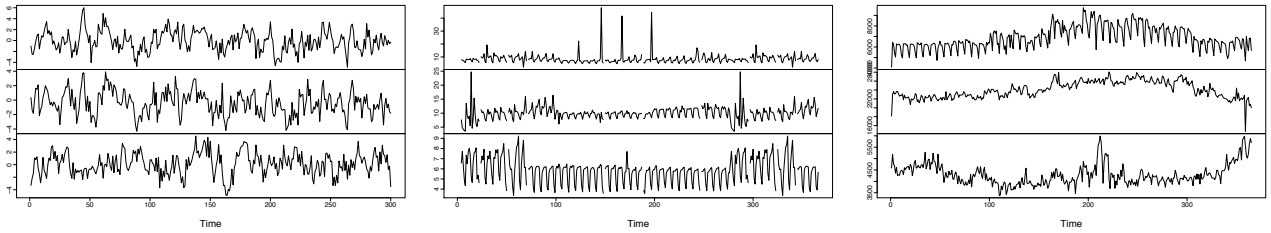


Figure 2: The hierarchical structure of the Wikipedia data.

[1] https://archive.ics.uci.edu/ml/datasets/PEMS-SF
[2] https://www.kaggle.com/c/web-traffic-time-series-forecasting/data



(a) Simulated data (b) Road traffic: 3 sensors (c) Wikipedia data: 3 websites

Figure 3: Visualization of the various time series considered in the experiments.

5.2 Methodology

In order to ensure stationarity, we remove the trend and seasonal component using the Seasonal and Trend decomposition using Loess (STL). We also apply a log transformation to stabilize the variance. Finally, we replace any outliers using linear interpolation, as implemented in the tsclean function in the forecast package for R.

For each series in each hierarchy, we produce one-step-ahead rolling-origin base forecasts for both the validation and test sets from a model fitted to the last T_1 observations. We use $T_1 = 400, 86, 120$ for the simulated, road traffic and Wikipedia data sets, respectively. We consider both ARIMA and Exponential Smoothing (ETS) forecasting algorithms fitted using the automatic model selection algorithm of [9], which is implemented in the forecast package for R. While the simulated data are generated using an ARMA process, using an ETS model allows us to simulate what happens in practice, where the true data generating processes is unknown and misspecified.

We consider the following hierarchical forecasting methods:

BASE. The base forecasts defined in (4) produced as described in the previous paragraph. Recall that the base forecasts are (probably) incoherent.

BU. The bottom-up forecasts defined in (3), i.e. where (5) is computed using (6).

MinTsam. The MinT forecasts given in (10) where (11) is computed with $\alpha = 1$.

MinTols. The MinT forecasts given in (10) where (11) is computed with $\hat{\boldsymbol{W}}_d = \hat{\sigma}^2 \boldsymbol{I}_n$ and $\alpha = 0$.

MinTshr. The MinT forecasts given in (10) where (11) is with an optimal shrinkage factor α, as described in [15].

ERM. The ERM forecasts where (5) is computed using (25).

ERMreg. The ERMreg forecasts where (5) is computed using (26). We use cross-validation to select the best value of λ, as implemented in the glmnet package in R.

ERMregbu. The ERMreg forecasts where (5) is computed using (27).

6 RESULTS

For each data set, we compute the one-step-ahead mean squared forecast errors for all series in the hierarchy, i.e.

$$\frac{1}{T_{\text{test}}} \sum_{t=T}^{T+T_{\text{test}}-1} \|\boldsymbol{y}_{t+1} - \boldsymbol{S}\boldsymbol{P}\hat{\boldsymbol{y}}_t(1)\|_2^2 ,$$

and the bottom series only, i.e.

$$\frac{1}{T_{\text{test}}} \sum_{t=T}^{T+T_{\text{test}}} \|\boldsymbol{b}_{t+1} - \boldsymbol{P}\hat{\boldsymbol{y}}_t(1)\|_2^2 .$$

where T_{test} is the number of observations in the test set. We have $T_{\text{test}} = 200, 120, 120$ for the simulated, road traffic and Wikipedia data, respectively. We then average the results for the 100 hierarchies.

We start by analyzing the results for the simulated data. Table 1 and 2 give the forecasting errors for the small and the large hierarchy, respectively.

	ARIMA		**ETS**	
	All	Bottom	All	Bottom
BASE	39.32 (0.33)	**8.20** (0.05)	43.64 (0.42)	9.09 (0.07)
BU	**39.11** (0.32)	8.20 (0.05)	43.59 (0.42)	9.09 (0.07)
ERM	40.23 (0.41)	8.43 (0.07)	41.26 (0.43)	8.59 (0.07)
ERMreg	39.45 (0.33)	8.29 (0.05)	40.83 (0.40)	8.53 (0.07)
ERMregbu	39.29 (0.33)	8.24 (0.05)	**40.76** (0.40)	**8.51** (0.07)
MINTsam	40.97 (0.48)	8.55 (0.08)	44.90 (0.71)	9.38 (0.14)
MINTols	39.14 (0.33)	8.20 (0.05)	43.63 (0.42)	9.10 (0.07)
MINTshr	39.15 (0.32)	8.20 (0.05)	43.61 (0.42)	9.09 (0.07)

Table 1: Forecast errors for the simulated data with a small hierarchy. Standard errors are given in parentheses.

Table 1 and 2 show that when using an ETS model, i.e when the base model is misspecified, our ERMreg methods outperform all the other methods. This confirms the robustness of our methods to misspecification of the base forecasts. When an ARMA model is

	ARIMA		ETS	
	All	Bottom	All	Bottom
BASE	11.03 (0.05)	2.05 (0.00)	12.21 (0.06)	2.28 (0.00)
BU	10.87 (0.05)	2.05 (0.00)	12.16 (0.06)	2.28 (0.00)
ERM	18.41 (0.13)	3.53 (0.02)	22.01 (0.19)	4.18 (0.02)
ERMreg	11.69 (0.05)	2.28 (0.00)	12.37 (0.06)	2.37 (0.00)
ERMregbu	10.88 (0.05)	2.06 (0.00)	**11.69** (0.05)	**2.22** (0.00)
MINTsam	12.13 (0.10)	2.30 (0.01)	13.04 (0.07)	2.44 (0.01)
MINTols	11.00 (0.05)	2.06 (0.00)	12.21 (0.06)	2.28 (0.00)
MINTshr	**10.85** (0.05)	**2.05** (0.00)	12.16 (0.06)	2.28 (0.00)

Table 2: Forecast errors ($\times 10^2$) for the simulated data with a large hierarchy. Standard errors are given in parentheses.

	ARIMA		ETS	
	All	Bottom	All	Bottom
BASE	34.47 (0.35)	1.72 (0.01)	36.97 (0.36)	1.82 (0.01)
BU	32.85 (0.31)	1.72 (0.01)	**32.95** (0.31)	1.82 (0.01)
ERMreg	31.37 (0.53)	1.78 (0.01)	36.78 (0.78)	**1.81** (0.02)
ERMregbu	**31.24** (0.32)	1.72 (0.01)	36.68 (0.61)	1.88 (0.01)
MINTols	34.18 (0.34)	1.73 (0.01)	36.66 (0.36)	1.84 (0.01)
MINTshr	32.67 (0.31)	**1.72** (0.01)	33.10 (0.31)	1.81 (0.01)

Table 3: Forecast errors for the road traffic data. Standard errors are given in parentheses.

	ARIMA		ETS	
	All	Bottom	All	Bottom
BASE	70.98 (0.68)	10.94 (0.09)	74.01 (0.71)	**11.10** (0.10)
BU	69.71 (0.65)	10.94 (0.09)	**72.43** (0.63)	11.10 (0.10)
ERMreg	72.22 (0.69)	11.19 (0.10)	73.87 (0.73)	11.21 (0.10)
ERMregbu	69.61 (0.64)	11.01 (0.09)	72.95 (0.65)	11.13 (0.10)
MINTols	70.29 (0.67)	10.96 (0.09)	73.69 (0.70)	11.11 (0.10)
MINTshr	**69.31** (0.66)	**10.93** (0.09)	72.78 (0.65)	11.10 (0.10)

Table 4: Forecast errors for the Wikipedia data. Standard errors are given in parentheses.

used, i.e. the base model is well-specified, our methods still provide good forecasts although this is the ideal scenario for MINT.

In Table 2, for the large hierarchy, we can see that MINTsam and ERM have larger forecast errors than the other methods. This is not surprising since both MINTsam and ERM do not use regularization, and as a result suffer from the accumulation of errors in high dimensions. In fact, the larger number of parameters to estimate significantly increases the forecast variance which leads to poor forecast accuracy. We can also see that ERM is more affected by high dimensionality than MINTsam. With MINTsam, the accumulation of errors will indirectly affect the estimation of the combination weights. In fact, in order to compute the combination weights, MINTsam plugs $\hat{\boldsymbol{W}}_s$ given in (11) into (10). However, since ERM solves the high-dimensional regression problem given in (15), it will be directly affected by the accumulation of errors. Nevertheless, we can see that MINTshr and ERMreg significantly reduce the forecast errors thanks to regularization which reduces forecast variance. The robustness of our ERMreg methods to misspecification can also be seen in Table 2.

Table 3 and 4 give the results for the road traffic and Wikipedia data, respectively. For clarity, given that MINTsam and ERM provide very poor forecasts (due to high-dimensionality, as discussed above), we omit their results from the tables. For both data sets, our method ERMregbu provides competitive results compared to the other methods.

Figure 4 plots the weight matrix of different methods computed for one hierarchy of the road traffic data set. Each entry of the matrix gives $|\boldsymbol{P}_{ij}|$, i.e. the absolute value of the ith row and the jth column of matrix $\boldsymbol{P}$. A white cell has zero value, and a darker cell has a higher value. In addition to differences in forecast accuracy, Figure 4 shows that the different methods provide combination weights with different structures. In fact, we can see that MINTshr and ERM have a dense weight matrix while ERMreg and ERMregbu provide sparse matrices, as excepted. The diagonal line for some matrices shows that the base forecast for the series under consideration has been selected in the forecast combination.

7 CONCLUSION

We considered the problem of forecasting multiple time series under hierarchical constraints. We proposed a new forecasting method which does not assume or constrain the forecasts to be unbiased but aim to find the best tradeoff between bias and forecast variance. Our method involves solving a multi-response regression problem for which we provided a closed-form solution. To deal with high-dimensional hierarchies, we also proposed a regularization method that has a theoretical justification, and proved its asymptotic properties. We also showed that our problems can be divided into multiple independent problems which can be solved efficiently in parallel. We compared our methods both theoretically and empirically with the state-of-the-art methods for hierarchical forecasting. Using both simulated and two real-world time series data, we showed our methods is competitive with the state-of-the-art methods in various conditions.

One direction for future work is the improvement of the computational complexity of our methods, as well as the development of new regularization methods which exploit the hierarchical structure of the data. Another important problem is the computation of probabilistic forecasts for high-dimensional hierarchies [1]. Finally, it is also possible to extend and adapt our methods to the problem of multi-source learning/forecasting with hierarchical structures [18–20].

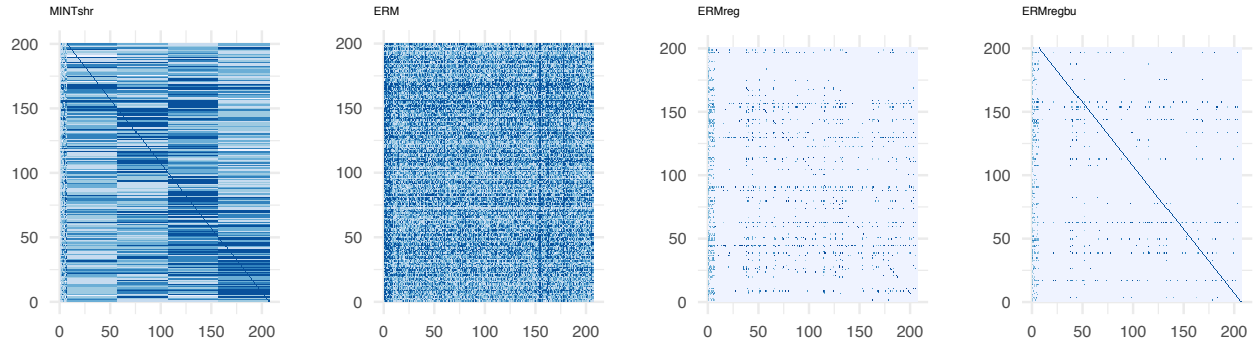


Figure 4: Matrix combination weights for different hierarchical forecasting methods.

8 ACKNOWLEDGEMENTS

This work is partially supported by the Australian Research Council under Grant No. DP150104292 and Grant No. DE170100713. We thank anonymous reviewers and Shanika Wickramasuriya for their insightful and constructive comments.

REFERENCES

[1] Ben Taieb, James Taylor, and Rob Hyndman. 2017. Coherent Probabilistic Forecasts for Hierarchical Time Series. In *Proceedings of The 34st International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Vol. 70. 3348–3357.

[2] Peter J Bickel and Elizaveta Levina. 2008. Regularized Estimation of Large Covariance Matrices. *Annals of statistics* 36, 1 (2008), 199–227.

[3] Peter J Bickel, Elizaveta Levina, and Others. 2008. Covariance regularization by thresholding. *Annals of statistics* 36, 6 (2008), 2577–2604.

[4] Marco Cuturi. 2011. Fast global alignment kernels. In *Proceedings of the 28th international conference on machine learning (ICML-11)*. 929–936.

[5] Jerome Friedman, Trevor Hastie, Holger Höfling, and Robert Tibshirani. 2007. Pathwise coordinate optimization. *The Annals of Applied Statistics* 1, 2 (Dec. 2007), 302–332.

[6] Tilmann Gneiting. 2011. Making and evaluating point forecasts. *J. Amer. Statist. Assoc.* 106, 494 (June 2011), 746–762.

[7] T Hastie, R Tibshirani, and M Wainwright. 2015. *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press.

[8] Trevor J Hastie, Robert Tibshirani, and Jerome H Friedman. 2008. *The elements of statistical learning*. Vol. 18. Springer-Verlag. 764 pages.

[9] Rob Hyndman and Yeasmin Khandakar. 2008. Automatic Time Series Forecasting: The forecast Package for R. *Journal of Statistical Software* 27, 1 (2008), 1–22.

[10] Rob J Hyndman, Roman A Ahmed, George Athanasopoulos, and Han Lin Shang. 2011. Optimal combination forecasts for hierarchical time series. *Computational statistics & data analysis* 55, 9 (Sept. 2011), 2579–2589.

[11] Rob J Hyndman, Alan J Lee, and Earo Wang. 2016. Fast computation of reconciled forecasts for hierarchical and grouped time series. *Computational Statistics & Data Analysis* 97 (May 2016), 16–32.

[12] Keith Knight. 2008. Shrinkage estimation for nearly singular designs. *Econometric Theory* 24, 2 (April 2008), 323–337.

[13] Mirko Kremer, Enno Siemsen, and Douglas J Thomas. 2016. The Sum and Its Parts: Judgmental Hierarchical Forecasting. *Management Science* 62, 9 (2016), 2745–2764.

[14] Vitaly Kuznetsov and Mehryar Mohri. 2015. Learning Theory and Algorithms for Forecasting Non-stationary Time Series. In *Advances in Neural Information Processing Systems 28*, C Cortes, N D Lawrence, D D Lee, M Sugiyama, R Garnett, and R Garnett (Eds.). Curran Associates, Inc., 541–549.

[15] Juliane Schäfer and Korbinian Strimmer. 2005. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology* 4 (Nov. 2005), 32.

[16] A Timmermann. 2006. Forecast combinations. In *Handbook of Economic Forecasting*. Vol. 1. Elsevier, 135–196.

[17] Shanika L Wickramasuriya, George Athanasopoulos, and Rob J Hyndman. 2018. Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *J. Amer. Statist. Assoc.* (March 2018), 1–45.

[18] Pei Yang and Jingrui He. 2015. Model Multiple Heterogeneity via Hierarchical Multi-Latent Space Learning. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1375–1384.

[19] Chenwei Zhang, Sihong Xie, Yaliang Li, Jing Gao, Wei Fan, and Philip S Yu. 2016. Multi-source Hierarchical Prediction Consolidation. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 2251–2256.

[20] Liang Zhao, Jieping Ye, Feng Chen, Chang-Tien Lu, and Naren Ramakrishnan. 2016. Hierarchical Incomplete Multi-source Feature Learning for Spatiotemporal Event Forecasting. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2085–2094.

A REGULARITY CONDITIONS

A.1 Definitions

In what follows, $O_p(\cdot)$ and $o_p(\cdot)$ denote the usual big O and little o in probability. $\xrightarrow{p}$ and $\xrightarrow{d}$ denote convergence in probability and convergence in distribution respectively.

Definition 9. *$U_T(s) \xrightarrow{e-d} U(s)$ if for any closed rectangles $R_1, ..., R_k$ with open interiors $R_1^0, ..., R_k^0$ and any real numbers $a_1, ..., a_k$,*

$$\begin{aligned}
&P\{\inf_{s\in R_1} U(s) > a_1, ..., \inf_{s\in R_k} U(s) > a_k\} \\
\leq &\liminf_{T\to\infty} P\{\inf_{s\in R_1} U(s) > a_1, ..., \inf_{s\in R_k} U(s) > a_k\} \\
\leq &\limsup_{T\to\infty} P\{\inf_{s\in R_1^0} U(s) > a_1, ..., \inf_{s\in R_k^0} U(s) > a_k\} \\
\leq &P\{\inf_{s\in R_1^0} U(s) > a_1, ..., \inf_{s\in R_k^0} U(s) > a_k\}
\end{aligned}$$

A.2 Regularity conditions for Theorem 6

D1 (near singularity)

(i) Let $\boldsymbol{M}_q = 1/q \sum_{t=1}^{T} \boldsymbol{x}_t \boldsymbol{x}_t^{'}$. $\boldsymbol{M}_q$ is nonsingular for all q whereas its probability limit, $\boldsymbol{M}$ is singular.

(ii) As $q \to \infty$, $\boldsymbol{x}_t$ satisfies

$$\frac{1}{q} \max_{1\leq t\leq q} \boldsymbol{x}_t \boldsymbol{M}_q^{-1} \boldsymbol{x}_t \to 0.$$

(iii) For some sequence of constants, $\{c_q\}$ tending to infinity,

$$c_q(\boldsymbol{M}_q - \boldsymbol{M}) \xrightarrow{p} \boldsymbol{D}$$

where $\boldsymbol{D}$ is a positive definite matrix on the null space of $\boldsymbol{M}$.

D2 (Shrinkage parameter and the rate of convergence) Let the shrinkage parameter λ be such that $\lambda/r_q \to \lambda_0$ where λ_0 is a positive constant and $r_q = \sqrt{q/c_q}$.

A.3 Regularity conditions for Corollary 8

To ensure that the sample objective function (15) converges to the limit objective function (31) and the minimizer of the former matches with that of the latter, we assume the following.

B1 (Data) $\{\boldsymbol{y}_t\}_{t=1}^{T}$ is strictly stationary, ergodic with exponentially decaying mixing coefficients.

B2 (Parameter Space) The true parameter $\boldsymbol{P}_0$ involving the Moore-Penrose generalized inverse belongs to the interior of a compact convex subset $\mathcal{P}$ of the Euclidean space, $\mathbb{R}^{m\times n}$.

B3 (Objective function) The nonrandom limit objective $\mathcal{L}(\boldsymbol{P})$ exists, is minimized uniquely at $\boldsymbol{P}_0$ involving the Moore-Penrose generalized inverse and is equicontinuous on $\boldsymbol{P}$, i.e. it is well behaved in the neighborhood around $\boldsymbol{P}_0$, i.e. $\|\boldsymbol{P} - \boldsymbol{P}_0\| \leq \varepsilon$ at some ε. For any $\eta > 0$, there exists $\varepsilon > 0$,

$$\liminf_{T\to\infty} \mathrm{P}\left(\sup_{\|\boldsymbol{P}-\boldsymbol{P}_0\|\geq\eta} (L_T(\boldsymbol{P}) - L_T(\boldsymbol{P}_0))) \leq -\varepsilon\right) = 1,$$

and for any $\eta > 0$, there exists $\varepsilon > 0$, s.t.

$$\inf_{\|\boldsymbol{P}-\boldsymbol{P}_0\|>\eta} \mathcal{L}(\boldsymbol{P}) - \mathcal{L}(\boldsymbol{P}_0) \geq \varepsilon.$$

B PROOFS

B.1 Proof of Theorem 3

If condition. First, we show that $\boldsymbol{P}_{\mathrm{BU}}$ is at least one of solutions of the objective function in (28). Note that $\boldsymbol{P}_{\mathrm{BU}}\boldsymbol{S} = \boldsymbol{I}$ and

$$\sum_{i=1}^{m}\sum_{j=1}^{n} |P_{\mathrm{BU}[ij]}|^0 = m.$$

Therefore, in order to show that $\boldsymbol{P}_{\mathrm{BU}}$ could be a solution, we must show that

$$\min(\sum_{i=1}^{m}\sum_{j=1}^{n} |P_{ij}|^0) = m$$

for any $\boldsymbol{P}$ such that $\boldsymbol{PS} = \boldsymbol{I}$.

Note that for $\boldsymbol{P}$ such that $\boldsymbol{PS} = \boldsymbol{I}$, $rank(\boldsymbol{PS}) = rank(\boldsymbol{I}) = m \leq \min(rank(\boldsymbol{P}), rank(\boldsymbol{S}))$, and $rank(\boldsymbol{S}) = m$. So, $rank(\boldsymbol{P}) \geq m$. Since $rank(\boldsymbol{I}) = m$, $rank(\boldsymbol{P}) \geq m$. This implies that the L_0 norm of $\boldsymbol{P}$ must be greater or equal to m. This is because the rank of $\boldsymbol{P}$ could not be greater than or equal to m if the number of non-zero values of $\boldsymbol{P}$ is lower than m. As a result, $\boldsymbol{P}_{\mathrm{BU}}$ is at least one of solutions of the above objective function.

Only if condition. Now, we turn our attention to showing that $\boldsymbol{P}_{\mathrm{BU}}$ is the unique solution for (28). Suppose that $\boldsymbol{P}_{\mathrm{BU}}$ is only one of many solutions for the above objective function. Then, there exists $\boldsymbol{P}^*$ such that $\boldsymbol{P}^* = \boldsymbol{P}_{\mathrm{BU}} + \boldsymbol{C}$ with a non-zero $m \times n$ sized matrix, $\boldsymbol{C}$ and $\boldsymbol{P}^*\boldsymbol{S} = \boldsymbol{I}$. Since $\min(\sum_{i=1}^{m}\sum_{j=1}^{n} |P_{ij}^*|^0) = m$, there are two cases we need to consider.

The first case is when one of the non-zero values of $\boldsymbol{P}^*$ is different from the non-zero values of $\boldsymbol{P}_{\mathrm{BU}}$ in the same location. This is contradictory since there is no such solution because in that case $\boldsymbol{P}^*\boldsymbol{S} \neq \boldsymbol{I}$, i.e. it should be $\boldsymbol{P}_{\mathrm{BU}} = \boldsymbol{P}^*$.

The second case is when one of the non-zero values of $\boldsymbol{P}_{\mathrm{BU}}$ becomes zero and one of zero values of $\boldsymbol{P}_{\mathrm{BU}}$ becomes non-zero. This should be satisfied to ensure that $\min(\sum_{i=1}^{m}\sum_{j=1}^{n} |P_{ij}^*|^0) = m$. Note that

$$\begin{aligned}
&\boldsymbol{P}_{\mathrm{BU}}\boldsymbol{S} + \boldsymbol{CS} = \boldsymbol{I} \implies \boldsymbol{CS} = \boldsymbol{0} \implies \\
&\boldsymbol{C}_a\boldsymbol{A} + \boldsymbol{C}_b = \boldsymbol{0} \implies \boldsymbol{C}_b = -\boldsymbol{C}_a\boldsymbol{A}
\end{aligned}$$

Then, one of the diagonal values of $\boldsymbol{C}_b$ should be -1 and all the others are zeros while one of the values of $\boldsymbol{C}_a$ should be non-zero and others are zeros.

Note that $\boldsymbol{C}_{b,ii} = -\sum_{j=}^{n} \boldsymbol{C}_{a,ij}\boldsymbol{A}_{ji}$ for $j = 1, ..., n$. Suppose that $i = k$ for some fixed k. Then, $\boldsymbol{C}_{b,kk} = -1$ and only one non-zero value belongs to $\boldsymbol{C}_{a,kj}$ and it should be 1. However, if any of $\boldsymbol{A}_{jk}$ contains at least one non-zero value except for $j = k$, then the equality will not hold. But this is the case because $\boldsymbol{A}$ is the aggregation matrix. This is contradictory since $\boldsymbol{C}_b = -\boldsymbol{C}_a\boldsymbol{A}$. Therefore, $\boldsymbol{C} = \boldsymbol{0}$.

B.2 Proof of Theorem 6

Recall that for $\boldsymbol{y}_t = (\boldsymbol{a}_t\ \boldsymbol{b}_t)'$, $\boldsymbol{y}_t = \boldsymbol{S}\boldsymbol{b}_t$ where $\boldsymbol{S} = \begin{bmatrix}\boldsymbol{A}' & \boldsymbol{I}_m\end{bmatrix}'$. That is, $\boldsymbol{a}_t$ can be expressed as linear combinations of $\boldsymbol{b}_t$. This implies that our regression problem is subject to the issue of nearly singular design. As seen from Remark 2, this is the natural consequence of hierarchical time series forecasting given that $\boldsymbol{Y}'\boldsymbol{Y}$ is singular whereas $\hat{\boldsymbol{Y}}'\hat{\boldsymbol{Y}}$ can be nonsingular due to estimation error.

Recall that

$$\boldsymbol{Q}_q(\boldsymbol{\beta}) = \|\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2/q + \lambda\|\boldsymbol{\beta}\|_1, \tag{41}$$

where $\boldsymbol{X} = (\boldsymbol{S} \otimes \hat{\boldsymbol{Y}})$, $\boldsymbol{z} = \mathrm{vec}(\boldsymbol{Y}) - \boldsymbol{X}\,\mathrm{vec}(\boldsymbol{P}_{\mathrm{BU}}')$, $\boldsymbol{\beta} = \mathrm{vec}(\boldsymbol{P}' - \boldsymbol{P}_{\mathrm{BU}}')$, $N = T - T_1 - h + 1$, and $q = Nm$.

It is worth noting that $\boldsymbol{Q}_q(\cdot)$ is convex by construction. Here we borrow the concept of near collinearity or near singularity introduced in [12], i.e. for a regressor matrix $\boldsymbol{X}$, $\boldsymbol{M}_q$ is nonsingular but its limit, $\boldsymbol{M}$ is singular where

$$\boldsymbol{M}_q = \frac{1}{q}\sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{x}_t'$$
$$\boldsymbol{M} = \mathrm{plim}_{q\to\infty}\boldsymbol{M}_q$$

Define $\boldsymbol{g} = r_q(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$. Note that via reparameterization, (41) can be rewritten as

$$\boldsymbol{Q}_q(\boldsymbol{g}) = (\|\boldsymbol{\varepsilon} - \boldsymbol{X}\boldsymbol{g}/r_q\|_2^2 - \|\boldsymbol{\varepsilon}\|_2^2)/q + \lambda(\|\boldsymbol{\beta} + \boldsymbol{g}/r_q\|_1 - \|\boldsymbol{\beta}\|_1) \tag{42}$$

where $\boldsymbol{\varepsilon} = \mathrm{vec}(\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta})$ and λ and r_q satisfies the assumption **D2**. Note that if $\hat{\boldsymbol{\beta}}$ minimizes $\boldsymbol{Q}_q(\boldsymbol{\beta})$, $\hat{\boldsymbol{g}} :\equiv r_q(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ minimizes $\boldsymbol{Q}_q(\boldsymbol{g})$.

Let $\boldsymbol{Q}_0(\boldsymbol{g}) = \lim_{q\to\infty}\boldsymbol{Q}_q(\boldsymbol{g})$. Due to convexity of $\boldsymbol{Q}_q(\cdot)$, $\boldsymbol{Q}_0(\cdot)$ and $\boldsymbol{Q}_q(\cdot)$ is stochastically equi-lower semicontinuous, which implies that $\boldsymbol{Q}_q(\boldsymbol{g}) \overset{e-d}{\to} \boldsymbol{Q}_0(\boldsymbol{g})$ (see [12]). Trivially, $r_q(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = O_p(1)$ due to the convexity of $\boldsymbol{Q}_q(\boldsymbol{g})$ and the usual asymptotic theory related to the LASSO estimation.

Combined with $\hat{\boldsymbol{g}} = O_p(1)$, the epi-convergence $\boldsymbol{Q}_q(\boldsymbol{g}) \overset{e-d}{\to} \boldsymbol{Q}_0(\boldsymbol{g})$, implies that the limiting distribution of $\hat{\boldsymbol{g}}$ is equivalent to that of the minimizer of the limit objective function of the sample objective function, i.e. $\hat{\boldsymbol{g}} \overset{d}{\to} \mathrm{argmin}\,\boldsymbol{Q}_0(\boldsymbol{g})$.

The remainder is to obtain the finite-dimensional limit of $\boldsymbol{Q}_q(\boldsymbol{g})$. Note that

$$(\|\boldsymbol{\varepsilon} - \boldsymbol{X}\boldsymbol{g}/c_q\|_2^2 - \|\boldsymbol{\varepsilon}\|_2^2)/q \overset{d}{\to} -2\boldsymbol{g}'\boldsymbol{V} + \boldsymbol{g}'\boldsymbol{D}\boldsymbol{g}$$
$$\lambda(\|\boldsymbol{\beta} + \boldsymbol{g}/r_q\|_1 - \|\boldsymbol{\beta}\|_1) \overset{d}{\to} \lambda_0\sum_{j=1}^{m\times n}\{g_j\mathrm{sgn}(\beta_j) + |g_j|I(\beta_j = 0)\}$$

where $\boldsymbol{V}$ is defined as a zero-mean multivariate normal random vector such that $\mathrm{Var}(\boldsymbol{g}'\boldsymbol{V}) = \sigma^2\boldsymbol{g}'\boldsymbol{D}\boldsymbol{g}$ for $\boldsymbol{g}$ satisfying $\boldsymbol{M}\boldsymbol{g} = \boldsymbol{0}$. This completes the proof.

B.3 Proof of Corollary 8

We show that $\hat{\boldsymbol{P}}_{\mathrm{ERM}}$ computed by solving (15) and $\hat{\boldsymbol{P}}_{MinT}$ are asymptotically equivalent to $\boldsymbol{P}_U$ where $\boldsymbol{P}_U$ is defined in (37).

Consider $\hat{\boldsymbol{W}}$ defined in (11), and assume $\alpha \to 0$ when $T \to \infty$. Then, by the usual laws of large numbers for the stationary and ergodic data, $\hat{\boldsymbol{W}} \overset{p}{\to} \boldsymbol{W}_u$. Furthermore, using Slutsky theorem, we have

$$\hat{\boldsymbol{P}}_{MinT} \overset{p}{\to} \boldsymbol{P}_U.$$

In order to show that $\hat{\boldsymbol{P}}_{\mathrm{ERM}} \overset{p}{\to} \boldsymbol{P}_U$, we note that

$$\sup_{\boldsymbol{P}\in\mathcal{P}}|L_T(\boldsymbol{P},\boldsymbol{y}) - \mathcal{L}(\boldsymbol{P})| \le \sup_{\boldsymbol{P}\in\mathcal{P}}|L_T(\boldsymbol{P},\boldsymbol{y}) - \mathrm{E}L_T(\boldsymbol{P},\boldsymbol{y})| + \sup_{\boldsymbol{P}\in\mathcal{P}}|\mathrm{E}L_T(\boldsymbol{P},\boldsymbol{y}) - \mathcal{L}(\boldsymbol{P})|$$

By construction, $L_T(\boldsymbol{P},\boldsymbol{y})$ is continuous in $\boldsymbol{P}\in\mathcal{P}$ for all $\boldsymbol{y}$ and is a measurable function of $\boldsymbol{y}$ for all $\boldsymbol{P}\in\mathcal{P}$. Due to the assumptions **B1**-**B3**, $L_T(\boldsymbol{P},\boldsymbol{y})$ is stochastic equicontinuous such that

$$\sup_{\boldsymbol{P}\in\mathcal{P}}\sup_{\boldsymbol{P}^*\in B(\boldsymbol{P},\epsilon)}|L_T(\boldsymbol{P},\boldsymbol{y}) - L_T(\boldsymbol{P}^*,\boldsymbol{y})| = o_p(1)$$

where $B(\boldsymbol{P},\epsilon)$ is a closed ball in $\mathcal{P}$ of radius $\epsilon \ge 0$ centered at $\boldsymbol{P}$. Due to the pointwise convergence of $L_T(\boldsymbol{P},\boldsymbol{y})$ to $\mathcal{L}(\boldsymbol{P})$ on $\boldsymbol{P}$ and uniform stochastic equicontinuity of $L_T(\boldsymbol{P},\boldsymbol{y})$,

$$\sup_{\boldsymbol{P}\in\mathcal{P}}|L_T(\boldsymbol{P},\boldsymbol{y}) - \mathcal{L}(\boldsymbol{P})| = o_p(1).$$

Since we define $\boldsymbol{P}_0$ as the true parameter involving the Moore-Penrose generalized inverse in **B2**, $\boldsymbol{P}_U$ is the unique one for $\boldsymbol{P}_0$ in a sense that $\boldsymbol{P}_U$ minimizes $\mathcal{L}(\boldsymbol{P})$ uniquely in the mean squared error sense. Note also that $\hat{\boldsymbol{P}}_{\mathrm{ERM}}$ is the unique minimizer of $L_T(\boldsymbol{P},\boldsymbol{y})$ involving the Moore-Penrose generalized inverse and shares the same criterion as $\boldsymbol{P}_U$. Therefore, due to the assumption **B3** and the argmax theorem, $\hat{\boldsymbol{P}}_{\mathrm{ERM}} \overset{p}{\to} \boldsymbol{P}_U$.