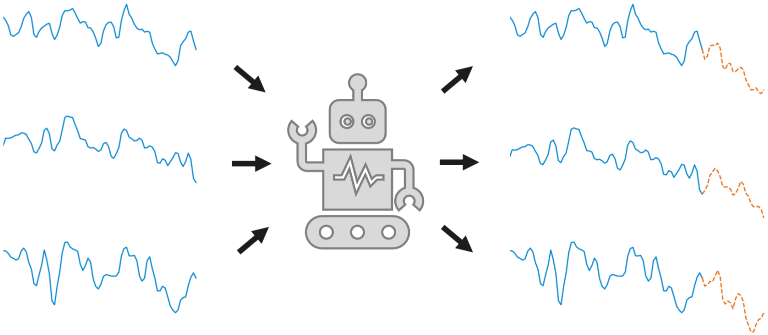
Theoretically Analysing Multi-Task Regression with Application to Time Series Forecasting

Romain Ilbert, Malik Tiomoko, Cosme Louart, Ambroise Odonnat, **Vasilii Feofanov**, Themis Palpanas, Ievgen Redko

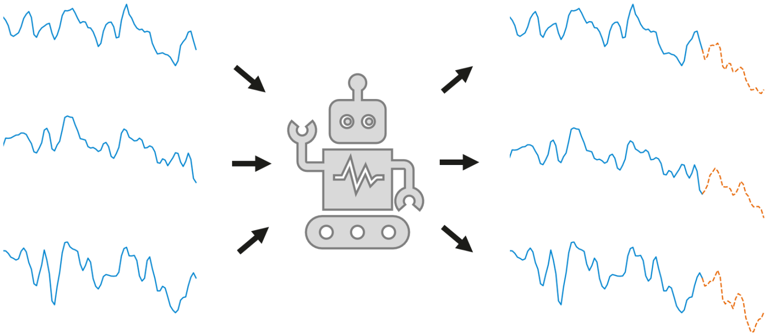
Huawei Paris Noah's Ark Lab
Paris Descartes University
The Chinese University of Hong Kong

NeurIPS 2024, Spotlight
NeurIPS @ Paris, December 5

Multivariate Time Series Forecasting

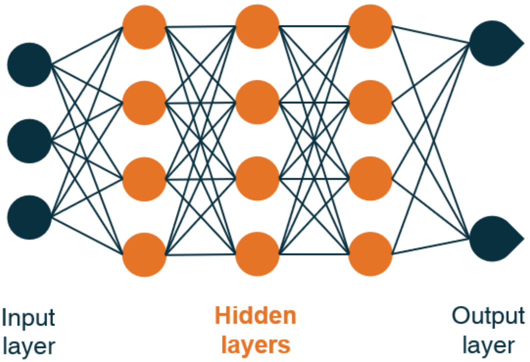


Multivariate Time Series Forecasting



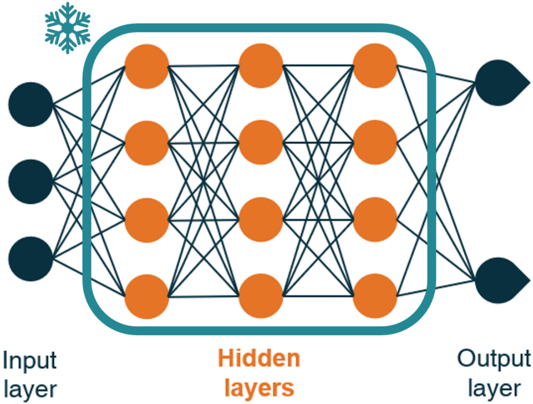
- Captures complex dependencies enhancing forecasting quality.
- Key for fields like economics, climate, and finance.
- Motivation of the paper: study it theoretically

Linear Setting



- Modern forecasters are deep-learning-based.
- For theoretical derivations, we consider linear forecasting,

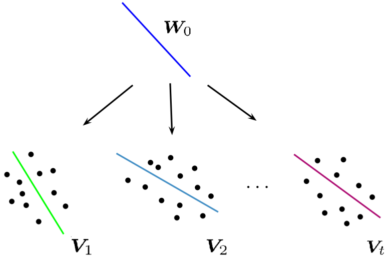
Linear Setting



- Modern forecasters are deep-learning-based.
- For theoretical derivations, we consider linear forecasting,
- Or a deep model with a frozen feature extractor.

Our Framework: Multivariate as Multi-Task

$$\boldsymbol{W}_t = \boldsymbol{W}_0 + \boldsymbol{V}_t$$

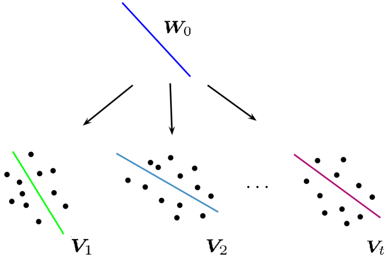


- We view multivariate forecasting as a **multi-task** problem.

[1] (Xu et al., 2013) Multi-output least-squares support vector regression machines.

Our Framework: Multivariate as Multi-Task

$$\boldsymbol{W}_t = \boldsymbol{W}_0 + \boldsymbol{V}_t$$



- We view multivariate forecasting as a **multi-task** problem.
- Soft Parameter Sharing approach[1]:
  - $\boldsymbol{W}_0$ catches the common part, reducing task overfitting.
  - $\boldsymbol{V}_t$ are task-specific terms for individual biases.

[1](Xu et al., 2013) Multi-output least-squares support vector regression machines.

Multi-Task Forecasting Framework

- **Problem Setup.** Time series channel $t \in \{1, \ldots, T\}$ is viewed as a distinct task:

$$\text{Training data: } \boldsymbol{X}^{(t)} \in \mathbb{R}^{d \times n_t}, \quad \text{Responses: } \boldsymbol{Y}^{(t)} \in \mathbb{R}^{q \times n_t},$$
$$d \text{ is seq. length}, q \text{ is pred. horizon}, n_t \text{ is sample size}$$

Multi-Task Forecasting Framework

- **Problem Setup.** Time series channel $t \in \{1, \ldots, T\}$ is viewed as a distinct task:

$$\text{Training data: } \boldsymbol{X}^{(t)} \in \mathbb{R}^{d \times n_t}, \quad \text{Responses: } \boldsymbol{Y}^{(t)} \in \mathbb{R}^{q \times n_t},$$
$$d \text{ is seq. length}, q \text{ is pred. horizon}, n_t \text{ is sample size}$$

- **Linear Signal-Plus-Noise Model:**

$$\boldsymbol{Y}^{(t)} = \frac{{\boldsymbol{X}^{(t)}}^\top \boldsymbol{W}_t}{\sqrt{Td}} + \boldsymbol{\epsilon}^{(t)}, \quad \forall t,$$

  - $\boldsymbol{\epsilon}^{(t)}$ is noise,
  - $\boldsymbol{W}_t = \boldsymbol{W}_0 + \boldsymbol{V}_t$ combines shared $\boldsymbol{W}_0$ and task-specific components $\boldsymbol{V}_t$:

Optimization Objective

- **Objective.** We aim to estimate the shared component $\boldsymbol{W}_0$ and task-specific components $\{\boldsymbol{V}_t\}_{t=1}^{T}$ by solving:

$$\min \frac{1}{2\lambda}\|\boldsymbol{W}_0\|_F^2 + \frac{1}{2}\sum_{t=1}^{T}\frac{\|\boldsymbol{V}_t\|_F^2}{\gamma_t} + \frac{1}{2}\sum_{t=1}^{T}\left\|\boldsymbol{Y}^{(t)} - \frac{{\boldsymbol{X}^{(t)}}^\top \boldsymbol{W}_t}{\sqrt{Td}}\right\|_F^2$$

  - λ controls impact of the common part on a final prediction.

Optimization Objective

- **Objective.** We aim to estimate the shared component $\boldsymbol{W}_0$ and task-specific components $\{\boldsymbol{V}_t\}_{t=1}^T$ by solving:

$$\min \frac{1}{2\lambda}\|\boldsymbol{W}_0\|_F^2 + \frac{1}{2}\sum_{t=1}^T \frac{\|\boldsymbol{V}_t\|_F^2}{\gamma_t} + \frac{1}{2}\sum_{t=1}^T \left\|\boldsymbol{Y}^{(t)} - \frac{{\boldsymbol{X}^{(t)}}^\top \boldsymbol{W}_t}{\sqrt{Td}}\right\|_F^2$$

  - λ controls impact of the common part on a final prediction.
  - γ_t controls overfitting strength to the task t.

Optimization Objective

- **Objective.** We aim to estimate the shared component $\boldsymbol{W}_0$ and task-specific components $\{\boldsymbol{V}_t\}_{t=1}^T$ by solving:

$$\min \frac{1}{2\lambda}\|\boldsymbol{W}_0\|_F^2 + \frac{1}{2}\sum_{t=1}^{T}\frac{\|\boldsymbol{V}_t\|_F^2}{\gamma_t} + \frac{1}{2}\sum_{t=1}^{T}\left\|\boldsymbol{Y}^{(t)} - \frac{{\boldsymbol{X}^{(t)}}^\top \boldsymbol{W}_t}{\sqrt{Td}}\right\|_F^2$$

  - λ controls impact of the common part on a final prediction.
  - γ_t controls overfitting strength to the task t.
  - Closed-form solution.

Optimization Objective

- **Objective.** We aim to estimate the shared component $\boldsymbol{W}_0$ and task-specific components $\{\boldsymbol{V}_t\}_{t=1}^T$ by solving:

$$\min \frac{1}{2\lambda}\|\boldsymbol{W}_0\|_F^2 + \frac{1}{2}\sum_{t=1}^{T}\frac{\|\boldsymbol{V}_t\|_F^2}{\gamma_t} + \frac{1}{2}\sum_{t=1}^{T}\left\|\boldsymbol{Y}^{(t)} - \frac{{\boldsymbol{X}^{(t)}}^\top \boldsymbol{W}_t}{\sqrt{Td}}\right\|_F^2$$

  - λ controls impact of the common part on a final prediction.
  - γ_t controls overfitting strength to the task t.
  - Closed-form solution.

- **Questions:**
  - What are generalization guarantees of the model?
  - How to balance the shared and task-specific components?

Random Matrix Analysis

Three types of asymptotic analysis:

- Traditional: d is fixed, n is large ($n \gg d, n \to \infty$).

Random Matrix Analysis

Three types of asymptotic analysis:

- Traditional: d is fixed, n is large ($n \gg d, n \to \infty$).
- Small-data: n is small, d is large, ($n \ll d$, $d \to \infty$).

Random Matrix Analysis

Three types of asymptotic analysis:

- Traditional: d is fixed, n is large ($n \gg d, n \to \infty$).
- Small-data: n is small, d is large, ($n \ll d$, $d \to \infty$).
- Large-dimensional: n, d are both large ($c_0 = \frac{d}{n} = \mathcal{O}(1)$, $(n, d) \to \infty$).

Random Matrix Analysis

Three types of asymptotic analysis:

- Traditional: d is fixed, n is large ($n \gg d, n \to \infty$).
- Small-data: n is small, d is large, ($n \ll d$, $d \to \infty$).
- Large-dimensional: n, d are both large ($c_0 = \frac{d}{n} = \mathcal{O}(1)$, $(n, d) \to \infty$).

Assumptions on data distribution:

- Noise is randomly sampled from a fixed distribution with 0-mean and covariance $\mathbf{\Sigma}_N \in \mathbb{R}^{q \times q}$.

Random Matrix Analysis

Three types of asymptotic analysis:

- Traditional: d is fixed, n is large ($n \gg d, n \to \infty$).
- Small-data: n is small, d is large, ($n \ll d$, $d \to \infty$).
- Large-dimensional: n, d are both large ($c_0 = \frac{d}{n} = \mathcal{O}(1)$, $(n, d) \to \infty$).

Assumptions on data distribution:

- Noise is randomly sampled from a fixed distribution with 0-mean and covariance $\mathbf{\Sigma}_N \in \mathbb{R}^{q \times q}$.
- Concentrated Data (Louart, C. and Couillet, R., 2018): variance of $\mathbf{x}^\top \boldsymbol{W}_t$ does not grow with dimension d,

Random Matrix Analysis

Three types of asymptotic analysis:

- Traditional: d is fixed, n is large ($n \gg d, n \to \infty$).
- Small-data: n is small, d is large, ($n \ll d$, $d \to \infty$).
- Large-dimensional: n, d are both large ($c_0 = \frac{d}{n} = \mathcal{O}(1)$, $(n, d) \to \infty$).

Assumptions on data distribution:

- Noise is randomly sampled from a fixed distribution with 0-mean and covariance $\mathbf{\Sigma}_N \in \mathbb{R}^{q \times q}$.
- Concentrated Data (Louart, C. and Couillet, R., 2018): variance of $\mathbf{x}^\top \boldsymbol{W}_t$ does not grow with dimension d,

  Particular cases:

  - Standard Gaussian distribution,
  - Lipschitz transformation of Gaussian (e.g., GAN images),
  - Open question: features learned by DNN?

Asymptotic Train and Test Risk

We want to evaluate the asymptotic train and test risk:

$$\mathcal{R}^{\infty}_{train} = \frac{1}{Tn}\sum_{t=1}^{T}\mathbb{E}\left[\|\boldsymbol{Y}^{(t)} - g(\boldsymbol{X}^{(t)})\|_2^2\right], \quad \mathcal{R}^{\infty}_{test} = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\boldsymbol{y}^{(t)} - g(\boldsymbol{x}^{(t)})\|_2^2].$$

Theorem (Asymptotic Train and Test Risk)

Under the large-dimension regime and concentration assumption, the asymptotic train and test risks are explicitly derived, with analytical curves in closed form depending on the hyperparameters, signal-generating hyperplane, and noise level.

Sketch of Proof.

- Following the notion of deterministic equivalents of a random matrix, we compute for the test risk:

$$\mathbb{E}_{\mathbf{x},\mathbf{X}^{(t)}}\left[\|\mathbf{y}^{(t)} - \boldsymbol{\omega}^{*}(\mathbf{X}^{(t)})^{\top}\mathbf{x}\|_2^2\right],$$

Particular Case: Two Tasks

For two tasks ($T = 2$) with identity covariance and $\gamma_1 = \gamma_2 = \gamma$, the asymptotic test risk simplifies to:

$$\mathcal{R}_{\text{test}}^{\infty} = \underbrace{\boldsymbol{D}_{ST}(\|\boldsymbol{W}_1\|_2^2 + \|\boldsymbol{W}_2\|_2^2)}_{\textit{Signal Term}} + \underbrace{\boldsymbol{C}_{CTT}\boldsymbol{W}_1^{\top}\boldsymbol{W}_2}_{\textit{Cross-Task Term}} + \underbrace{\boldsymbol{N}_{NT}\operatorname{tr}\boldsymbol{\Sigma}_n}_{\textit{Noise Term}},$$

- $\boldsymbol{D}_{ST}$, $\boldsymbol{C}_{CTT}$ and $\boldsymbol{N}_{NT}$ are functions of λ, γ, n and d.

Particular Case: Two Tasks

For two tasks ($T = 2$) with identity covariance and $\gamma_1 = \gamma_2 = \gamma$, the asymptotic test risk simplifies to:

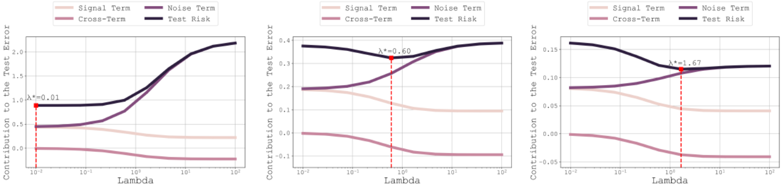
$$\mathcal{R}_{\text{test}}^{\infty} = \underbrace{\boldsymbol{D}_{ST}(\|\boldsymbol{W}_1\|_2^2 + \|\boldsymbol{W}_2\|_2^2)}_{\textit{Signal Term}} + \underbrace{\boldsymbol{C}_{CTT}\boldsymbol{W}_1^{\top}\boldsymbol{W}_2}_{\textit{Cross-Task Term}} + \underbrace{\boldsymbol{N}_{NT}\operatorname{tr}\boldsymbol{\Sigma}_n}_{\textit{Noise Term}},$$

- $\boldsymbol{D}_{ST}$, $\boldsymbol{C}_{CTT}$ and $\boldsymbol{N}_{NT}$ are functions of λ, γ, n and d.

Optimal balance between signal and noise terms:

$$\lambda^{\star} = \frac{n}{d}\left(\frac{\|\boldsymbol{W}_1\|_2^2 + \|\boldsymbol{W}_2\|_2^2}{\operatorname{tr}\boldsymbol{\Sigma}_N} + \frac{\boldsymbol{W}_1^{\top}\boldsymbol{W}_2}{\operatorname{tr}\boldsymbol{\Sigma}_N}\right) - \frac{\gamma}{2}.$$

Contribution of Each Term to Tesk Risk



(a) $\frac{n}{d} = 0.5$ (b) $\frac{n}{d} = 1.5$ (c) $\frac{n}{d} = 2.5$

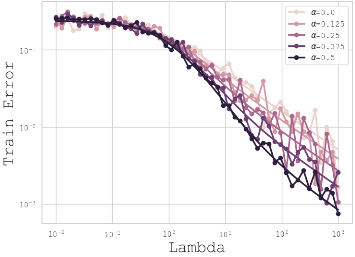
- Observations. As lambda increases, the cross-term and signal term decrease, while the noise term increases.
- Explanation. A large lambda forces tasks to interact, leveraging their relationships (decreasing cross term) but risking to increase noise and create non-existent patterns.

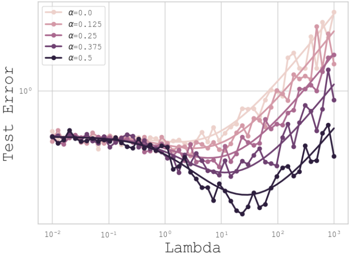
Theoretical vs Empirical Risk

- Experimental Setup:
  - Two-task setting ($T = 2$):
    $\boldsymbol{W}_1 \sim \mathcal{N}(0, I_d), \;\; \boldsymbol{W}_2 = \alpha\, \boldsymbol{W}_1 + \sqrt{1-\alpha^2}\, \boldsymbol{W}_1^{\perp}$.
  - $\alpha \in [0, 1]$ controls task similarity, $\boldsymbol{W}_1^{\perp}$ is orthogonal to $\boldsymbol{W}_1$.

Theoretical vs Empirical Risk

- Experimental Setup:
  - Two-task setting ($T = 2$):
    $\boldsymbol{W}_1 \sim \mathcal{N}(0, I_d), \;\; \boldsymbol{W}_2 = \alpha\, \boldsymbol{W}_1 + \sqrt{1-\alpha^2}\, \boldsymbol{W}_1^{\perp}$.
  - $\alpha \in [0, 1]$ controls task similarity, $\boldsymbol{W}_1^{\perp}$ is orthogonal to $\boldsymbol{W}_1$.
- Results:
  - We compare theoretical asymptotic error with empirical one by varying λ and α.
  - Theoretical curves align well the empirical ones $\Rightarrow$ potential for model selection.





Forecasting as Multi-Task Regression

- Idea: use multi-task loss to train univariate model for multivariate forecasting.
  - 3 Forecasters: `PatchTST`, `DLinearU`, `Transformer`.
  - 3 Multivariate SOTA: `SAMformer`, `DLinearM`, `iTransformer`.
  - λ and γ_t are hyperopted.
- This easy trick to learn channel interactions improves all the 3 considered models.

Dataset	H	with MTL regularization			without MTL regularization					
		PatchTST	DLinearU	Transformer	PatchTST	DLinearU	DLinearM	Transformer	SAMformer†	iTransformer†
ETTh1	96	0.385	**0.367***	0.368	0.387	0.397	0.386	0.370	0.381	0.386
	192	0.422	**0.405***	0.407*	0.424	0.422	0.437	0.411	0.409	0.441
	336	0.433*	0.431	0.433	0.442	0.431	0.481	0.437	**0.423**	0.487
	720	0.430*	0.454	0.455*	0.451	0.428	0.519	0.470	**0.427**	0.503
ETTh2	96	0.291	**0.267***	0.270	0.295	0.294	0.333	0.273	0.295	0.297
	192	0.346*	**0.331***	0.337	0.351	0.361	0.477	0.339	0.340	0.380
	336	**0.332***	0.367	0.366*	0.342	0.361	0.594	0.369	0.350	0.428
	720	**0.384***	0.412	0.405*	0.393	0.395	0.831	0.428	0.391	0.427
Weather	96	**0.148**	0.149*	0.154*	0.149	0.196	0.196	0.170	0.197	0.174
	192	**0.190**	0.206*	0.198*	0.193	0.243	0.237	0.214	0.235	0.221
	336	**0.242***	0.249*	0.258	0.246	0.283	0.283	0.260	0.276	0.278
	720	**0.316***	0.326*	0.331	0.322	0.339	0.345	0.326	0.334	0.358

Thank you for your attention!

Paper

Paris Noah's Ark Lab