Data-secure transfer learning in heterogeneous low-rank and sparse panel VAR models

George Michailidis

gmichail@ucla.edu

Joint work with Yuchen Xu

CRETE Conference - Tinos 2025

Setting under Consideration- I

Suppose

1. one has collected time series data on p variables for T_0 time points and plans to fit a Vector Autoregressive (VAR) model; i.e.,

$$X_t^0 = A_0 X_{t-1}^0 + \epsilon_t^0, \quad \mathbb{E}(\epsilon_t^0) = 0, \ \text{Cov}(\epsilon_t^0) = \Sigma_\epsilon \tag{1}$$

2. one has access to the estimated parameters $\hat{A}^m$ from a collection (panel) of M VAR models on the same p variables The M entities can correspond to countries, states, patients, etc.

Setting under Consideration - II

A few questions of interest:

- Is it possible to "borrow strength" from selected $\hat{A}^m$ to improve the estimation of $\hat{A}^0$?
- If so, which $\hat{A}^m$'s are "most relevant" and how can the process of borrowing strength be effectively operationalized?
- Is "borrowing strength" always beneficial, or can it be detrimental in certain scenarios?

Remarks

- The setting corresponds to parameter-based transfer learning, where a model is first trained on source tasks (typically with large data sets), and its learned parameters are then used to aid training on a target task (often with limited data)
- Our focus is on high-dimensional settings – $p \sim T_0$
- In the existing literature (e.g., transfer learning for sparse regression via Lasso –Li, Cai & Li, 2022), it is commonly assumed that the auxiliary data sets used to train the source tasks are fully accessible

  In contrast, we do not make this assumption

How are the $\hat{A}^m$'s obtained?

Digression - A flexible and effective panel VAR model

Consider a panel linear VAR model

$$X^{t,m} = A^m X^{t-1,m} + \varepsilon^{t,m}, \qquad m = 1, \cdots, M, \ t = 1, \cdots, T, \ X^{t,m} \in \mathbb{R}^p \tag{2}$$

The model in (2) has many parameters: $(M \times p^2)$

To reduce the number of parameters in model (2) while preserving interpretability we consider the following Heterogeneous low-rank + sparse panel VAR (H(L+S)PVAR) model

A Heterogeneous Low Rank plus Sparse Panel VAR Model

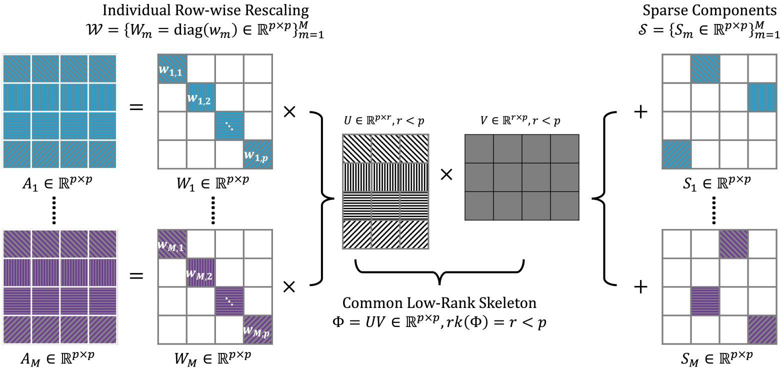


Figure: Depiction of H(L+S)PVAR model

H(L+S) PVAR Formulation

The Panel Vector Auto-Regressive (H(L+S) PVAR) model takes the form

$$X_t^m = A_m X_{t-1}^m + \epsilon_t^m, \quad \mathbb{E}(\epsilon_t^m) = 0,\ \mathrm{Cov}(\epsilon_t^m) = \Sigma_m, \quad m = 1, \ldots, M, \tag{3}$$

where $\{A_m \in \mathbb{R}^{p \times p}\}_{m=1}^M$ are modeled as

$$A_m = W_m \Phi + S_m, \tag{4}$$

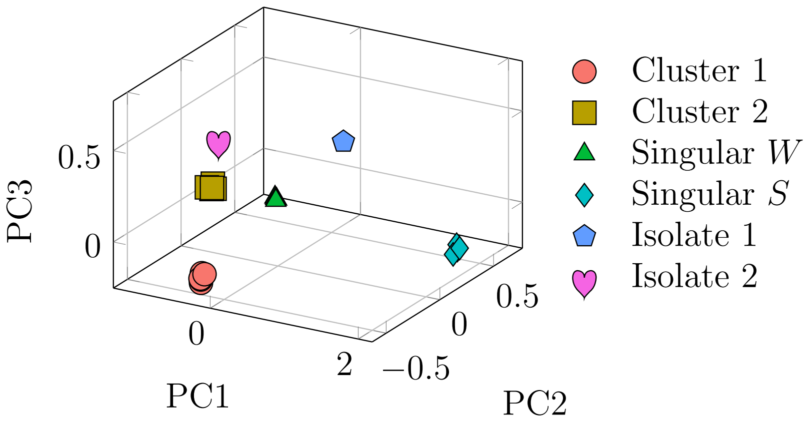
- $\Phi \in \mathbb{R}^{p \times p}$ with $\mathrm{rk}(\Phi) = r < p$ is the shared low-rank basis;
- $W_m = \mathrm{diag}(w_m)$ are rescaling factors;
- $S_m \in \mathbb{R}^{p \times p}$ with $\|S_m\|_0 \ll p^2$ are idiosyncratic sparse components

Illustrating H(L+S)PVAR Generality and Versatility

Recall $A_m = W_m \Phi + S_m$.

Consider a collection of entities $m = 1, \cdots, M$ for which:

- Subsets of A^m form clusters (similar W^m, S^m)
- Some A^m are totally different than all other entities – Isolates
- There exist some entities that have a purely sparse model: $W_m = 0$
- There exist some entities that have a purely low rank model: $S_m = 0$


PC3
0.5
0
PC1
0
2
PC2
−0.5
0
0.5
Cluster 1
Cluster 2
Singular W
Singular S
Isolate 1
Isolate 2

Properties of the H(L+S)PVAR Model

The parameters of the H(L+S)PVAR model can be estimated by an ADMM algorithm

Under certain regularity conditions—such as model stability, parameter identifiability, and restricted strong convexity—the parameters can be estimated consistently under high-dimensional scaling

Key working assumption for transfer learning

The transition matrix of the target task VAR model can be expressed as

$$A^0 = W_0 \Phi + S_0 \tag{5}$$

Key working assumption for transfer learning

The transition matrix of the target task VAR model can be expressed as

$$A^0 = W_0 \Phi + S_0 \tag{5}$$

Objective:
To estimate $\widehat{A}^0$, we leverage $\hat{\Phi}$ from the pretrained H(L+S)PVAR model

However, a transfer learning procedure must be developed for estimating the W_0 and S_0 components

Estimation with Known Transferrable Source Tasks/Models

Let

$$W_m = W_0 + \delta_m^W, \qquad S_m = S_0 + \delta_m^S \tag{6}$$

i.e., the parameters of the target task model are sparse "perturbations" of selected ones in the source tasks models

Then, for thresholds $h_W > 0, h_S > 0$, one can identify the indices of informative W_m's and S_m's from the source panel defined as

$$\mathcal{A}_W := \{m : \|\delta_m^W\|_2 \leq h_W\} \subset \{1, \cdots, M\}, \tag{7}$$
$$\mathcal{A}_S := \{m : \|\delta_m^S\|_2 \leq h_S\} \subset \{1, \cdots, M\}$$

VAR-TL - I

Consider a H(L+S)PVAR model with

1. pretrained parameters $\widehat{\Phi}, \{\widehat{W}_m, \widehat{S}_m\}_{m=1}^{M}$,
2. informative models determined by index sets $\mathcal{A}_W$ and $\mathcal{A}_S$

The TL procedure comprises the following two steps

VAR-TL - II

1. Calculate Pooled Averages:

$$\begin{aligned} \bar{W}_0 &= \text{mean}\{\widehat{W}_m : m \in \mathcal{A}_W\}, \\ \bar{S}_0 &= \mathit{\Pi}_{\mathcal{P}_S}(\text{mean}\{\widehat{S}_m : m \in \mathcal{A}_S\}), \\ \mathcal{P}_S &= \text{supp}(\text{median}\{|\widehat{S}_m| : m \in \mathcal{A}_S\}) \end{aligned} \tag{8}$$

where the operator $\mathit{\Pi}_{\mathcal{P}}(C)$ projects the matrix C onto the support $\mathcal{P}$

Set $\bar{W}_0$ ($\bar{S}_0$) to be zero, if $\mathcal{A}_W$ ($\mathcal{A}_S$) is an empty set

VAR-TL - III

2. Bias correction fit:

$$(\widehat{\delta}^W, \widehat{\delta}^S) = \underset{\delta^W, \delta^S}{\arg\min} \frac{1}{2T_0} \| \left((\bar{W}_0 + \delta^W)\widehat{\Phi} + (\bar{S}_0 + \delta^S) \right) X_0 - Y_0 \|_2^2 \tag{9}$$
$$+ p_\lambda(\delta^W, \delta^S),$$

where p_λ is a regularizer with tuning parameters $\lambda = (\lambda^W, \lambda^S)$

Given the diagonal nature of W_0 and the sparse nature of S_0, an ℓ_1 regularizer is used in (9)

Theoretical guarantees for VAR-TL

1. Under certain regularity conditions on the pretrained H(L+S)PVAR model,
2. Assuming the sets $(\mathcal{A}_W, \mathcal{A}_S)$ are known,
3. An appropriate choice of the tuning parameters λ,
4. $\hat{\Phi}$ is estimated accurately (rate $\frac{pr}{MT}$),
5. $h_W \leq \sqrt{\frac{p}{T_0}}$, $h_S \leq \sqrt{\frac{s \log p}{T_0}}$,

then with high probability

$$\|\widehat{W_0} - W_0\|_F^2 + \|\widehat{S}_0 - S_0\|_F^2 \leq c\left[\frac{p}{|\mathcal{A}_W|T} + \frac{s \log p}{|\mathcal{A}_S|T}\right] \quad (10)$$

Comparison with a L+S target task model

Suppose that one uses only the T_0 samples and estimates a low-rank+sparse VAR model with $A^0 = L_0 + S_0$

Then, with high probability we get (Basu, Li and M, 2019)

$$\|\widehat{L}_0 - L_0\|_F^2 + \|\widehat{S}_0 - S_0\|_F^2 \leq c\left(\frac{pr}{T_0} + \frac{s \log p}{T_0}\right) \quad (11)$$

When is TL beneficial? - I

The previous results show that TL is beneficial when

1. $|\mathcal{A}_W|T \geq T_0$ $(|\mathcal{A}_S|T \geq T_0)$
2. M is reasonably large (for accurate estimation of Φ)
3. The thresholds h_W, h_S are tight enough;

$$h_W \leq \sqrt{\frac{p}{T_0}}, \quad h_S \leq \sqrt{\frac{s \log p}{T_0}}$$

i.e., the parameters of some source task models are within the "confidence neighborhood" of the target task parameters

When is TL beneficial? - II

- Condition 1 is relatively mild; typically, the pretrained model is assumed to be estimated based on a large data set
  Moreover, transfer learning is particularly beneficial when the target task suffers from limited data, making it difficult to train an accurate model from scratch
- Condition 3 requires that at least one member of the pretrained panel model exhibits dynamics similar to those of the target task model

The VAR-TL Mechanism in Practice - I

Key idea:

- Obtain a rough estimate of $\widehat{A}_0$ – i.e., $\widehat{W}_0, \widehat{S}_0$ using the available $\hat{\Phi}, \widehat{W}_m, \widehat{S}_m$
- Statistically determine whether a pretrained $\widehat{W}_m$ (or $\widehat{S}_m$) significantly contributes to it, accounting for the noises and the source's informative level h_W (or h_S)

The VAR-TL Mechanism in Practice - II

1. Compute

$$\mu_m := \| \operatorname{diag}\left(\frac{Q_\gamma^\gamma(Y_0 X_0^+ - \widehat{W}_m \widehat{\Phi}) X_0 X_0' \widehat{\Phi}'}{T_0}\right)\|_2, \qquad (12)$$

$$\kappa_m := \frac{\langle |\widehat{S}_m|, |\frac{(Y_0 X_0^+ - \widehat{S}_m - \tilde{W}_m \widehat{\Phi}) X_0 X_0'}{T_0}| \rangle}{\|\widehat{S}_m\|_F},$$

where $X_0^+ = X_0'(X_0 X_0')^{-1}$ is a pseudo-inverse of the design matrix X_0, Q_γ^γ truncates at the lower and upper $\gamma = \frac{\log(p^2)}{p^2}$-quantile, and $\widetilde{W}_m = \arg\min_{W=\operatorname{diag}(w)} \| Y_0 X_0^+ - \widehat{S}_m - W\widehat{\Phi}\|_1$

Remarks

Some involved algebra shows that the tails of the quantities $Y_0 X_0^+ - \widehat{W}_m \widehat{\Phi}$ are primarily contributed by the sparse S_0; hence, the trimming operator $^{\gamma}_{\gamma}$

Then, regressing them on $\hat{\phi} X_0$ gives a surrogate for $\widehat{W_0}$

Analogous type calculations lead to a surrogate for $\widehat{S}_0$

The VAR-TL Mechanism in Practice - III

2. Setting $\hat{s}, \hat{r}$ the sparsity level and the rank detected in the source data, determine the estimated informative sets as:

$$\widehat{\mathcal{A}}_W = \left\{ m : \mu_m \precsim \sqrt{\frac{\hat{r}p}{MT}} + \sqrt{\frac{p}{T}} + \sqrt{\frac{p\log(p)}{T_0}} \right\},$$
$$\widehat{\mathcal{A}}_S = \left\{ m : \kappa_m \precsim \sqrt{\frac{\hat{r}\hat{s}}{MT}} + \sqrt{\frac{\hat{s}\log(p)}{T}} + \sqrt{\frac{\hat{s}\log(p)}{T_0}} \right\}. \tag{13}$$

Performance Evaluation on Synthetic Data - I

Setting: correctly specified H(L+S)PVAR model and target task model

- pretrained H(L+S)PVAR model with $(p, M, T) = (30, 10, 800)$.
- Target task sample size: $T_0 \in \{100, 200, 400, 800, 1600, 6400\}$.
- Transfer structure: $(\omega_W, \omega_S) = (0.4, 0.3)$, where $\omega_W = \frac{|\mathcal{A}_W|}{M}$ and $\omega_S = \frac{|\mathcal{A}_S|}{M}$.

Performance Evaluation on Synthetic Data - II

Comparators:

- VAR-TL (full TL)
- Nominal TL (uses only pretrained $\widehat{\Phi}$), estimates W_0, S_0 from T_0 data
- H(L+S)PVAR Refit (joint model with full access to all data, including the T_0 samples)
- Oracle: based on true values of Φ, W_0, S_0
- Local Low-rank and Sparse VAR (Basu, Li & M, 2019) (target task data only), baseline

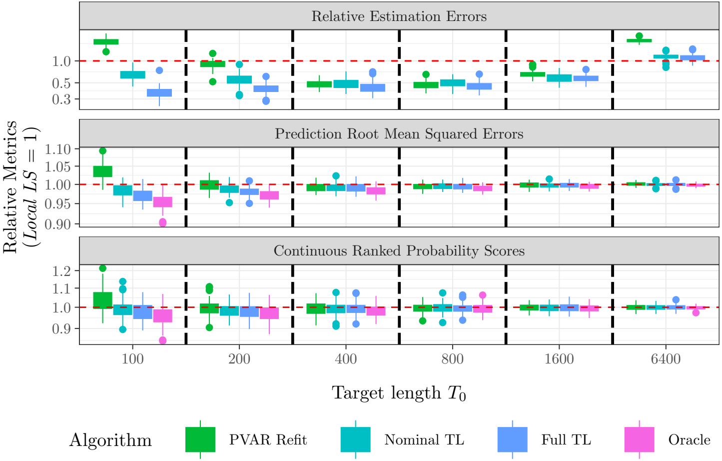


Figure: Comparison of methods across different metrics based on 50 data replicates; $p = 30$

Performance Evaluation on Synthetic Data - III

Setting: mis-specified H(L+S)PVAR model and target task model

Mis-specification Mechanisms

1. Dense VAR model

$$X_t^m = A_m X_{t-1}^m + \epsilon_t^m,\ \epsilon_t^m \sim N(0, \Sigma_m);\quad A_m = B_m + S_m, \tag{14}$$

   where $B_m \in \mathbb{R}^{p \times p}$ is dense and $S_m \in \mathbb{R}^{p \times p}$ is sparse

2. Dynamic factor model

$$\begin{aligned} F_t &= A_f F_{t-1} + (I + S_f) f_t, & f_t &\sim N(0, I), \\ X_t^m &= B_m F_t + \epsilon_t^m, & \epsilon_t^m &\sim N(0, \Sigma_m), \end{aligned} \tag{15}$$

   where $A_f \in \mathbb{R}^{r \times r}$ is the coefficient matrix of the factor VAR model, $S_f \in \mathbb{R}^{r \times r}$ introduces some sparse contemporaneous dependence among factor innovations, and $B_m \in \mathbb{R}^{p \times r}$ is the loading matrix for the m-th entity

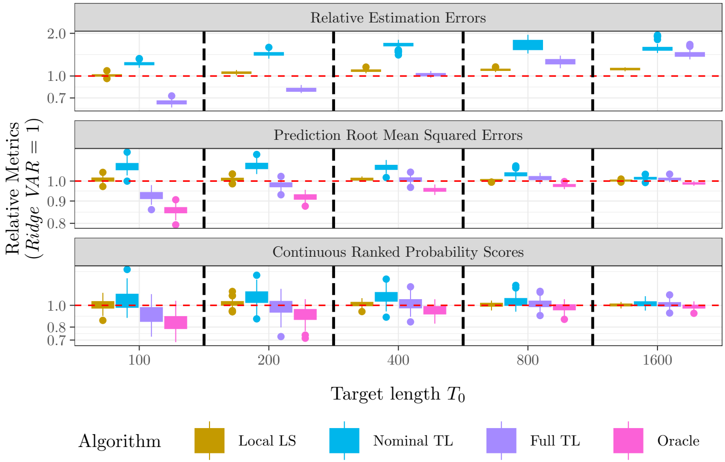


Figure: Comparison of methods across different metrics based on 50 data replicates with $\omega_W = \omega_S = 0.2$, $p = 40$, $M = 20$ for Mechanism 1

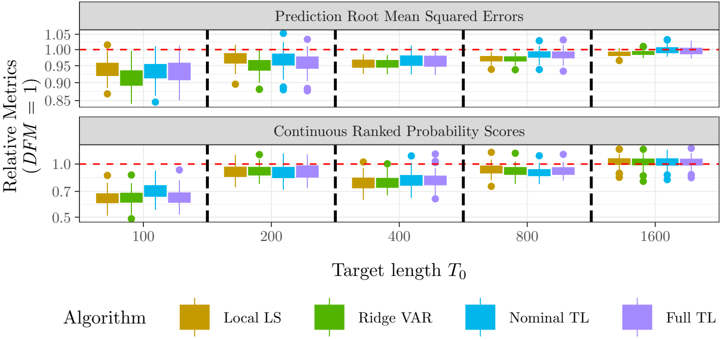


Figure: Comparison of methods across different metrics based on 50 data replicates with $\omega_W = \omega_S = 0.2$, $p = 40$, $M = 20$ for Mechanism 2

Applications to Economic Data

1. Focuses on $p = 35$ key macroeconomic indicators variables from countries within the Eurozone
   H(L+S)PVAR model built on data for $M = 8$ countries –Belgium, Finland, France, Ger- many, Netherlands, Norway, Spain, and Sweden– and assess the performance of VAR-TL on Austria, Denmark, and Italy
2. Analyzes $p = 18$ labor-market variables (unemployment rates, hourly earnings, weekly working hours, etc.)
   H(L+S)PVAR model built on data for $M = 35$ US states, and assess the performance of VAR-TL on Arkansas (AR), Colorado (CO), Connecticut (CT), Idaho (ID), and West Virginia (WV).

	Algorithm \ Country	Austria	Finland	Italy
Relative Prediction RMSE (Ridge VAR = 1)	Local LS	1.004	0.980	1.003
	Nominal TL	0.969	0.985	0.947
	Full TL	0.964	0.979	0.981
Relative Prediction CRPS (Ridge VAR = 1)	Local LS	0.995	0.994	1.006
	Nominal TL	0.890	0.942	0.838
	Full TL	0.888	0.881	0.943

	Algorithm \ State	AR	CO	CT	ID	WV
Relative Prediction RMSE (Ridge VAR = 1)	Local LS	1.002	0.989	0.994	0.984	0.997
	Nominal TL	0.992	0.981	0.967	0.956	0.985
	Full TL	0.992	0.975	0.957	0.948	0.986
Relative Prediction CRPS (Ridge VAR = 1)	Local LS	0.978	1.016	1.079	1.079	1.064
	Nominal TL	1.189	1.034	1.082	1.072	1.073
	Full TL	1.189	1.026	1.075	1.017	1.063

Concluding Remarks

- Proposed a data-secure transfer learning framework for panel VARs that selectively leverages low-rank and sparse structures
- TL mechanism supported by theoretical guarantees
- The methodology demonstrates strong empirical performance—especially with short time series—and robustness to model mis-specification
- Application to real-world economic data also demonstrates the effectiveness of the proposed TL methodology