High-Dimensional Matrix Autoregressions (MAR): A Unified Framework for Structured Regularization

George Michailidis

gmichail@ucla.edu

Joint work with Yuchen Xu, Partha Sarkar and Kshitij Khare

CRETE Conference
Ios, July 2026

Structured High-Dimensional MAR: Setting and Goal

We observe multivariate time series for a collection of related entities:

$$X_t \in \mathbb{R}^{p \times M}, \qquad t = 1, \ldots, T,$$

where rows are variables and columns are entities.

Examples

- macroeconomic indicators across countries;
- brain signals across subjects or regions;
- environmental variables across locations.

Scientific goals

- learn lead-lag dependence between variables;
- borrow information across entities;
- preserve interpretability through Granger causal structure.

Panel VAR as a Natural Competitor

A standard VAR (lag-1) for a single entity is given by

$$X_t = AX_{t-1} + \varepsilon_t, \qquad X_t \in \mathbb{R}^p.$$

The entry A_{ij} measures the lagged effect of variable j on variable i.

$$A_{ij} \neq 0 \quad \Longleftrightarrow \quad j \text{ Granger-causes } i.$$

With M entities, a panel VAR model becomes

$$X_t^{(m)} = A_m X_{t-1}^{(m)} + \varepsilon_t^{(m)}, \qquad m = 1, \ldots, M.$$

Core challenge:

How should we link $\{A_m\}_{m=1}^M$ to exploit commonalities while preserving entity-level heterogeneity?

Panel VAR Strategies: Interpretability vs. Parsimony

Constrained aggregation

- equality constraints across entities;
- random effects: $A_m(i,j) = \theta_{ij} + u_m(ij)$;
- fused/group penalties or shrinkage priors.

Why interpretable?

The model stays in the original VAR coordinate system: each $A_m(i,j)$ remains an observed-variable Granger effect.

Latent representations

$A_m = W_m L + S_m$, L low-rank, S_m sparse, W_m diag

or $A_m = U G_m V^\top$.

- shared low-dimensional dynamic structure;
- entity-specific scaling or latent transmission;
- Granger effects are recovered from the reconstructed A_m.

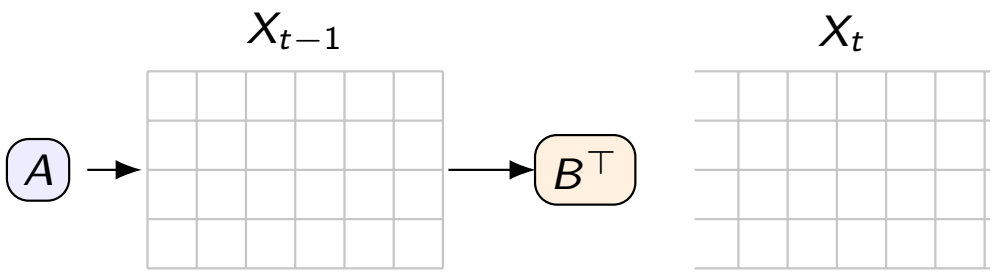
Matrix-Variate Dynamics and the MAR Model

Instead of estimating one VAR matrix per entity, arrange the panel as a matrix:

$$X_t \in \mathbb{R}^{p \times M}.$$

Use a separable bilinear transition:

$$X_t = A X_{t-1} B^\top + E_t, \qquad A \in \mathbb{R}^{p \times p}, \quad B \in \mathbb{R}^{M \times M}.$$



Interpretation

A captures variable-to-variable dynamics; B captures entity-to-entity transmission.

MAR Model: Origins and Recent Extensions

Original MAR framework

Chen, Xiao and Yang (2021) introduced the bilinear matrix autoregressive model

$$X_t = AX_{t-1}B^\top + E_t.$$

They showed that the model preserves the row/column structure of matrix-valued time series, reduces dimensionality, and can be estimated by alternating least squares methods in the fixed-dimensional regime, with p, M fixed and $T \to \infty$.

Recent high-dimensional extensions

- **Jiang et al. (2024):** banded and entrywise sparse MAR estimation under diverging dimensions.
- **Wu et al. (2025):** Global-VAR-inspired structural MAR extensions.
- **Ghosh, Roy and Chakraborty (2026):** additive MAR-type model

  $$X_t = AX_{t-1} + X_{t-1}B^\top + E_t,$$

  with low-rank plus sparse coefficient matrices.

This work: develops a unified two-stage framework for sparse, group-sparse, low-rank, and mixed-structure regularized MAR models, with both algorithmic and statistical guarantees.

MAR as a Kronecker-Structured VAR

Vectorizing the MAR model gives

$$\mathrm{vec}(X_t) = (B \otimes A)\,\mathrm{vec}(X_{t-1}) + \mathrm{vec}(E_t).$$

Thus, MAR is a VAR with transition matrix

$$C = B \otimes A.$$

Dimensional reduction

Unrestricted VAR:

$$p^2 M^2 \text{ parameters.}$$

MAR:

$$p^2 + M^2 \text{ parameters.}$$

Stability

$$\rho(C) = \rho(B \otimes A) = \rho(A)\rho(B),$$
$$\rho(\cdot) \equiv \text{spectral radius.}$$

The MAR(1) process is stable when

$$\rho(A)\rho(B) < 1.$$

High-dimensional gap

When the time series length T is limited, $p^2 + M^2$ can be large; structured regularization is needed.

Granger Effects in the MAR Model

For target entity m, the m-th column satisfies

$$X_{\cdot,m,t} = \sum_{\ell=1}^{M} B_{m\ell} A X_{\cdot,\ell,t-1} + E_{\cdot,m,t}.$$

Entry-wise,

$$X_{i,m,t} = \sum_{\ell=1}^{M} \sum_{j=1}^{p} \underbrace{B_{m\ell} A_{ij}}_{\text{Granger effect from } (j,\ell) \text{ to } (i,m)} X_{j,\ell,t-1} + E_{i,m,t}.$$

Granger causal effect

$$(j,\ell) \to (i,m) \quad \Longleftrightarrow \quad B_{m\ell} A_{ij} \neq 0.$$

The effect separates into a variable-side mechanism A_{ij} and an entity-side transmission $B_{m\ell}$.

Why Structured Regularization?

We estimate

$$\min_{A,B} \frac{1}{2T} \sum_{t=1}^{T} \left\| X_t - A X_{t-1} B^\top \right\|_F^2 + \lambda_A P_A(A) + \lambda_B P_B(B),$$

where $P_A(\cdot)$ and $P_B(\cdot)$ are regularizers that encode the assumed structure of A and B.

Key modeling point.
Because every Granger effect is a product

$$B_{m\ell} A_{ij},$$

regularizing A or B changes entire families of Granger causal effects.

Entrywise Sparsity: Useful but Rigid

If both A and B are entrywise sparse, then

$$(j, \ell) \to (i, m) \quad \Longleftrightarrow \quad A_{ij} \neq 0 \quad \text{and} \quad B_{m\ell} \neq 0.$$

Equivalently,

$$\text{supp}(B \otimes A) = \text{supp}(B) \times \text{supp}(A).$$

- If $A_{ij} = 0$, the variable effect $j \to i$ disappears for all entity pairs (m, ℓ).
- If $B_{m\ell} = 0$, no variable effect travels from entity ℓ to entity m.
- If both survive, the selected variable link is transmitted through the selected entity channel.

Takeaway

Pure sparsity is interpretable, but rigid: it selects isolated entries in A and B, while the actual Granger causal graph is their cross-product. With limited data, isolated selected entries may be hard to distinguish from noise.

Beyond Sparsity: Structured MAR Regularization

Group sparsity

- A zero group in A removes a block of variable-to-variable effects across all entities.
- A zero group in B removes a block of entity-to-entity transmissions across all variables.

Low rank

If $A = U_A V_A^\top$, then

$$B_{m\ell} A_{ij} = B_{m\ell}(u_{A,i}^\top v_{A,j}).$$

Granger causal effects are driven by a few shared low-dimensional latent dynamic patterns.

Mixed structure

Example: sparse A, low-rank B:

$$B_{m\ell} A_{ij} = (u_{B,m}^\top v_{B,\ell}) A_{ij}.$$

One layer is selective; the other captures shared latent transmission.

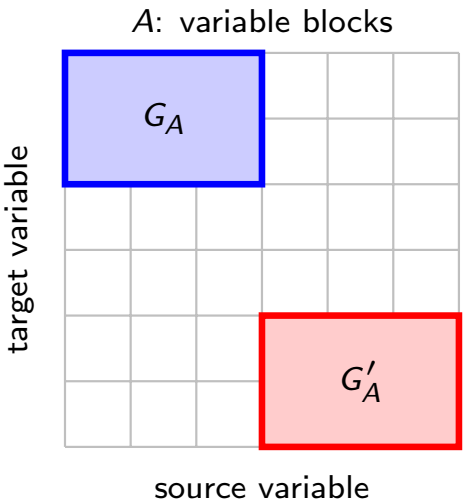
Takeaway

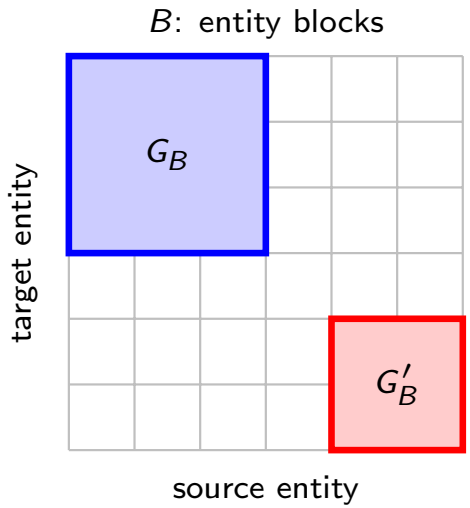
MAR lets the variable layer and entity layer play different scientific roles.

Illustration of Group Sparsity: Block-Level Granger Effects

Instead of selecting entries one at a time, use group penalties:

$$P_A(A) = \sum_{g \in \mathcal{G}_A} \sqrt{|g|}\, \|A_g\|_F, \qquad P_B(B) = \sum_{h \in \mathcal{G}_B} \sqrt{|h|}\, \|B_h\|_F.$$





Interpretation

A zero group in A removes a block of variable-to-variable effects across all entities. A zero group in B removes a block of entity-to-entity transmissions across all variables.

Regularized MAR Objective and Optimization Challenge

$$g(A,B) = f(A,B) + \lambda_A P_A(A) + \lambda_B P_B(B), \qquad f(A,B) = \frac{1}{2T}\sum_{t=1}^{T} \left\| X_t - AX_{t-1}B^\top \right\|_F^2 .$$

Supported structures

- sparse: $P(A) = \|A\|_1$;
- group-sparse: $P(A) = \sum_g \sqrt{|g|}\,\|A_g\|_F$;
- low-rank: $P(A) = \|A\|_*$;
- mixed: different structures for A and B.

Key technical challenge:

The objective is biconvex:
convex in A for fixed B, and convex in B for fixed A, but not jointly convex.

Two-Stage Estimation Strategy

A natural algorithm is alternating regularized least squares. However, because the objective is not jointly convex, initialization matters.

Alternating minimization may converge to a poor stationary point if started far from the true parameters.

Two-Stage Estimation Strategy

A natural algorithm is alternating regularized least squares. However, because the objective is not jointly convex, initialization matters.

Alternating minimization may converge to a poor stationary point if started far from the true parameters.

Stage 1: vectorized warm start

Estimate the Kronecker transition matrix $C = B \otimes A$, then extract $(\widehat{A}_0, \widehat{B}_0)$ by nearest Kronecker product approximation.

Stage 2: local refinement

Starting from $(\widehat{A}_0, \widehat{B}_0)$, alternate regularized least-squares updates for A and B.

Warm start gets close to a good stationary solution; alternating updates refine locally.

Stage 1: Estimate the Kronecker Transition

Step 1: structured estimate of C

$$\widehat{C} = \arg\min_{C} \left\{ \frac{1}{T} \sum_{t=1}^{T} \|\mathrm{vec}(X_t) - C\mathrm{vec}(X_{t-1})\|_2^2 + \lambda_C P_C(C) \right\}.$$

Step 2: Nearest Kronecker Product (NKP) approximation

$$(\widehat{A}_0, \widehat{B}_0) = \arg\min_{A,B} \left\| \widehat{C} - B \otimes A \right\|_F^2.$$

This is computed by rearranging $\widehat{C}$ and taking the leading singular vectors of the rearranged matrix (Van Loan – Pitsianis (1993) algorithm).

Scale ambiguity

Since $B \otimes A = (cB) \otimes (A/c)$, we impose a normalization so that A and B are numerically identifiable.

Stage 2: Alternating Regularized Least Squares

For $k = 0, 1, 2, \ldots$, update

$$\widehat{A}_{k+1} = \arg\min_{A} \{ f(A, \widehat{B}_k) + \lambda_A P_A(A) \},$$

$$\widehat{B}_{k+1} = \arg\min_{B} \{ f(\widehat{A}_{k+1}, B) + \lambda_B P_B(B) \}.$$

Computational benefit

- each subproblem is convex;
- proximal gradient methods apply;
- proximal maps are standard.

Proximal maps

Regularizer	Proximal step
ℓ_1	soft thresholding
Group lasso	group shrinkage
Nuclear norm	singular value thresholding

Algorithmic Convergence: What Needs to be Shown?

Because of biconvexity, the goal is not to prove convergence to the global minimizer. Instead, we prove convergence to a joint stationary point:

$$0 \in \partial_A g(\widehat{A}, \widehat{B}), \qquad 0 \in \partial_B g(\widehat{A}, \widehat{B}).$$

Key ingredients

1. **Sufficient descent:** every update decreases the objective by an amount proportional to the squared step size.
2. **Subgradient control:** once the successive iterate differences become small, the first-order optimality violation also becomes small.
3. **Kurdyka–Łojasiewicz (KL) property of the objective function:** prevents infinitely many tiny oscillations and gives full-sequence convergence.

Takeaway

The algorithm settles at a stationary point of the nonsmooth regularized objective.

Algorithmic Convergence Guarantee

Sufficient descent

Under stability, bounded structured spaces, bounded covariance, and restricted strong convexity,

$$g(\widehat{A}_k, \widehat{B}_k) - g(\widehat{A}_{k+1}, \widehat{B}_{k+1}) \geq c \left(\left\| \widehat{A}_{k+1} - \widehat{A}_k \right\|_F^2 + \left\| \widehat{B}_{k+1} - \widehat{B}_k \right\|_F^2 \right).$$

Relative-error control

$$\operatorname{dist}\{0, \partial g(\widehat{A}_{k+1}, \widehat{B}_{k+1})\} \leq L \left\| (\widehat{A}_{k+1}, \widehat{B}_{k+1}) - (\widehat{A}_k, \widehat{B}_k) \right\|_F.$$

Role of the KL property

The smooth loss is polynomial in A and B, and the regularizers used here are semi-algebraic, so g has the KL property. Descent + relative error + KL imply convergence of the full sequence to a stationary point.

Statistical Consistency: Roadmap and Additional Assumptions

The proof controls the final estimator relative to the true parameters

$$(\bar{A}, \bar{B}), \qquad \bar{C} = \bar{B} \otimes \bar{A}.$$

Proof roadmap

1. warm-start consistency of $(\widehat{A}_0, \widehat{B}_0)$;
2. local contraction of alternating refinement;
3. final error decomposition.

New ingredients for statistical rates

- deviation bounds: empirical noise terms are controlled in the dual norms associated with the regularizers;
- local Schur complement condition: the A- and B-blocks of the Hessian are not too strongly coupled.

Main message

$$\text{final error} = \text{oracle statistical error} + \text{optimization / nonconvexity error}.$$

Stage 1: Warm Start from the Kronecker Transition

Start from the vectorized MAR representation:

$$\mathrm{vec}(X_t) = C\,\mathrm{vec}(X_{t-1}) + \mathrm{vec}(E_t), \qquad C = B \otimes A.$$

First estimate C, then project it onto the Kronecker form:

$$(\widehat{A}_0, \widehat{B}_0) = \arg\min_{A,B} \|\widehat{C} - B \otimes A\|_F^2.$$

Warm-start guarantee

If λ_C is large enough to dominate the stochastic error, then

$$\|\widehat{C} - \bar{C}\|_F \lesssim \frac{\lambda_C \Psi_A \Psi_B}{\alpha\gamma}.$$

After the nearest Kronecker product step and normalization,

$$M\kappa_B^2 \|\widehat{A}_0 - \bar{A}\|_F^2 + p\kappa_A^2 \|\widehat{B}_0 - \bar{B}\|_F^2 \lesssim \frac{\lambda_C^2 \Psi_A^2 \Psi_B^2}{\alpha^2\gamma^2}, \text{ where:}$$

- Ψ_A, Ψ_B: effective complexity of A and B;
  e.g., $\Psi^2 = s$ for sparse, $\Psi^2 = mg$ for group sparse, $\Psi^2 = r$ for low rank.
- $\alpha\gamma$: effective curvature of the loss function.
- κ_A, κ_B: scale normalizations for A and B, needed because $B \otimes A = (cB) \otimes (A/c)$.
- ϕ: noise and temporal-dependence scale; it enters the high-probability choices of $\lambda_A, \lambda_B, \lambda_C$; larger ϕ for more persistent time series.

Interpreting the Warm-Start Bound

The warm start is accurate when noise is small relative to curvature and model complexity:

$$\text{estimation error} \approx \frac{\text{noise level} \times \text{model complexity}}{\text{curvature of the objective function}}.$$

The weights $M\kappa_B^2$ and $p\kappa_A^2$ appear because A is learned through M entity channels, while B is learned through p variable channels.

Final Consistency Bound: Oracle Error plus Nonconvexity Price

After the warm start, alternating regularized least squares stays in a local neighborhood of $(\bar{A}, \bar{B})$. The final error is measured in the curvature-weighted scale

$$\mathcal{E}(\widehat{A}, \widehat{B}) = \kappa_B^2 M \|\widehat{A} - \bar{A}\|_F^2 + \kappa_A^2 p \|\widehat{B} - \bar{B}\|_F^2.$$

Unified bound

For suitable regularization parameters (λ_A, λ_B),

$$\mathcal{E}(\widehat{A}, \widehat{B}) \lesssim \frac{\lambda_A^2 \Psi_A^2}{\alpha^2 \gamma^2 \kappa_B^2 M} + \frac{\lambda_B^2 \Psi_B^2}{\alpha^2 \gamma^2 \kappa_A^2 p}.$$

After plugging in high-probability choices of λ_A, λ_B,

$$\mathcal{E}(\widehat{A}, \widehat{B}) \lesssim \underbrace{\text{oracle statistical error}}_{\text{intrinsic estimation cost}} + \underbrace{\text{optimization error}}_{\text{price of nonconvexity}}.$$

Structure	**Oracle statistical component**	**Complexity**
Sparse / group sparse	$\dfrac{\phi^2(\Psi_A^2 + \Psi_B^2)\log(pM)}{\alpha^2\gamma^2 T}$	$\Psi^2 = s$ or mg
Low-rank	$\dfrac{\phi^2(r_A p + r_B M)}{\alpha^2\gamma^2 T}$	$\Psi_A^2 = r_A,\ \Psi_B^2 = r_B$
Mixed	$\dfrac{\phi^2\{\Psi_A^2 \log(pM) + r_B M\}}{\alpha^2\gamma^2 T}$	sparse/group A, low-rank B

Takeaway Message

The statistical component matches the oracle rate of an ideal procedure that operates in the correct local neighborhood of the true parameters.

The extra term is the price of using a practical algorithm for a nonconvex bilinear estimation problem.

Takeaway Message

The statistical component matches the oracle rate of an ideal procedure that operates in the correct local neighborhood of the true parameters.

The extra term is the price of using a practical algorithm for a nonconvex bilinear estimation problem.

Does the Optimization Error Vanish?

Not automatically. IT vanishes only when the structural complexity of A and B is small relative to the growing matrix dimensions p and M, and when T grows fast enough relative to the structural complexity.

Synthetic Experiments: Design

Representative settings

1. Sparse/group-sparse: A is group-sparse and B is sparse.
2. Mixed constraints: A is group-sparse and B is low-rank.

- vary temporal dimension T;
- vary joint spectral radius $\rho(A)\rho(B) \in \{0.5, 0.9\}$;
- compare NKP initialization, identity initialization, and adversarial initialization.

Metrics

- support recovery F1 score;
- relative estimation error $\left\|\widehat{G} - G\right\|_F / \|G\|_F$;
- forecasting accuracy: MSFE, RMSE, CRPS.

Simulation: Initialization and Support Recovery

		F1 of group sparse $A \in \mathbb{R}^{64\times 64}$			F1 of sparse $B \in \mathbb{R}^{48\times 48}$		
ρ	T	Refine	Identity	Adversarial	Refine	Identity	Adversarial
0.5	64	0.850 (0.366)	0.000 (0.000)	0.000 (0.000)	0.732 (0.316)	0.000 (0.000)	0.000 (0.000)
	128	1.000 (0.000)	0.337 (0.027)	0.340 (0.028)	0.907 (0.004)	0.584 (0.119)	0.595 (0.123)
	256	1.000 (0.001)	0.591 (0.308)	0.651 (0.324)	0.932 (0.004)	0.790 (0.108)	0.799 (0.134)
	512	1.000 (0.000)	1.000 (0.000)	0.968 (0.145)	0.949 (0.004)	0.949 (0.004)	0.938 (0.049)
0.9	64	0.997 (0.005)	0.946 (0.156)	0.997 (0.005)	0.878 (0.011)	0.865 (0.041)	0.878 (0.011)
	128	0.999 (0.003)	0.999 (0.003)	0.999 (0.003)	0.895 (0.006)	0.895 (0.006)	0.895 (0.006)
	256	1.000 (0.001)	1.000 (0.001)	1.000 (0.001)	0.908 (0.006)	0.908 (0.006)	0.908 (0.006)
	512	1.000 (0.002)	1.000 (0.002)	1.000 (0.002)	0.914 (0.005)	0.914 (0.005)	0.914 (0.005)

Empirical message

NKP initialization is crucial for accurate support recovery, especially when T is small relative to p and M. Identity and adversarial starts can fail badly in the low-sample regime.

Simulation: Prediction and Structural Tradeoff

		Prediction RMSE			Prediction CRPS		
ρ	T	Refine	Identity	Adversarial	Refine	Identity	Adversarial
0.5	64	1.034 (0.046)	1.139 (0.003)	1.139 (0.003)	92.488 (8.204)	111.080 (0.498)	111.080 (0.498)
	128	1.007 (0.002)	1.004 (0.002)	1.004 (0.002)	53.001 (0.146)	52.465 (0.168)	52.484 (0.178)
	256	1.003 (0.003)	1.002 (0.003)	1.002 (0.003)	41.573 (0.092)	41.516 (0.100)	41.513 (0.088)
	512	1.002 (0.004)	1.002 (0.004)	1.002 (0.004)	39.633 (0.141)	39.634 (0.141)	39.631 (0.140)
0.9	64	1.028 (0.004)	1.026 (0.006)	1.028 (0.004)	91.239 (0.626)	90.923 (1.157)	91.238 (0.628)
	128	1.014 (0.003)	1.014 (0.003)	1.014 (0.003)	53.589 (0.171)	53.551 (0.209)	53.549 (0.208)
	256	1.007 (0.003)	1.007 (0.003)	1.007 (0.003)	41.767 (0.091)	41.766 (0.089)	41.764 (0.089)
	512	1.004 (0.003)	1.004 (0.003)	1.004 (0.003)	39.696 (0.111)	39.696 (0.110)	39.698 (0.115)

Empirical message

The method can trade a very small loss in predictive accuracy for substantially cleaner structure and Granger-causal interpretability.

Simulation: Empirical Rate Check

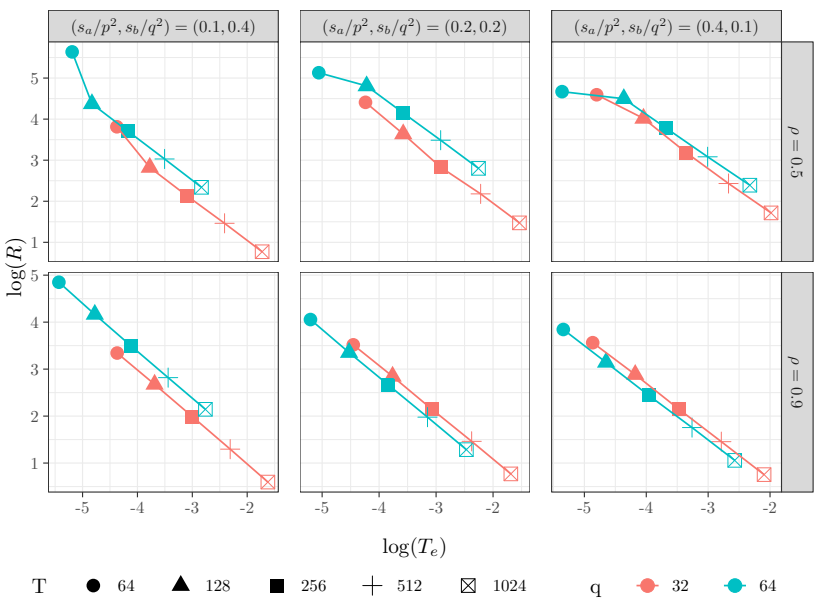
The theory motivates the weighted error

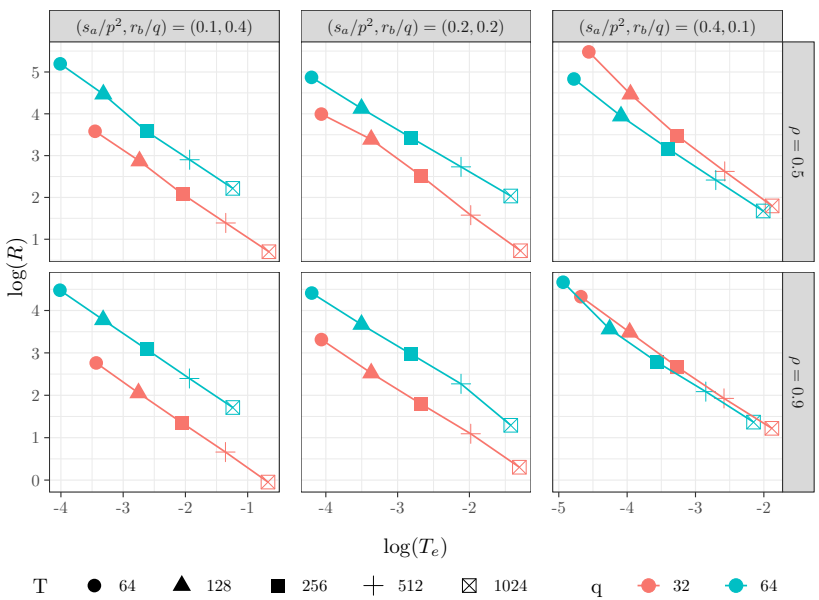
$$R = M \left\| A - \widehat{A} \right\|_F^2 + p \left\| B - \widehat{B} \right\|_F^2,$$

up to the scale-normalization constants κ_A, κ_B. Define the effective sample length

$$T_e = \begin{cases} T/[(s_A + s_B)\log(pM)], & \text{sparse/group-sparse,} \\ T/[s_A \log(pM) + r_B M], & \text{mixed.} \end{cases}$$

Ignoring optimization error suggests $\log R$ decreases approximately linearly with slope -1 in $\log T_e$.





Eurozone Macroeconomic Indicators: Data and Benchmarks

Data layout

- 11 European countries, 2001–2019.
- 10 monthly indicators and 5 quarterly indicators.
- Monthly variables are expanded into three within-quarter variables to align frequencies.

$$p = 10 \times 3 + 5 = 35, \qquad M = 11.$$

Benchmarks

- ridge-penalized individual VARs;
- stacked/vectorized ridge VAR;
- unpenalized ALS MAR;
- NKP initialization;
- proposed regularized MAR structures.

Eurozone Results: Predictive Accuracy

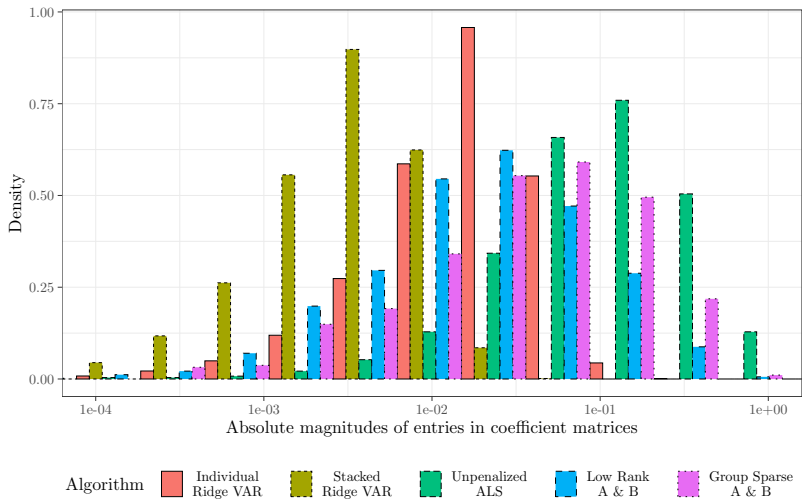
Model Structure	Individual Ridge VAR	Stacked Ridge VAR	Unpenalized ALS	NKP Init.	LR A LR B	LR A Sparse B	LR A GS B	GS A LR B	GS A GS B
RMSE	**0.76**	0.77	1.06	0.78	0.77	0.77	0.78	0.77	0.77
CRPS	**10.55**	10.69	14.87	10.87	10.87	10.86	10.83	10.75	10.81

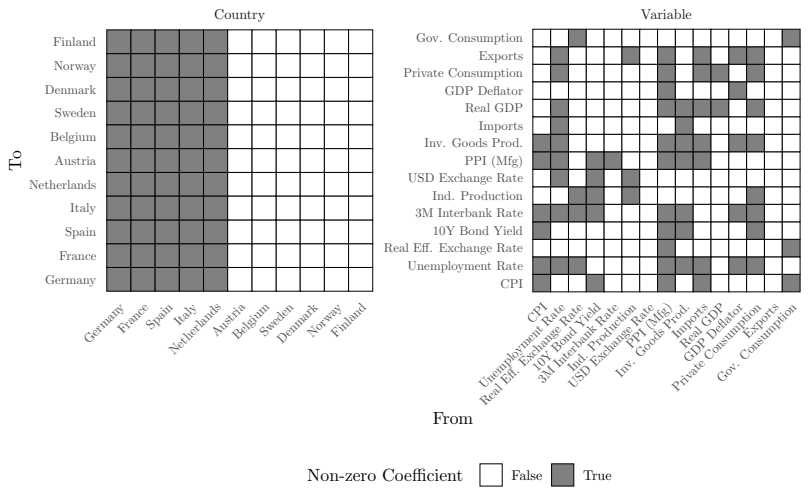
LR = low-rank, GS = group-sparse.

Interpretation

Ridge VARs are slightly best in pure prediction, but regularized MAR is competitive while preserving a two-layer variable/entity interpretation.

Eurozone Results: Parsimony and Interpretability





Takeaway

The ridge benchmarks can be marginally better in prediction, but their coefficients are tiny and hard to interpret. Penalized MAR yields more interpretable coefficient structure for both country and indicator lead-lag effects.

Concluding Remarks

Main contribution

The two-stage regularized MAR framework unifies sparse, group-sparse, low-rank, and mixed structures for matrix-valued time series.

- The MAR Granger effect is separable: $B_{m\ell}A_{ij}$.
- NKP warm starts stabilize a biconvex optimization problem.
- The theoretical results separate oracle statistical estimation error from optimization/nonconvexity error.
- Simulations validate structural recovery and empirical rate behavior.
- Eurozone data illustrate the tradeoff between prediction, parsimony, and interpretability.

Looking ahead: MAR versus flexible panel VARs

The MAR model is most attractive when the dynamics are approximately separable: a common variable-level mechanism A is transmitted across entities through an entity network B, giving Granger effects of the form

$$(j, \ell) \to (i, m) \quad \Longleftrightarrow \quad B_{m\ell} A_{ij} \neq 0.$$

This yields strong parsimony and interpretability, but it can also be restrictive.

For very large or highly heterogeneous panels, flexible panel VAR models with entity-specific transition matrices A_m, linked through shrinkage, factors, clustering, or hierarchical structure, may capture richer entity-level heterogeneity.