Bivariate Matrix-valued Linear Regression (BMLR)
Finite-sample performance under Identifiability and Sparsity

Nayel Bettache

Cornell University & Capital Fund Management

NeurIPS 2025

Problem & Model

Motivation

- Modern data often has matrix structure on both sides:
  - spatiotemporal signals, dynamic imaging
  - multivariate longitudinal data, dynamic networks
- Rows / columns encode different dimensions (time, space, conditions, . . .).

Bivariate Matrix-valued Linear Regression (BMLR)

$$X_t \in \mathbb{R}^{m\times q}, \quad Y_t \in \mathbb{R}^{n\times p}, \quad Y_t = A^* X_t B^* + E_t, \quad t = 1, \dots, T.$$

- $A^* \in \mathbb{R}_+^{n\times m}$: rows have ℓ_1-norm 1 $\Rightarrow$ identifiability, mixing/attention interpretation.
- $B^* \in \mathbb{R}^{q\times p}$; E_t: i.i.d. Gaussian noise.

Goal: estimate A^*, B^* with explicit formulas and finite-sample guarantees.

Naive Approaches vs BMLR Structure

Naive 1: independent trace regressions

$$[Y_t]_{ij} = \mathrm{Tr}(X_t^\top M_{ij}^*) + [E_t]_{ij}, \quad M_{ij}^* \in \mathbb{R}^{m\times q}.$$

- Treats the np coordinates independently.
- Ignores correlations and shared structure across entries of Y_t.

Naive 2: vectorize everything

$$\mathrm{vec}(Y_t) = (B^*)^\top \otimes A^* \ \mathrm{vec}(X_t) + \mathrm{vec}(E_t).$$

- Standard multivariate regression on $\mathrm{vec}(X_t)$.
- Recovering A^*, B^* from $\hat{M}$ is a non-convex Kronecker factorization.
- High variance when $nmpq \gg T$; matrix structure is not used explicitly.

Our strategy: stay in matrix form and exploit the bilinear structure $Y_t = A^* X_t B^*$ directly.

Main Idea: Explicit Estimators via a Factorization

Population (noiseless) model:
$$M_t = A^* X_t B^*, \quad t = 1, \dots, T.$$

Stack vectorized matrices:
$$\mathbb{M} = \begin{pmatrix} \mathrm{vec}(M_1)^\top \\ \vdots \\ \mathrm{vec}(M_T)^\top \end{pmatrix} \in \mathbb{R}^{T \times np}, \quad \mathbb{X} = \begin{pmatrix} \mathrm{vec}(X_1)^\top \\ \vdots \\ \mathrm{vec}(X_T)^\top \end{pmatrix} \in \mathbb{R}^{T \times mq}.$$

Assume $\mathbb{X}$ has full column rank and define $\mathbb{C} := (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \mathbb{M} \in \mathbb{R}^{mq \times np}$.

Key factorization:
$$[\mathbb{C}]_{(k,l)}^{(i,j)} = [A^*]_{ik} \, [B^*]_{lj},$$

hence A^* and B^* can be recovered as simple averages over rows / columns of $\mathbb{C}$.

Sample version:
$$\widehat{\mathbb{C}} := (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \mathbb{Y} = \mathbb{C} + \mathbb{D}.$$

Plug-in:
$$[\hat{B}]_{lj} = \frac{1}{n} \sum_{i,k} [\widehat{\mathbb{C}}]_{(k,l)}^{(i,j)}, \quad [\hat{A}]_{ik} = (\text{average ratio}) \ \text{clipped to } [0,1].$$

No alternating minimization, no iterative optimization.

Finite-sample Guarantees (Dense & Sparse)

Assume an orthogonal design:

$$\mathbb{X}^\top \mathbb{X} = T I_{mq}.$$

Then $\widehat{\mathbb{C}}$ is matrix-normal with independent Gaussian entries.

Dense case

- Explicit non-asymptotic bounds for $\left\|\hat{B} - B^*\right\|_F$, $\left\|\hat{B} - B^*\right\|_{\mathrm{op}}$, and $\left\|\hat{A} - A^*\right\|_+$.
- Qualitative behavior:
  - errors shrink with sample size T;
  - $\hat{B}$ improves with larger row dimension n ("blessing of dimensionality");
  - $\hat{A}$ benefits from larger p, q, but deteriorates as n, m grow.

Sparse case

- Hard-thresholded estimators $\hat{B}^S$, $\hat{A}^S$.
- Frobenius error scales with the sparsity level $\|B^*\|_0$, $\|A^*\|_0$.
- Under signal-strength conditions, we recover the exact support with high probability.

Experiments & Takeaways

Synthetic experiments

- Randomly generated A^*, B^*, X_t; Gaussian noise.
- Study $\left\|\hat{A} - A^*\right\|$, $\left\|\hat{B} - B^*\right\|$ vs. T and dimensions.
- Empirical trends match theory, including the asymmetric roles of A^* and B^*.

CIFAR-10 image denoising

- $32 \times 32 \times 3$ images; apply noisy linear transforms via $(A^*)^{-1}$ and $(B^*)^{-1}$.
- Learn $\hat{A}, \hat{B}$ on training images, correct noisy test images via $X_{\mathrm{corr}} = \hat{A} X_{\mathrm{noisy}} \hat{B}$.
- Corrected images are substantially closer to originals (Frobenius distance) across noise levels.

Takeaways

- BMLR couples both sides of matrix-structured data: $Y_t = A^* X_t B^* + E_t$.
- We provide explicit, optimization-free estimators with finite-sample guarantees, dense and sparse.
- Simple procedures already work well on synthetic data and real images.