Mixed Linear Regression via Approximate Message Passing

UNIVERSITY OF CAMBRIDGE

Nelvin Tan   Ramji Venkataramanan

Mixed Linear Regression (MLR)

Model. Heterogenous unlabeled data:

Each observation $Y_1, \dots, Y_n$ comes from one of L regressors $\beta^{(1)}, \dots, \beta^{(L)} \in \mathbb{R}^p$, but we don't know which one.

$$Y_i = \langle X_i, \beta^{(1)} \rangle c_{i1} + \cdots + \langle X_i, \beta^{(L)} \rangle c_{iL} + \epsilon_i , \quad i \in [n].$$

For each i, exactly one of $(c_{i1}, \dots, c_{iL})$ is 1, the rest are 0.

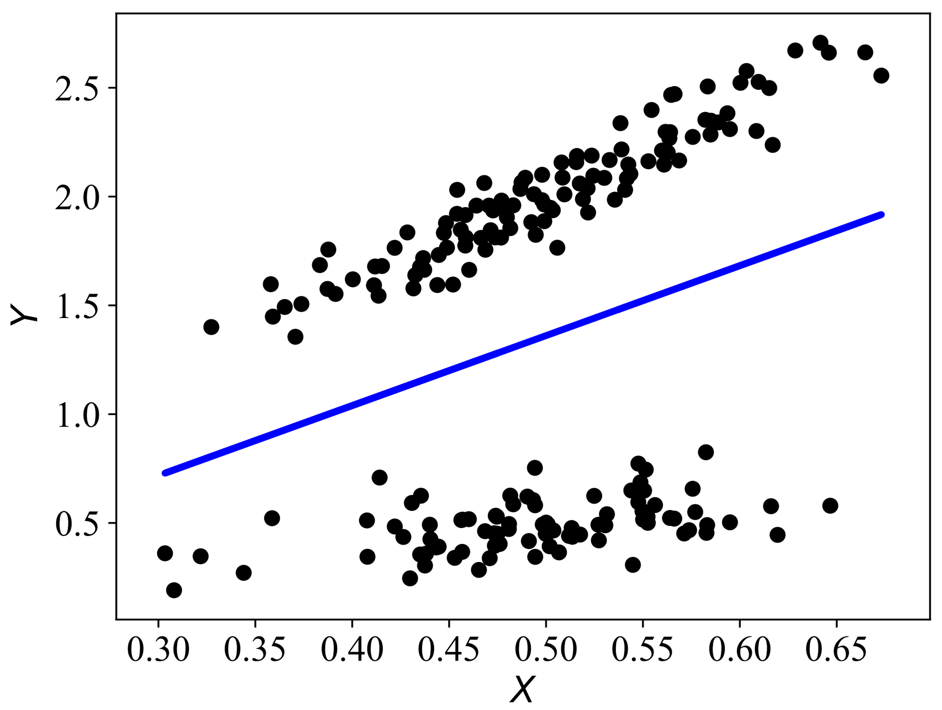

Figure 1. MLR example with two components and $p = 1$. The line, obtained via ordinary least squares, shows that standard linear regression is inadequate here.

Goal. Estimate multiple signals (regressors) $\beta^{(1)}, \dots, \beta^{(L)} \in \mathbb{R}^p$ from observations $Y_1, \dots, Y_n$ and feature vectors $X_1, \dots, X_n \in \mathbb{R}^p$.

Main contributions

1. Novel approximate message passing **(AMP)** algorithm for MLR
   - Can be tailored to take advantage of prior information on signals
   - Per-iteration complexity $\mathcal{O}(np)$
2. Rigorous performance characterization for AMP via ***state evolution*** in the high-dimensional limit as $n, p \to \infty$, with $n/p \to \delta$
   - Precise asymptotics for MSE and correlation of AMP iterates with signals

Approximate Message Passing

AMP algorithm: Starting with initial guess $\widehat{B}^0$, for $k \geq 0$, compute:

$$\Theta^k = X\widehat{B}^k - \widehat{R}^{k-1}(F^k)^\top, \quad \widehat{R}^k = g_k(\Theta^k, Y),$$
$$B^{k+1} = X^\top \widehat{R}^k - \widehat{B}^k (C^k)^\top, \quad \widehat{B}^{k+1} = f_{k+1}(B^{k+1}).$$

- Iteratively produces estimates $\widehat{B}^k$ of B
- Denoisers g_k and f_{k+1} are Lipschitz and applied component-wise
- $C^k = \frac{1}{n}\sum_{i=1}^n g_k'(\Theta_i^k, Y_i)$ and $F^{k+1} = \frac{1}{n}\sum_{j=1}^p f_{k+1}'(B_j^{k+1})$

State Evolution Theorem

Main assumptions:

- Features $X_i \sim_{\text{i.i.d.}} \mathcal{N}(0, I_p/n)$. As $n, p \to \infty$, we have $n/p \to \delta$
- Limiting distribution of the rows of B exists and follows $\bar{B}$

Theorem. For any pseudo-Lipschitz ϕ,

$$\frac{1}{p}\sum_{j=1}^p \phi(B_j^{k+1}, B_j) \to \mathbb{E}[\phi(\mathrm{M}_B^{k+1}\bar{B} + G_B^{k+1}, \bar{B})],$$

where $G_B^{k+1} \sim \mathcal{N}(0, \mathrm{T}_B^k)$, and the parameters $\mathrm{M}_B^k, \mathrm{T}_B^k \in \mathbb{R}^{L\times L}$ are computed via a deterministic state evolution recursion.

Normalized squared correlation. By choosing a suitable ϕ, we can compute the asymptotic correlation between each signal and its AMP estimate. For iteration k and signal $\beta^{(l)}$:

$$\underbrace{\frac{\langle \hat{\beta}^{(l),k}, \beta^{(l)} \rangle^2}{\|\hat{\beta}^{(l),k}\|_2^2 \, \|\beta^{(l)}\|_2^2}}_{\text{empirical}} \to \underbrace{\frac{\mathbb{E}[f_{k,l}(\mathrm{M}_B^k \bar{B} + G_B^k)\bar{B}_l]^2}{\mathbb{E}[f_{k,l}(\mathrm{M}_B^k \bar{B} + G_B^k)^2 \, \mathbb{E}[\bar{B}_l^2]}}_{\text{theoretical}}.$$

Choice of denoisers. The state evolution parameters depend on choice of f_k, g_k. We propose:

$$f_k(s) = \mathbb{E}[\bar{B} \mid \mathrm{M}_B^k \bar{B} + G_B^k = s]$$
$$g_k(u, y) = \mathrm{Cov}[Z \mid Z^k = u]^{-1}\big(\mathbb{E}[Z \mid Z^k = u, \bar{Y} = y] - \mathbb{E}[Z \mid Z^k = u]\big),$$

which minimizes the effective noise in each iteration.

Proof idea

MLR is an instance of a **Matrix Generalized Linear Model** with latent variables.

Let $B = [\beta^{(1)}, \dots, \beta^{(L)}]$ be the signal matrix and Ψ_i an auxiliary vector. The matrix GLM model is:

$$Y_i = q(B^\top X_i, \Psi_i), \quad i \in [n],$$

where q is a known output function.

To get the MLR, take $B = (\beta^{(1)}, \dots, \beta^{(L)})$ and $\Psi_i = (c_{i1}, \dots, c_{iL}, \epsilon_i)$.

Proof idea. Establish state evolution result for AMP for matrix GLM via reduction to an abstract AMP [1].

Numerical Simulations

Model choice. We look at the two-component case:

$$Y_i = \langle X_i, \beta^{(1)} \rangle c_i + \langle X_i, \beta^{(2)} \rangle (1 - c_i) + \epsilon_i,$$

where $c_i \sim_{\text{i.i.d.}}$ Bernoulli(α), with $\alpha \in (0,1)$, and $\epsilon_i \sim_{\text{i.i.d.}} \mathcal{N}(0, \sigma^2)$.

Standard Gaussian prior. Independent signals with

$$\beta_j^{(1)}, \beta_j^{(2)} \sim_{\text{i.i.d.}} \mathcal{N}(0,1), \quad j \in [p].$$

Gaussian Prior Plots

For each setting, we plot:

- Empirical normalized squared correlation (labeled 'AMP')
- Theoretical normalized squared correlation (labeled 'SE')

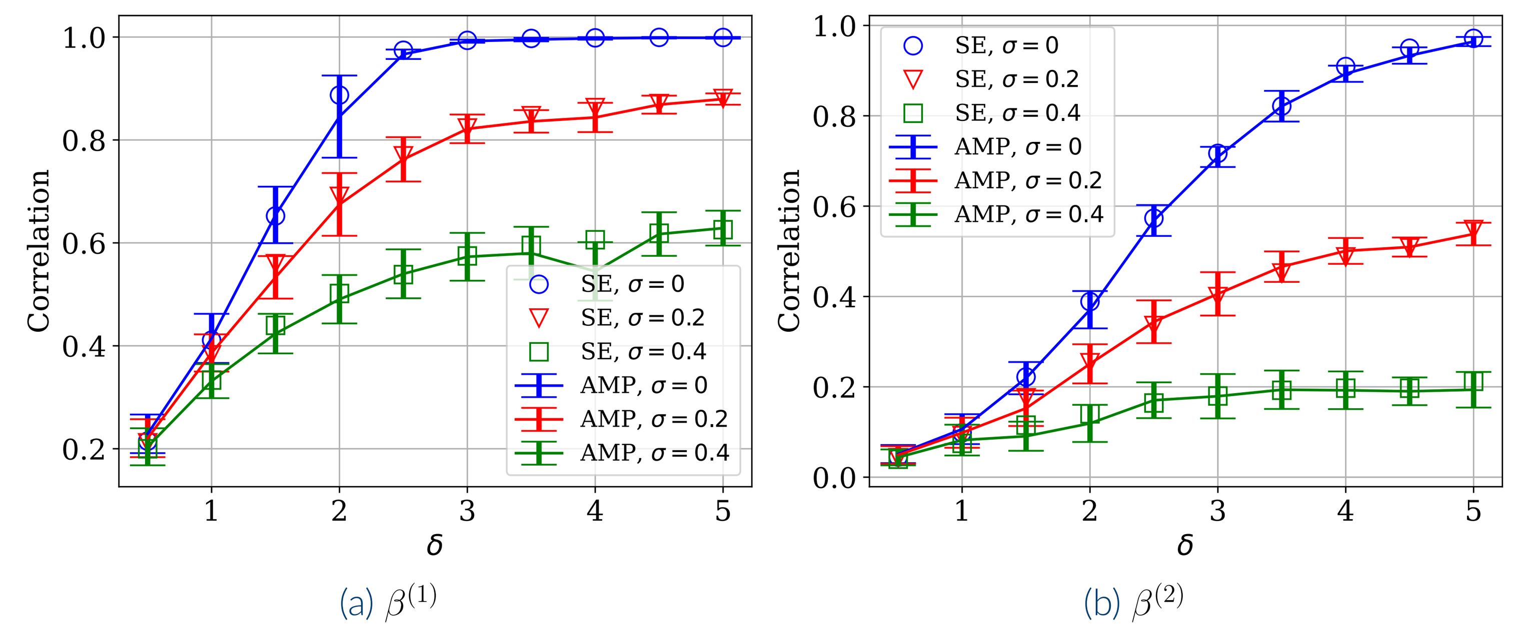

Figure 2. Normalized squared correlation vs. δ for various noise levels σ, with $\alpha = 0.7$

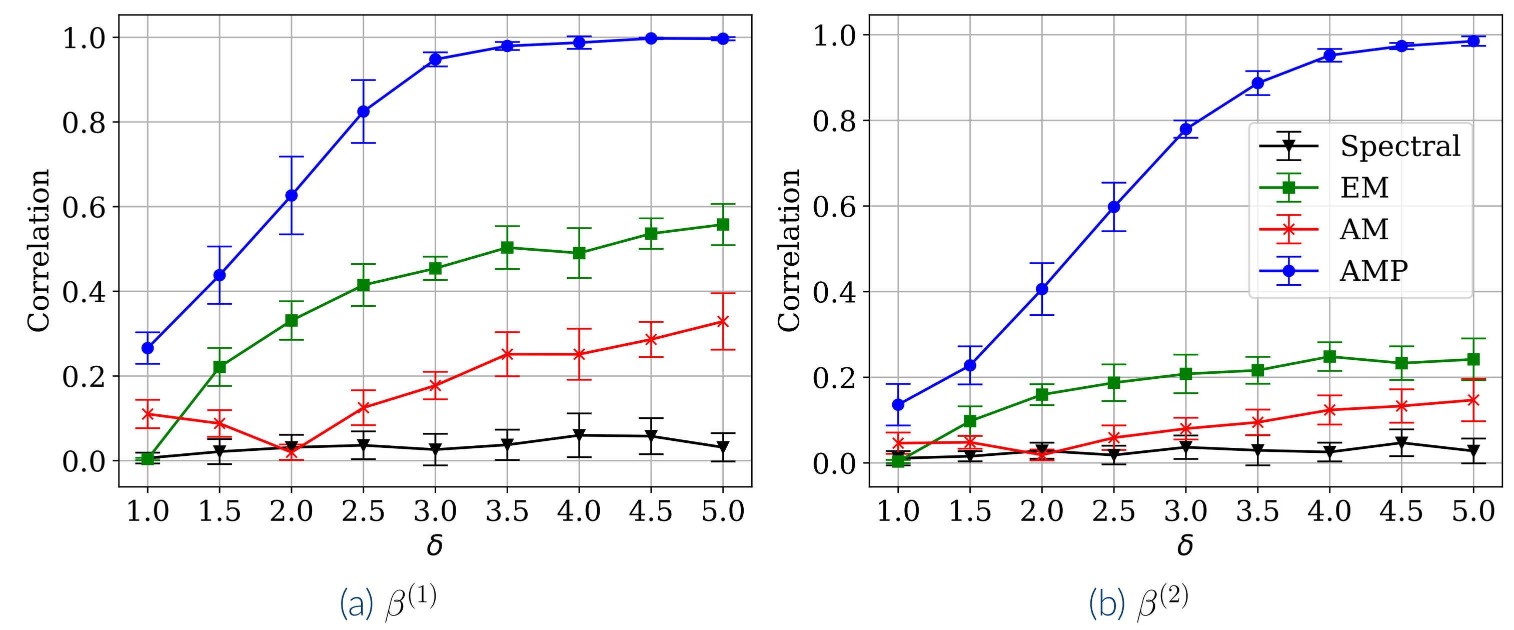

Figure 3. Comparison of different estimators; Normalized squared correlation vs. δ, with $\alpha = 0.6$, and $\sigma = 0$.

AMP significantly outperforms other popular techniques:

Spectral estimator, Alternating Minimization, Expectation-Maximization

Summary

- **Novel AMP algorithm** for mixed linear regression
- Sharp asymptotic guarantees via **state evolution**
- Algorithm and guarantees can be generalized to any instance of matrix GLM, e.g., max-affine regression, mixture-of-experts [2]

References

[1] Oliver Y. Feng, Ramji Venkataramanan, Cynthia Rush, and Richard J. Samworth. A unifying tutorial on approximate message passing. *Foundations and Trends in Machine Learning*, 2022.

[2] Nelvin Tan and Ramji Venkataramanan. Mixed regression via approximate message passing, 2023. https://arxiv.org/abs/2304.02229.