Regularization for Shuffled Data Problem via Exponential Family Prior on the Permutation Group

Zhenbang Wang

George Mason University
Department of Statistics

joint work with Emanuel Ben-David and Martin Slawski

September 30, 2021

Probability 101

- $\Pi_{n\times n}$ is a permutation matrix if $\forall \pi_{ij} \in \{0,1\}$, $\sum_{i=1}^{n} \pi_{ij} = 1$ for $j = 1, ..., n$ and $\sum_{j=1}^{n} \pi_{ij} = 1$ for $i = 1, ...n$.

  e.g. Identity matrix

- Permutation group $\mathcal{P}(n)$ is the group of one-one mappings from a finite set $\{1, ..., n\}$ into itself

$$\text{e.g. } \begin{pmatrix} 1 & 1 \\ 2 & 2 \\ 3 & 3 \\ \vdots & \\ n & n \end{pmatrix} \Longleftrightarrow \begin{bmatrix} 1 & \cdots & 0 \\ 0 & \ddots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & 1 \end{bmatrix} \quad d_{\mathrm{H}}\left(\begin{pmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 3 \\ \vdots & \\ n & n \end{pmatrix}, \mathrm{I}_n \right) = 2$$

- Hamming distance - the number of positions at which the corresponding numbers are different

Example : "Shuffled Linear Regression"

Considering the following linear regression model

$$\underset{n\times 1}{\mathsf{Y}} = \underset{n\times n}{\Pi^*}\ \underset{n\times 1}{\mathsf{Y}^*}, \quad \mathsf{Y}^* = \mathsf{X}\beta^* + \sigma\epsilon, \quad \epsilon \sim \mathcal{N}_n(0, I_n)$$

Objective : Given Y and X, estimate Π^* (Permutation Recovery) and β^* (Unlabeled Sensing)

Example : d = 1, sign of β^* is positive

Consider the maximum likelihood estimator of $\Pi*$

$$\begin{aligned}\hat{\Pi}_{\mathrm{ML}} &= \underset{\Pi\in\mathcal{P}(n)}{\operatorname{argmin}} \|\mathrm{Y} - \Pi\mathrm{X}\beta^*\|_2^2 \\ &= \underset{\Pi\in\mathcal{P}(n)}{\operatorname{argmin}} \|\mathrm{Y}\|_2^2 + \|\Pi\mathrm{X}\beta^*\|_2^2 - 2\langle \mathrm{Y}, \Pi\mathrm{X}\beta^*\rangle \\ &= \underset{\Pi\in\mathcal{P}(n)}{\operatorname{argmax}} \langle \mathrm{Y}, \Pi\mathrm{X}\beta^*\rangle \\ &= \underset{\Pi\in\mathcal{P}(n)}{\operatorname{argmax}} \langle \mathrm{Y}, \Pi\mathrm{X}\rangle = \underset{\Pi\in\mathcal{P}(n)}{\operatorname{argmax}} \sum_{i=1}^{n} x_{\pi(i)} y_i\end{aligned}$$

So $\hat{\Pi}_{\mathrm{ML}}$ is given by the following

$$\sum_{i=1}^{n} x_{\hat{\pi}_{\mathrm{ML}}(i)} y_i = \sum_{i=1}^{n} x_{(i)} y_{(i)}$$

Example : d = 1, sign of β^* is positive

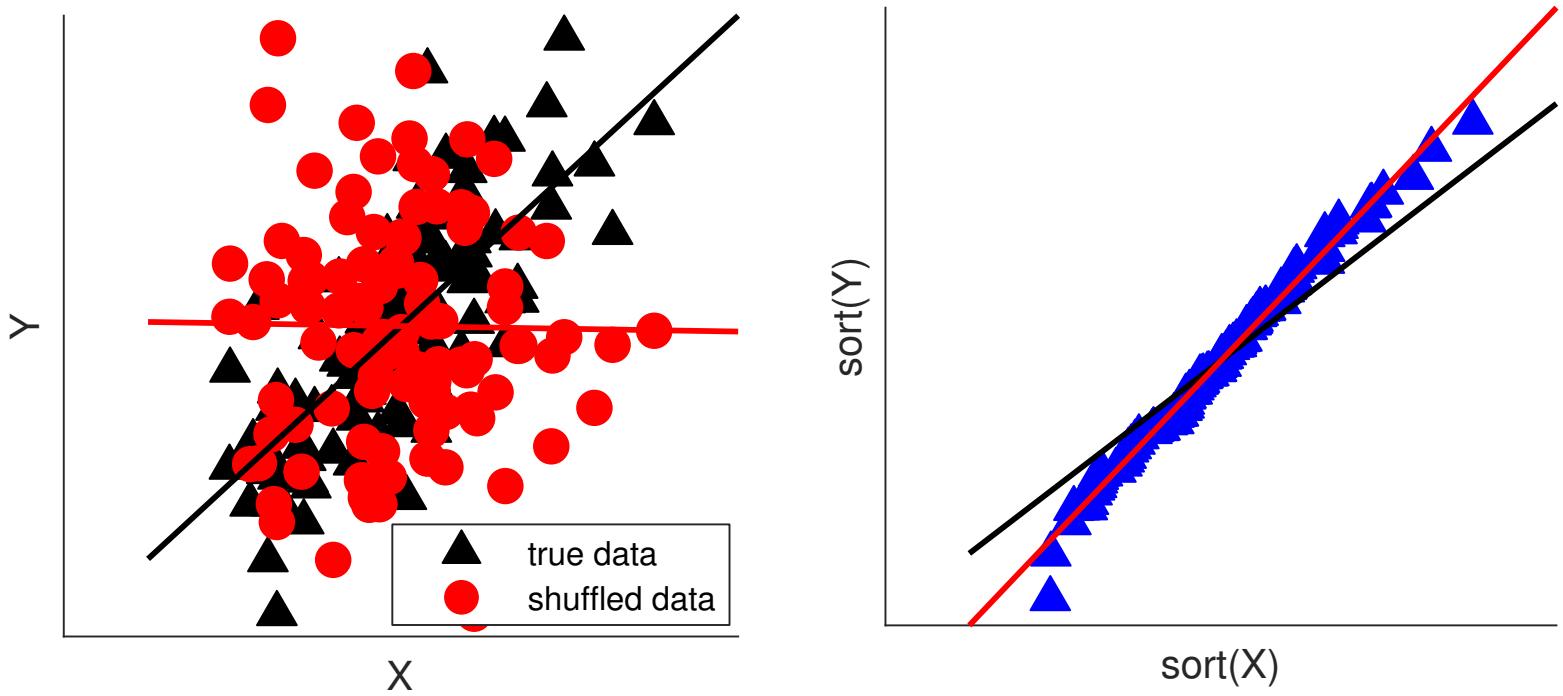

Figure: Samples from the model $y_i = x_{\pi^*(i)}\beta^* + \varepsilon_i$, $n = 100$, with a random permutation (Left). Estimation of β^* based on $\hat{\Pi}_{\mathrm{ML}}$ (Right).

Issues with ML approach

$$\widehat{\beta}_{\mathrm{ML}} = \sum_{i=1}^{n} x_{(i)} y_{(i)} \Big/ \sum_{i=1}^{n} x_i^2, \qquad \widehat{\sigma}^2_{\mathrm{ML}} = \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i \widehat{\beta}_{\mathrm{ML}})^2$$

1. Due to combinatorial setup, there are few approaches that are both computationally inexpensive and statistically promising.
2. Assuming a uniform prior (non-informative) on Π, the ML approach suffer from critical overfitting ($\widehat{\sigma}^2_{\mathrm{ML}} \to 0$ in probability as $n \to \infty$).

Exponential family prior

One possible way of dealing with overfitting is by imposing additional regularization on the permutation which is given by

$$p(\Pi) \propto \exp(\gamma \cdot \mathrm{tr}(C^\top \Pi)) = \exp(\gamma \sum_{1 \le i,j \le n} c_{ij}\pi_{ij})$$

where γ is a positive constant.

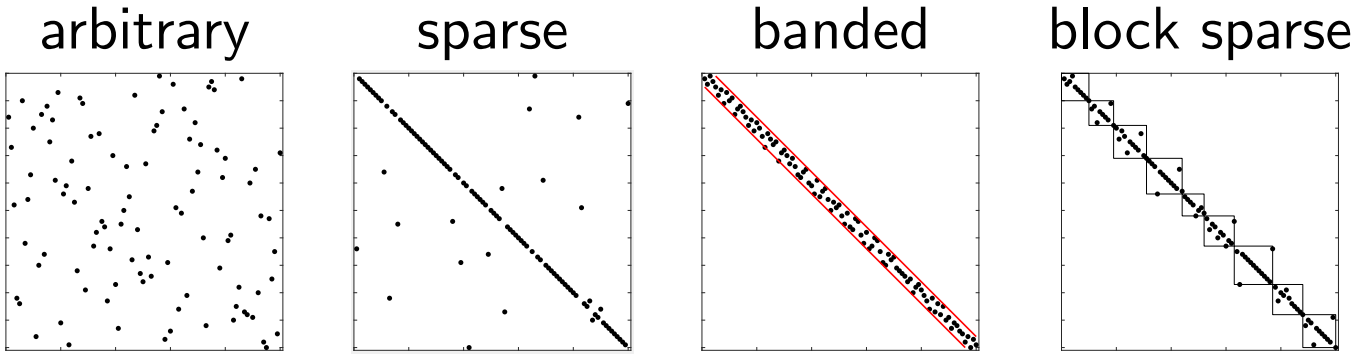

Figure: Example of structure of permutation

Exponential family prior

$$p(\Pi) \propto \exp(\gamma \cdot \mathrm{tr}(C^\top \Pi)) = \exp(\gamma \sum_{1 \le i,j \le n} c_{ij}\pi_{ij})$$

- If $\gamma = 0$, then $p(\Pi) = \frac{1}{n!}$.
- If $C = \mathrm{I}_n$, then $\mathrm{tr}(C^\top \Pi) = n - d_{\mathrm{H}}(\Pi, \mathrm{I}_n)$, $p(\Pi) \propto \exp(-\gamma d_{\mathrm{H}}(\Pi, \mathrm{I}_n))$.
- $C = \{M \in \mathbb{R}^{n \times n} : M_{i,j} = 0, |i-j| > r : M_{i,j} = \phi(|i-j|)\ o.w\}$
- $C = \{M \in \mathbb{R}^{n \times n} : M_{i,j} = 0, \text{if i,j are not in the same block}\}$

Set up

The conditional likelihood of the observed data $\mathcal{D} = \{x_i, y_i\}_{i=1}^n$ is given by:

$$L(\boldsymbol{\theta}, \pi|\mathcal{D}) = \prod_{i=1}^{n} p(y_i, x_{\pi(i)}|\boldsymbol{\theta}) = L(\boldsymbol{\theta}, \Pi|\mathcal{D}) = \prod_{i=1}^{n}\prod_{j=1}^{n} p(y_i, x_j|\boldsymbol{\theta})^{\pi_{ij}}$$

where Π is the matrix representation of π and $\pi_{ij} = I(\pi(i) = j)$.

Inference of $\boldsymbol{\theta}$ with π missing (EM approach)

Consider the estimating θ by maximizing the integrated likelihood

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\text{argmax}}\, L(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\text{argmax}}\, \mathsf{E}_\pi[L(\boldsymbol{\theta}, \pi|\mathcal{D})] = \underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\text{argmin}} - \log \mathsf{E}_\pi[L(\boldsymbol{\theta}, \pi|\mathcal{D})]$$

So instead, we minimize the surrogate of $-\log \mathsf{E}_\pi[L(\boldsymbol{\theta}, \pi|\mathcal{D})]$ which is given by Jensen inequality

$$\begin{aligned} -\log \mathsf{E}_\pi[L(\boldsymbol{\theta}, \pi|\mathcal{D})] &\leq \mathsf{E}_{\pi|\mathcal{D},\boldsymbol{\theta}}[-\log L(\boldsymbol{\theta}, \pi|\mathcal{D})] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathsf{E}_{\pi|\mathcal{D},\boldsymbol{\theta}}[\pi_{ij}]\{-\log p(y_i, \mathrm{x}_j|\boldsymbol{\theta})\} \end{aligned}$$

EM scheme

So the EM algorithm is given by

$$\text{E} - \text{Step} : Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)}) = \sum_{i=1}^{n}\sum_{j=1}^{n} \text{E}_{\pi|\mathcal{D},\boldsymbol{\theta}^{(t-1)}}[\pi_{ij}]\{-\log p(y_i, \mathrm{x}_j|\boldsymbol{\theta}^{(t-1)})\}$$

$$\text{M} - \text{Step} : \boldsymbol{\theta}^{(t)} = \text{argmin}\ \ Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)})$$

Main Challenge

In E-step of EM algorithm,

$$\mathsf{E}_{\pi|\mathcal{D},\boldsymbol{\theta}}[\pi_{ij}] = \sum_{\Pi\in\mathcal{P}(n)} p(\pi_{ij}|\mathcal{D},\boldsymbol{\theta})\pi_{ij} \propto \sum_{\Pi\in\mathcal{P}(n)} p(\mathcal{D},\boldsymbol{\theta}|\pi_{ij})p(\pi_{ij})\pi_{ij}$$

Note : Taking the sum over $n!$ permutations is computationally impossible. ($10! \approx 3$ million)

$$\hat{\mathsf{E}}_{\pi|\mathcal{D},\boldsymbol{\theta}}[\pi_{ij}] = \frac{1}{m-b}\sum_{k=b+1}^{m} \pi_{ij}^{(k)}, \quad (i,j) \in [n]^2.$$

Markov Chain Monte Carlo (Metropolis-Hastings)

Objective : Generate random samples $\pi^{(1)}, \ldots, \pi^{(k)}, \ldots, \pi^{(m)}$ from posterior distribution $p(\pi|\mathcal{D}, \theta) \propto p(\mathcal{D}, \boldsymbol{\theta}|\pi)p(\pi)$

- proposal distribution : Fisher–Yates shuffle

$$\begin{pmatrix} 1 & 1 \\ 2 & 2 \\ 3 & 3 \\ \vdots & \\ n & n \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 3 \\ \vdots & \\ n & n \end{pmatrix} \rightarrow \begin{pmatrix} 1 & n \\ 2 & 1 \\ 3 & 3 \\ \vdots & \\ n & 2 \end{pmatrix} \rightarrow \ldots$$

- acceptance probability : $\min\left\{\frac{p(\widetilde{\pi}|\mathcal{D},\theta;\gamma)}{p(\pi^{(k)}|\mathcal{D},\theta;\gamma)}, 1\right\}$

Monte Carlo EM scheme

Input: $\mathcal{D} = \{\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n\}$, γ, EM-iter
Initialize $\theta^{(0)} \leftarrow \widehat{\theta}_{\text{init}}$.
for $t = 0, \ldots, ,$EM-iter

$\widehat{\pi}_{\text{init}} \leftarrow \text{argmax}_{\pi \in \mathcal{P}(n)}\, p(\pi|\mathcal{D}, \theta^{(t)})$.

$\widehat{\mathrm{E}}[\pi|\mathcal{D}, \theta^{(t)}] \leftarrow \mathtt{MH}(\mathcal{D}, \theta^{(t)}, \widehat{\pi}_{\text{init}}, \gamma, m)$.

$\theta^{(t+1)} \leftarrow \min_\theta \left\{ \sum_{i=1}^n \sum_{j=1}^n \widehat{\mathrm{E}}[\pi_{ij}|\mathcal{D}, \theta^{(t)}]\{-\log p(x_j, y_i; \theta)\} \right\}$.

$t \leftarrow t + 1$

end for

Metropolis-Hastings algorithm

Input: $\mathcal{D}, \theta, \widehat{\pi}_{\text{init}}, \gamma, m$
Initialize $\pi^{(0)} \leftarrow \widehat{\pi}_{\text{init}}$.
for $k = 0, \ldots, m$

Sample $(i, j) \in [n]^2$

$\widetilde{\pi}(i) \leftarrow \pi^{(k)}(j)$, $\widetilde{\pi}(j) = \pi^{(k)}(i)$.

$r(\widetilde{\pi}, \pi^{(k)}) \leftarrow \min\left\{\frac{p(\widetilde{\pi}|\mathcal{D},\theta;\gamma)}{p(\pi^{(k)}|\mathcal{D},\theta;\gamma)}, 1\right\}$.

Draw $u \sim U([0,1])$.

if $r(\widetilde{\pi}, \pi^{(k)}) > u$: $\pi^{(k+1)} \leftarrow \widetilde{\pi}$.

else: $\pi^{(k+1)} \leftarrow \pi^{(k)}$.

$k \leftarrow k + 1$.

end for
return $\widehat{\mathrm{E}}[\pi|\mathcal{D}, \theta]$

Models

- Linear Regression : $\theta = (\beta, \sigma^2)$, $-\log p(\mathrm{x}, y; \beta, \sigma^2) = \frac{1}{2\sigma^2}(y - \mathrm{x}^\top \beta)^2$
- Generalized linear model : $\theta = (\beta, \phi)$,
  $-\log p(\mathrm{x}, y; \beta, \phi) = \frac{\psi(\mathrm{x}^\top \beta) - y\mathrm{x}^\top \beta}{a(\phi)} + c(y, \phi)$
- Multivariate Normal model : $\theta = \Omega$, $\mathrm{z} = [\mathrm{x}^\top \ \mathrm{y}^\top]^\top$ and
  $-\log p(\mathrm{x}, \mathrm{y}; \Omega) = -\log\det\Omega + \mathrm{tr}(\Omega \mathrm{z}\mathrm{z}^\top)$

Evaluation Metric : Relative Estimation Error

- Linear Regression : $\|\beta^{\text{est}} - \beta^*\|_2 / \|\beta^*\|_2$
- Generalized linear model : Same as Linear Regression
- Multivariate Normal model : $\|\text{Corr}^{\text{est}} - \text{Corr}^*\|_F$

Sparse Permutation

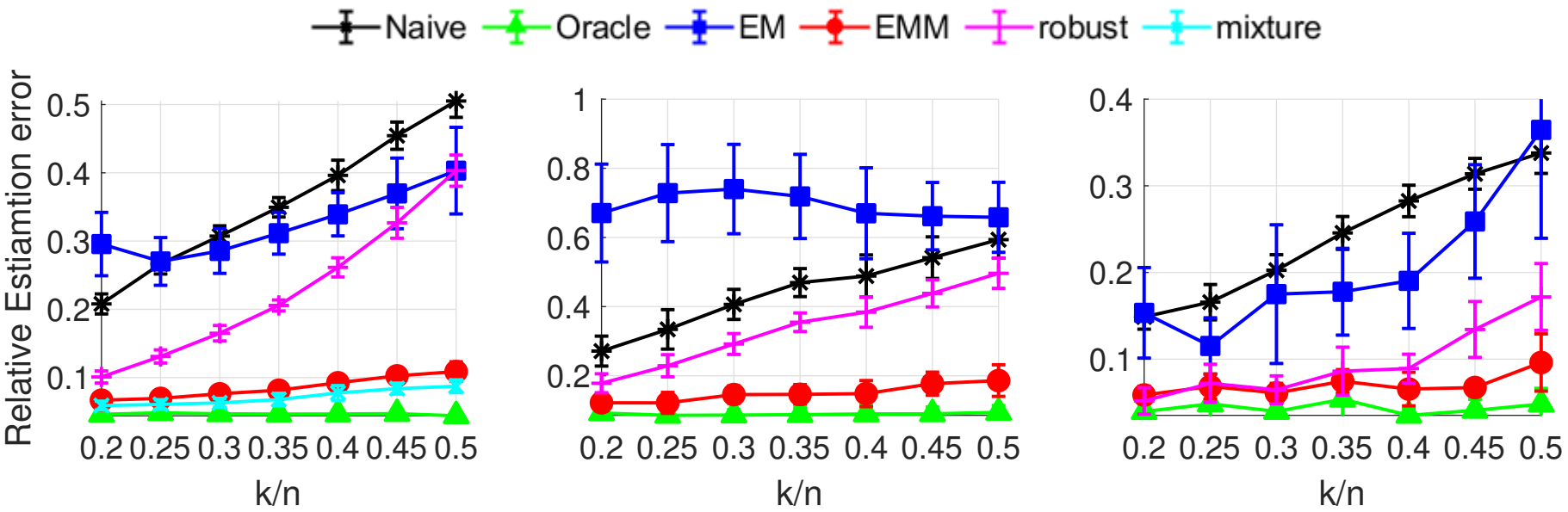

Figure: Simulation of synthetic data with Sparse Permutation (Linear regression[Left], GLM[Middle], Covariance[Right]). Each curves are average over 100 replications. The error bar represent $\pm 5 \times$ standard error.

Block Sparse Permutation

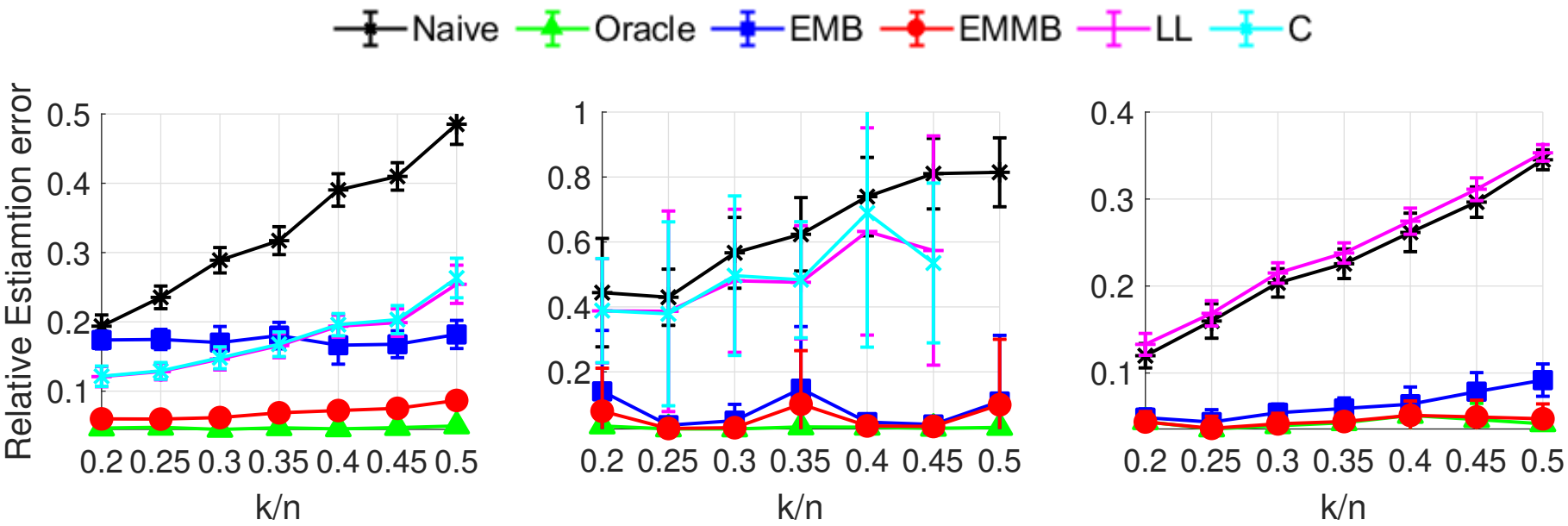

Figure: Simulation of synthetic data with Block Sparse Permutation (Linear regression[Left], GLM[Middle], Covariance[Right]). Each curves are average over 100 replications. The error bar represent $\pm 5 \times$ standard error.

Banded Permutation

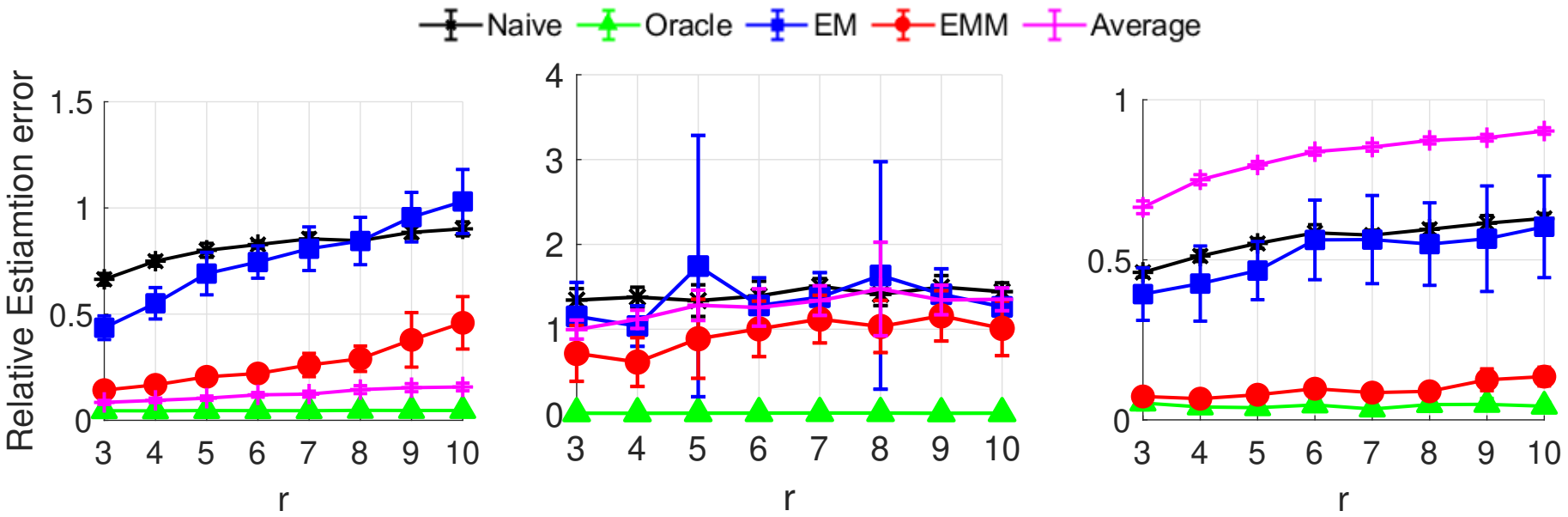

Figure: Simulation of synthetic data with Banded Permutation (Linear regression[Left], GLM[Middle], Covariance[Right]). Each curves are average over 100 replications. The error bar represent $\pm 5 \times$ standard error.

Take away

- Propose a framework for estimation in shuffled data problems
- Propose a prior distribution on Permutation Group that can be integrated into previous framework and prevent over-fitting issue of ML and unregularized EM approaches
- Future work : Full Bayesian Method (e.g. Gibbs sampler, Data augmentation and etc)

Thank you for your attention

Example : How to choose regularization parameter γ

Recall the hamming prior (Mallows model)

$$p(\pi) \propto \exp(-\gamma\, d_{\mathrm{H}}(\pi, \mathrm{id}))$$

Theorem

Suppose that π follows the Hamming prior p with parameter γ. Then for all $2 \leq k < n$

$$\mathrm{P}_{\pi\sim p}\left(d_H(\pi, id) \geq k\right) \begin{cases} \leq \exp(-k\delta \log n) & \text{if } \gamma \geq (1+\delta)\log n,\ \delta > 0, \\ \geq c(k,n) & \text{if } \gamma \leq \log(n-k), \end{cases}$$

where $c(k,n) \to \frac{1}{4}\frac{!k}{k!}$ as $n \to \infty$, with $!k$ denoting the number of derangements of k elements (i.e., the number of permutations without fixed point).

Example : How to choose regularization parameter γ