Regularization for Shuffled Data Problems via Exponential Family Priors on the Permutation Group

Zhenbang Wang
George Mason University

Emanuel Ben-David
U.S. Census Bureau

Martin Slawski
George Mason University

Abstract

In the analysis of data sets consisting of $(\mathbf{x}, \mathbf{y})$-pairs, a tacit assumption is that each pair corresponds to the same observational unit. If, however, such pairs are obtained via record linkage of two files, this assumption can be violated as a result of mismatch error rooting, for example, in the lack of reliable identifiers in the two files. Recently, there has been a surge of interest in this setting under the term "Shuffled Data" in which the underlying correct pairing of $(\mathbf{x}, \mathbf{y})$-pairs is represented via an unknown permutation. Explicit modeling of the permutation tends to be associated with overfitting, prompting the need for suitable methods of regularization. In this paper, we propose an exponential family prior on the permutation group for this purpose that can be used to integrate various structures such as sparse and local shuffling. This prior turns out to be conjugate for canonical shuffled data problems in which the likelihood conditional on a fixed permutation can be expressed as product over the corresponding $(\mathbf{x}, \mathbf{y})$-pairs. Inference can be based on the EM algorithm in which the E-step is approximated by sampling, e.g., via the Fisher-Yates algorithm. The M-step is shown to admit a reduction from n^2 to n terms if the likelihood of $(\mathbf{x}, \mathbf{y})$-pairs has exponential family form. Comparisons on synthetic and real data show that the proposed approach compares favorably to competing methods.

1 Introduction

Shuffled data problems refer broadly to situations in which the goal is to perform inference for a functional of the joint distribution of a pair of random variables $(\mathbf{x}, \mathbf{y})$ (such as, e.g., their covariance) based on separate samples $\{\mathbf{x}_i\}_{i=1}^n$ and $\{\mathbf{y}_i\}_{i=1}^m$ that involve matching pairs $\{(\mathbf{x}_{\pi^*(i)}, \mathbf{y}_i)\}_{i=1}^m$ pertaining to the same statistical unit, where the map $\pi^* : \{1, \ldots, m\} \to \{1, \ldots, n\}$ may only be observed incompletely. This is a rather common scenario in data integration problems in which different pieces of information about a shared set of entities reside in multiple data sources that need to be combined in order to perform a given data analysis task. The process of identifying matching parts across two or more files is often far from trivial in the absence of unique identifiers, and has thus grown into a vast and active field of research known as *record linkage*, e.g., Binette and Steorts (2020). The above shuffled data model represents a direct approach to account for mismatches in record linkage and the impact on downstream data analysis. Shuffled data problems were first systematically discussed in DeGroot et al. (1971), with little progress until a few years ago given advances in computation (Gutman et al., 2013). Recently, shuffled data problems have generated widespread interest, fueled by applications in signal processing (Unnikrishnan et al., 2018; Pananjady et al., 2018), correspondence problems in computer vision (Pananjady et al., 2017; Li et al., 2021) and NLP (Grave et al., 2019; Shi et al., 2021), biomedical data analysis (Ma et al., 2021a; Abid and Zou, 2018), and data privacy (Domingo-Ferrer and Muralidhar, 2016; Gordon et al., 2021).

On the theoretical side, several papers have investigated the statistical limits of signal estimation and permutation recovery in *unlabeled sensing* in which the goal is to recover a signal θ^* from n noisy linear measurements $y_i = \langle \mathbf{x}_{\pi^*(i)}, \theta^* \rangle + \epsilon_i$, $1 \leq i \leq n$, where π^* is an unknown permutation, i.e., $m = n$ and π^* is one-to-one (Unnikrishnan et al., 2018; Pananjady et al., 2018; Hsu et al., 2017; Abid et al., 2017; Tsakiris and Peng, 2019). Another line of research has studied the setting in which $\mathbf{x}$ and $\mathbf{y}$ are scalar and related by a monotone map (Carpentier and Schlüter, 2016; Rigollet and Weed, 2019; Flammarion et al., 2019; Ma et al., 2020; Balabdaoui et al., 2021).

A common conclusion from these works is that shuffled data problems are generally plagued by both statistical and computational challenges. First, the combinatorial nature of π^* makes it hard to devise computationally tractable approaches with provable guarantees. Existing algorith-

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

mic "solutions" involve integer programming (Tsakiris and Peng, 2019; Peng and Tsakiris, 2020; Mazumder and Wang, 2021), the EM algorithm (Gutman et al., 2013; Abid and Zou, 2018; Tsakiris et al., 2020), sampling and approximate inference (McVeigh et al., 2019; Steorts et al., 2016; Klami, 2012). Regardless of the computational challenges, shuffled data problems tend to be highly susceptible to noise and prone to overfitting. In fact, statistical guarantees typically involve unrealistically stringent signal-to-noise requirements (Pananjady et al., 2018; Hsu et al., 2017). Loosely speaking, this issue results from the fact that the set of permutations grows rapidly in size with n. This observation suggests that suitable forms of regularization hinging on prior information on π^* are needed to constrain the size of the parameter space under consideration. Several papers consider partial shufflings in which varying fractions of $(\mathbf{x}_i, \mathbf{y}_i)$-pairs are already observed with the correct correspondence (Slawski and Ben-David, 2019; Slawski et al., 2019, 2020; Zhang and Li, 2020; Peng et al., 2021), and only the remaining portion of the data is subject to shuffling. Another constraint commonly encountered in record linkage is that π^* is block-structured with known composition of the blocks based on auxiliary variables that are required to agree for matching records. In domains such as signal processing and computer vision, π^* is often constrained to act locally in the sense that indices are shuffled only within small time windows or image regions (Ma et al., 2021b; Abbasi et al., 2021).

The goal of the present paper is the development of a regularization framework for shuffled data problems that integrates those and other constraints in a unified way. To that end, we introduce an exponential family prior on the permutation group that is flexible enough to accommodate any kind of prior information that can be expressed solely in terms of index pairs (i, j). This prior turns out to be conjugate for canonical shuffled data problems in which the likelihood conditional on a fixed permutation can be expressed as product over the corresponding $(\mathbf{x}, \mathbf{y})$-pairs. Inference is based on the MC-EM algorithm considered in Wu (1998) and Abid and Zou (2018). We show that for exponential family likelihood, the resulting M-step is particularly scalable since it involves n instead of n^2 terms. Moreover, computation of the MAP estimator of π^* with the remaining parameters fixed reduces to a linear assignment problem, and hence remains computationally tractable. Theoretical results and a collection of experiments for various shuffled data setups demonstrate the usefulness of regularization based on the proposed prior in comparison to the unregularized counterpart and other baselines.

Notations. We denote by $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ the observed merged data, subject to shuffling. We use $(\mathbf{x}, \mathbf{y})$ for a generic pair of matching records. We use $\mathbf{X}$ and $\mathbf{Y}$ for the row-wise concatenation of $\{\mathbf{x}_i\}_{i=1}^n$ and $\{\mathbf{y}_i\}_{i=1}^n$, respectively. We let $p(\cdot)$ denote the density (PDF) of a list of variables in $(\cdot)$, and accordingly $p(\cdot \,|\, \cdot)$ is used for conditional PDFs. We write $u \sim p$ to express that random variable u has density p. The symbol $\mathbf{E}_{(\ldots)}[\cdot]$ is the expectation w.r.t. $(\ldots)$. The Hamming distance on the permutation group $\mathcal{P}(n)$ of $[n] = \{1, \ldots, n\}$ is denoted by d_{H}. The symbol tr is used for the matrix trace, and I_n denotes the identity matrix of dimension n. The cardinality of a set is denoted by $|\cdot|$, and $\mathbb{I}$ denotes indicator function.

Conventions. We often refer to a permutation via the underlying map π and the associated matrix $\Pi = (\pi_{ij})$ in an interchangeable fashion, and accordingly $\mathcal{P}(n)$ and subsets thereof may refer to both maps and matrices. Asterisked symbols such as π^*, θ^*, σ_* etc. refer to ground truth parameters; non-asterisked symbols such as π, θ, σ etc. refer to generic elements of the associated parameter spaces.

2 Approach

Our approach will be presented as follows: we start with a brief motivation, followed by a more formal systematic introduction, and conclude with technical details pertaining to computation and model fitting.

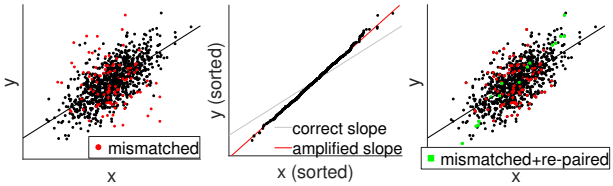

Figure 1: L: Samples from the model $y_i = x_{\pi^*(i)}\beta^* + \epsilon_i$, $i \in [n]$, $n = 1{,}000$, with 10% random mismatch. M: Re-paired data $(x_{\widehat{\pi}_{\mathrm{ML}}(i)}, y_i)_{i=1}^n = (x_{(i)}, y_{(i)})_{i=1}^n$ and corresponding amplified slope $\widehat{\beta}_{\mathrm{ML}}$. R: Re-paired data $(x_{\widehat{\pi}(i)}, y_i)_{i=1}^n$ based on the proposed Hamming prior.

2.1 Motivating examples

Consider the simple linear regression setup $y_i = x_{\pi^*(i)}\beta^* + \sigma_*\epsilon_i$, where x_i and ϵ_i are independent standard normal random variables, $1 \le i \le n$, and π^* permutes 10% of the indices uniformly at random. Suppose that the sign of β^* is known to be positive. Then the ML estimator of π^* (or equivalently, the MAP estimator under a uniform prior over $\mathcal{P}(n)$) is given by the permutation $\widehat{\pi}_{\mathrm{ML}}$ that matches the corresponding order statistics in $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$:

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

As shown in Figure 1, $\widehat{\pi}_{\mathrm{ML}}$ performs rather poorly. The scatterplot of the matching of corresponding order statistics is far from that of the underlying correct pairing. In fact, $\widehat{\pi}_{\mathrm{ML}}$ is associated with massive overfitting. Specifically, let

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

denote the resulting ML estimators of β^* and σ_*^2, respectively. It is straightforward to show that

$$\widehat{\sigma}^2_{\mathrm{ML}} \to 0, \quad n^{-1}\textstyle\sum_{i=1}^n (x_i\beta^* - x_i\widehat{\beta}_{\mathrm{ML}})^2 \to \sigma_*^2 \tag{2}$$

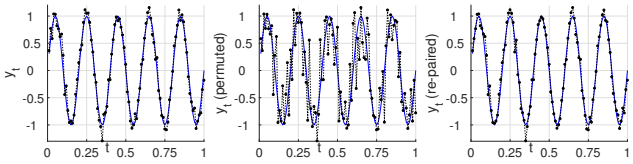

Figure 2: L: Data $y_{t_i} = \sin(10\pi t_i) + 0.1\epsilon_i$, $\epsilon_i \sim N(0,1)$, $i \in [n]$. M: Locally permuted data $(t_i, y_{t_{\pi^*(i)}})$ R: Corrected data $(t_i, y_{t_{\widehat{\pi}\circ\pi^*(i)}})$ based on the proposed prior.

in probability as $n \to \infty$. This is alarming since it implies that the least squares fit absorbs all the noise.

Figure 1 shows that the ML estimator is too aggressive in forming "corrected" pairs $(x_{\widehat{\pi}_{\mathrm{ML}}(i)}, y_i)$ given that only 10% of the observations are actually mismatched, and among those 10%, only a fraction contributes substantial mismatch that exceeds the noise inherent in the problem. Sparsity of π^* is often a reasonable assumption in post-linkage data analysis (Chambers and Diniz da Silva, 2020; Slawski and Ben-David, 2019), where sparsity here means that set of mismatches $\{i \in [n] : \pi^*(i) \neq i\}$ has significantly smaller cardinality than n. Given an upper bound on the number of mismatches, say k, it is appropriate to consider the following constrained ML estimator of π^*:

$$\max_{\pi\in\mathcal{P}(n)} \sum_{i=1}^n x_{\pi(i)} y_i \quad \text{subject to } d_{\mathsf{H}}(\pi, \mathsf{id}) \leq k, \quad (3)$$

where id is the identity map on $[n]$ and $d_{\mathsf{H}}(\pi,\pi') = \sum_{i=1}^n \mathbb{I}(\pi(i) \neq \pi'(i))$ denotes the Hamming distance on $\mathcal{P}(n)$. To the best of our knowledge, there is no efficient algorithm for computing the maximizer directly. However, there exists a Lagrangian multiplier $\gamma > 0$ such that (3) is equivalent to the optimization problem

$$\begin{aligned} &\max_{\pi\in\mathcal{P}(n)} \big\{\textstyle\sum_{i=1}^n x_{\pi(i)} y_i - \gamma d_{\mathsf{H}}(\pi, \mathsf{id})\big\} \quad (4)\\ &= \max_{\pi\in\mathcal{P}(n)} \big\{ \textstyle\sum_{i=1}^n \sum_{j=1}^n \pi_{ij}(x_j y_i - \gamma \mathbb{I}(i \neq j))\big\}, \end{aligned}$$

which is a linear assignment problem with cost matrix $C = \big(\gamma\mathbb{I}(i \neq j) - x_j y_i\big)$, which is computationally tractable according to the discussion following (7) below. The right panel of Figure 1 highlights the improvement that can be achieved by the resulting estimator which here only makes a small number of re-pairings of (x, y) capturing pairs that correspond to massive mismatch error in the left panel.

Figure 2 illustrates scenarios in which π^* is not sparse (with a mismatch rate exceeding 80%), but constrained to be a "local shuffling" in the sense that $\max_{i\in[n]} |\pi^*(i) - i| \leq r$, i.e., the corresponding permutation matrix is a band matrix with bandwidth at most r. This scenario is particularly relevant when the data is recorded sequentially (e.g., over different time points) or across a spatial domain endowed with a notion of distance, and it is known that π^* can only mix up the order of data inside a specific time window or within a local neighborhood. There are numerous applications in which π^* is locally constrained such as genome sequencing (Abid et al., 2017), signal processing (Balakhrisnan, 1962; Abbasi et al., 2021), or computer vision (Ma et al., 2021b). The illustrative example in Figure 2 can be thought of a regression problem in which the signal is a sine with known frequency but unknown (positive) amplitude β^*, i.e., $y_{t_i} = \beta^* \sin(10\pi t_i) + 0.1\epsilon_i$, $1 \leq i \leq n$ (left panel). However, the observed data is of the form $(y_{t_{\pi^*(i)}})_{i=1}^n$ for some unknown (local) permutation π^* (middle panel). If β^* is known to be positive, then the (unconstrained) ML estimator $\widehat{\pi}_{\mathrm{ML}}$ of π^* matches the order statistics $\{\mu_{(i)}\}_{i=1}^n$ and $\{y_{(i)}\}_{i=1}^n$, where $\mu_i = \sin(10\pi t_i)$, $i \in [n]$. In order to improve over the ML estimator using the prior knowledge of local shuffling, we impose the constraint that the alternative estimator $\widehat{\pi}$ does not pair any indices more than $r = 3$ apart. This estimator can be obtained as solution of the optimization problem

$$\begin{aligned} &\max_{\pi\in\mathcal{P}(n)} \textstyle\sum_{i=1}^n \mu_i y_{\pi(i)} \ \text{ subject to } |\pi(i) - i| \leq r,\ i \in [n]\\ &= \max_{\pi\in\mathcal{P}(n)} \big\{ \textstyle\sum_{i=1}^n \sum_{j=1}^n \pi_{ij}(\mu_i y_j - c_{ij})\big\}, \quad (5) \end{aligned}$$

where $c_{ij} = 0$ if $|i - j| \leq r$ and $c_{ij} = +\infty$ otherwise. As in (4), the problem in (5) side is a linear assignment problem and hence computationally tractable, and corresponds to MAP estimation under the family of priors considered below. The corrected, i.e., repaired data $(t_i, y_{\widehat{\pi}\circ\pi^*(i)})_{i=1}^n$ based on this approach are depicted in Figure 2 (R).

2.2 Exponential family prior on $\mathcal{P}(n)$

The priors discussed in the two examples of the previous subsection can be understood as specific instances of a more general family of prior distributions over $\mathcal{P}(n)$. Specifically, we consider the family of priors

$$p(\pi) \propto \exp(\gamma \operatorname{tr}(\Pi^\top M)), \quad M \in \mathbb{R}^{n\times n}, \quad \gamma > 0, \quad (6)$$

where $\gamma > 0$ is the concentration parameter, and the matrix M (which is not required to have any specific properties) defines the mode(s) of the distribution $\operatorname{argmax}_{\Pi\in\mathcal{P}(n)}\langle\Pi, M\rangle$, where $\langle\cdot,\cdot\rangle$ here represents the trace inner product on the space of matrices. In the same vein, the mode(s) of the distribution correspond to the set of matrices closest to M with respect to the same norm. Moreover, the distribution specified by (6) is of exponential family form with respect to the trace inner product (Wainwright and Jordan, 2008).

Linear Assignment Problems (LAPs). Linear assignment problems are a class of optimization problems for computing optimal one-to-one matchings of two sets of items (Burkard et al., 2009). LAPs are of the form

$$\min_{\Pi\in\mathcal{P}(n)}\langle\Pi, C\rangle, \quad (7)$$

where C is a given cost matrix. By the Birkhoff-von Neumann theorem (Ziegler, 1995), the minimum over $\mathcal{P}(n)$ can be replaced by the minimum over $\mathcal{DS}(n)$, the set of n-by-n

doubly stochastic matrices. Therefore, (7) reduces to a linear program in n^2 variables and $n^2 + 2n$ linear constraints. This implies that computing a mode of (6) is tractable.

Specific examples. Below, we consider a few examples of interest that are special cases of (6).

(I) **Hamming prior**. Consider the choice $M = I_n$. For any $\Pi \in \mathcal{P}(n)$, we then have $\langle \Pi, I_n \rangle = n - \sum_{i=1}^n \mathbb{I}(\Pi_{ii} \neq 1) = n - d_{\mathsf{H}}(\pi, \mathsf{id})$, where $d_{\mathsf{H}}(\pi, \pi') = \sum_{i=1}^n \mathbb{I}(\pi(i) \neq \pi'(i))$ denotes the Hamming distance on $\mathcal{P}(n)$. Since n does not depend on π, Eq. (6) reduces to

$$p(\pi) \propto \exp(-\gamma\, d_{\mathsf{H}}(\pi, \mathsf{id})), \tag{8}$$

which appeared in the first example of the preceding section, cf. (4), in which the goal was to take into account the underlying low rate of mismatches. The prior (8) is a specific Mallows' prior $p(\pi) \propto \exp(-\gamma d(\pi, \pi_0))$ for a base permutation π_0 and a metric d on $\mathcal{P}(n)$ (Mallows, 1957).

(II) **Local shuffling prior**. As in the second example in §2.1, suppose we want to have the prior p place most of its mass on permutations that move indices within small windows, i.e., $|\pi(i) - i|$ tends to be small. This can be achieved by choosing the entries of M in (6) as $M_{ij} = -\phi(|i - j|)$ for a non-decreasing function ϕ. The choice $\phi(u) = 0$ if $|u| \leq r$ for a positive integer r and $\phi(u) = +\infty$ otherwise yields the approach (5) that underlies the example in Fig. 2 above.

(III) **Block prior**. In record linkage, it is common that $\Pi^* = \mathrm{bdiag}(\Pi_1^*, \ldots, \Pi_B^*)$ is block diagonal with known block composition. For example, suppose that gender, ethnicity, and age group are used as matching variables, and that these three categorical variables are free of errors. In this case, mismatches can only involve pairs (i, j) falling into the same block corresponding to a specific combination of the above variables. Such known block structure can be encoded via prior (6) by choosing $M_{ij} = -\infty$ if (i, j) are not contained in the same block and $M_{ij} = 0$ otherwise. This corresponds to a uniform prior for each block, i.e., $p(\pi) = \prod_{b=1}^B p(\pi_b)$ with $p(\pi_b) \propto 1$, $b \in [B]$. The prior for each block does not have to be uniform; e.g., a Hamming prior as in Example (I) above can be used instead. Moreover, the hard block constraint can be relaxed.

(IV) **Lahiri-Larsen prior**. In their seminal work on linear regression in the presence of mismatch errors, Lahiri and Larsen (2005) and Chambers (2009) assume that $\pi^* \sim p(\pi)$ whose expectation $\mathbf{E}_{p(\pi)}[\Pi^*] = Q$ is known to the (post-linkage) data analyst. In the framework considered here, it is convenient to use $M = Q$ in (6). An example for Q is the so-called exchangeable linkage model (Chambers, 2009; Zhang and Tuoto, 2021) with $Q = (1-\alpha)I_n + \frac{\alpha}{n-1}\mathbf{1}_n\mathbf{1}_n^\top$. In this case, the resulting prior is equivalent to the Hamming prior considered in Example (I). More complex priors are obtained depending on the structure of Q.

2.3 Integration in Shuffled Data Problems

We now outline how the above prior can be integrated into generic shuffled data problems. The proposed Monte-Carlo EM (Wei and Tanner, 1990) framework builds upon the paper by Wu (1998) that has been rediscovered in the more recent work Abid and Zou (2018). The MC-EM scheme in Wu (1998) was further developed in Gutman et al. (2013) based on the concept of data augmentation (Tanner and Wong, 1987). None of Wu (1998); Abid and Zou (2018); Gutman et al. (2013) consider informative priors for π.

Conditional & Integrated Likelihood. Suppose we are given data $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ potentially contaminated by mismatch error. Let $p(\mathbf{x}_j, \mathbf{y}_i; \theta)$ be the likelihood (depending on a parameter θ) for the pair $(\mathbf{x}_j, \mathbf{y}_i)$, $(i, j) \in [n]^2$. The likelihood for θ resulting from $\mathcal{D}$ conditional on a specific $\pi \in \mathcal{P}(n)$ is given by

$$L(\theta|\pi) = \prod_{i=1}^n p(\mathbf{x}_{\pi(i)}, \mathbf{y}_i; \theta) = \prod_{i=1}^n \prod_{j=1}^n p(\mathbf{x}_j, \mathbf{y}_i; \theta)^{\pi_{ij}} \tag{9}$$

Conjugacy. It is worth noting that under (9), the posterior $p(\pi|\mathcal{D}, \theta)$ is a member of the family of distributions specified by $p(\pi)$, i.e., the latter is a conjugate prior. This follows from the observation that

$$\begin{aligned} p(\pi|\mathcal{D}, \theta) &\propto p(\mathcal{D}|\pi, \theta) \cdot p(\pi) = L(\theta|\pi) \cdot p(\pi) \\ &= \exp\left(\textstyle\sum_{i=1}^n \sum_{j=1}^n \pi_{ij} \left[\log\{p(\mathbf{x}_j, \mathbf{y}_i|\theta)\} + \gamma M_{ij}\right]\right) \\ &= \exp(\mathrm{tr}(\Pi^\top M_{\mathcal{D},\theta,\gamma})), \end{aligned} \tag{10}$$

with $M_{\mathcal{D},\theta,\gamma} = \big(\log(p(\mathbf{x}_j, \mathbf{y}_i|\theta)) + \gamma M_{ij}\big)$.

The (conditional) likelihood (9) can be maximized with respect to both θ and π as, e.g., in Pananjady et al. (2018); Abid et al. (2017); Slawski and Ben-David (2019). Alternatively, θ is considered as the quantity of primary interest, which suggests the *integrated likelihood*

$$L(\theta) = \mathbf{E}_\pi[L(\theta|\pi)] = \textstyle\sum_{\pi \in \mathcal{P}(n)} L(\theta|\pi) p(\pi). \tag{11}$$

As seen in §2.1, maximizing the conditional likelihood is prone to overfitting, prompting a need for regularization. The use of the integrated likelihood mitigates that problem at best slightly, but not substantially (cf. supplement for details), hence regularization remains relevant.

MC-EM scheme. The Expectation-Maximization (EM) algorithm (Dempster et al., 1977) is an established heuristic for minimizing the negative log-likelihood $\ell(\theta) = -\log L(\theta)$ corresponding to (11) via a sequence of surrogates $\{\widetilde{\ell}^{(t)}(\theta; \theta^{(t)})\}_{t \geq 0}$ that are minimized successively:

$$-\log \mathbf{E}_\pi[L(\theta|\pi)] \quad \rightsquigarrow \quad \mathbf{E}_{\pi|\mathcal{D},\theta^{(t)}}[-\log L(\theta|\pi)],$$

where $\widetilde{\ell}^{(t)}(\theta; \theta^{(t)}) := \mathbf{E}_{\pi|\mathcal{D},\theta^{(t)}}[-\log L(\theta|\pi)]$ is equal to

$$\textstyle\sum_{i=1}^n \sum_{j=1}^n \mathbf{E}[\pi_{ij}|\mathcal{D}, \theta^{(t)}]\{-\log p(\mathbf{x}_j, \mathbf{y}_i; \theta)\}, \tag{12}$$

the so-called *expected complete data* negative log-likelihood. The surrogates $\{\widetilde{\ell}^{(t)}(\cdot\,;\theta^{(t)})\}$ tend to be easier to minimize since they are linear combinations of standard likelihood terms as encountered for fixed and known π. Surrogates are updated according to

$$\theta^{(t+1)} \leftarrow \operatorname*{argmin}_{\theta} \widetilde{\ell}^{(t)}(\theta;\theta^{(t)}) \;\; \rightsquigarrow \;\; \widetilde{\ell}^{(t+1)}(\theta;\theta^{(t+1)}).$$

Here, the main challenge of this scheme is the E-step, i.e, the calculation of the expectation on the right term in (12). For any pair (i,j), we have

$$\begin{aligned}&\mathbf{E}[\pi_{ij}|\mathcal{D},\theta^{(t)}] \propto \textstyle\sum_{\pi\in\mathcal{P}(n)} p(\mathcal{D}|\pi,\theta^{(t)})p(\pi)\pi_{ij}\\ &= \textstyle\sum_{\pi\in\mathcal{P}(n)} \big\{\prod_{i=1}^n p(\mathbf{x}_{\pi(i)},\mathbf{y}_i;\theta^{(t)})\big\}\, p(\pi)\pi_{ij}.\end{aligned}$$

Since the summation over $\mathcal{P}(n)$ is not computationally tractable, the expectation needs to be approximated, e.g., via Monte Carlo simulation. Since for the same reason, the posterior $p(\pi|\mathcal{D},\theta^{(t)})$ is only accessible up to an unknown constant (cf. (10)), it is appropriate to resort to Markov Chain Monte Carlo (MCMC) (Gelman et al., 2013). The Metropolis-Hastings (MH) algorithm can be used to generate a Markov Chain $\{\pi^{(k)}\}_{k\geq 1}$ whose stationary distribution equals $p(\pi|\mathcal{D},\theta^{(t)})$. This yields the approximation

$$\widehat{\mathbf{E}}[\pi_{ij}|\mathcal{D},\theta^{(t)}] = \tfrac{1}{m-b}\textstyle\sum_{k=b+1}^m \pi_{ij}^{(k)},\;\; (i,j)\in[n]^2. \quad (13)$$

where b denotes the length of the "burn-in" period, and m denotes the total length of the Markov chain. Substituting (13) into (12) then yields what is known as MC-EM scheme, cf. Algorithm 1. Conveniently, there is a proposal distribution for the MH algorithm that is easy to work with, known as Fisher-Yates sampling: it generates a new permutation from the current one by swapping the assignments of a pair of indices (cf. Algorithm 2).

Initialization. The choice of the initial iterate $\theta^{(0)}$ can critically impact the quality of the solution that is returned by EM schemes given that the latter is a local strategy that finds a stationary point of a (in general) non-convex objective near the initial iterate. Several consistent initial estimators are known for regression setups depending on the structure of π (Lahiri and Larsen, 2005; Chambers and Diniz da Silva, 2020; Slawski et al., 2021; Peng et al., 2021), and those naturally lend themselves as initial iterate.

Careful initialization of the MH subroutine is important in order to ensure that $p(\pi|\mathcal{D},\theta^{(t)})$ is explored well given that $|\mathcal{P}(n)| = n!$ while the number of MCMC iterations m is limited. Fortunately, under the prior (6), computing the mode $\operatorname{argmax}_\pi p(\pi|\mathcal{D},\theta^{(t)})$ reduces to an LAP of the form (7) in virtue of (10). Initialization via the mode has the advantage that the Markov chain is started in a high density region. The hope is that the resulting iterates (generated according to a localized proposal distribution) will pick up most of the mass of $p(\pi|\mathcal{D},\theta^{(t)})$ so that (13) will well approximate the underlying expectation.

Algorithm 1 Monte Carlo EM (MC-EM) algorithm

Input: $\mathcal{D} = \{\{\mathbf{x}_i\}_{i=1}^n, \{\mathbf{y}_i\}_{i=1}^n\}$, γ, `EM_iter`
Initialize $\theta^{(0)} \leftarrow \widehat{\theta}_{\text{init}}$.
for $t = 0,\ldots,$ `EM_iter`
$\widehat{\pi}_{\text{init}} \leftarrow \operatorname{argmax}_{\pi\in\mathcal{P}(n)} p(\pi|\mathcal{D},\theta^{(t)})$.
$\widehat{\mathbf{E}}[\pi|\mathcal{D},\theta^{(t)}] \leftarrow \texttt{MH}(\mathcal{D},\theta^{(t)},\widehat{\pi}_{\text{init}},\gamma,m)$.
$\theta^{(t+1)} \leftarrow \min_\theta \Big\{\sum_{i,j=1}^n \widehat{\mathbf{E}}[\pi_{ij}|\mathcal{D},\theta^{(t)}]\{-\log p(\mathbf{x}_j,\mathbf{y}_i;\theta)\}\Big\}$.
$t \leftarrow t+1$; **end for**

Algorithm 2 `MH` sub-routine

Input: $\mathcal{D},\theta,\widehat{\pi}_{\text{init}},\gamma,m$; **Initialize** $\pi^{(0)} \leftarrow \widehat{\pi}_{\text{init}}$.
for $k = 0,\ldots,m$
Sample $(i,j)\in[n]^2$. $\widetilde{\pi}(i) \leftarrow \pi^{(k)}(j)$, $\widetilde{\pi}(j) = \pi^{(k)}(i)$.
$r(\widetilde{\pi},\pi^{(k)}) \leftarrow \min\Big\{\frac{p(\widetilde{\pi}|\mathcal{D},\theta;\gamma)}{p(\pi^{(k)}|\mathcal{D},\theta;\gamma)},1\Big\}$.
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** $\widehat{\mathbf{E}}[\pi|\mathcal{D},\theta]$ as in (13)

Reduction under exponential family likelihood. For a variety of exponential family models, the expected complete data negative log-likelihood (12) involves n instead of n^2 terms. Specifically, (12) will be $\sum_{i=1}^n r\{\mathbf{x}_i,\mathbf{y}_i,(\mathbf{E}[\Pi|\mathcal{D},\theta^{(t)}]^\top\mathbf{Y})_i;\theta)\}$ for a function r depending at most on $\{\mathbf{x}_i,\mathbf{y}_i,(\mathbf{E}[\Pi|\mathcal{D},\theta^{(t)}]^\top\mathbf{Y})_i\}_{i=1}^n$. Examples of interest are presented in the sequel.

(i) Least squares regression. In this case, we take $-\log p(\mathbf{x},y;\beta,\sigma^2) = (y-\mathbf{x}^\top\beta)^2/(2\sigma^2)$, which corresponds to the negative likelihood of a linear regression model with i.i.d. zero-mean Gaussian errors with variance σ^2. This yields the following expression for the expected complete data negative log-likelihood:

$$\begin{aligned}&(2\sigma^2)^{-1}\textstyle\sum_{i=1}^n\sum_{j=1}^n \mathbf{E}[\pi_{ij}|\mathcal{D},\theta^{(t)}](y_i-\mathbf{x}_j^\top\beta)^2\\ &= \sigma^{-2}\big\{\tfrac{1}{2}\|\mathbf{X}\beta\|_2^2 - \langle\mathbf{E}[\Pi|\mathcal{D},\theta^{(t)}]^\top\mathbf{Y},\mathbf{X}\beta\rangle\big\},\end{aligned}$$

which is identical to a least squares objective with design matrix $\mathbf{X}$ and response vector $\mathbf{E}[\Pi^\top|\mathcal{D},\theta^{(t)}]\mathbf{Y}$.

(ii) Generalized linear models. In this case, we have $-\log p(\mathbf{x},y;\beta,\phi) = \frac{\psi(\mathbf{x}^\top\beta)-y\mathbf{x}^\top\beta}{a(\phi)} + c(y,\phi)$, where a, ψ and c denote scale, cumulant, and partition function, respectively. Similar to above, one shows that

$$\begin{aligned}&\tfrac{1}{a(\phi)}\textstyle\sum_{i,j=1}^n \mathbf{E}[\pi_{ij}|\mathcal{D},\theta^{(t)}]\{\psi(\mathbf{x}_j^\top\beta)-y_i\mathbf{x}_j^\top\beta\} + c(\mathbf{Y},\phi)\\ =&\tfrac{1}{a(\phi)}\textstyle\sum_{i=1}^n\{\psi(\mathbf{x}_i^\top\beta) - (\mathbf{E}[\Pi|\mathcal{D},\theta^{(t)}]^\top\mathbf{Y})_i\,\mathbf{x}_i^\top\beta\} + c(\mathbf{Y},\phi).\end{aligned}$$

While the canonical link is assumed above, this is not necessary to achieve the reduction from n^2 to n terms.

(iii) Precision matrix estimation & multivariate normal data. Let $(\mathbf{x},\mathbf{y})\sim N(\mu_*,\Omega_*^{-1})$ with precision matrix Ω_*. Estimation of μ_* is unaffected by π^*; w.l.o.g. $\mu_* = 0$. Up to constants, $-\log p(\mathbf{x},\mathbf{y};\Omega) = -\log\det\Omega + \operatorname{tr}(\Omega\mathbf{z}\mathbf{z}^\top)$,

where $\mathbf{z} = [\mathbf{x}^\top \mathbf{y}^\top]^\top$; $\mathbf{z}\mathbf{z}^\top$ has diagonal blocks $\mathbf{x}\mathbf{x}^\top$, $\mathbf{y}\mathbf{y}^\top$ and off-diagonal blocks $\mathbf{x}\mathbf{y}^\top$, $\mathbf{y}\mathbf{x}^\top$. Thus $\mathrm{tr}(\Omega\mathbf{z}\mathbf{z}^\top) = \mathrm{tr}(\Omega_{\mathbf{xx}}\mathbf{x}\mathbf{x}^\top) + \mathrm{tr}(\Omega_{\mathbf{yy}}\mathbf{y}\mathbf{y}^\top) + 2\mathrm{tr}(\Omega_{\mathbf{yx}}\mathbf{x}\mathbf{y}^\top)$, where $\Omega_{\mathbf{xx}}$, $\Omega_{\mathbf{yy}}$ etc. are the corresponding sub-matrices of Ω. Hence,

$$\begin{aligned}&\textstyle\sum_{i=1}^n \sum_{j=1}^n \mathbf{E}[\pi_{ij}|\mathcal{D}, \Omega^{(t)}]\{-\log p(\mathbf{x}_j, \mathbf{y}_i; \Omega)\}\\ &= -n(\log\det\Omega + \mathrm{tr}(\Omega \mathbf{S}_{\mathbf{E}[\Pi|\mathcal{D},\Omega^{(t)}]}),\end{aligned}$$

where $\mathbf{S}_{\mathbf{E}[\Pi|\mathcal{D},\Omega^{(t)}]}$ consists of blocks $\mathbf{X}^\top\mathbf{X}/n$, $\mathbf{Y}^\top\mathbf{Y}/n$, $\mathbf{X}^\top\,\mathbf{E}[\Pi|\mathcal{D},\Omega^{(t)}]^\top\mathbf{Y}/n$ and $\mathbf{Y}^\top\,\mathbf{E}[\Pi|\mathcal{D},\Omega^{(t)}]\mathbf{X}/n$.

Computational complexity of Algorithm 1. For exponential family models benefitting from the above reduction, updating θ in the M-step involves n terms, and is computationally equivalent to a standard estimation problem. Apart from the initialization of the Markov chain, the approximate E-step has complexity $O(m)$, where m denotes the length of the Markov chain. Computing the acceptance probability, updating $\pi^{(k)}$, and keeping track of $\widehat{\mathbf{E}}[\Pi|\mathcal{D},\theta^{(t)}]^\top\mathbf{Y}$ within Algorithm 2 can be done in time $O(1)$ since the proposal distribution only changes $\pi^{(k)}$ at two positions. However, m is recommended to be of the order $\Omega(n)$, heuristically justified by the fact that in the worst case a permutation is the product of $n-1$ transpositions.

Remarks. **(i)** Following Tanner and Wong (1987) and Gutman et al. (2013), the MC-EM approach can be converted into a fully Bayesian approach: given that MC-EM involves sampling from $p(\pi|\mathcal{D},\theta)$, one can as well sample from $p(\theta|\mathcal{D},\pi)$ in an alternating fashion, which yields a Gibbs sampler for the joint posterior $p(\theta,\pi|\mathcal{D})$. **(ii)** The framework herein is not limited to permutations: we may be given $\mathcal{D}_{\mathbf{x}} = \{\mathbf{x}_i\}_{i=1}^N$ and $\mathcal{D}_{\mathbf{y}} = \{\mathbf{y}_i\}_{i=1}^n$, $N > n$ (w.l.o.g.), and then consider maps $\pi : [n] \to [N]$ represented by a matrix $\Pi \in \{0,1\}^{n\times N}$ with unit row sums. Priors of the form (6) given a mode $M \in \mathbb{R}^{n\times N}$ as well as conditional and integrated likelihoods can be defined analogously to (9) and (11). **(iii)** We think of the sampling scheme as a template rather than an efficient approach; improving efficiency, e.g., along the lines of Zanella (2020); Grathwohl et al. (2021), is left for future work.

3 Theoretical Insights

In this section, we present some analysis of the proposed prior from a regularization perspective and provide guidance on the choice of the hyperparameter γ. Data-driven selection of γ based on Empirical and Hierarchical Bayes approaches are detailed in the supplement.

Hamming prior. Our first results concerns the MAP estimator of Π^* under the Hamming prior (8). Specifically, we consider the linear regression setup

$$\begin{aligned}&y_i = \mu_{\pi^*(i)}(\mathbf{x}) + \sigma_*\epsilon_i, \quad \mu_i(\mathbf{x}) = \mathbf{x}_i^\top\beta^*,\\ &\mathbf{x}_i \sim N(0, I_d), 1 \le i \le n, \{\epsilon_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} N(0,1),\end{aligned} \tag{14}$$

as considered in prior work on shuffled linear regression (Pananjady et al., 2018; Hsu et al., 2017). The theorem below considers the sparse setting in which the underlying π^* satisfies the constraint $d_{\mathsf{H}}(\pi^*, \mathsf{id}) \le k$ for k "small enough" as made precise below. For simplicity, it is assumed that β^* and σ_*^2 are known; various estimators for the regression parameter in this scenario have been proposed (Zhang and Li, 2020; Slawski et al., 2021; Peng et al., 2021).

Theorem 3.1. *Suppose the setting* (14) *holds true. Let* $\widehat{\Pi}$ *denote the resulting MAP estimator of* Π^* *with* $d_{\mathsf{H}}(\Pi^*, I_n) \le k$. *Then, if* $\gamma \ge 3\gamma_0$, *where* $\gamma_0 = 72\sqrt{\mathsf{SNR}}\log(en/k)$, *we have* $d_{\mathsf{H}}(\widehat{\Pi}, I_n) \le 2k$ *and*

$$\|(\widehat{\Pi} - \Pi^*)\boldsymbol{\mu}\|_2 \le \sigma_*\big(17\sqrt{k\log(en/3k)} + \sqrt{2\gamma}\big).$$

with probability at least $1 - 2/n$ *and* $1 - 3/n$, *respectively, where* $\boldsymbol{\mu} = (\mu_i(\mathbf{x}))_{i=1}^n$ *and* $\mathsf{SNR} = \|\beta^*\|_2^2/\sigma_*^2$.

Theorem 3.1 implies that if γ is chosen larger than the threshold γ_0, the MAP estimator $\widehat{\Pi}$ will be $2k$-sparse, which matches the sparsity of Π^* up to the factor 2. By the triangle inequality, $d_{\mathsf{H}}(\Pi^*, \widehat{\Pi}) \le 3k$, i.e., $\widehat{\Pi}$ and Π^* will be close in Hamming distance. Moreover, for values γ such that $3\gamma_0 \le \gamma \le C\gamma_0$ for $C > 3$, we obtain

$$\|(\widehat{\Pi} - \Pi^*)\boldsymbol{\mu}\|_2 \lesssim \sigma_*(\sqrt{k\log(en/k)} + \mathsf{SNR}^{1/4}\sqrt{k\log(en/k)}).$$

The dependence on the signal-to-noise ratio SNR is improved compared to the naive estimator $\widehat{\Pi}_0 = I_n$, whose error scales as $\sigma_*\sqrt{k\log(en/k)}\mathsf{SNR}^{1/2}$; for small SNR, one cannot hope for improvements over $\widehat{\Pi}_0$ in general. In light of the discussion in §2.1, the improvement over the maximum likelihood estimator $\widehat{\Pi}_{\mathrm{ML}}$ whose error scales as $\sigma_*\sqrt{n}$, is substantial as long as k is small relative to n.

The next result yields a lower bound on γ ensuring a prescribed level of sparsity k based on the prior only.

Proposition 3.2. *Suppose that* π *follows the Hamming prior* (8) *with parameter* γ. *Then for all* $2 \le k < n$

$$\mathbf{P}_{\pi\sim p}\left(d_{\mathsf{H}}(\pi, \mathsf{id}) \ge k\right) \begin{cases} \le \exp(-k\delta\log n), \\ \quad \textit{if } \gamma \ge (1+\delta)\log n,\ \delta > 0, \\ \ge c(k,n), \quad \textit{if } \gamma \le \log(n-k), \end{cases}$$

where $c(k,n) \to \frac{1}{4}\frac{!k}{k!}$ *as* $k \to \infty$, *with* $!k$ *denoting the number of derangements of* k *elements.*

Proposition 3.2 asserts that the hyperparameter γ of the prior (8) should be chosen proportional to $\log(n-k) \sim \log n$ as n gets large in order to ensure that the prior places essentially no mass outside the Hamming ball $\{\pi : d_{\mathsf{H}}(\pi, \mathsf{id}) \le k\}$. The threshold $\gamma \sim \log n$ is sharp in the sense that if $\gamma \le \log(n-k)$, the prior will place at least mass $\Omega(1) = \frac{1}{4}!k/k! \sim \frac{1}{4e}$ for not too small k outside that Hamming ball. The likelihood $p(\mathcal{D}|\pi)$ favors permutations with best fit to the given data, so that the posterior mass $\mathbf{P}_{\pi|\mathcal{D}}(\{\pi : d_{\mathsf{H}}(\pi, \mathsf{id}) \le k\})$ will be less than the prior mass. It is thus natural to consider $\gamma \sim \log n$ as initial point.

Local shuffling prior. The next statement addresses the scenario in Fig. 2 for Lipschitz functions. In particular, the level of penalty needed for the MAP solution to satisfy the condition $\max_i |\widehat{\pi}(i) - i| \leq r$ is provided.

Proposition 3.3. *Suppose that* $y_i = \mu_{\pi^*(i)} + \sigma_* \epsilon_i$, $i \in [n]$, *with* $\{\epsilon_i\}_{i=1}^n \overset{i.i.d.}{\sim} N(0,1)$, *where* $\mu_i = f^*(i/n)$, $i \in [n]$, *for a function* $f^* : [0,1] \to \mathbb{R}$ *that is L-Lipschitz. Let further the matrix* M *in the prior* (6) *have entries* $M_{ij} = \mathbb{I}(|i-j| > r)$, $1 \leq i, j \leq n$, *for a given bandwidth* $r = \max_{i \in [n]} |\pi^*(i) - i|$. *If* $\gamma > \frac{2L}{\sigma^*}\left(\sqrt{\log n} + \sqrt{2r}\right)$, *the resulting MAP estimator* $\widehat{\pi}$ *satisfies* $|\widehat{\pi}(i) - i| \leq r$, $i \in [n]$, *with probability at least* $1 - \exp(-(\sqrt{2}-1)^2/2) - 2/n$. *Under the same event,*

$$\frac{1}{n}\sum_{i=1}^n (\mu_{\widehat{\pi}(i)} - \mu_{\pi^*(i)})^2 \leq \frac{4\sqrt{2}\sigma^* L \cdot r}{n}.$$

In theory, the assertion of the above proposition can be achieved by setting $\gamma = \infty$. Established solvers of LAPs require the entries of the cost matrix to be finite. In addition, solver accuracy can degrade with the magnitude of the entries (Bernhard, 2021).

4 Experiments

In this section, we present the results of experiments conducted with synthetic and real data. In the supplement, we also show an example demonstrating the use of the data augmentation approach discussed at the end of §2.3 as an alternative to the Monte-Carlo EM scheme.

Synthetic data. We consider data generation according to the following three models ($1 \leq i \leq n$):

Linear Regression (`LR`): $y_i | \mathbf{x}_{\pi^*(i)} \sim N(\mathbf{x}_{\pi^*(i)}^\top \beta^*, \sigma_*^2)$,
Poisson (`GLM`): $y_i | \mathbf{x}_{\pi^*(i)} \sim \text{Poisson}(\exp(\mathbf{x}_{\pi^*(i)}^\top \beta^* + \beta_0^*))$,
Multi**V**ariate**N**ormal: $\mathbf{z}_i = (\mathbf{x}_{\pi^*(i)}, \mathbf{y}_i) \sim N(\mu_*, \Omega_*^{-1})$.

The $\{\mathbf{x}_i\}_{i=1}^n$ and $\{\epsilon_i\}_{i=1}^n$ are i.i.d random samples from the $N(0, I_d)$ and $N(0,1)$ distributions, respectively. The regression parameter β^* is sampled uniformly from the sphere $\{\beta \in \mathbb{R}^d : \|\beta\|_2 = 3\}$, and $\beta_0^* \sim N(0,1)$. For `MVN`, we let $\mu_* = 0$ and $\Omega_*^{-1} = (1-\rho_*) I_{p+q} + \rho_* \mathbf{1}_{p+q} \mathbf{1}_{p+q}^\top$. Finally, π^* is a permutation selected uniformly at random from one of the following constraint sets:

$$\text{(i) } k\text{-}\textbf{Sparse}\text{: } \left\{\pi \in \mathcal{P}(n) : \sum_{i=1}^n \mathbb{I}(\pi(i) \neq i) \leq k\right\}, \tag{15}$$

$$\text{(ii) } r\text{-}\textbf{Banded}\text{: } \left\{\pi \in \mathcal{P}(n) : \max_{1 \leq i \leq n} |\pi(i) - i| \leq r\right\},$$

$$\text{(iii) } k\text{-}\textbf{SparseBlock}\text{: } \left\{\pi \in \mathcal{P}(B) : \sum_{i=1}^n \mathbb{I}(\pi(i) \neq i) \leq k \cdot B\right\},$$

where $\mathcal{P}(B)$ denotes the set of block-structured permutations corresponding to B blocks of uniform size n/B, i.e., $\{1, \ldots, n/B\}, \ldots, \{(B-1)(n/B)+1, \ldots, n\}$. Note that in (i), k refers to the number total of mismatches, whereas in (iii) k refers to the number of mismatches per block. We fix $n = 1,000$, $d = 20$, $\sigma_* = 1$, $\rho_* = 0.8$, $p = q = 5$, $B = 50$. The mismatch rates k/n and $k \cdot B/n$ in (15)(i) and (15)(iii), respectively, are varied between 0.2 and 0.5 in steps of 0.05, and the bandwidth r in (15)(ii) is varied between 3 and 10. For each setup and each value of k and r, 100 independent replications are performed. The following approaches are compared:

(I) **naive**. Standard maximum likelihood estimation as used for parameter estimation in the absence of mismatches, which corresponds to fixing $\pi = \mathsf{id}$ as the identity.

(II) **oracle**. The unknown permutation π^* is considered as known, and standard maximum likelihood estimation is used for parameter estimation after fixing $\pi = \pi^*$.

(III) **robust** [for setting k-**Sparse** only]. For setup `LR`, the regression parameter is estimated on the `robustfit` function in (MATLAB, 2019). For setup `GLM`, the regression parameter is estimated based on the robust GLM estimation method (Wang et al., 2020) that uses observation-specific dummy variables and penalization. For setup `MVN`, Ω_*^{-1} is estimated according to the `robustcov` function in `MATLAB` which implements the minimum covariance determinant estimator (Rousseeuw and Driessen, 1999).

(IV) **EM, EMH, EML, EMB**. Algorithm 1 using uniform, Hamming, local shuffling, and block-Hamming prior, respectively, which reflect the constraint sets (i) to (iii) in (15). The EM iterations are initialized by setting $\pi = \mathsf{id}$, and the number of EM iterations is limited to 400. The number of MCMC iterations per EM iteration is set to 8k, half of which are counted towards the "burn-in period". We note that a modified MH algorithm is used for EML (cf. supplement); for EMB, the MH scheme in Algorithm 2 is applied blockwise. For the **Sparse** and **SparseBlock** settings, the hyperparameter γ is chosen based on Proposition 3.2, which suggests $\gamma \propto \log(n)$. For the **Banded** setting, the choice $\gamma = 1$ was found to yield good performance.

(V) **Lahiri & Larsen (LL), Chambers (C)** [for setting k-**SparseBlock** only]. The approaches described in Chambers (2009) and Lahiri and Larsen (2005) with the choice $Q = \mathbf{E}[\Pi^*] = I_B \otimes Q_0$, where $Q_0 = (1-\alpha_*) I_{n/B} + \alpha_* \mathbf{1}_{n/B} \mathbf{1}_{n/B}^\top$, $\alpha_* = (k \cdot B)/n$. For setup `MVN`, the LL approach amounts to estimation of Ω_* by the inverse of the modified sample covariance matrix $\widetilde{S}$ with blocks $\widetilde{S}_{\mathbf{xx}} = \mathbf{X}^\top \mathbf{X}/n$, $\widetilde{S}_{\mathbf{xy}} = \mathbf{X}^\top Q^\top \mathbf{Y}/n$, and $\widetilde{S}_{\mathbf{yy}} = \mathbf{Y}^\top \mathbf{Y}/n$.

(VI) **Averaging** [for setting r-**Banded** only]. We compute (componentwise) running averages of the $\{\mathbf{x}_i\}_{i=1}^n$ and $\{\mathbf{y}_i\}_{i=1}^n$ within sliding windows of size r, and estimate the parameters β^* or Ω_* from these local averages as usual (i.e., as if these were the original, uncontaminated data).

For better comparison across experimental configurations, we visualize the relative estimation error (REE) $\|\beta^{\text{est}} - \beta^*\|_2 / \|\beta^*\|_2$ and $\|\text{Corr}^{\text{est}} - \text{Corr}^*\|_{\mathsf{F}}$, where β^{est} and Corr^{est} are placeholders for the aforementioned estimators; "Corr" refers to the correlation matrix corresponding to Ω_*^{-1}. Selected results are shown in Figure 3, which displays aver-

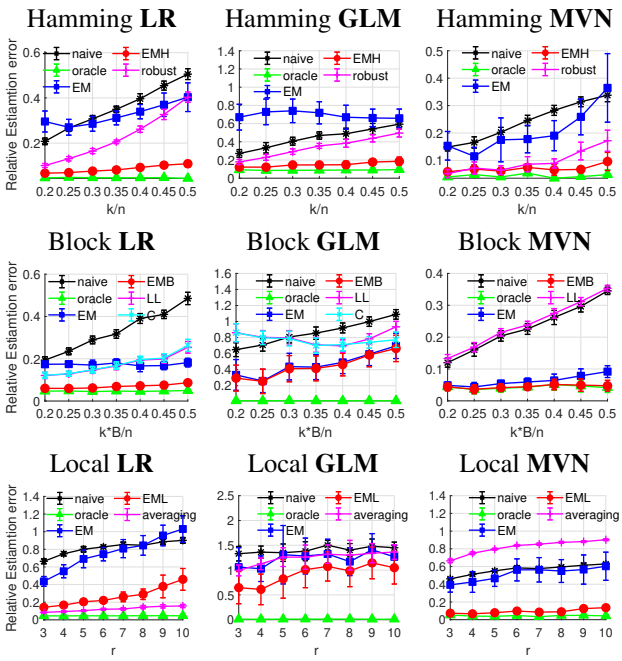

Figure 3: Results of the synthetic data experiments. The corresponding error bar represent $\pm 3 \times$ standard error. The figure captions represent the prior and model (bold) under consideration.

ages of the REE over 100 replications for each model and each permutation. Overall, it can be seen that EMB, EMH, EML achieve significant improvements over their unregularized counterpart and the other baselines.

data	model	prior
Italian survey data(ISD)	LR	hamming
El Nino data(END)	LR	block
CPS wages(CPS)	LR	hamming
Bike sharing data(BSD)	GLM	block
Flight Ticket Prices(FTP)	MVN	hamming
Supply Chain Management(SCM)	MVN	hamming
Beijing Air Quality data(BAQD)	MVN	local

Real data. We consider seven benchmark data sets for shuffled data problems. The data sets are preprocessed versions of their original counterparts (details on data processing can be found in the supplement). Even though the data sets themselves are real, the permutations that scramble the given matching pairs $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ are synthetic; for each data set, we consider 100 independent random permutations depending on the underlying setting. We consider the same list of competitors and associated configurations as for the synthetic data experiments. Asterisked ground truth parameters here refer to oracle estimates based on knowledge of π^*, and relative estimation error (REE) is defined accordingly in terms of those ground truth parameters.

Hamming & Block prior. As shown in Fig. 4, the proposed approach consistently improves over naive least squares once the fraction of mismatches exceeds 0.2, and yields substantial improvements as that fraction increases.

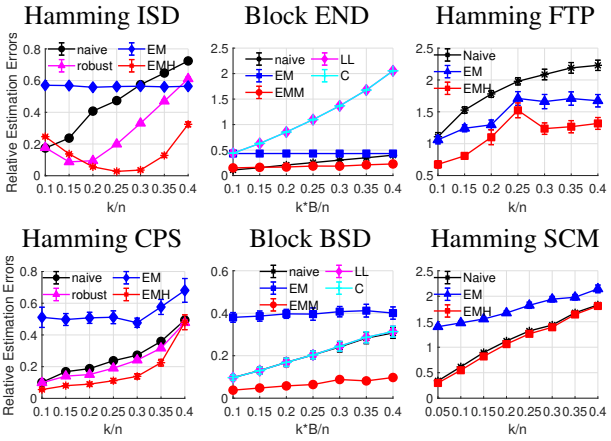

Figure 4: Results of the real data experiments. The corresponding error bars represent $\pm 3 \times$ standard error. The figure captions represent the prior and data set (cf. table above) under consideration.

The regularized EM approach based on the priors in EMB, EMH, EML noticeably reduces error induced by shuffling.

Local shuffling prior. As shown in Table 1, the EM approach with local shuffling prior achieves significant error reductions compared to the naive approach and the EM approach without regularization.

Table 1: Results of the real data experiment (Beijing Air Quality Data) with local shuffling permutation. Each number in the table is the average REE over 100 replications.

Methods	naive	EM	EML
$\|\text{Corr}^{\text{est}} - \text{Corr}^*\|_F$	0.76	1.97	0.34
standard error	0.0012	0.0111	0.0010

5 Conclusion

In this paper, we have proposed a framework for regularized estimation in shuffled data problems by means of an exponential family prior on the permutation group. The exponential family form is convenient for computational purposes yet sufficiently rich to incorporate common forms of prior knowledge. The proposed prior is not tailored to specific data analysis problems, but can be applied generically. The results in this paper confirm the importance of regularization in shuffled data problems given the inherent danger of overfitting already with little noise. While the approach covers various constraints that can be imposed on the underlying permutation, it is not exhaustive. For example, suppose we have information on the cycles of the permutation (numbers or lengths). Such information cannot be expressed in terms of index pairs, and hence requires a different paradigm. Kondor et al. (2007); Huang et al. (2009) use Fourier analysis on the permutation group (Diaconis, 1988) to facilitate learning of permutations, and it is an interesting direction of future research to study how that approach can be leveraged for the type of shuffled data problems considered in the present paper.

Acknowledgements

We thank four reviewers and a meta-reviewer for their constructive comments and suggestions.

Zhenbang Wang and Martin Slawski were partially supported by grants NSF-CCF-1849876 and NSF-SES-2120318.

References

A. Abbasi, A. Tasissa, and S. Aeron. R-local sensing: A novel graph matching approach for multiview unlabeled sensing under local permutations. *IEEE Open Journal of Signal Processing*, 2:309–317, 2021.

A. Abid and J. Zou. Stochastic EM for shuffled linear regression. In *Allerton Conference on Communication, Control, and Computing*, pages 470–477, 2018.

A. Abid, A. Poon, and J. Zou. Linear regression with shuffled labels. arXiv:1705.01342, 2017.

M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government printing office, 1964.

F. Balabdaoui, C. R. Doss, and C. Durot. Unlinked monotone regression. *Journal of Machine Learning Research*, 22:172, 2021.

A. Balakhrisnan. On the problem of time jitter in sampling. *IRE Transactions on Information Theory*, 8:226–236, 1962.

F. Bernhard. Fast Linear Assignment Problem using Auction Algorithm, October 2021.

O. Binette and R. Steorts. (Almost) All of Entity Resolution. arXiv:2008.04443, 2020.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

R. Burkard, M. Dell'Amico, and S. Martello. *Assignment Problems: Revised Reprint*. SIAM, 2009.

A. Carpentier and T. Schlüter. Learning relationships between data obtained independently. In *Proceedings of the International Conference on Artifical Intelligence and Statistics (AISTATS)*, pages 658–666, 2016.

R. Chambers. Regression analysis of probability-linked data. Technical report, Statistics New Zealand, 2009.

R. Chambers and A. Diniz da Silva. Improved secondary analysis of linked data: a framework and an illustration. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(1):37–59, 2020.

M. DeGroot, P. Feder, and P. Goel. Matchmaking. *The Annals of Mathematical Statistics*, 42:578–593, 1971.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

P. Diaconis. Group representations in probability and statistics. *Lecture Notes-Monograph Series*, 11, 1988.

J. Domingo-Ferrer and K. Muralidhar. New directions in anonymization: permutation paradigm, verifiability by subjects and intruders, transparency to users. *Information Sciences*, 337:11–24, 2016.

N. Flammarion, C. Mao, and P. Rigollet. Optimal Rates of Statistical Seriation. *Bernoulli*, 25:623–653, 2019.

M. A. Fligner and J. S. Verducci. Distance based ranking models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 48(3):359–369, 1986.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. CRC press, 2013.

D. Gordon, J. Katz, M. Liang, and J. Xu. Spreading the privacy blanket: Differentially oblivious shuffling for differential privacy. *Cryptology ePrint Archive*, 2021.

W. Grathwohl, K. Swersky, M. Hashemi, D. Duvenaud, and C. Maddison. Oops I took a gradient: Scalable sampling for discrete distributions. In *International Conference on Machine Learning (ICML)*, pages 3831–3841, 2021.

E. Grave, A. Joulin, and Q. Berthet. Unsupervised alignment of embeddings with wasserstein procrustes. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1880–1890, 2019.

R. Gutman, C. Afendulis, and A. Zaslavsky. A Bayesian Procedure for File Linking to Analyze End-of-Life Medical Costs. *Journal of the American Statistical Association*, 108:34–47, 2013.

D. Hsu, K. Shi, and X. Sun. Linear regression without correspondence. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1531–1540, 2017.

J. Huang, C. Guestrin, and L. Guibas. Fourier-Theoretic Probabilistic Inference over Permutations. *Journal of Machine Learning Research*, 10(5), 2009.

A. Klami. Variational bayesian matching. In *Asian Conference on Machine Learning (ACML)*, pages 205–220, 2012.

R. Kondor, A. Howard, and T. Jebara. Multi-object tracking with representations of the symmetric group. In *Artificial Intelligence and Statistics*, pages 211–218, 2007.

P. Lahiri and M. D. Larsen. Regression analysis with linked data. *Journal of the American Statistical Association*, 100(469):222–230, 2005.

F. Li, K. Fujiwara, F. Okura, and Y. Matsushita. Generalized shuffled linear regression. In *Proceedings of the*

IEEE/CVF International Conference on Computer Vision, pages 6474–6483, 2021.

R. Ma, T. Cai, and H. Li. Optimal permutation recovery in permuted monotone matrix model. *Journal of the American Statistical Association*, 116:1358–1372, 2020.

R. Ma, T. Cai, and H. Li. Optimal estimation of bacterial growth rates based on a permuted monotone matrix. *Biometrika*, 108(3):693–708, 2021a.

Y. Ma, P. Boufounos, H. Mansour, and S. Aeron. Multi-view Sensing with Unknown Permutations: an Optimal Transport Approach. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1440–1444, 2021b.

C. Mallows. Non-null ranking models. *Biometrika*, 44: 114–130, 1957.

MATLAB. *Version 9.7 (R2019b)*. The MathWorks Inc., Natick, Massachusetts, 2019.

R. Mazumder and H. Wang. Linear Regression with Mismatched Data: A Provably Optimal Local Search Algorithm. In *Integer Programming and Combinatorial Optimization: 22nd International Conference, IPCO 2021, Atlanta, GA, USA, May 19–21, 2021, Proceedings 22*, pages 443–457. Springer, 2021.

B. McVeigh, B. Spahn, and J. Murray. Scaling Bayesian Probabilistic Record Linkage with Post-Hoc Blocking: An Application to the California Great Registers. arXiv:1905.05337, 2019.

A. Pananjady, M. Wainwright, and T. Cortade. Denoising linear models with permuted data. arXiv:1704.07461, 2017.

A. Pananjady, M. Wainwright, and T. Cortade. Linear regression with shuffled data: Statistical and computational limits of permutation recovery. *IEEE Transactions on Information Theory*, 3826–3300, 2018.

L. Peng and M. Tsakiris. Linear Regression without Correspondences via Concave Minimization. *IEEE Signal Processing Letters*, 27:1580–1584, 2020.

L. Peng, B. Wang, and M. Tsakiris. Homomorphic sensing: Sparsity and noise. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8464–8475, 2021.

G. Peyré and M. Cuturi. Computational Optimal Transport: With Applications to Data Science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

Y. Plan and R. Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: a convex programming approach. *IEEE Transactions on Information Theory*, 59:482–494, 2013.

P. Rigollet and J. Weed. Uncoupled isotonic regression via minimum Wasserstein deconvolution. *Information and Inference*, 8:691–717, 2019.

P. J. Rousseeuw and K. V. Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999.

X. Shi, X. Lu, and T. Cai. Spherical Regresion under Mismatch Corruption with Application to Automated Knowledge Translation. *Journal of the American Statistical Association*, 116:1953–1964, 2021.

M. Slawski and E. Ben-David. Linear regression with sparsely permuted data. *Electronic Journal of Statistics*, 13:1–36, 2019.

M. Slawski, M. Rahmani, and P. Li. A Sparse Representation-Based Approach to Linear Regression with Partially Shuffled Labels. In *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI)*, 2019.

M. Slawski, E. Ben-David, and P. Li. Two-stage approach to multivariate linear regression with sparsely mismatched data. *Journal of Machine Learning Research*, 21(204):1–42, 2020.

M. Slawski, G. Diao, and E. Ben-David. A Pseudo-Likelihood Approach to Linear Regression with Partially Shuffled Data. *Journal of Computational and Graphical Statistics*, pages 1–13, 2021.

R. Steorts, R. Hall, and S. Fienberg. A Bayesian approach to graphical record linkage and deduplication. *Journal of the American Statistical Association*, 111(516):1660–1672, 2016.

A. Tancredi and B. Liseo. Regression analysis with linked data: problems and possible solutions. *Statistica*, 75(1): 19–35, 2015.

M. Tanner and W. Wong. The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association*, 82:528–540, 1987.

M. Tsakiris and L. Peng. Homomorphic sensing. In *International Conference on Machine Learning (ICML)*, pages 6335–6344, 2019.

M. Tsakiris, L. Peng, A. Conca, L. Kneip, Y. Shi, and H. Choi. An Algebraic-Geometric Approach to Shuffled Linear Regression. *IEEE Transactions on Information Theory*, 66:5130–5144, 2020.

J. Unnikrishnan, S. Haghighatshoar, and M. Vetterli. Unlabeled sensing with random linear measurements. *IEEE Transactions on Information Theory*, 64:3237–3253, 2018.

R. Vershynin. *High-Dimensional Probability. An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Z. Wang, E. Ben-David, and M. Slawski. Estimation in exponential family regression based on linked data contaminated by mismatch error. arXiv:2010.00181, 2020.

G. C. Wei and M. A. Tanner. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411):699–704, 1990.

Y. N. Wu. A note on broken sample problem. Technical report, Department of Statistics, University of Michigan, 1998.

G. Zanella. Informed proposals for local MCMC in discrete spaces. *Journal of the American Statistical Association*, 115(530):852–865, 2020.

H. Zhang and P. Li. Optimal estimator for unlabeled linear regression. In *Proceedings of the 37th International Conference on Machine Learning*, pages 11153–11162, 2020.

L.-C. Zhang and T. Tuoto. Linkage-data linear regression. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(2):522–547, 2021.

G. Ziegler. *Lectures on polytopes*. Graduate Texts in Mathematics. Springer, 1995. Updated 7th edition of first priting.

A Detailed derivations of the expected complete data negative log-likelihood for selected models

In this section, we provide detailed steps for deriving the expected complete data negative likelihood for least squares regression and precision matrix estimation for multivariate Normal data.

Recall that the expected complete data negative likelihood is given by

$$\mathbf{E}_{\pi|\mathcal{D},\theta^{(t)}}[-\log L(\theta|\pi)] = \sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{E}[\pi_{ij}|\mathcal{D},\theta^{(t)}]\{-\log p(\mathbf{x}_j,\mathbf{y}_i;\theta)\}$$

(i) Least squares regression. In this case, we take $-\log p(\mathbf{x},y;\beta,\sigma^2) = (y-\mathbf{x}^\top\beta)^2/(2\sigma^2)$, and we have that

$$\begin{aligned}
\frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{E}[\pi_{ij}|\mathcal{D},\theta^{(t)}](y_i-\mathbf{x}_j^\top\beta)^2 &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{E}[\pi_{ij}|\mathcal{D},\theta^{(t)}]\left\{\frac{1}{2}(\mathbf{x}_j^\top\beta)^2 - y_i\mathbf{x}_j^\top\beta\right\} + c \\
&= \frac{1}{2\sigma^2}\sum_{i=1}^{n}\mathbf{E}\left[\sum_{j=1}^{n}\pi_{ij}(\mathbf{x}_j^\top\beta)^2\Big|\mathcal{D},\theta^{(t)}\right] \\
&\quad - \frac{1}{\sigma^2}\sum_{j=1}^{n}\mathbf{x}_j^\top\beta\sum_{i=1}^{n}\mathbf{E}[\pi_{ij}|\mathcal{D},\theta^{(t)}]y_i \\
&= \frac{1}{\sigma^2}\left\{\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i^\top\beta)^2 - \sum_{i=1}^{n}\left\{\mathbf{E}[\Pi|\mathcal{D},\theta^{(t)}]^\top\mathbf{Y}\right\}_i(\mathbf{x}_i^\top\beta)\right\} \\
&= \frac{1}{\sigma^2}\left\{\frac{1}{2}\|\mathbf{X}\beta\|_2^2 - \langle\mathbf{E}[\Pi|\mathcal{D},\theta^{(t)}]^\top\mathbf{Y},\mathbf{X}\beta\rangle\right\}.
\end{aligned}$$

(ii) Precision matrix estimation for multivariate normal data. In this case, $-\log p(\mathbf{x},\mathbf{y};\Omega) = -\log\det\Omega + \mathrm{tr}(\Omega\mathbf{z}\mathbf{z}^\top)$, where $\mathbf{z} = [\mathbf{x}^\top\ \mathbf{y}^\top]^\top$; $\mathbf{z}\mathbf{z}^\top$ has diagonal blocks $\mathbf{x}\mathbf{x}^\top$, $\mathbf{y}\mathbf{y}^\top$ and off-diagonal blocks $\mathbf{x}\mathbf{y}^\top$, $\mathbf{y}\mathbf{x}^\top$.

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{E}[\pi_{ij}|\mathcal{D},\Omega^{(t)}]\{-\log p(\mathbf{x}_j,\mathbf{y}_i;\Omega)\} \\
&= -n\log\det\Omega + \mathrm{tr}\left(\Omega_{\mathbf{xx}}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top\right) + \mathrm{tr}\left(\Omega_{\mathbf{yy}}\sum_{i=1}^{n}\mathbf{y}_i\mathbf{y}_i^\top\right) \\
&\quad + \mathrm{tr}\left(\Omega_{\mathbf{yx}}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{E}[\pi_{ij}|\mathcal{D},\Omega^{(t)}]\mathbf{x}_j\mathbf{y}_i^\top\right) \\
&= -n\log\det\Omega + \mathrm{tr}(\Omega_{\mathbf{xx}}\mathbf{X}^\top\mathbf{X}) + \mathrm{tr}(\Omega_{\mathbf{xx}}\mathbf{Y}^\top\mathbf{Y}) + \mathrm{tr}(\Omega_{\mathbf{yx}}\mathbf{X}^\top\mathbf{E}[\Pi|\mathcal{D},\Omega^{(t)}]^\top\mathbf{Y}) \\
&= -n(\log\det\Omega + \mathrm{tr}(\Omega\mathbf{S}_{\mathbf{E}[\Pi|\mathcal{D},\Omega^{(t)}]}).
\end{aligned}$$

B Derivation of the claims in Eq. (2)

Let $\mu_i = x_i\beta^*$, $1 \le i \le n$, and let P_μ^n and P_y^n be the probability measures with mass $1/n$ at the $\{\mu\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$, respectively. Then the squared Wasserstein-2 distance W_2^2 between P_μ^n and P_y^n is given by (cf. Peyré and Cuturi, 2019)

$$\begin{aligned}\mathsf{W}_2^2(P_\mu^n, P_y^n) = \min_{\pi\in\mathcal{P}(n)} \frac{1}{n}\sum_{i=1}^n \{y_i - \mu_{\pi(i)}\}^2 &= \frac{1}{n}\sum_{i=1}^n y_i^2 + \frac{1}{n}\sum_{i=1}^n \mu_i^2 - \frac{2}{n}\left\{\max_{\pi\in\mathcal{P}(n)}\sum_{i=1}^n x_i y_i\right\}\beta^* \\ &= \frac{1}{n}\sum_{i=1}^n y_i^2 + \frac{1}{n}\sum_{i=1}^n \mu_i^2 - \frac{2}{n}\sum_{i=1}^n x_{(i)}y_{(i)}\beta^* \end{aligned} \tag{S.1}$$

We have that $\mathsf{W}_2^2(P_\mu^n, P_y^n) \to \mathsf{W}_2^2(P_\mu, P_y) = (\sqrt{(\beta^*)^2+\sigma_*^2} - \beta^*)^2$ in probability as $n\to\infty$, where P_μ and P_y denote the Gaussian measures $N(0,(\beta^*)^2)$ and $N(0,(\beta^*)^2+\sigma_*^2)$, respectively (Peyré and Cuturi, 2019, Remark 2.31). At the same time, $\frac{1}{n}\sum_{i=1}^n y_i^2 \to (\beta^*)^2+\sigma_*^2$ and $\frac{1}{n}\sum_{i=1}^n \mu_i^2 \to (\beta^*)^2$ in probability as $n\to\infty$. Substitution into (S.1) and invoking Slutsky's theorem, we have that

$$\frac{1}{n}\sum_{i=1}^n x_{(i)}y_{(i)} \to \sqrt{(\beta^*)^2+\sigma_*^2}.$$

in probability as $n\to\infty$. The first result in (2) then follows immediately from Slutsky's Theorem and the fact that $n^{-1}\sum_{i=1}^n x_i^2 \to 1$, and observe that the third result in (2) is obtained as a direct consequence with the same reasoning. The result $\widehat{\sigma}^2_{\text{ML}} \to 0$ is obtained by expanding the square

$$\frac{1}{n}\sum_{i=1}^n y_i^2 - \frac{2}{n}\sum_{i=1}^n y_i x_i \widehat{\beta}_{\text{ML}} + \frac{1}{n}\sum_{i=1}^n x_i^2(\widehat{\beta}_{\text{ML}})^2,$$

and analyzing each of the terms accordingly.

C Proof of Theorem 3.1

In light of relation (10), straightforward manipulations and omission of terms not depending on Π show that the MAP estimator $\widehat{\Pi}$ is the minimizer of the optimization problem

$$\min_{\Pi\in\mathcal{P}(n)} \left\{ -\langle \mathbf{Y}, \Pi\boldsymbol{\mu}\rangle + \sigma_*^2\gamma d_{\mathsf{H}}(\Pi, I_n)\right\}. \tag{S.2}$$

Since $\widehat{\Pi}$ minimizes (S.2), the following basic inequality holds true:

$$-\langle \mathbf{Y}, \widehat{\Pi}\boldsymbol{\mu}\rangle + \sigma_*^2\gamma d_{\mathsf{H}}(\widehat{\Pi}, I_n) \le -\langle \mathbf{Y}, \Pi^*\boldsymbol{\mu}\rangle + \sigma_*^2\gamma d_{\mathsf{H}}(\Pi^*, I_n) \tag{S.3}$$

Decomposing $\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\xi}$ with $\boldsymbol{\xi} = \sigma_*\Pi^*\boldsymbol{\epsilon}$ and re-arranging terms in the above inequality yields that

$$\langle \Pi^*\boldsymbol{\mu}, (\Pi^* - \widehat{\Pi})\boldsymbol{\mu}\rangle - \langle \boldsymbol{\xi}, (\widehat{\Pi} - \Pi^*)\boldsymbol{\mu}\rangle + \sigma_*^2\gamma d_{\mathsf{H}}(\widehat{\Pi}, I_n) \le \sigma_*^2\gamma k,$$

where we have substituted $d_{\mathsf{H}}(\Pi^*, I_n) = k$. By the Cauchy-Schwarz inequality, $\langle \Pi^*\boldsymbol{\mu}, \widehat{\Pi}\boldsymbol{\mu}\rangle \le \|\Pi^*\boldsymbol{\mu}\|_2^2$, which implies that the first term in the previous inequality is non-negative. This in turn yields that

$$-\langle \boldsymbol{\xi}, (\widehat{\Pi} - \Pi^*)\boldsymbol{\mu}\rangle + \sigma_*^2\gamma d_{\mathsf{H}}(\widehat{\Pi}, I_n) \le \sigma_*^2\gamma k. \tag{S.4}$$

In the sequel, we will derive a probabilistic lower bound on the first term on the left hand side.

For any integer $1 \le s \le n$, consider the event $\mathcal{E}_s = \{d_{\mathsf{H}}(\widehat{\Pi}, I_n) \le s\}$, and let $\overline{v} = \frac{(\widehat{\Pi}-\Pi^*)\boldsymbol{\mu}}{2\|\beta^*\|_2}$. We have that

$$\|\overline{v}\|_2 = \sup_{\|u\|_2\le 1}\left\langle u, \frac{(\widehat{\Pi} - \Pi^*)\boldsymbol{\mu}}{2\|\beta^*\|_2}\right\rangle = \sup_{\|u\|_2\le 1}\left\langle \frac{(\widehat{\Pi} - \Pi^*)u}{2}, \frac{\boldsymbol{\mu}}{\|\beta^*\|_2}\right\rangle$$

Observe that conditional on $\mathcal{E}_s$, for any vector $u \in \mathbb{R}^n$, $(\widehat{\Pi} - \Pi^*)u$ can have at most $m_s = s+k$ non-zero entries. Moreover, $\|(\widehat{\Pi} - \Pi^*)u\|_2 \leq 2\|u\|_2$. Finally, note that in light of the setting (14) under consideration, $\boldsymbol{\mu}/\|\beta^*\|_2 \sim N(0, I_n)$. It follows that for any $t > 0$

$$\mathbf{P}(\|\overline{v}\|_2 \geq t|\mathcal{E}_s) \leq \mathbf{P}\left(\sup_{u \in \mathcal{B}_0(m_s)} \langle u, g\rangle > t\right), \quad g \sim N(0, I_n). \tag{S.5}$$

where for any integer $1 \leq \ell \leq n$, $\mathcal{B}_0(\ell)$ here denotes the set of all unit vectors having at most ℓ non-zero entries. Denote by $w(\mathcal{B}(\ell)) = \mathbf{E}_{g\sim N(0,I_n)}[\sup_{u\in\mathcal{B}(\ell)}\langle u, g\rangle]$ the corresponding Gaussian width (Vershynin, 2018, §7.5). Choosing $t = w(\mathcal{B}(m_s)) + c_1\sqrt{2\log n}$ in (S.5) for $c_1 \geq 1$, standard tail bounds for the suprema of Gaussian processes (Boucheron et al., 2013, Theorem 5.8) yield

$$\mathbf{P}(\|\overline{v}\|_2 \geq w(\mathcal{B}_0(m_s)) + c_1\sqrt{2\log n}\ |\mathcal{E}_s) \leq n^{-c_1^2}.$$

Using that $w(\mathcal{B}_0(m_s)) \leq 4\sqrt{m_s \log(en/m_s)}$ (e.g., Plan and Vershynin, 2013, Lemma 2.3) and the fact that $s\log(en/s) \geq \log n$ for any $n \geq s \geq 1$, we have with $c_1 = \sqrt{2}$

$$\mathbf{P}(\|\overline{v}\|_2 \geq 6\sqrt{m_s \log(en/m_s)}\ |\mathcal{E}_s) \leq 1/n^2. \tag{S.6}$$

Combining this with the definition of $\overline{v}$ yields that

$$\mathbf{P}(\|(\widehat{\Pi} - \Pi^*)\boldsymbol{\mu}\|_2 \geq 12\|\beta^*\|_2\sqrt{m_s \log(en/m_s)}\ |\mathcal{E}_s) \leq 1/n^2.$$

Let now $\tau_s = 12\|\beta^*\|_2\sqrt{m_s\log(en/m_s)}$ and $\mathcal{F}_s = \{\|(\widehat{\Pi} - \Pi^*)\boldsymbol{\mu}\|_2 \leq \tau_s\}$, and note that

$$\mathbf{P}(\langle \boldsymbol{\xi}, (\widehat{\Pi} - \Pi^*)\boldsymbol{\mu}\rangle > t|\mathcal{F}_s \cap \mathcal{E}_s) \leq \mathbf{P}\left(\sup_{v\in\mathcal{B}_0(m_s)} \langle g, v\rangle \sigma\tau_s > t\right), \quad g \sim N(0, I_n).$$

Using the same argument as above, we choose $t = \sigma_*\tau_s\{w(\mathcal{B}_0(m_s)) + c_2\sqrt{2\log n}\}$ with $c_2 = \sqrt{2}$. Putting together the pieces as above, we obtain

$$\mathbf{P}(\langle \boldsymbol{\xi}, (\widehat{\Pi} - \Pi^*)\boldsymbol{\mu}\rangle > \underbrace{12\cdot 6}_{=72}\cdot\sigma_*\|\beta^*\|_2 m_s\log(en/m_s)|\mathcal{F}_s \cap \mathcal{E}_s) \leq 1/n^2.$$

Now let $\gamma_0 = 72\sqrt{\mathsf{SNR}}\log(en/k) \geq 72\sqrt{\mathsf{SNR}}\log(en/m_s)$ and define the event

$$\mathcal{G}_s = \left\{\frac{\left|\langle \boldsymbol{\xi}, (\widehat{\Pi} - \Pi^*)\boldsymbol{\mu}\rangle\right|}{\sigma_*^2} \leq \gamma_0 m_s\right\}$$

Observe that conditional on $\mathcal{E}_s \cap \mathcal{G}_s$, the earlier inequality (S.4) implies that (recalling that $m_s = s + k$)

$$-\gamma_0\sigma_*^2(s+k) + \sigma_*^2\gamma s \leq -\langle \boldsymbol{\xi}, (\widehat{\Pi} - \Pi^*)\boldsymbol{\mu}\rangle + \sigma_*^2\gamma d_{\mathsf{H}}(\widehat{\Pi}, I_n) \leq \sigma_*^2\gamma k.$$

Combination of the left and right hand sides and re-arranging terms implies the inequality

$$\gamma_0 \geq \gamma\frac{s-k}{s+k}$$

Now for any $s \geq 2k$, the right hand side is lower bounded by $(1/3)\gamma$. This in turn yields a contradiction whenever γ is chosen such that $\gamma > 3\gamma_0$. In order to conclude that $d_{\mathsf{H}}(\widehat{\Pi}, I_n) \leq 2k$ with the stated probability in that case, it remains to provide a corresponding lower bound on the probability of the event $\bigcup_{s=1}^n(\mathcal{E}_s \cap \mathcal{G}_s)$, i.e., at least one of the events $\{\mathcal{E}_s \cap \mathcal{G}_s\}_{s=1}^n$ occurs. Since the events inside the union are disjoint, we obtain that

$$\begin{aligned}\mathbf{P}\left(\bigcup_{s=1}^n(\mathcal{E}_s \cap \mathcal{G}_s)\right) = \sum_{s=1}^n \mathbf{P}(\mathcal{E}_s \cap \mathcal{G}_s) \geq \sum_{s=1}^n \mathbf{P}(\mathcal{E}_s \cap \mathcal{G}_s \cap \mathcal{F}_s) &= \sum_{s=1}^n \underbrace{\mathbf{P}(\mathcal{G}_s|\mathcal{E}_s\cap\mathcal{F}_s)}_{\geq 1-1/n^2}\underbrace{\mathbf{P}(\mathcal{F}_s|\mathcal{E}_s)}_{\geq 1-1/n^2}\mathbf{P}(\mathcal{E}_s) \\ &\geq \sum_{s=1}^n(1-2/n^2)\,\mathbf{P}(\mathcal{E}_s) \geq 1 - 2/n.\end{aligned}$$

In order to prove the second part of Theorem 3.1, we first invoke the following basic inequality equivalent to (S.3)

$$\|\widehat{\Pi}\boldsymbol{\mu} - \mathbf{Y}\|_2^2 + 2\sigma_*^2\gamma d_{\mathsf{H}}(\widehat{\Pi}, I_n) \leq \|\Pi^*\boldsymbol{\mu} - \mathbf{Y}\|_2^2 + 2\sigma_*^2\gamma d_{\mathsf{H}}(\Pi^*, I_n).$$

Expanding the squares and re-arranging yields conditional on $\bigcup_{s=1}^n(\mathcal{E}_s \cap \mathcal{G}_s)$

$$\begin{aligned}\|\widehat{\Pi}\boldsymbol{\mu} - \Pi^*\boldsymbol{\mu}\|_2^2 &\leq 2\langle \boldsymbol{\xi}, \widehat{\Pi}\boldsymbol{\mu} - \Pi^*\boldsymbol{\mu}\rangle + 2\sigma_*^2\gamma(k-s) \\ &\leq 2 \sup_{u \in \mathcal{B}_0(3k)} \langle \boldsymbol{\xi}, u\rangle \|\Pi^*\boldsymbol{\mu} - \widehat{\Pi}\boldsymbol{\mu}\|_2 + 2\sigma_*^2\gamma(k-s),\end{aligned}$$

where in the second inequality, we have used that if $\gamma > 3\gamma_0$, conditional on on $\bigcup_{s=1}^n(\mathcal{E}_s \cap \mathcal{G}_s)$, it holds that $d_{\mathsf{H}}(\widehat{\Pi}, I_n) \leq 2k$. The latter inequality is of the form

$$x^2 - 2bx - c \leq 0, \qquad x := \|\widehat{\Pi}\boldsymbol{\mu} - \Pi^*\boldsymbol{\mu}\|_2, \quad b := \sup_{u \in \mathcal{B}_0(3k)} \langle \boldsymbol{\xi}, u\rangle, \quad c = 2\sigma_*^2\gamma(k-s).$$

After elementary manipulations, we obtain the inequality $x \leq \sqrt{b^2 + c} + b \leq 2b + \sqrt{c}$, which translates to

$$\|\widehat{\Pi}\boldsymbol{\mu} - \Pi^*\boldsymbol{\mu}\|_2 \leq 2 \sup_{u \in \mathcal{B}_0(3k)} \langle \boldsymbol{\xi}, u\rangle + \sigma_*\sqrt{2\gamma} \leq \sigma_*\left(17\sqrt{k\log(en/3k)} + \sqrt{2\gamma}\right),$$

with probability at least $1 - 2/n - 1/n = 1 - 3/n$, where the term $\sup_{u \in \mathcal{B}_0(3k)}$ is controlled similarly to (S.6).

D Proof of Proposition 3.2

The probability mass function of (8) is given by (Fligner and Verducci, 1986):

$$p(\pi) = \frac{\exp(-\gamma d_{\mathsf{H}}(\pi, \mathsf{id}))}{\psi(\gamma)}, \qquad \psi(\gamma) := n!\exp(-\gamma n)\sum_{k=0}^n \frac{(\exp(\gamma)-1)^k}{k!}. \tag{S.7}$$

In the sequel, let us write $\{D(\pi) = d\}$ as a shortcut for the event $\{d_{\mathsf{H}}(\pi, \mathsf{id}) = d\}$. We then have

$$\mathbf{P}_{\pi\sim p}(D(\pi) \geq k) = \sum_{d=k}^n \frac{\exp(-\gamma d)}{\psi(\gamma)}\binom{n}{d}!d,$$

where $!d$ denotes the number of derangements of $\{1, \ldots, d\}$, i.e., the number of permutations τ of d objects such that $\tau(j) \neq j$ for all $1 \leq j \leq d$. Straightforward manipulations yield

$$\mathbf{P}(D(\pi) \geq k) = \sum_{d=k}^n \frac{\exp(-\gamma d)\frac{n!}{d!(n-d)!}!d}{n!\exp(-\gamma n)\sum_{\ell=0}^n \frac{(\exp(\gamma)-1)^\ell}{\ell!}} = \sum_{d=k}^n \frac{\exp(\gamma(n-d))}{(n-d)!\sum_{\ell=0}^n \frac{(\exp(\gamma)-1)^\ell}{\ell!}}\frac{!d}{d!}. \tag{S.8}$$

For $x \geq 0$ and integer $m \geq 1$, define the (upper) incomplete Gamma function and its "normalized" counterpart by

$$\Gamma(m, x) = \int_x^\infty t^{n-1}e^{-t}\,dt, \qquad \widetilde{\Gamma}(m, x) = \Gamma(m, x)/\Gamma(m),$$

where $\Gamma(m) = \Gamma(m, 0) = (m-1)!$ denotes the Gamma function. It can be shown that (Abramowitz and Stegun, 1964, §6.5)

$$\sum_{k=0}^m \frac{x^k}{k!} = e^x\widetilde{\Gamma}(m+1, x). \tag{S.9}$$

Further note that for $k \leq d \leq n$, we have that $\frac{!k}{k!} \leq !d/d! \leq !n/n! \leq e^{-1}$. Accordingly, for $\frac{!k}{k!} \leq c_0(n,k) \leq !n/n!$, we obtain the following for the right hand side of (S.8):

$$\begin{aligned}c_0(n,k)\sum_{d=k}^n \frac{\exp(\gamma(n-d))}{(n-d)!\sum_{\ell=0}^n \frac{(\exp(\gamma)-1)^\ell}{\ell!}} &= c_0(n,k)\sum_{i=0}^{n-k}\frac{(\exp(\gamma))^i}{i!}\frac{1}{\sum_{\ell=0}^n \frac{(\exp(\gamma)-1)^\ell}{\ell!}} \\ &= c_0(n,k)\frac{e^{\exp(\gamma)}\,\widetilde{\Gamma}(n-k+1, \exp(\gamma))}{e^{\exp(\gamma)-1}\,\widetilde{\Gamma}(n+1, \exp(\gamma)-1)} \\ &= c_0(n,k)\frac{\widetilde{\Gamma}(n-k+1, \exp(\gamma))}{\widetilde{\Gamma}(n+1, \exp(\gamma)-1)}.\end{aligned}$$

At this point, we consider the upper bound on the probability of interest as stated in the proposition. We have

$$\frac{\widetilde{\Gamma}(n-k+1,\exp(\gamma))}{\widetilde{\Gamma}(n+1,\exp(\gamma)-1)} = \frac{\int_{\exp(\gamma)}^{\infty} t^{n-k}e^{-t}dt}{\int_{\exp(\gamma)-1}^{\infty} t^{n}e^{-t}dt}\frac{\Gamma(n+1)}{\Gamma(n-k+1)} \leq \frac{\int_{\exp(\gamma)}^{\infty} t^{n-k}e^{-t}dt}{\int_{\exp(\gamma)}^{\infty} t^{n}e^{-t}dt}\frac{n!}{(n-k)!}$$
$$\leq \exp(-\gamma k)n^k = \exp(-\delta k \log n).$$

provided $\gamma \geq (1+\delta)\log n$, which concludes the proof of the upper bound.

Regarding the lower bound, observe that in view of relation (S.9), the ratio of normalized incomplete Gamma functions can be expressed via the ratio of CDFs of two independent Poisson random variables, that is

$$\frac{\widetilde{\Gamma}(n-k+1,\exp(\gamma))}{\widetilde{\Gamma}(n+1,\exp(\gamma)-1)} = \frac{\mathbf{P}(X_1 \leq n-k)}{\mathbf{P}(X_2 \leq n)},$$

where X_1 and X_2 are two independent Poisson random variables with parameters $\exp(\gamma)$ and $\exp(\gamma)-1$, respectively. Setting $\gamma = \log(n-k)$ yields that the right hand side is a function of the form $c_1(n,k)$ that is lower and upper bounded by $1/4$ and 1, respectively, as $n \to \infty$. Taking $c(k,n) = c_0(k,n) \cdot c_1(k,n)$ yields the assertion.

E Proof of Proposition 3.3

Similar to Eq. (S.3) in the proof of Proposition 3.1, we have the basic inequality

$$-\langle \mathbf{Y}, \widehat{\Pi}\boldsymbol{\mu}\rangle + \sigma_*^2\gamma \sum_{(i,j):|i-j|>r} \widehat{\Pi}_{ij} \leq -\langle \mathbf{Y}, \Pi^*\boldsymbol{\mu}\rangle. \tag{S.10}$$

In the sequel, we will show that under the stated conditions, the left hand side must exceed the right hand side unless $\widehat{\Pi}_{ij} = 0$ for all (i,j) such that $|i-j| > r$. Expanding $\mathbf{Y} = \Pi^*\boldsymbol{\mu} + \sigma_*\boldsymbol{\epsilon}$ and re-arranging terms yields the inequality

$$\sigma_*^2\gamma \sum_{(i,j):|i-j|>r} \widehat{\Pi}_{ij} \leq \sigma_*\langle \widehat{\Pi}\boldsymbol{\mu} - \Pi^*\boldsymbol{\mu}, \boldsymbol{\epsilon}\rangle = \sigma_* \sum_{i:|\widehat{\pi}(i)-i|>r} \epsilon_i(\mu_{\widehat{\pi}(i)} - \mu_{\pi^*(i)}) + \sigma_* \sum_{i:|\widehat{\pi}(i)-i|\leq r} \epsilon_i(\mu_{\widehat{\pi}(i)} - \mu_{\pi^*(i)}), \tag{S.11}$$

where we have used that $\|\Pi^*\boldsymbol{\mu}\|_2^2 - \langle \widehat{\Pi}\boldsymbol{\mu}, \Pi^*\boldsymbol{\mu}\rangle \geq 0$. For the second term on the right hand side, the triangle inequality yields that for all indices i that are summed over, we have $|\widehat{\pi}(i) - \pi^*(i)| \leq 2r$. Using the Cauchy-Schwarz inequality in combination with the Lipschitz property of the underlying function, we obtain that

$$\sum_{i:|\widehat{\pi}(i)-i|\leq r} \epsilon_i(\mu_{\widehat{\pi}(i)} - \mu_{\pi^*(i)}) \leq \left(\sum_{i:|\widehat{\pi}(i)-i|\leq r} \epsilon_i^2\right)^{1/2} \left(\sum_{i:|\widehat{\pi}(i)-i|\leq r} (\mu_{\widehat{\pi}(i)} - \mu_{\pi^*(i)})^2\right)^{1/2}$$
$$\leq \frac{2r \cdot L}{\sqrt{n}}\|\boldsymbol{\epsilon}\|_2 \tag{S.12}$$

By standard concentration results (e.g., Wainwright, 2019, § 2.3), the event $\mathcal{E}_1 = \{\|\boldsymbol{\epsilon}\|_2 \leq \sqrt{2n}\}$ holds with probability at least $1 - \exp((\sqrt{2}-1)^2/2)$. We now turn to the first term on the right hand side of (S.11). We have the upper bound

$$\sum_{i:|\widehat{\pi}(i)-i|>r} \epsilon_i(\mu_{\widehat{\pi}(i)} - \mu_{\pi^*(i)}) \leq L\|\boldsymbol{\epsilon}\|_\infty \cdot \mathsf{card}(\{i : |\widehat{\pi}(i) - i| > r\}), \tag{S.13}$$

where we have used that $\max_{i\neq j} |\mu_i - \mu_j| \leq L$. Standard concentration results yield that the event $\mathcal{E}_2 = \{\|\boldsymbol{\epsilon}\|_\infty \leq 2\sqrt{\log n}\}$ holds with probability at least $1 - 2/n$. Combining (S.12) and (S.13) yields that conditional on $\mathcal{E}_1$ and $\mathcal{E}_2$, the right hand side of (S.11) is upper bounded by

$$2\sigma_* L\left(\sqrt{\log n} \cdot \mathsf{card}(\{i : |\widehat{\pi}(i) - i| > r\}) + \sqrt{2}r\right).$$

At the same time, the left hand side of (S.11) evaluates as $\sigma_*^2\gamma \cdot \mathsf{card}(\{i : |\widehat{\pi}(i) - i| > r\})$. If the expression $\mathsf{card}(\ldots)$ is zero, the claim follows trivially. Otherwise, the condition $\gamma > \frac{2L(\sqrt{\log n}+\sqrt{2}r)}{\sigma_*}$ ensures that the left hand side exceeds the right hand side, which is a contradiction, and hence it must hold that $|\widehat{\pi}(i) - i| \leq r$, $1 \leq i \leq n$.

Observe that conditional on the event $\{|\widehat{\pi}(i) - i| \leq r,\ 1 \leq i \leq n\}$, the basic inequality (S.10) reduces to

$$-\langle \mathbf{Y}, \widehat{\Pi}\boldsymbol{\mu}\rangle \leq -\langle \mathbf{Y}, \Pi^*\boldsymbol{\mu}\rangle \iff \|\mathbf{Y} - \widehat{\Pi}\boldsymbol{\mu}\|_2^2 \leq \|\mathbf{Y} - \Pi^*\boldsymbol{\mu}\|_2^2$$

Substituting $\mathbf{Y} = \Pi^*\boldsymbol{\mu} + \sigma\boldsymbol{\epsilon}$ in the inequality of the right hand side and expanding squares, we obtain that conditional on $\{|\widehat{\pi}(i) - i| \leq r,\ 1 \leq i \leq n\}$ and $\mathcal{E}_1$

$$\|\widehat{\Pi}\boldsymbol{\mu} - \Pi^*\boldsymbol{\mu}\|_2^2 \leq 2\sigma_* \sum_{i=1}^{n} \epsilon_i(\mu_{\widehat{\pi}(i)} - \mu_{\pi^*(i)}) \leq 4\sigma_* \frac{r \cdot L}{\sqrt{n}} \|\boldsymbol{\epsilon}\|_2 \leq 4\sqrt{2}\sigma_*(r \cdot L),$$

with the same arguments as used for (S.11) and (S.12). Dividing both sides by n yields the assertion.

F Metropolis-Hastings scheme for local permutations

Algorithm 3 Monte Carlo EM Algorithm for local permutations

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

Sample $i \in [n]$ uniformly at random.

Sample j uniformly from $\{\max\{i - r, 1\}, ..., \min\{i + r, n\}\}$.

If $|\pi^{(k)}(i) - \pi^{(k)}(j)| > r$

`invalid-mcmc-steps` $\leftarrow$ `invalid-mcmc-steps` + 1; **continue**;

end If

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

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

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 $\widehat{\mathbf{E}}[\pi|\mathcal{D}, \theta]$ as in (13) with m replaced by $m -$ `invalid-mcmc-steps`.

G Additional information regarding real data analysis

In this section, we provide references of each data set and regression model used in the real data analysis. A summary of each data set is shown in Table S.1 below.

Table S.1: Overview of the data set used in the real data analysis. *refers to the total number of MCMC iterations after the burn-in period within each block.

data(abbreviation)	n	d/p	q	model	prior	MCMC Step
Italian survey data(ISD) (Slawski et al., 2021)	2011	2		LR	hamming	2k
El Nino data(END) (Slawski et al., 2021)	93935	5		LR	block	1.5k*
CPS wages(CPS) (Slawski et al., 2021)	534	11		LR	hamming	2k
Bike sharing data(BSD) (Wang et al., 2020)	731	16		GLM	block	1.5k*
Flight Ticket Prices(FTP) (Slawski et al., 2020)	335	30	6	MVN	hamming	2k
Supply Chain Management(SCM) (Slawski et al., 2020)	8966	35	16	MVN	hamming	4k
Beijing Air Quality data(BAQD) (Slawski et al., 2020)	9762	5	5	MVN	local shuffling	2k

H Integrated maximum likelihood estimator and overfitting

In this section, it is briefly explained that under a uniform prior $p(\pi) \propto 1$, the integrated maximum likelihood estimator based on (11) still exhibits a tendency to overfit, in a spirit similar to what is shown in §2.1 for the maximum likelihood estimator of π^*. To demonstrate this point, we consider the following setup:

$$\mathbf{Y}|\mathbf{X}, \Pi, \beta, \sigma_*^2 \sim N(\Pi\mathbf{X}\beta, \sigma_*^2), \qquad p(\Pi) \propto 1, \quad p(\beta) \propto 1, \tag{S.14}$$

and $\sigma_*^2 > 0$ fixed. The integrated likelihood corresponding to (11) is then given by

$$L(\beta) = p(\mathcal{D}|\beta) = \int p(\mathcal{D}|\pi, \beta) p(\pi|\beta)\, d\pi = \int \frac{p(\pi, \beta|\mathcal{D}) p(\mathcal{D})}{p(\pi|\beta) p(\beta)} p(\pi|\beta)\, d\pi \propto \int p(\beta|\pi, \mathcal{D}) p(\pi|\mathcal{D})\, d\pi.$$

Observe that under (S.14)

$$p(\beta|\pi, \mathcal{D}) \sim N((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^\top\Pi\mathbf{X}\beta^*, \sigma_*^2(\mathbf{X}^\top\mathbf{X})^{-1}), \qquad p(\pi|\mathcal{D}) \propto \exp\left(-\frac{\|\mathtt{P}_{\Pi\mathbf{X}}^{\perp}\mathbf{Y}\|_2^2}{2\sigma_*^2}\right),$$

where $\mathtt{P}_{\Pi\mathbf{X}}^{\perp}$ denotes the projection on the orthogonal complement of the column space of $\Pi\mathbf{X}$. Note that $p(\pi|\mathcal{D})$ is high for permutations achieving good fit (overfit) to the data, and the optimization problem $\max_\beta p(\mathcal{D}|\beta)$ will be dominated by the modes of those distributions $p(\beta|\pi, \mathcal{D})$ for which the corresponding weight $p(\pi|\mathcal{D})$ is high. In particular, in regimes with $\mathsf{SNR} = \|\beta^*\|_2^2/\sigma_*^2$ large, the maximizer of the integrated likelihood will not substantially differ from the estimator returned by $\max_{\pi,\beta} p(\mathcal{D}|\beta, \pi)$ (the MLE in §2.1), which is known to overfit dramatically.

I Data Augmentation example

In this paragraph we present a brief illustration of the proposed approach when used in conjunction with data augmentation, i.e., both the parameter and the permutation are sampled in an alternating fashion (cf. Remark **(i)** at the end of §2.3). For this purpose, we consider the Italian household survey discussed in Tancredi and Liseo (2015), see also Table S.1. This data set involves a simple linear regression problem in which the household income (in 1k Euros) in 2010 is is regressed on the same quantity in 2008, including an intercept term.

The process of file linkage subject to mismatch error involving the income data from the two years under consideration is simulated by generating a permutation π^* uniformly at random from the Hamming ball of radius k around the identity permutation, where $k/n = 0.4$.

We follow the paradigm of data augmentation in Tanner and Wong (1987) by considering π^* as missing data. This yields the following scheme that alternates between sampling of a permutation π and sampling of regression parameters $\beta = (\beta_0, \beta_1)$ and σ^2 given responses $\mathbf{Y} = (y_i)_{i=1}^n$ (income from 2010) and design matrix $\mathbf{X} = [\mathbf{1}_n\ (x_i)_{i=1}^n]$ (intercept and income from 2008).

$$\begin{aligned}
&\text{(I) Augmentation Step:} && \text{Sample } \pi^{(j)} \text{ from } p(\pi|\mathbf{Y}, \mathbf{X}, \beta^{(k-1)}, {\sigma^2}^{(k-1)}),\ \ j = 1, \dots, m,\\
&\text{(II) Posterior Step:} && \text{(a) Sample } \beta^{(k)} \text{ from } \frac{1}{m}\sum_{j=1}^m p(\beta|{\sigma^2}^{(k-1)}, \pi^{(j)}, \mathbf{Y}, \mathbf{X}),\\
& && \text{(b) Sample } {\sigma^2}^{(k)} \text{ from } \frac{1}{m}\sum_{j=1}^m p(\sigma^2|\beta^{(k)}, \pi^{(j)}, \mathbf{Y}, \mathbf{X}),
\end{aligned}$$

where m denotes the number of samples in the augmentation step, and k denotes the iteration counter for the parameters (β, σ^2).

Sampling in step (I) is implemented according to the MH procedure shown in Algorithm 2. Furthermore, under the usual non-informative prior distribution for (β, σ^2), i.e., $p(\beta, \sigma^2) \propto \sigma^{-2}$, the full conditional distributions appearing in step (II) are given by

$$\begin{aligned}
&\beta|\sigma^2, \Pi, \mathbf{Y}, \mathbf{X} \sim N(\widetilde{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}), \quad \sigma^2|\beta, \Pi, \mathbf{Y}, \mathbf{X} \sim \text{Inv-}\chi^2(n-d, s^2),\\
&\widetilde{\beta} := (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\Pi^\top\mathbf{Y}, \quad s^2 := \frac{1}{n-d}\|\mathbf{Y} - \Pi\mathbf{X}\widetilde{\beta}\|_2^2,
\end{aligned}$$

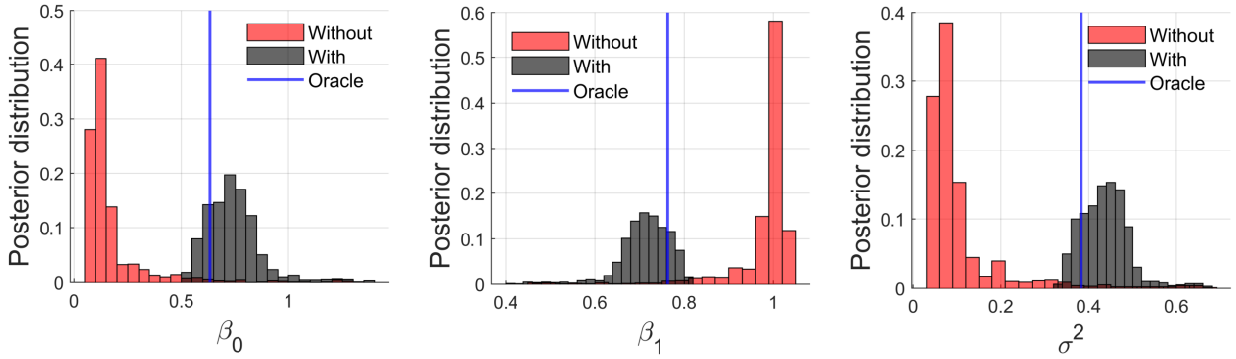

Figure S.1: Comparison between the posterior distributions for the parameters $(\beta_0, \beta_1, \sigma^2)$ for the Italian household survey data (i) with regularization based on the proposed Hamming prior for π ($\gamma = \log n$, grey histogram) and without regularization, i.e., uniform prior for π (red histogram). "Oracle" refers to the least squares estimator in the absence of mismatch error.

where Inv-$\chi^2(\nu, a^2)$ refers to the scaled inverse χ^2-distribution with scale parameter $a > 0$ and ν degrees of freedom (cf. §14 in Gelman et al. (2013) for more details on Bayesian inference for linear regression models).

For this illustration, we use $m = 100$, where each sequence $\{\pi^{(j)}\}$ is generated by uniform thinning of Markov chains of length $4,000$ generated by Algorithm 2. The number of samples $(\beta^{(k)}, {\sigma^2}^{(k)})$ obtained via the above scheme is taken as 1,000. The sampling procedure is initialized from step (II) with the identity permutation. We compare both the unregularized case with the uniform prior for π as well as the regularized case with the Hamming prior (8) ($\gamma = \log n$ in view of Proposition 3.2).

Figure S.1 confirms that the proposed approach achieves visible improvements over the unregularized approach which suffers from serious amplification bias affecting the slope parameter β_1 and serious underestimation of the error variance, as predicted by the brief analysis accompanying the first introductory example in §2.1.

J Empirical and Hierarchical Bayes approaches

In this section, we outline how the hyperparameter γ of the proposed prior on π can be selected based on Empirical and Hierarchical Bayes approaches. For simplicity, these approaches are presented for the linear regression model (14) in an exemplary fashion.

J.1 Hierarchical Bayes

Consider the following hierarchical model specification:

$$\begin{aligned} &p(\beta, \sigma^2) \propto \sigma^{-2}, \\ &p(\gamma) \propto \text{Gamma}(a, b), \\ &p(\pi|\gamma) \propto \exp(-\gamma\, d_{\text{H}}(\pi, \mathsf{id}))/\psi(\gamma), \\ &p(\mathbf{Y}|\mathbf{X}, \beta, \sigma^2, \pi, \gamma) \propto \exp\left(-\frac{\|\mathbf{Y} - \Pi\mathbf{X}\beta\|_2^2}{2\sigma^2}\right)., \end{aligned} \tag{S.15}$$

where $\psi(\gamma)$ denotes the terms in the normalization constant in the prior $p(\pi|\gamma)$ that depend on γ.

The data augmentation approach in the previous section can be extended as follows:

$$\begin{aligned}
&\text{(I) Augmentation Step:} \quad \text{Sample } \pi^{(j)} \text{ from } p(\pi|\mathbf{Y},\mathbf{X},\beta^{(k-1)},\sigma^{2\,(k-1)},\gamma^{(k-1)}),\ \ j=1,\ldots,m,\\
&\text{(II) Posterior Step:} \quad \text{(a) Sample } \beta^{(k)} \text{ from } \frac{1}{m}\sum_{j=1}^{m} p(\beta|\sigma^{2\,(k-1)},\gamma^{(k-1)},\pi^{(j)},\mathbf{Y},\mathbf{X}),\\
&\qquad\qquad\qquad\quad \text{(b) Sample } \sigma^{2\,(k)} \text{ from } \frac{1}{m}\sum_{j=1}^{m} p(\sigma^2|\beta^{(k)},\gamma^{(k-1)},\pi^{(j)},\mathbf{Y},\mathbf{X}),\\
&\qquad\qquad\qquad\quad \text{(c) Sample } \gamma^{(k)} \text{ from } \frac{1}{m}\sum_{j=1}^{m} p(\gamma|\beta^{(k)},\sigma^{2\,(k)},\pi^{(j)},\mathbf{Y},\mathbf{X}).
\end{aligned}$$

Compared to the data augmentation scheme in the previous section, the only addition is given by part (c), which requires sampling from the full conditional distribution of γ. Under (S.15), this full conditional can be expressed as follows.

$$\begin{aligned}
p(\gamma|\beta,\sigma^2,\pi,\mathbf{Y},\mathbf{X}) &\propto \frac{p(\beta,\sigma^2,\pi,\gamma,\mathbf{Y},\mathbf{X})}{p(\beta,\sigma^2,\pi,\mathbf{Y},\mathbf{X})}\\
&\propto \frac{p(\mathbf{Y}|\beta,\sigma^2,\pi,\gamma,\mathbf{X})\times p(\beta,\sigma^2)\times p(\pi|\gamma)\times p(\gamma)}{\int p(\mathbf{Y}|\beta,\sigma^2,\pi,\gamma,\mathbf{X})\times p(\beta,\sigma^2)\times p(\pi|\gamma)\times p(\gamma)d\gamma}\\
&\propto \frac{p(\pi|\gamma)\times p(\gamma)}{\int p(\pi|\gamma)\times p(\gamma)d\gamma}\\
&\propto \frac{\frac{\exp(-\gamma d_{\mathrm{H}}(\pi,\mathsf{id}))}{\psi(\gamma)}\times\frac{b^a}{\Gamma(a)}\gamma^{a-1}e^{-b\gamma}}{\int \frac{\exp(-\gamma d_{\mathrm{H}}(\pi,\mathsf{id}))}{\psi(\gamma)}\times\frac{b^a}{\Gamma(a)}\gamma^{a-1}e^{-b\gamma}d\gamma}
\end{aligned}$$

Since we cannot obtain a closed form expression for the full conditional γ (because of the term $\psi(\gamma)$), it is necessary to resort to rejection sampling, which is straightforward here since γ is one-dimensional.

J.2 Empirical Bayes

In the Empirical Bayes approach, γ is considered as the second parameter to be optimized in the M-step (in addition to the primary parameter of interest θ). This yields the following scheme.

Algorithm 4 Monte Carlo EM Empirical Bayes (MC-EM-EB) algorithm

Input: $\mathcal{D}=\{\{\mathbf{x}_i\}_{i=1}^n,\{\mathbf{y}_i\}_{i=1}^n\}$, γ, `EM_iter`
Initialize $\theta^{(0)}\leftarrow\widehat{\theta}_{\text{init}}$, $\gamma^{(0)}\leftarrow\widehat{\gamma}_{\text{init}}$.
for $t=0,\ldots,$ `EM_iter`
$\widehat{\pi}_{\text{init}}\leftarrow\operatorname{argmax}_{\pi\in\mathcal{P}(n)}p(\pi|\mathcal{D},\theta^{(t)},\gamma^{(t)})$.
$\widehat{\mathbf{E}}[\pi|\mathcal{D},\theta^{(t)},\gamma^{(t)}]\leftarrow\texttt{MH}(\mathcal{D},\theta^{(t)},\widehat{\pi}_{\text{init}},\gamma^{(t)},m)$.
$\theta^{(t+1)}\leftarrow\min_\theta\left\{\sum_{i,j=1}^n\widehat{\mathbf{E}}[\pi_{ij}|\mathcal{D},\theta^{(t)},\gamma^{(t)}]\{-\log p(\mathbf{x}_j,\mathbf{y}_i;\theta)\}\right\}$
$\gamma^{(t+1)}\leftarrow\min_\gamma\left\{\sum_{i,j=1}^n\widehat{\mathbf{E}}[\pi_{ij}|\mathcal{D},\theta^{(t)},\gamma^{(t)}]\{-\log p(\pi_{ij},\gamma)\}\right\}$
$t\leftarrow t+1$
end for

Note that the M-step decouples into two separate optimization problems since the likelihood does not depend on γ. For the same reason, the M-step update for θ remains unchanged compared to the case where γ is treated as fixed. In the following, we elaborate on the M-step update for γ. We have

$$\begin{aligned}
\widehat{\gamma} &= \operatorname*{argmin}_\gamma\left\{-\sum_{i,j=1}^n\widehat{\mathbf{E}}[\pi_{ij}|\mathcal{D},\theta^{(t+1)},\gamma^{(t)}]\log p(\pi_{ij},\gamma)\right\}\\
&= \operatorname*{argmin}_\gamma\left\{\log(\psi(\gamma))+\gamma\times(n-\operatorname{tr}(\widehat{\mathbf{E}}[\Pi|\mathcal{D},\theta^{(t)},\gamma^{(t)}]))\right\},
\end{aligned}\tag{S.16}$$

where we recall that

$$\psi(\gamma) := n! \exp(-\gamma n) \sum_{k=0}^{n} \frac{(\exp(\gamma) - 1)^k}{k!}. \tag{S.17}$$

In the sequel, we will demonstrate that the above minimization problem in γ has a rather simple (approximate) closed-form update. We will use the approximation

$$\begin{aligned}\exp(\exp(\gamma) - 1) &= \sum_{k=0}^{n} \frac{(\exp(\gamma) - 1)^k}{k!} + \sum_{k=n+1}^{\infty} \frac{(\exp(\gamma) - 1)^k}{k!} \\ &\approx \sum_{k=0}^{n} \frac{(\exp(\gamma) - 1)^k}{k!},\end{aligned}$$

assuming that n is sufficiently large. Accordingly, we have that

$$\begin{aligned}\widehat{\gamma} &= \operatorname{argmin}\{\log(\psi(\gamma)) + \gamma \times (n - \operatorname{tr}(\mathbf{E}[\Pi|\mathcal{D}, \theta^{(t)}, \gamma^{(t)}])\} \\ &\approx \operatorname{argmin}\{\exp(\gamma) - 1 - \gamma \cdot n + n \cdot \gamma - \gamma \cdot \operatorname{tr}(\mathbf{E}[\Pi|\mathcal{D}, \theta^{(t)}, \gamma^{(t)}])\} \\ &\approx \operatorname{argmin}\{\exp(\gamma) - 1 - \gamma \cdot \operatorname{tr}(\mathbf{E}[\Pi|\mathcal{D}, \theta^{(t)}, \gamma^{(t)}])\}\end{aligned}$$

Note that the terms inside the curly brackets are convex in γ. Therefore, taking the derivative with respect to γ and setting the result equal to zero, we have that

$$\widehat{\gamma} \approx \log(\operatorname{tr}(\mathbf{E}[\Pi|\mathcal{D}, \theta^{(t)}, \gamma^{(t)}])).$$