Statistics and Its Interface Volume 16 (2023) 379–396

Estimation in exponential family regression based on linked data contaminated by mismatch error*

Zhenbang Wang, Emanuel Ben-David, and Martin Slawski†

Identification of matching records in multiple files can be a challenging and error-prone task. Linkage error can considerably affect subsequent statistical analysis based on the resulting linked file. Several recent papers have studied post-linkage linear regression analysis with the response variable in one file and the covariates in a second file from the perspective of the "Broken Sample Problem" and "Permuted Data". In this paper, we present an extension of this line of research to exponential family response given the assumption of a small to moderate number of mismatches. A method based on observation-specific offsets to account for potential mismatches and ℓ_1-penalization is proposed, and its statistical properties are discussed. We also present sufficient conditions for the recovery of the correct correspondence between covariates and responses if the regression parameter is known. The proposed approach is compared to established baselines, namely the methods by Lahiri-Larsen and Chambers, both theoretically and empirically based on synthetic and real data. The results indicate that substantial improvements over those methods can be achieved even if only limited information about the linkage process is available.

AMS 2000 subject classifications: Primary 62F35, 62J07, 62J12, 62D10; secondary 62D99.

Keywords and phrases: Record linkage, Broken sample problem, Generalized linear models, Penalized estimation, Permutation.

1. INTRODUCTION

A tacit assumption in regression is that response-predictor pairs correspond to the same statistical unit. In practice, this assumption is often violated at least in part when different subsets of variables are collected in an asynchronous fashion and are subsequently combined into a single data set. Roughly speaking, the latter amounts to merging multiple data sets given agreement on a set of matching variables shared across those data sets; Figure 1 serves as an illustration.

*The first and the last author are partially supported by NSF award CRII: CIF: 1849876.

†Corresponding author.

This setting is well-studied in the field of *record linkage*, e.g., [1, 2, 3, 4]. A principal reason for its importance is considerable potential in reducing efforts for data collection in the situation that a research question of interest can be answered simply by combining existing databases. In a nutshell, probabilistic record linkage is concerned with the identification of matching records, i.e., pieces of information contained in different data sets belonging to the same statistical unit, given only approximate identifiers. The uncertainty associated with those introduces two types of linkage errors, (i) *missed matches* and (ii) *mismatches*. The former refers to two matching records not being linked, while (ii) refers to two records erroneously linked in the sense that those records belong to different statistical units. The present work is concerned with the consequences of mismatches on subsequent regression analysis. Since the work of Neter [5] in 1965, it is well known that mismatches can negatively affect model fit and parameter estimation, specifically bearing a strong potential for attenuation bias. Following up on [5], a variety of papers discuss strategies for bias correction in linear regression with mismatches in the response variable given specific information about the linkage process, including work by Scheuren and Winkler [6, 7], Lahiri and Larsen [8] and Chambers [9]. Generalizations of this line of research beyond one-to-one matching and linear models are discussed in [10] and [11] via estimating equation-based approaches.

A somewhat more direct paradigm for dealing with mismatches in the response variable originates in the "Broken Sample Problem", a term used in a series of papers by DeGroot and collaborators, e.g., [12, 13]. In brief, matching pairs are represented via an unknown permutation (cf. Figure 1); the latter is typically regarded as a nuisance parameter, but inference for it might be of interest for the purpose of pinpointing and correcting errors in the linkage process. While this paradigm was widely regarded as infeasible in the record literature due to the combinatorial nature and the associated computational challenges, it has recently experienced a surge of interest, fueled by advances in computing and an array of engineering and machine learning problems that can be cast as linear regression with unknown permutation [14, 15, 16, 17, 18, 19, 20, 21].

In this paper, we adopt the "Broken Sample" formulation for generalized linear regression models with exponential family response [22] as an alternative to the methods in Lahiri-Larsen [8, 10] and Chambers [9, 11] to account

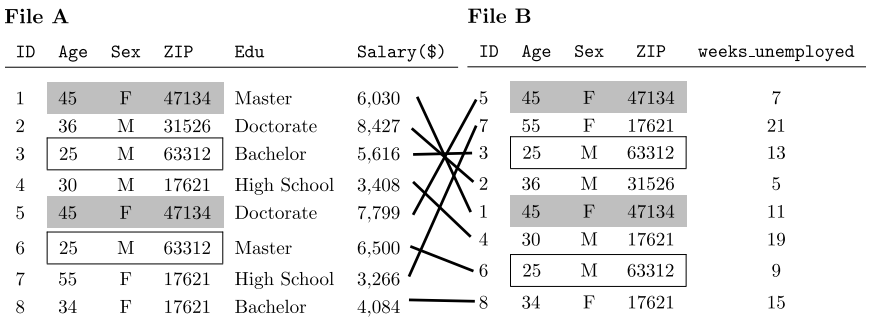

Figure 1. Schematic illustration of the setting studied in this paper. Two files A and B are linked to study how a response variable contained in one file (here: duration of unemployment in weeks) depends on covariates (here: education level and previous monthly salary) contained in another file. Linkage based on quasi-identifiers, here given by the triple (`Age`, `Gender`, `ZIP`), can be error-prone due to ambiguities (highlighted by boxes and grey color, respectively), and bear a potential for mismatch error affecting post-linkage regression. Matching records are connected by lines.

for potential mismatches in the response variable. The approach taken herein arises as a natural generalization of work by Slawski and Ben-David [23] for Gaussian response. An appealing property of the approach is that no information about the linkage process is required, in contrast to the methods put forth by Lahiri & Larsen and Chambers. This can be an important advantage if file linkage has been performed by a third party, and the data analyst is only provided the linked file, a situation that is not uncommon in practice given that file linkage is often based on sensitive personal information such as names or addresses.

In return, the proposed approach requires a limit on the fraction of mismatches. Empirical studies suggest that the critical threshold ranges between 20 and 30%, and a significant degradation in performance for larger fractions. While such a limit indicates room for improvement, parameter estimation in the setting of arbitrary mismatch contamination becomes in general infeasible from both computational and statistical viewpoints [24]. In practice, the underlying fraction of mismatches can be estimated based on the statistical modeling framework used for record linkage [25], or via clerical review of a random sample of linked pairs prior to secondary analysis.

Contributions. In this paper, we study the "Broken Sample" problem, also known as "Regression with Unknown Permutation" or "Regression with Shuffled Data", for generalized linear models. While the work herein is in the same spirit as prior work on this subject, the mechanism generating mismatches is not required to be a permutation; instead, we work with the sample-to-register linkage paradigm in [11, 26] in which the number of responses (observed sample) is allowed to be smaller than the number of predictors (contained in the register). As outlined above, primary interest concerns the regime of "sparsely mismatched" data in which the fraction of mismatches is subject to specific limits as elaborated in the sequel. The main technical contributions herein concern (1) restoration of the correctly matching records ("permutation recovery") for known regression parameter, (2) estimation of the regression parameter via computationally tractable schemes. In combination, (1) and (2) pave the way for "plug-in" estimation of the (generalized) permutation, thereby sidestepping the computational barriers that are associated with joint estimation. Specifically, (1) is shown to be reducible to sorting, and recovery results are derived under certain separability conditions naturally extending those for linear regression [27, 23]; regarding (2), we follow the route taken in [23] in which sparse mismatch contamination is captured by observation-specific dummy variables and ℓ_1-penalization whose statistical analysis is based on techniques in [28]. The proposed approach is compared to the Lahiri-Larsen-type method in [10] theoretically as well as empirically by means of simulations and the bike sharing data from [29].

Related work. There is a rapidly growing body of literature on regression with unknown permutation, starting from [14, 30]. The paper [30] presents necessary and sufficient conditions for permutation recovery for linear models with Gaussian design. Extensions to multivariate linear models are considered in [19, 27]. The papers [16, 24] show that consistent estimation of the regression parameter is impossible without substantial additional assumptions. Tsakiris and collaborators [20, 21] have studied important theoretical aspects such as well-posedness from an algebraic perspective, and have also put forth practical computational schemes such as a branch-and-bound algorithm (cf. also [31]) and concave maximization [32]. An approximate EM scheme with a Markov-Chain-Monte-Carlo (MCMC) approximation of the E-step is discussed in [17, 33]. The latter work in turn bears a relationship with the Bayesian approach in [34] and its implementation via Gibbs sampling. Approaches to linear and multivariate linear regression with sparsely mismatched data are studied in [18, 23, 35, 36].

In comparison, relatively few papers have considered regression with permuted data outside the standard linear model: examples include spherical regression [37], univariate isotonic regression and statistical seriation [38, 39, 40, 41, 42], and binary regression [43].

The method for regression parameter estimation considered herein has been studied in prior work to deal with generic contaminations in linear regression [44], logistic regression [45] and other generalized linear models [46]. Unlike the present paper, none of these works contain a rigorous statistical analysis.

Lastly, as indicated at the beginning of the introduction, there is a separate line of research focusing on parameter estimation under mismatch error in the response given at least a fair amount of knowledge about the linkage process. We refer to the surveys [10, 47]. Recovery of the underlying permutation is not considered in those works.

1.1 Problem statement

We consider a regression setup in which the response and the predictor variables are contained in two files $F_{\mathbf{x}} = \{\mathbf{x}_j\}_{j=1}^N \subset \mathbb{R}^d$ and $F_y = \{y_i\}_{i=1}^n \subset \mathbb{R}$, respectively, $n \leq N$. Record linkage yields a merged file $F_{\mathbf{x}\bowtie y} = \{(\mathbf{x}_{\ell_i}, y_i)\}_{i=1}^n$ with $\ell_i \neq \ell_j$ for $i \neq j$, i.e., resulting from complete and one-to-one linkage of $F_{\mathbf{x}}$ and F_y. Linkage is typically based on additional contextual information (matching variables); however, we generally assume herein that only the merged file $F_{\mathbf{x}\bowtie y}$ is given and no further information about the linkage process is available. The case $N = n$ applies to "sample-to-sample" linkage, with two pieces of information pertaining to the same set of entities collected via two separate samples; the case $N > n$ applies to "sample-to-register" linkage [26], which occurs, e.g., when linking population surveys conducted on a sample of individuals to an administrative database (see, e.g., [48] for an example of contemporary interest). Following [9, 11, 10, 26], we assume that each $\mathbf{x}_j$ is associated with a corresponding latent response variable y_j^*, $1 \leq j \leq N$, while $y_i = y^*_{\pi^*(i)}$, $1 \leq i \leq n$, for a map $\pi^* : \{1, \ldots, n\} \to \{1, \ldots, N\}$. For simplicity, we refer to π^* as "permutation" even if $N > n$. The linked pair $(\mathbf{x}_{\ell_i}, y_i)$ is called a *mismatch* if $\pi^*(i) \neq \ell_i$, $1 \leq i \leq n$. Without loss of generality, we assume that $\ell_i = i$, $1 \leq i \leq n$, so that $F_{\mathbf{x}\bowtie y} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$.

In this paper, we assume that the distribution of $y_j^*|\mathbf{x}_j$, $1 \leq j \leq N$, follows a generalized linear model (GLM) [22], i.e., the corresponding conditional density is given by

$$(1) \qquad f_j(y; \vartheta_j) = \exp\left\{\frac{y\vartheta_j - \psi(\vartheta_j)}{a(\phi)} + c(y, \phi)\right\},$$

where ϑ_j and ϕ are referred to as natural parameter and scale parameter, respectively, and $a(\cdot)$, $\psi(\cdot)$, and $c(\cdot)$ are all known functions referred to as scale function, cumulant, and partition function, respectively; unless stated otherwise, we assume GLMs with canonical link or canonical parameterization, i.e., $\vartheta_j = \eta_j := \mathbf{x}_j^\top \beta^*$, $1 \leq j \leq N$. The $\{\eta_j\}_{j=1}^N$ are referred to as linear predictors based on a regression parameter β^* of interest. To simplify notation, the intercept is typically absorbed into the $\{\mathbf{x}_j\}_{j=1}^N$ even though occasionally we spell out its presence by writing $\eta_j = \beta_0^* + \mathbf{x}_j^\top \beta^*$, $1 \leq j \leq N$. If $\{(\mathbf{x}_j, y_j^*)\}_{j=1}^N$ were given, an estimate for β^* could be obtained by minimizing the following negative log-likelihood corresponding to (1):

$$(2) \qquad \min_{\beta \in \mathbb{R}^d} \ell^*(\beta), \qquad \ell^*(\beta) := -\sum_{j=1}^N \{y_j^* \mathbf{x}_j^\top \beta - \psi(\mathbf{x}_j^\top \beta)\}.$$

However, inference for β^* based on the merged file $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ is in general far from straightforward due to the presence of mismatches. It is well known that the naïve approach that amounts to substitution of $\{(\mathbf{x}_j, y_j^*)\}_{j=1}^N$ in (2) by $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ can exhibit massive bias even if the fraction of mismatched pairs is small. A natural alternative is to consider the joint negative log-likelihood of both β^* and the unknown permutation π^*:

$$(3) \qquad \min_{\beta \in \mathbb{R}^d,\, \pi \in \mathcal{P}(n,N)} \ell(\beta, \pi), \; \ell(\beta, \pi) := -\sum_{i=1}^n \{y_i \mathbf{x}_{\pi(i)}^\top \beta - \psi(\mathbf{x}_{\pi(i)}^\top \beta)\},$$

where $\mathcal{P}(n, N) = \{\pi : \{1, \ldots, N\} \to \{1, \ldots, n\}\}$, π is injective; we shall use $\mathcal{P}(n)$ as a shorthand for all permutations of $\{1, \ldots, n\}$. Note that the symbol π refers to a generic element of $\mathcal{P}(n, N)$, whereas π^* is used for the underlying parameter.

Recent results on the linear model, which corresponds to $\psi(z) = z^2/2$, imply that the optimization problem (3) is intractable [30, 24]. Moreover, putting computational intractability aside, the minimizer of (3) fails to yield consistent estimators of β^* or π^* without suitable lower bounds on $\|\beta^*\|_2^2/\phi^2$ [30, 24, 23, 16].

An alternative viewpoint is to think of π^* as a (latent) random variable depending on contextual information used for linking $F_{\mathbf{x}}$ and F_y, and to focus on inference for β^*. At a high level, this is the strategy adopted in [11, 8, 10]. The success of this line of work shows that it is well possible to obtain accurate estimators of β^* if, loosely speaking, the distribution of π^* is concentrated on a subset of $\mathcal{P}(n, N)$ of manageable size. As elaborated below, the effectiveness of this approach can be particularly well understood in the situation that π^* is known to "block-structured" into a good number of blocks, where the blocks arise from contextual information (cf. Figure 1). In the absence of the latter, a similar reduction can be achieved under the assumption that mismatches occur sparsely in $F_{\mathbf{x}\bowtie y}$ in the sense that $\pi^*(i) \neq i$ is the exception rather than the rule, that is the fraction k_n/n is "small", where $k_n = |\{1 \leq i \leq n : \pi^*(i) \neq i\}|$ denotes the number of mismatches. This assumption is often justifiable given that record linkage tends to provide largely accurate albeit not perfect matchings, particularly if rich contextual information is available when linking $F_{\mathbf{x}}$ and F_y

even though such information may not be available to the analyst of $F_{\mathbf{x}\bowtie y}$. At the same time, it is well known that even small fractions k_n/n can significantly disrupt subsequent regression [5].

While past work on the subject has predominantly focused on estimation of the regression parameter, there is a clear path towards estimating the permutation π^*. In fact, for any fixed β, the optimization problem in π only, i.e.,

$$\begin{aligned} &\min_{\pi\in\mathcal{P}(n,N)} -\sum_{i=1}^{n}\{y_i\mathbf{x}_{\pi(i)}^\top\beta - \psi(\mathbf{x}_{\pi(i)}^\top\beta)\} \\ &= \min_{\pi\in\mathcal{P}(n,N)} -\sum_{i=1}^{n}\{y_i\mathbf{x}_{\pi(i)}^\top\beta - \Psi_{\pi(i)}\}, \\ &\Psi := \big(\psi(\mathbf{x}_i^\top\beta)\big)_{i=1}^{N}, \end{aligned} \tag{4}$$

is a specific instance of a *linear assignment problem (LAP)*, a well-studied class that can be solved efficiently by linear programming as well as tailored algorithms [49].

This observation suggests the estimation of π^* based on (4) with β replaced by an estimator of β^*. The accuracy of this scheme has not been studied for generalized linear models with the exception of the Gaussian linear model [27, 23]. Below, we present some first insights into this questions for other selected generalized linear models.

Outline. In §2, we present our approach for estimation of the regression parameter, and investigate its properties theoretically as well as empirically via simulations. Section §3 is devoted to permutation recovery for known regression parameter, i.e., (4) above with $\beta = \beta^*$. A comparison of the proposed approach and the methods by Lahiri-Larsen and Chambers is provided in §4 and §5, which also contains a case study on real data. We conclude with a summary and an overview on potential directions of future research in §6.

1.2 Notation

For the convenience of the reader, we here collect essential notations used in this paper. For a positive integer ℓ, $\mathbf{1}_\ell$ and I_ℓ denote the vector of ones and the identity matrix, respectively, of dimension ℓ. The n-by-d design matrix associated with covariates $\{\mathbf{x}_i\}_{i=1}^n$ contained in the merged file $F_{\mathbf{x}\bowtie y}$ is denoted by $\mathbf{X}$; unless noted otherwise, we assume that $\mathbf{X}$ includes the column for the intercept (i.e., $\mathbf{X} = [\mathbf{1}_n\ \mathbf{X}_0]$). Likewise, the values for the response in $F_{\mathbf{x}\bowtie y}$ are collected in a vector $\mathbf{y} = (y_i)_{i=1}^n$. The function $\mathbb{I}(\cdot)$ represents the indicator function with value one if its argument is true and zero else. With some abuse of notation, if f is a function of a single argument and $\mathbf{v}$ is a vector of dimension ℓ, we write $f(\mathbf{v})$ for $(f(v_1),\dots,f(v_\ell))^\top$. We let $a\vee b = \max\{a,b\}$ and $a\wedge b = \min\{a,b\}$. We make use of the usual Landau notation in terms of O, o, Ω and Θ. We often use $a\lesssim b$, $b\gtrsim a$, and $a\asymp b$ as shortcuts for $a = O(b)$, $b = \Omega(a)$ and $a = \Theta(b)$, respectively. Numerical constants are denoted by C, C', C_1, c, c_1 etc. whose values may change from line to line. We use the symbols η and μ (with varying subscripts) to refer to the linear and conditional expectation of the response y given covariates $\mathbf{x}$. The associated mappings (the link function and its inverse) are denoted by g and h, respectively, as depicted in the diagram below.

$$\eta = \mathbf{x}^\top\beta \underset{g(\cdot)=h^{-1}(\cdot)}{\overset{h(\cdot)=\psi'(\cdot)}{\rightleftarrows}} \mu = \mathbf{E}[y|\mathbf{x}]$$

2. ESTIMATION OF THE REGRESSION PARAMETER

In this section, we formulate our approach for estimating the regression parameter in GLMs in the presence of mismatch error, i.e., in the setting outlined in §1.1. A bound on the ℓ_2-estimation error of the proposed approach is presented subsequently, which is complemented by numerical results based on simulated data.

2.1 Approach

Defining $o_i^* = (\mathbf{x}_{\pi^*(i)} - \mathbf{x}_i)^\top\beta^*$, we have that $y_i|\mathbf{x}_i, o_i^*$ follows a GLM with linear predictor $\eta_{\pi^*(i)} = \mathbf{x}_i^\top\beta^* + o_i^*$, $1\le i\le n$. Clearly, $\pi^*(i) = i$ implies that $o_i^* = 0$ and in turn $\eta_{\pi^*(i)} = \eta_i = \mathbf{x}_i^\top\beta^*$, $1\le i\le n$. Accordingly, the underlying idea is to a fit a generalized linear model in which each linear predictor is augmented by an individual "dummy variable" or "offset" in order to account for potential mismatches. Without additional constraints or regularization, such approach is not meaningful since it is overparameterized and trivially achieves perfect fit. However, in a sparse mismatch regime with $\pi^*(i) = i$ holding for all except for k_n indices, the use of sparsity-promoting penalties like the ℓ_1-penalty becomes a natural choice. This gives rise to the following formulation: for $\theta\in\mathbb{R}^{d+n}$, we consider the partitioning $\theta = [\beta^\top\ \xi^\top]^\top$ with $\beta\in\mathbb{R}^d$ and $\xi\in\mathbb{R}^n$. We then consider estimation based on minimizing the penalized negative log-likelihood given by

$$\begin{aligned} &\ell_{\mathrm{pen}}(\theta) = \ell(\theta) + \lambda\|\xi\|_1, \\ &\ell(\theta) := \frac{1}{n}\left\{-\langle\mathbf{X}\beta + \sqrt{n}\xi, \mathbf{y}\rangle + \sum_{i=1}^{n}\psi(\mathbf{x}_i^\top\beta + \sqrt{n}\xi_i)\right\}, \end{aligned} \tag{5}$$

where $\mathbf{X}$ is the usual design matrix with rows $\{\mathbf{x}_i^\top\}_{i=1}^n$, $\mathbf{y} = (y_i)_{i=1}^n$, and $\lambda\ge 0$ is a tuning parameter whose choice will be discussed below. In (5), the dummy variables $\xi = (\xi_i)_{i=1}^n$ are in correspondence to the $\{o_i^*\}_{i=1}^n$ above; re-scaling by $n^{-1/2}$ is done exclusively to ensure that the ℓ_2-norms of the columns of the augmented design matrix $[\mathbf{X}\ \ \sqrt{n}I_n]$ are all of the order $\sqrt{n}$, which is beneficial to the presentation of the theoretical analysis of (5) in §2.4 below[1].

[1]The requirement that the column norms scale as $\sqrt{n}$ is standard in the analysis of estimators for high-dimensional (generalized) linear models, cf. [15, §7.1.2].

Algorithm 1 Block coordinate descent algorithm for (5)

Initialize $\widehat{\xi}^{(0)} = 0$ and $\widehat{\beta}^{(0)}$ as the ordinary GLM estimate based on $(\mathbf{X}, \mathbf{y})$.

1. Update for ξ:

$$\widehat{\xi}_i^{(t+1)} \leftarrow \mathbb{I}\left\{\frac{|y_i - \widehat{\mu}_i^{(t)}|}{\sqrt{n}} - \lambda > 0\right\} \cdot \frac{\left((\psi')^{-1}(y_i - s_i\lambda) - \widehat{\eta}_i^{(t)}\right)}{\sqrt{n}}, \; s_i = \operatorname{sign}(y_i - \widehat{\mu}_i^{(t)})\sqrt{n}, \quad 1 \le i \le n.$$

2. Update for β :

$$\widehat{\beta}^{(t+1)} \leftarrow \widehat{\beta}^{(t)} + (\mathbf{X}^\top W^{(t)}\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{y} - \psi'(\mathbf{X}\widehat{\beta}^{(t)} + \sqrt{n}\widehat{\xi}^{(t+1)}))$$

where $\widehat{\eta}_i^{(t)} = \mathbf{x}_i^\top \widehat{\beta}^{(t)}$, $\widehat{\mu}_i^{(t)} = \psi'(\widehat{\eta}_i^{(t)})$, $V_i^{(t)} = \psi''(\widehat{\eta}_i^{(t)} + \sqrt{n}\widehat{\xi}_i^{(t+1)})$, $1 \le i \le n$, and $W^{(t)} = \operatorname{diag}\{V_i^{(t)}\}_{i=1}^n$.

Formulations of the form (5) or similar have been considered in prior work in different contexts. The use of dummy variables to deal with data contamination in (ordinary) linear models has been discussed in She & Owen [44], Laska et al. [50], Nguyen & Tran [51], and more recently in Bhatia et al. [52]. Extensions to generalized linear models have been proposed in [45, 46, 53]. Slawski & Ben-David [23] study and analyze this approach in detail for mismatch contamination in linear models, and the present paper arises as a direct extension of their work. It is worth emphasizing that despite prior work on the formulation (5), the latter has not been studied specifically for mismatch contamination. Moreover, none of the earlier works [45, 46, 53] contain a complete theoretical analysis as provided herein.

We note in passing that (5) can be applied broadly in situations beyond linkage of $F_{\mathbf{x}}$ and F_y. For example, rather common situations are (i) a subset of the covariates is contained in the same file in the response, (ii) file linkage involves more than two files, each containing different subsets of the covariates and/or the response. Both (i) and (ii) can be addressed via (5) based on suitable choices of the variables $\{o_i^*\}_{i=1}^n$ (cf. supplement for a detailed discussion (http://intlpress.com/site/pub/files/_supp/sii/2023/0016/0003/SII-2023-0016-0003-s001.pdf)).

2.2 Computation

There are various ways of solving the convex optimization problem (5). A particularly suitable approach that exploits structure specific to (5) is block coordinate descent with blocks formed by β and ξ, respectively. The key observation is that for any fixed value of β, minimization over ξ can be performed in closed form via a soft thresholding-type update [54]. On the other hand, note that for any fixed ξ minimization with respect to β amounts to fitting the underlying GLM with offset $\sqrt{n}\xi_i$ for observation i, $1 \le i \le n$. Since the proposed algorithm is already iterative, alternating between updates of β and ξ, we only perform a single iteration of weighted least squares (aka Fisher Scoring) when updating β; this is equivalent to minimizing the quadratic Taylor approximation of the objective (with ξ treated as fixed) around the current iterate $\widehat{\beta}^{(t)}$. A schematic description of the algorithm is provided below.

Given extensive numerical experience, Algorithm 1 converges in practice after few iterations. In order to establish convergence theoretically, the two updates above would need to be combined with a suitable mechanism for step size selection [55]. Since the latter is standard in the optimization literature, we refrain from discussing this aspect in detail to avoid digressions.

2.3 Incorporating blocking variables

Recall that $o_i^* = (\mathbf{x}_{\pi^*(i)} - \mathbf{x}_i)^\top \beta^*$, $1 \le i \le n$. Note that if $N = n$ so that π^* is a permutation, it is easy to see that $\sum_{i=1}^n o_i^* = 0$. As a result, the additional constraint $\sum_{i=1}^n \xi_i = 0$ may be added to the optimization problem (5). This observation can be put to much more use if π^* is known to be "block-structured" in the sense that the data set can be partitioned into disjoint groups $G_1, \ldots, G_K \subset \{1, \ldots, n\}$ such that $i \in G_j$ for some $j \in \{1, \ldots, K\}$ implies that $\pi^*(i) \in G_j$, $1 \le i \le n$; in other words, the permutation only moves indices within, but not across groups. With the same reasoning as above, we then have $\sum_{i \in G_j} o_i^* = 0$, which accordingly yields the constraints $\sum_{i \in G_j} \xi_i = 0$, $1 \le j \le K$, to be added to (5). Specifically, this yields the optimization problem

$$\min_\theta \ell(\theta) + \lambda\|\xi\|_1 \;\text{ subject to } \mathbf{C}\xi = \mathbf{0}, \tag{6}$$

where $\mathbf{C} \in \mathbb{R}^{K \times n}$ has entries $C_{ji} = 1$ if $i \in G_j$ and zero otherwise, $1 \le j \le K$, $1 \le i \le n$. In particular, in the case of singleton groups with $G_j = \{i\}$ for $i \in \{1, \ldots, n\}$, it immediately follows that $\xi_i = 0$ and the corresponding variable can be eliminated in (6). If K is large, this yields a substantial number of extra constraints whose integration in (6) can considerably boost performance relative to the unconstrained minimizer not taking any advantage of the block structure of π^*. The latter arises when additional knowledge about the linkage process is available. More specifically, it is common that matching records are known to agree on certain combinations of variables present for the records in both $F_{\mathbf{x}}$ and F_y (e.g., demographic variables such as gender, age, and/or race, approximate geographical location based on ZIP code, approximate time stamps, etc.). Those variables are typically referred to as blocking variables, and the

corresponding groups $\{G_j\}_{j=1}^K$ are given by subsets of observations sharing the same values for all blocking variables.

In summary, while the approach (5) works without any knowledge about the linkage process and the existence of blocking variables, it is possible to achieve enhancements if such information is available. This aspect is investigated in detail in §5.

2.4 Analysis

In the sequel, we derive a non-asymptotic upper bound on the ℓ_2-error $\|\widehat{\theta} - \theta^*\|_2$, where $\theta^* = [\beta^{*\top} \ \xi^{*\top}]^\top$ with $\xi_i^* = \frac{1}{\sqrt{n}}(\mathbf{x}_{\pi^*(i)} - \mathbf{x}_i)^\top \beta^*$ and $\widehat{\theta} = [\widehat{\beta}^\top \ \widehat{\xi}^\top]^\top$ denotes a minimizer of ℓ_{pen} in (5). We henceforth drop the dependence on the number of mismatches k on n and simply write k instead of k_n. Before stating the final result in Theorem 1, we present and discuss the assumptions underlying that result. Our analysis generalizes results in [23] concerning the ordinary linear model.

Assumptions and Conditions.

(**A**) The rows $\{\mathbf{x}_i\}_{i=1}^n$ of $\mathbf{X}$ are i.i.d. copies of a random vector $\mathbf{x}$ with the following properties: 1) there exists a positive definite matrix Σ with uniformly bounded eigenvalues, i.e., $\sigma_{\min} I_d \preceq \Sigma \preceq \sigma_{\max} I_d$, such that $\langle v, \Sigma^{-1/2}\mathbf{x}\rangle$ is a sub-Gaussian random variable with sub-Gaussian norm at most γ for all $v \in \mathbb{R}^d$*, 2) there exists a constant $r > 0$ such that $\mathbf{P}(\|\mathbf{x}\|_2 \le r) = 1$.

We note that (**A**) allows for the inclusion of an intercept by requiring that the first component of $\mathbf{x}$ equals one with probability one. In addition, the entries of $\mathbf{x}$ neither need to be mean zero nor uncorrelated; with $\Sigma = \mathbf{E}[\mathbf{x}\mathbf{x}^\top]$ chosen as the population second moment matrix, only uniform upper and lower bounds for its eigenvalues are required. The condition that the essential support of $\mathbf{x}$ is contained in an ℓ_2-ball of bounded radius is imposed to ensure bounded linear predictors, as further elaborated below. Note that if $\mathbf{x}$ is sub-Gaussian with unbounded support (e.g., $\mathbf{x} \sim N(0, I_d)$), the truncation $T(\mathbf{x}) := \mathbf{x}\mathbb{I}(\|\mathbf{x}\|_2 \le r) + \frac{r\mathbf{x}}{\|\mathbf{x}\|_2}\mathbb{I}(\|\mathbf{x}\|_2 > r)$ conforms with (**A**).

(**C1**) There exist sequences $\nu_n, \epsilon_n = o(1)$ as $n \to \infty$ such that $\|\nabla \ell(\theta^*)\|_\infty \le \nu_n$ with probability at least $1 - \epsilon_n$.

(**C2**) There exists a sequence $\epsilon_n' = o(1)$ as $n \to \infty$ and constants $R > 0$, $0 < \lambda_R \le \Lambda_R < \infty$ such that with probability at least $1 - \epsilon_n'$

$$\min_{1\le i\le n} \min_{u:\|u\|_2\le R} \psi_i''(\theta^* + u) \ge \lambda_R,$$
$$\frac{32}{\sigma_{\min}\lambda_R}(\lambda + \nu_n)\sqrt{d+k} \le R, \quad (\#)$$
$$\max_{1\le i\le n} \max_{u:\|u\|_2\le R} \psi_i''(\theta^* + u) \le \Lambda_R.$$

where $\theta \mapsto \psi_i(\theta) := \psi(\mathbf{x}_i^\top \beta + \sqrt{n}\xi_i)$ and $\psi_i''(\theta) := \frac{d^2}{dz^2}\psi(z)\Big|_{z=\mathbf{x}_i^\top\beta+\sqrt{n}\xi_i}$, $1 \le i \le n$.

Condition (**C1**) holds under assumption (**A**) with $\nu_n = C\sqrt{\frac{\log(d+n)}{n}}$ and $\epsilon_n = c/n$ if additionally at least one of the following two properties holds [28]:

(i) ψ'' is uniformly bounded,

(ii) $\mathbf{E}[\max_{|u|\le 1} \psi''(\mathbf{x}^\top\beta^* + u)^\alpha] \le B$ for some $\alpha \ge 2$ and $0 < B < \infty$.

In particular, property (i) is satisfied for logistic regression, while property (ii) is satisfied for Poisson regression if $\|\beta^*\|_2$ is uniformly bounded, which together with the boundedness assumption $\mathbf{P}(\|\mathbf{x}\|_2 \le r) = 1$ in (**A**) implies that $\mathbf{x}_i^\top\beta^* + \sqrt{n}\xi_i^* = \mathbf{x}_{\pi^*(i)}^\top\beta^*$, $1 \le i \le n$, are uniformly bounded. Boundedness of the $\{\mathbf{x}_i\}_{i=1}^n$ and of β^* is also needed for (**C2**) to hold since ψ'' can generally only be lower and upper bounded on a compact interval; as a result, similar boundedness assumptions are commonly imposed in the literature (e.g., [57, Definition 8.1], [58, Example 1]).

Note that in (**C2**), a valid radius R needs to satisfy the condition (#) for the error bound in Theorem 1 below not to be vacuous. With the choice $\lambda \asymp \nu_n$ as indicated by that theorem, and the scaling $\nu_n \lesssim \sqrt{\log(d+n)/n}$ as discussed above, (#) is of the form

$$\frac{C'}{\sigma_{\min}\lambda_R}\sqrt{\frac{\log\{d+n\}(d+k)}{n}} \le R, \quad (7)$$

which translates to a condition on the sample size of the form $n \ge C_R\, \sigma_{\min}^2 \log\{d+n\}\,(d+k)$, where C_R is a constant depending on R. In turn, the latter condition restricts the number of mismatches k to be at most of the order $n/\log n$. The left hand side of (7) coincides with the bound on the ℓ_2-estimation error stated in the theorem below, which asserts consistent estimation given $\frac{\log\{d+n\}\,(d+k)}{n} \to 0$ and ν_n, λ scaling as discussed above.

Theorem 1. *Suppose that assumption* (**A**) *and conditions* (**C1**), (**C2**) *hold. Consider any minimizer of* $\widehat{\theta}$ *of* ℓ_{pen} *in with* $\lambda = \lambda_n$ *chosen such that* $2\nu_n \le \lambda \le C\nu_n$ *for some* $C > 0$. *Then there exists constants* $C' > 0$, $c \in (0,1)$, *so that if* $n \ge C' \frac{\sigma_{\max}}{\sigma_{\min}} \cdot \left(\frac{\Lambda_R}{\lambda_R}\right)^2 \left\{(d+k)\log\left(\frac{n}{d+k}\right) \vee \log n\right\}$, *it holds that*

$$\|\widehat{\theta} - \theta^*\|_2 \le \inf\left\{0 < r \le R : \ r > \frac{32}{\sigma_{\min}\lambda_R}(\lambda + \nu_n)\sqrt{d+k}\right\}$$

with probability at least $1 - \epsilon_n - \epsilon_n' - 2/n$.

In addition to the condition on the sample size implied by (7), the above statement involves a second condition on n which is slightly less stringent in terms of the required ratio $n/(d+k)$, but additionally involves a dependency on the

*cf., e.g.,[56, §2.5] for a concise discussion of sub-Gaussian random variables and their properties.

condition number of Σ and the Hessian of ψ in a neighborhood around the true parameter, none of which, however, entails any additional requirements beyond (**A**), (**C1**), and (**C2**).

Theorem 1 qualitatively aligns with earlier results in [23] concerning the linear model. While the latter is covered by the analysis above, the underlying assumptions and conditions can be considerably relaxed or simplified in this case.

2.5 Simulations

We here present selected simulation results that are intended to corroborate and complement the analyis of the previous section; additional simulation results can be found in the supplement [59]. Data is generated according to model (1) with $n = 1000$, $d = 50$, and the entries of β^* are drawn i.i.d. from the $N(0, 1)$-distribution, and subsequently normalized so that the ℓ_2-norm equals a specific value (see below). The entries of the design matrix are drawn from the uniform distribution on the interval $[-\sqrt{3}, \sqrt{3}]$[2]. The following distributions for the response are considered.

- *Poisson*: intercept $\beta_0^* = 2$, $\|\beta^*\|_2 = 2$,
- *Binomial*: the number of Bernoulli trials per observation is fixed as $m = 25$, $\beta_0^* = 2$, and $\|\beta^*\|_2 = 4$; the case of binary response ($m = 1$) is discussed in a dedicated paragraph.
- *Gamma*: the shape parameter is fixed as $\nu = 50$ (or equivalently, the dispersion parameter ϕ is set to $1/\nu$), $\beta_0^* = 8$ and $\|\beta^*\|_2 = 2$.

We do not present results on Gaussian response, and instead refer to [23].

The map π^* is drawn uniformly at random from the set of permutations on $\{1, \dots, n\}$ that move exactly k indices, where $k/n \in \{0.05, 0.1, \dots, 0.4\}$[3].

In alignment with Theorem 1, the regularization parameter λ is chosen as $\lambda = C \cdot \sigma_y \cdot \sqrt{\frac{\log(n+d)}{n}}$, where C is referred to as "tuning constant" and σ_y is a calibration factor depending on the distribution of the response. The calibration factor is chosen as an approximation of the expected standard deviation of the response variables $\{y_i\}_{i=1}^n$, where the expectation is taken with respect to the random predictors[4]. The use of σ_y is motivated by results on linear regression in [23] and an analysis of the ℓ_2-estimation error for ξ^* if β^* were *known* (omitted for space reasons). In practice, σ_y can be approximated by taking the average of the variance function of the corresponding GLM evaluated at the $\{y_i\}_{i=1}^n$; in order to enable Figure 2 that is specifically dedicated to the selection of λ, the factor σ_y is fixed so that it does not vary across different randomly generated data sets.

For each triplet $(\beta_0^*, \|\beta^*\|_2, k/n)$, 100 independent replications are considered. The following approaches are compared.

"naive". Plain GLM estimation based on $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ without adjustment for mismatches. Note that the fitted values of this approach coincide with those of the proposed approach (see below) if $\lambda \in (\lambda_{\max}, \infty)$, where $\lambda_{\max} = \|\nabla_\xi \ell(\widehat{\theta}^{\text{naive}})\|_\infty$ and $\widehat{\theta}^{\text{naive}} = ((\widehat{\beta}^{\text{naive}})^\top \ \mathbf{0}_n^\top)^\top$ with $\widehat{\beta}^{\text{naive}}$ denoting the plain GLM estimate; this is an immediate consequence of the KKT conditions of (5).

oracle. Plain GLM estimation based on the mismatch-free data $\{(\mathbf{x}_i, y_i^*)\}_{i=1}^n$.

"proposed". $\widehat{\theta}(\lambda) = (\widehat{\beta}^\top \ \widehat{\xi}^\top)^\top$ is estimated according to Algorithm 1 with λ chosen as explained above. The tuning constant C is varied over a logarithmically spaced grid. When presenting the results in Figure 3, we display the oracle selection of $\lambda = \lambda(C)$ minimizing $\|\widehat{\theta}(\lambda) - \theta^*\|_2$ as well as the range of a quantity of interest for $C \in [C_{\text{Lower}}, C_{\text{Upper}}]$, where C_{Lower} and C_{Upper} are fixed numbers. The resulting upper and lower bounds over this range are complemented by confidence bars with height $5\times$ standard error. Figure 2 displays explicitly how the (normalized) estimation error $\|\widehat{\theta} - \theta^*\|_2 / \|\widehat{\theta}^{\text{naive}} - \theta^*\|_2$ depends on the choice of C; the division by $\|\widehat{\theta}^{\text{naive}} - \theta^*\|_2$ with $\widehat{\theta}^{\text{naive}}$ defined under "naive" makes the results interpretable across different settings.

Lahiri-Larsen (LL) and Chambers' (C) method. These are well-established baselines for the given problem; a detailed discussion and comparison is postponed to §4. Both methods are provided with the correctly specified expected permutation matrix $\mathbf{Q} = \mathbf{E}[\Pi^*] = (1 - k/n)I_n + \frac{k/n}{n-1}\mathbb{1}_n$, where $\mathbb{1}_n$ denotes an n-by-n matrix of ones. For Poisson regression, in about 30% of the cases, the two estimators could not be obtained since roots of the underlying estimating equations could not be found (this issue is not unexpected, and elaborated on in the supplement); these problematic instances are excluded when reporting results. For binary response (cf. Figure 4), the same issue persists in the vast majority of instances, and we hence refrain from a comparison altogether.

The above four approaches are evaluated in terms of their ℓ_2-estimation error for the regression parameter, i.e., $\|\beta^{\text{est}} - \beta^*\|_2$, and the deviance (Kullback-Leibler divergence) between $\mu^* = \mathbf{E}[\mathbf{y}^*|\mathbf{x}]$ and $\mu^{\text{est}} = \big(h(\mathbf{x}_i^\top \beta^{\text{est}})\big)_{i=1}^n$, where we recall that $h(\cdot)$ denotes the response function in GLMs. The notation β^{est} represents a placeholder for any of the three estimators introduced above. The results shown in Figures 2, 3 and 4 are averages over the 100 replications obtained for each setting.

Figure 2 confirms that as the tuning constant C increases, the error ratio approaches one as expected since $\widehat{\theta} \to \widehat{\theta}^{\text{naive}}$

[2]For space reasons, we here only present results for this specific class of random designs. Additional simulation results contained in the supplement show that the outcome is similar for a wide range of random designs.

[3]This can be achieved by first selecting a random subset of size k, and then generating a random permutation of that subset (if the resulting permutation happens to have a fixed point, it is rejected and drawn again).

[4]This expectation is evaluated using numerical integration, approximating the linear predictor by a $N(\beta_0^*, \|\beta^*\|_2^2)$-random variable justified by the central limit theorem.

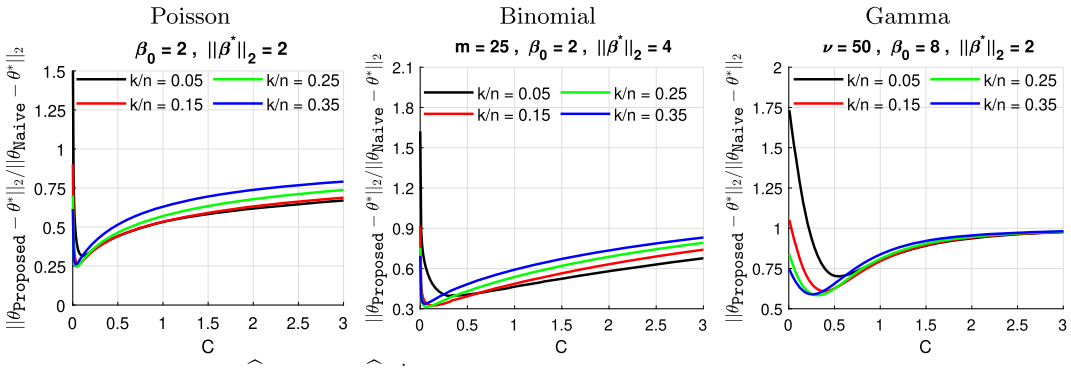

Figure 2. Estimation error ratios $\|\widehat{\theta} - \theta^*\|_2 / \|\widehat{\theta}^{naive} - \theta^*\|_2$ in dependence of the tuning constant C appearing in the tuning parameter λ.

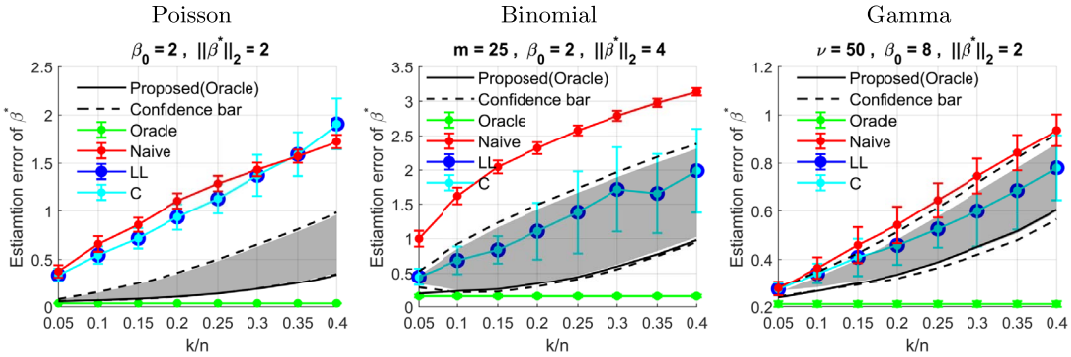

Figure 3. Average estimation errors $\|\beta^{est} - \beta^*\|_2$. The lower and upper boundary of the shaded area show the minimum and maximum error over all choices of the tuning constant $C \in [0.1, 2]$, and the corresponding dashed lines represent $\pm 5 \times$ standard error.

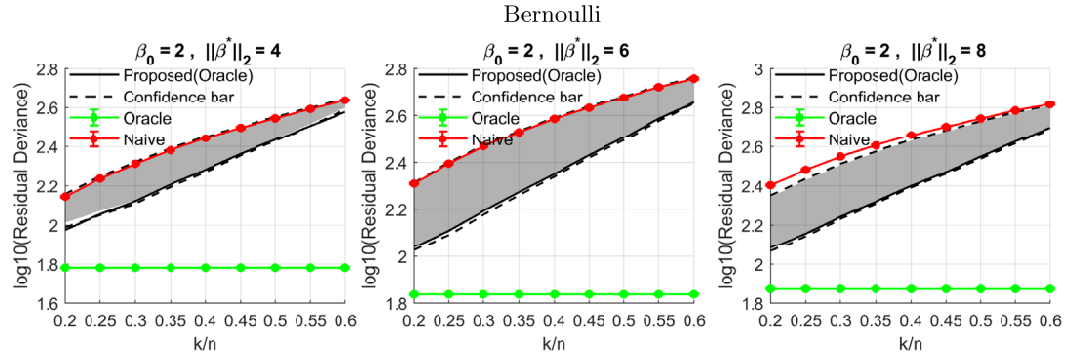

Figure 4. Average deviances between μ^* and μ^{est} for Bernoulli response. The lower and upper boundary of the shaded area show the minimum and maximum error over all choices of the tuning constant $C \in [0.5, 2]$, and the corresponding dashed lines represent $\pm 5 \times$ standard error.

as C grows. Second, we note that the error ratio increases sharply beyond one as $C \to 0$; this corresponds to a regime of overfitting. Note that as $C \to 0$ the proposed approach effectively yields an over-parameterized model achieving perfect fit on a given data set. This in agreement with Theorem 1 which requires a lower bound on λ for its results to hold.

Figure 2 indicates that $C \in [0.2, 1]$ typically yields satisfactory results independent of the specific setting or the specific distribution of the response variable.

Average estimation errors for the regression parameter are shown in Figure 3. Shaded areas are used to represent the error range for the proposed approach in dependence of

the tuning constant C: the upper and lower margins of the shaded areas represent the maximum and minimum error over $C \in [0.1, 2]$, while the dashed lines outside the shaded areas indicate $\pm 5 \times$ standard error. Overall, Figure 3 shows that the proposed estimator can achieve substantial improvements over the naive estimator in a variety of settings. The extent of the improvement generally increases with the fraction of mismatches and the signal level as measured by $\|\beta^*\|_2$. The Lahiri-Larsen (LL) and Chambers' methods (C) mostly improve over the naive estimator as well, with a performance that often still falls within the error range (grey region) of the proposed method even though the latter has a visible edge over these two baselines when evaluating the entirety of simulation settings (cf. supplement), particularly with a (near) optimal selection of the tuning parameter.

Binary response. We now address the case of binary logistic regression. Note that this case behaves differently from the three distributions for the response considered above, including binomial response with a significant number of trials: it is clear that a massive number of samples is required in order to estimate β^* accurately from binary response if $\|\beta^*\|_2$ is large; at the same time, if $\|\beta^*\|_2$ is small, the separation between the two classes corresponding to the two values of the response variable is weak and thus the inherent noise is scarcely distinguishable from mismatch error in the response variable, which amounts to what has been extensively studied in the machine learning literature under the terms "label noise" (e.g., [60]). For this reason, we adopt the simulation setup above with sufficiently strong signal, i.e., $\|\beta^*\|_2 \in \{4, 6, 8\}$ while $\beta_0^* = 2$, but the three competitors are evaluated in terms of the deviance between μ^* and μ^{est} rather than in terms of the estimation error for the regression parameter. Figure 4 shows that the proposed approach achieves improvements over the naive estimator, but the improvements are less pronounced than for the three other settings. The observed moderate improvement is in alignment with what is reported in the paper [45] that studies the empirical performance of the proposed estimator exclusively in the setting of binary response with noisy labels.

3. PERMUTATION RECOVERY

In this section, we study the maximum likelihood (ML) estimator $\widehat{\pi}$ of π^* for known β^* (4). While the assumption of known β^* may appear limiting, the results of this section can still be useful from at least two considerations: first, it is not unreasonable to expect that they continue to be valid if β^* is replaced by an accurate estimator; second, they provide some insights into what is at best achievable in practice. The first result states that ML estimation of π^* for known β^* is computationally tractable as already indicated in the introduction of this paper.

Proposition 1. *Consider ML estimation of π^*, i.e., optimization problem (4) for $\beta = \beta^*$.*

(a) *We have*

$$\begin{aligned}&\min_{\pi \in \mathcal{P}(n,N)} -\sum_{i=1}^{n} \{y_i \mathbf{x}_{\pi(i)}^\top \beta^* - \psi(\mathbf{x}_{\pi(i)}^\top \beta^*)\} \\ &= \min_{\Pi \in \mathbf{B}(n,N)} \operatorname{tr}(C^\top \Pi) = \min_{\Pi \in \mathbf{P}(n,N)} \operatorname{tr}(C^\top \Pi),\end{aligned}$$

where $\mathbf{B}(n, N)$ and $\mathbf{P}(n, N)$ denote the sets of binary and non-negative n-by-N matrices, respectively, with unit row sums and column sums bounded by one, and the entries of the matrix $C = (C_{ij})$ are given by $C_{ij} = -y_i \mathbf{x}_j^\top \beta^* + \psi(\mathbf{x}_j^\top \beta^*)$, $1 \le i \le n$, $1 \le j \le N$.

(b) *In particular, if $n = N$,*

$$\begin{aligned}&\min_{\pi \in \mathcal{P}(n)} -\sum_{i=1}^{n} \{y_i \mathbf{x}_{\pi(i)}^\top \beta^* - \psi(\mathbf{x}_{\pi(i)}^\top \beta^*)\} \\ &= \min_{\pi \in \mathcal{P}(n)} -\sum_{i=1}^{n} y_i \mathbf{x}_{\pi(i)}^\top \beta + c = -\sum_{i=1}^{n} y_{(i)} (\mathbf{x}^\top \beta^*)_{(i)} + c,\end{aligned} \tag{8}$$

where $c = \sum_{i=1}^n \psi(\mathbf{x}_i^\top \beta^*)$ and the subscript (i) refers to the i-th order statistic, i.e., for $v = (v_i)_{i=1}^n$, $v_{(1)} \le \ldots \le v_{(n)}$.

Proposition 1 **(a)** states that the MLE $\widehat{\pi}$ of π^* can be obtained by a solving a linear assignment problem, a specific linear program (cf., e.g., [49]). Part **(b)** on the case $n = N$ states that $\widehat{\pi}$ is given by the permutation that pairs the corresponding order statistics of $\{\mathbf{x}_i^\top \beta^*\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$. For later reference, we note that the conclusion of part **(b)** with regard to the form of $\widehat{\pi}$ continues to hold if the link function is not the canonical link. This is an immediate consequence of the fact that $\widehat{\pi}$ is invariant under monotonically increasing transformations of $\{\mathbf{x}_i^\top \beta^*\}_{i=1}^n$. The proof of Proposition 1 follows from existing results on linear assignment problems [49] and is omitted.

For the remainder of this section, we suppose that $N = n$ and present sufficient conditions for specific GLMs under which $\widehat{\pi}$ achieves perfect recovery in the sense of $\{\widehat{\pi} = \pi^*\}$[5]. We refrain from presenting a unified analysis applicable to an entire class of GLMs for two reasons: first, sharper results can be obtained from case-specific analysis; second, permutation recovery turns out to be entirely or at least largely infeasible for a variety of GLMs, e.g., i) binomial response with a small number of trials due to excessive ties among the $\{y_i\}_{i=1}^n$, ii) exponential response with canonical (i.e., reciprocal) link since in this case recovery fails for a wide range of random designs (cf. Theorem 3 below).

3.1 Recovery results

We start the presentation of our results by conditioning on the predictors $\{\mathbf{x}_i\}_{i=1}^n$, and hence for fixed conditional expectations of the responses. The extension to random predictors is considered subsequently.

[5]The assumption $N = n$ is not believed to be essential, and we expect that similar results can be established for the general case at the expense of a considerably more complex exposition.

In this subsection, it is appropriate to distinguish between random variables $\{Y_i\}_{i=1}^n$ and their realizations $\{y_i\}_{i=1}^n$. We let $\mu_i = \mathbf{E}[Y_i|\mathbf{x}_{\pi^*(i)}] = h(\mathbf{x}_{\pi^*(i)}^\top \beta^*)$, $1 \le i \le n$, where h denotes the inverse link function of the underlying GLM. Unless stated otherwise, h refers to the canonical link.

Theorem 2. *Suppose without loss generality that $\mu_1 \le \mu_2 \le \cdots \le \mu_n$, and consider the MLE $\widehat{\pi}$ given by the minimizer of (8). For any $\delta > 0$, we have $\mathbf{P}(\widehat{\pi} \neq \pi^*|\mathbf{X}) < \delta$ if*

(a) $Y_i \sim N(\mu_i, \sigma^2)$, $1 \le i \le n$:
$\min_{1\le i\le n-1}(\mu_{i+1} - \mu_i) > 2\sigma\sqrt{\log\frac{n-1}{\delta}}$,
(b) $Y_i \sim Poisson(\mu_i)$, $1 \le i \le n$:
$\min_{1\le i\le n-1}(\sqrt{\mu_{i+1}} - \sqrt{\mu_i}) > \sqrt{\log\frac{n-1}{\delta}}$,
(c) $Y_i \sim Gamma(\nu, \mu_i)$, $1 \le i \le n$:
$\min_{1\le i\le n-1}\frac{\mu_{i+1}}{\mu_i} > 4\left(\frac{n-1}{\delta}\right)^{1/\nu}$.

Part (a) already appears in similar form in [23]. Part (b) can be linked to (a) by noting that the standard deviation of a Poisson random variable with mean μ equals $\sqrt{\mu}$. Substituting σ in (a) by $\sqrt{\mu_i}$ and dividing both sides by this quantity then approximately yields (b). Part (c) can be understood according to a similar heuristic: observing that the standard deviation of $Y_i \sim \text{Gamma}(\nu, \mu_i)$ is given by $\mu_i/\sqrt{\nu}$, $1 \le i \le n$, substituting σ in (a) by μ_i yields a requirement on the ratio μ_{i+1}/μ_i. Note that for $\nu \asymp \log n$, the ratios need to exceed a constant factor $C > 1$, which still requires $\mu_n/\mu_1 = C^{n-1}$. For the exponential distribution, $\nu = 1$, and there is thus little hope that (c) can be satisfied in practice even for small n.

Building on Theorem 2, we next consider random predictors $\{\mathbf{x}_i\}_{i=1}^n$, thereby providing specific examples of designs in which the recovery conditions are satisfied with high probability.

Theorem 3. *Consider the MLE $\widehat{\pi}$ given by the minimizer of (8). Suppose that the $\{\mathbf{x}_i\}_{i=1}^n$ are i.i.d. random vectors with independent, unit variance entries whose densities are bounded by $M < \infty$ almost everywhere. For any $\delta > 0$, we have $\mathbf{P}(\widehat{\pi} \neq \pi^*) < \delta$ if*

(a) $Y_i \sim N(\mu_i, \sigma^2)$, $1 \le i \le n$:
$\|\beta^*\|_2^2 > \frac{8\sigma^2M^2n^2(n-1)^2}{\delta^2}\log\left(\frac{n(n-1)}{\delta}\right)$,
(b) $Y_i \sim Poisson(\mu_i)$, $1 \le i \le n$:
$\|\beta^*\|_2^2 > \frac{16M^2n^2(n-1)^2}{\delta^2}\log\left(\frac{n(n-1)}{\delta}\right)$,
and $\beta_0^*/\|\beta^*\|_2$ is such that
$\sup_{u:\|u\|_2=1}\mathbf{P}\left(\min_{1\le i\le n}\langle u, \mathbf{x}_i\rangle \le -\frac{\beta_0^*}{\|\beta^*\|_2}\right) < \delta/2$,
(c) $Y_i \sim Gamma(\nu, \mu_i)$, $\mu_i = \exp(\eta_i)$, $1 \le i \le n$:
$\|\beta^*\|_2^2 > \frac{M^2n^2(n-1)^2}{2\nu^2\delta^2}\left(\log 4\left(\frac{n(n-1)}{\delta}\right)^{1/\nu}\right)^2$.

Part (a) appears in similar form for isotropic Gaussian $\{\mathbf{x}_i\}_{i=1}^n$ in [23]. The statement here extends that result to a much broader class of designs, without imposing any condition on the tails of the distribution of the $\{\mathbf{x}_i\}_{i=1}^n$. Part (b) for the Poisson distribution involves an extra condition compared to (a) which in essence requires a lower bound on $\beta_0^*/\|\beta^*\|_2$ to ensure that all of the $\{\mu_i\}_{i=1}^n$ are sufficiently bounded away from one with high probability. For example, if the entries of the $\{\mathbf{x}_i\}_{i=1}^n$ are i.i.d. and symmetric around zero, and $\beta^* = (1, \ldots, 1)^\top$, say, the resulting linear predictor will assume negative values with probability $1/2$ which then translate to expectations between zero and one via the inverse link function (exponential). According to Theorem 2, we need the spacing between the $\{\mu_i\}_{i=1}^n$ to be at least proportional to the corresponding standard deviations, which is violated in the range $[0, 1]$ since the standard deviations are given by $\{\sqrt{\mu_i}\}_{i=1}^n$. Depending on the distribution of the $\{\mathbf{x}_i\}_{i=1}^n$, the condition on $\beta_0^*/\|\beta^*\|_2$ can be made explicit: in the simplest case with $\{\mathbf{x}_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} N(0, I_d)$, we have $\{\langle \mathbf{x}_i, u\rangle\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} N(0, 1)$ for any unit vector u, and it is then not hard to show that the extra condition in (b) holds if $\beta_0^*/\|\beta^*\|_2 \ge \sqrt{2\log n} + \sqrt{\log(2n/\delta)}$. Regarding part (c), let us emphasize that the result here concerns the log-link rather than the canonical (reciprocal) link. A recovery result for the latter appears out of reach since the use of the reciprocal link would lead to a clustering of the $\{\mu_i\}_{i=1}^n$ in $[0, 1]$ independent of β_0^* for many common choices for the distribution of the $\{\mathbf{x}_i\}_{i=1}^n$.

Incorporating blocking variables. Following up on the discussion in §2.3, it is worth pointing out that permutation recovery based on (8) decouples across blocks, and hence can be performed in a block-by-block fashion. The recovery conditions in Theorems 2 and 3 can be applied for each block in that n gets replaced by the number of elements belonging to the respective block, and are thus easier to satisfy. For example, if $n_j = n/K$, $j = 1, \ldots, K$, and $\delta = \delta_0/K$ for a given failure probability δ_0, it is easy to check that all terms involving $n - 1$ get replaced by $n/K - 1$. This yield substantial benefits particularly if $n/K = O(1)$, i.e., in the case of many small blocks.

Mismatch recovery. Note that (8) does not take advantage of a sparse mismatch regime if the latter is known to hold. Unfortunately, it turns out that replacing the minimum in (8) by the minimum over all permutations moving at most k indices gives rise to a considerably harder optimization problem unlike the simple solution via sorting obtained in the absence of such constraint. In spite of this, there is a natural workaround involving two steps: 1. identify the set of mismatches $S_* = \{i : \pi^*(i) \neq i\}$, 2. solve the minimization problem in (8) restricted to the observations in S_*. With step 2. set and analyzed according to the preceding theorems, it remains to consider step 1., which we refer to as "mismatch recovery". We suggest to address this task by assessing the fit of each y_i to its counterpart $\mu_i = h(\mathbf{x}_i^\top \beta^*)$. Conditional on $\pi^*(i) = i$, the distribution of

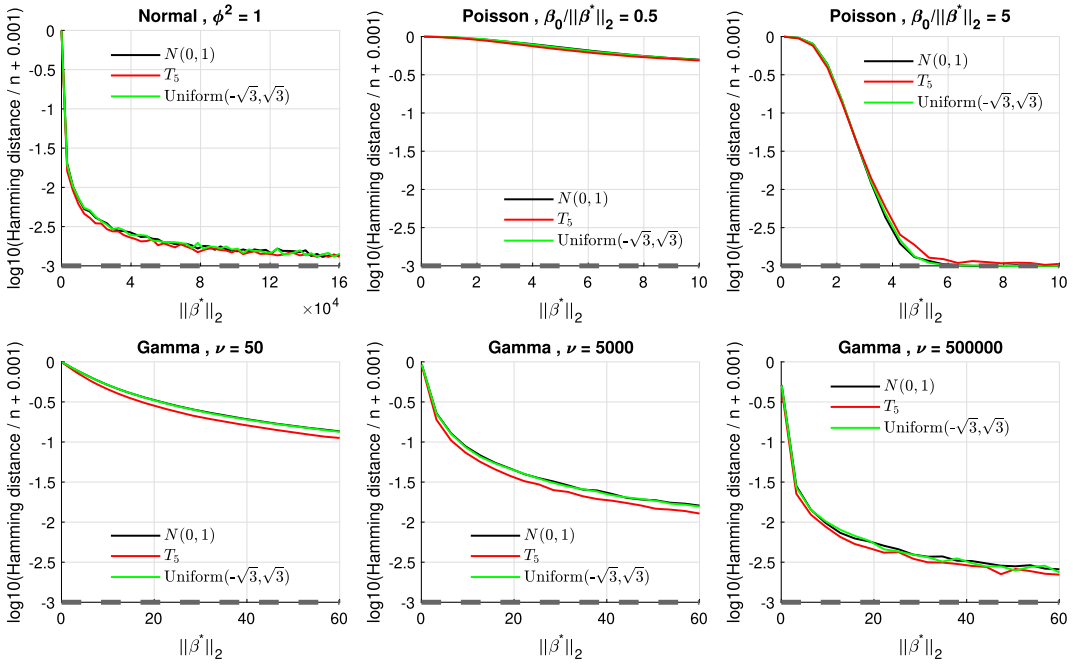

Figure 5. Hamming distance between $\widehat{\pi}(\beta^*)$ and π^* (on a $\log_{10}(\cdot+\epsilon)$ scale with $\epsilon = 0.001$), averaged over 1000 replications. Different curves correspond to different random design.

the y_i is known, and we may thus evaluate

$$p_i^* = \begin{cases} \mathbf{P}_{\pi^*(i)=i}(Y_i \geq y_i) & \text{if } y_i \geq \mu_i \\ \mathbf{P}_{\pi^*(i)=i}(Y_i \leq y_i) & \text{if } y_i \leq \mu_i, \end{cases} \quad i = 1, \ldots, n,$$

where the probability is with respect to the underlying random variables $\{Y_i\}_{i=1}^n$ conditional on $\pi_i^*(i) = i$, $1 \leq i \leq n$. Note that similar to the notion of p-value, small p_i^* can be considered as evidence against $\pi_i^*(i) = i$, $1 \leq i \leq n$. Accordingly, we may estimate S_* as the set of indices corresponding to the k smallest values among the $\{p_i^*\}_{i=1}^n$; alternatively, if k is unknown, we may estimate S_* by $\{i : \ p_i^* \leq \tau\}$ for a threshold $\tau \in (0,1)$.

As an illustration, let us consider linear regression with Gaussian errors. Observe that if $\pi^*(i) = i$, $1 \leq i \leq n$, we have for all $t > 0$

$$\mathbf{P}\left(y_i - \mathbf{x}_i^\top \beta^* > \sigma t\right) = \mathbf{P}\left(y_i - \mathbf{x}_i^\top \beta^* < \sigma t\right) = 1 - \Phi(t),$$

where Φ denotes the CDF of an $N(0,1)$-random variable, hence $p_i^* = 1 - \Phi(|y_i - \mathbf{x}_i^\top \beta^*|/\sigma)$, $1 \leq i \leq n$. In order to fix τ, a natural options is to require that $\sum_{i=1}^n p_i^* \leq \delta$ for $\delta \in (0,1)$ and thus $p_i^* \leq \delta/n$, $1 \leq i \leq n$. Using the standard Gaussian tail bound $1 - \Phi(t) \leq \exp(-t^2/2)$ for $t > 0$ yields that $p_i^* \leq \delta/n$ once $|y_i - \mathbf{x}_i^\top \beta^*| > \sigma\sqrt{2\log(n/\delta)}$, $1 \leq i \leq n$. Accordingly, in order for mismatches i with $\pi^*(i) \neq i$ to be detectable, it is required that $|(\mathbf{x}_{\pi^*(i)} - \mathbf{x}_i)^\top \beta^*| = |\mu_{\pi^*(i)} - \mu_i| \geq 2\sigma\sqrt{2\log(n/\delta)}$, which is almost identical to the requirement in Theorem 2 (a). We conclude that mismatch recovery, i.e., the estimation of S_*, and permutation recovery obey similar regimes.

3.2 Simulation

We complement Theorems 2 and 3 with simulation results. The entries of the design matrix are sampled i.i.d. from three unit-variance distributions: (i) standard Normal, (ii) the uniform distribution on $[-\sqrt{3}, \sqrt{3}]$, and (iii) (rescaled) t-distribution with five degrees of freedom, and responses are subsequently generated according to model (1) based on the Normal distribution with identity link, and the Poisson and Gamma distribution with log-link. The regression parameter β^* is drawn uniformly at random from spheres in dimension d of varying radii $\|\beta^*\|_2$. The intercept is taken as $\beta_0^* = c \cdot \|\beta^*\|_2$ with $c \in \{0.5, 5\}$. The sample size is fixed as $n = 200$. For each configuration consisting of the distribution of the design matrix, the radius $\|\beta^*\|_2$, and the value of β_0^*, 1000 replications are performed. In each replication, we evaluate the normalized Hamming distance $n^{-1}\sum_{i=1}^n \mathbb{I}(i \neq \widehat{\pi}(i))$, where $\widehat{\pi}$ is the minimizer of (8).

In light of Theorem 3, permutation recovery can be achieved if and only if $\|\beta^*\|_2$ is large enough, and specifically for Poisson case, if in addition $\beta_0^*/\|\beta^*\|_2$ exceeds a certain threshold. Figure 5 confirms this qualitatively. In particular, we observe that the ratio $\beta_0^*/\|\beta^*\|_2$ is crucial in the Poisson case unlike the other cases. Furthermore, we note that the distribution of the design does not have a significant impact on the results: all three random design can achieve (at least approximate) permutation recovery given sufficient signal as

quantified by $\|\beta^*\|_2$. Finally, note that for the Gamma distribution, recovery results improve as the shape parameter ν increases.

4. COMPARISON TO THE LAHIRI-LARSEN & CHAMBERS METHODS

In this section, as a follow-up to the simulation study in §2.5, we aim to provide a short comparison of the proposed method and an established method whose prototype was proposed in [8] and later extended to a wider class of regression models including generalized linear models [10], and will hence be referred to as the Lahiri-Larsen (LL) method. A closely related approach is due to Chambers [9, 26, 11]. The former turns out to be somewhat easier to analyze in the block-structured permutation setting outlined in §2.3 that will also be adopted in the sequel. The ultimate goal of our discussion is to delineate scenarios in which the proposed estimator tends to be superior and inferior, respectively, relative to the LL method. Both methods differ noticeably with regard to the assumptions on π^* and the required amount of knowledge about the linkage process, and the regimes of interest differ accordingly.

In a nutshell, the LL method assumes that $\mathbf{y} = \Pi^*\mathbf{y}^*$ with $\mathbf{y}_j^*|\mathbf{x}_j$, $1 \le j \le N$, following a GLM as specified in §1.1 and Π^* being a generalized random permutation matrix associated with the map $\pi^* : \{1, \dots, N\} \to \{1, \dots, n\}$ whose (i, j)-th entry equals one if $\pi^*(i) = j$ and zero otherwise, $1 \le i \le n$, $1 \le j \le N$. The LL method further assumes that Π^* is conditionally independent of $\mathbf{y}^*$ given $\{\mathbf{x}_j\}_{j=1}^N$ and that the corresponding conditional expectation of Π^* is given by $\mathbf{Q} \in \mathbb{R}^{n\times N}$. Let $\mathbf{X}_N$ denote the design matrix associated with the full set of covariates $\{\mathbf{x}_i\}_{i=1}^N$[6]. Equipped with the above assumptions, it is readily shown that

$$\begin{aligned} &\mathbf{X}_N^\top \mathbf{Q}^\top (\mathbf{y} - \mathbf{Q}\boldsymbol{\mu}^*(\beta)) = \mathbf{0} \;\Leftrightarrow \\ &\mathbf{X}_N^\top \mathbf{Q}^\top (\Pi^*\mathbf{y}^* - \mathbf{Q}\boldsymbol{\mu}^*(\beta)) = \mathbf{0}, \\ &\boldsymbol{\mu}^*(\beta) := \big(\psi'(\mathbf{x}_j^\top \beta)\big)_{j=1}^N \end{aligned} \tag{9}$$

is an unbiased estimating equation in the sense that

$$\begin{aligned} &\mathbf{E}_{\Pi^*,\mathbf{y}^*}[\mathbf{X}_N^\top \mathbf{Q}^\top (\Pi^*\mathbf{y}^* - \mathbf{Q}\boldsymbol{\mu}^*(\beta^*))] \\ &= \mathbf{E}_{\mathbf{y}^*}[\mathbf{X}_N^\top \mathbf{Q}^\top \mathbf{Q}(\mathbf{y}^* - \boldsymbol{\mu}(\beta^*))] = \mathbf{0}. \end{aligned}$$

The estimator is particularly easy to understand in the setting in which π^* is a block-structured permutation ($N = n$) as discussed in §2.3 and the additional assumption that for each block G_j, the corresponding permutation is chosen uniformly at random. Without loss of generality, let $\Pi^* = \mathrm{bdiag}(\Pi_1^*, \dots, \Pi_K^*)$ be the block diagonal matrix associated with π^*. It then follows that

$$\mathbf{Q} = \mathrm{bdiag}(\mathbb{1}_{n_1}, \dots, \mathbb{1}_{n_K}), \tag{10}$$

where for any integer m, the symbol $\mathbb{1}_m$ denotes an m-by-m matrix of ones, multiplied by $1/m$. Since each matrix block matrix is a projection (averaging operator), $\mathbf{Q} = \mathbf{Q}^\top = \mathbf{Q}^2$ is a projection as well. Note that even if the block-wise permutations are not chosen uniformly at random, we may still use (10) for $\mathbf{Q}$ in (9) to obtain an unbiased estimating equation since $\mathbf{Q}^\top\Pi^* = \mathbf{Q}$ for *any* permutation matrix Π^* block-structured as $\mathbf{Q}$. Letting $\widehat{\beta}^{\mathrm{LL}}$ denote a solution of this estimating equation, asymptotic theory implies that $\widehat{\beta}^{\mathrm{LL}}$ is a consistent estimator of β^* with asymptotic covariance matrix

$$\mathbf{E}_{\mathbf{y}}[J_\Gamma(\beta^*)]^{-1}\ \mathrm{Cov}_{\mathbf{y}}(\Gamma(\beta^*))\ \mathbf{E}_{\mathbf{y}}[J_\Gamma(\beta^*)^\top]^{-1}, \tag{11}$$

where $\Gamma : \mathbb{R}^d \to \mathbb{R}^d$ denotes the function defining the estimating equation, and J_Γ denotes the Jacobian of Γ. Straightforward calculations (cf. supplement for a derivation) yield that

$$\begin{aligned} &\mathbf{E}_{\mathbf{y}}[J_\Gamma(\beta^*)] = \mathbf{X}^\top\mathbf{Q}\mathbf{V}(\beta^*)\mathbf{X},\ \mathrm{Cov}_{\mathbf{y}}(\Gamma(\beta^*)) \\ &= \mathbf{X}^\top\mathbf{Q}\mathbf{V}(\beta^*)\mathbf{Q}\mathbf{X}, \end{aligned} \tag{12}$$

where $\mathbf{V}(\beta^*)$ is a diagonal matrix whose diagonal entries are given by the variances of y_i^* given by $\psi''(\mathbf{x}_i^\top\beta^*)$, $1 \le i \le n$. Note that for the inverse in (11) to exist, it is necessary that the rank of the first matrix in (12) is at least d, which is generally the case implied by the condition $K \ge d$.

In the sequel, we shall argue that for a wide range of random designs, the entries of (11) are of the order $O_{\mathbf{P}}(K^{-1})$, i.e., the estimator $\widehat{\beta}^{\mathrm{LL}}$ converges asymptotically at the same rate, and that rate will be recovered exactly for linear regression with Gaussian errors.

For this purpose, observe first that the rows of $\mathbf{QX}$ are given by n_j replications of $\overline{\mathbf{x}}_j = \frac{1}{n_j}\sum_{i\in G_j}\mathbf{x}_i$, $1 \le j \le K$. Accordingly, we have

$$\mathrm{Cov}_{\mathbf{y}}(\Gamma(\beta^*)) = \sum_{j=1}^K \sum_{i\in G_j} \psi''(\mathbf{x}_i^\top\beta^*)\,\overline{\mathbf{x}}_j\overline{\mathbf{x}}_j^\top,$$

which scales as $O_{\mathbf{P}}(K)$ for a wide range of random designs that satisfy (i) $\overline{\mathbf{x}}_j = O_{\mathbf{P}}(n_j^{-1/2})$, $1 \le j \le K$, and (ii) $\psi''(\mathbf{x}_i^\top\beta^*) = O_{\mathbf{P}}(1)$, $1 \le i \le n$. With a similar reasoning, one also obtains $\mathbf{E}_{\mathbf{y}}[J_\Gamma(\beta^*)] = O_{\mathbf{P}}(K)$ and thus in combination $O_{\mathbf{P}}(K^{-1})$ for the asymptotic covariance (11).

For linear regression with Gaussian errors, the same order additionally holds in a non-asymptotic fashion. Note that in this case, the underlying estimation equation has the closed form solution $\widehat{\beta}^{\mathrm{LL}} = (\mathbf{X}^\top\mathbf{Q}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Q}\mathbf{y}$ and hence $\mathrm{Cov}(\widehat{\beta}^{\mathrm{LL}}) = \phi^2(\mathbf{X}^\top\mathbf{Q}\mathbf{X})^{-1} = \phi^2\left(\sum_{j=1}^K n_j\overline{\mathbf{x}}_j\overline{\mathbf{x}}_j^\top\right)^{-1}$. The same expression is obtained from (11) and (12) by using $\mathbf{V}(\beta^*) = \phi^2 I_d$ and the fact that $\mathbf{Q} = \mathbf{Q}^2$.

While the above discussion does not provide a comprehensive analysis of the estimator $\widehat{\beta}^{\mathrm{LL}}$ given the focus on a specific choice of $\mathbf{Q}$ in the setting of a block-structured

[6] Relevant to the sample-to-register linkage setting only, cf. §1.1. Note that $\mathbf{X}_N = \mathbf{X}$ if $N = n$.

permutation, we feel that this very choice is presumably among the most relevant in practice. In their landmark paper [8], Lahiri and Larsen consider the entries of $\mathbf{Q}$ being taken as the probability that observation i in file $F_{\mathbf{x}}$ and observation j in File $F_{\mathbf{y}}$ are a match, $1 \leq i, j \leq n$, computed from the Fellegi-Sunter model [2] given a set of comparison variables. However, it is generally not guaranteed that this choice is misspecified. In addition, the LL method only asserts unbiasedness when averaging over random Π^*, whereas in practice, the data analyst has to deal with a merged data set arising from a single realization of Π^*. Choosing $\mathbf{Q}$ according to (10) based on a uniform-at-random model within blocks avoids these issues, producing an unbiased estimator from the comparison of blocking variables only.

In light of the above findings and discussions, we summarize advantages and disadvantages of the estimator $\widehat{\beta}^{\mathrm{LL}}$ in relation to the estimator (5) proposed herein.

Advantages.

- The estimator $\widehat{\beta}^{\mathrm{LL}}$ does not rely on sparsely occuring mismatches with k/n being bounded by a (small) constant. In fact, $\widehat{\beta}^{\mathrm{LL}}$ can tolerate a fraction of mismatches close to one, and is still guaranteed to be unbiased.
- The improvement of $\widehat{\beta}^{\mathrm{LL}}$ over the naive estimator is less tied to the "signal strength" in terms of $\|\beta^*\|_2$.
- Asymptotic confidence intervals can be obtained from the expression for the asymptotic covariance (11).
- The approach is free of tuning parameters.

Disadvantages.

- The approach can suffer from a high variance, and is generally not guaranteed to be consistent as the sample size n grows. Instead, consistency requires that the number of blocks $K \to \infty$. Even if the latter holds true, the asymptotic rate of convergence $O_{\mathbf{P}}(K^{-1})$ is suboptimal unless $n/K = O(1)$.
- In view of the previous bullet, $\widehat{\beta}^{\mathrm{LL}}$ hinges on the availability of additional information that gives rise to a sufficiently fine block partitioning (i.e., consisting of a good number of blocks).
- The approach does not account for potential errors in variables used to generate the block partitioning (cf., e.g., [61]).
- The estimation equation (9) is not guaranteed to have a unique root, and thus practical algorithms for its solution such as Newton's method may deliver roots that are not consistent. In some situations, no roots may exist (cf. supplement).

Chambers' method. For the sake of completeness, we present a brief account of the approach due to Chambers in relation to the LL method. Chambers' method is based on the estimating equation

$$\mathbf{X}^\top(\mathbf{y} - \mathbf{Q}\boldsymbol{\mu}^*(\beta)) = \mathbf{0}, \tag{13}$$

which differs from the LL estimation equation (9) only in that $\mathbf{QX}$ is replaced by $\mathbf{X}$. Estimation equation (13) is unbiased, but the resulting estimator tends to converge at a slower rate than the LL estimator in the block setting discussed above. Using similar arguments as before (cf. supplement), it can be shown that the asymptotic covariance of the Chambers estimator is given by

$$(\mathbf{X}^\top\mathbf{Q}^\top\mathbf{V}(\beta^*)\mathbf{X})^{-1} \times (\mathbf{E}_{\Pi^*}[\mathbf{X}^\top\Pi^*\mathbf{V}(\beta^*)\Pi^{*\top}\mathbf{X}] + \boldsymbol{\Xi}) \times \\ \times (\mathbf{X}^\top\mathbf{V}(\beta^*)\mathbf{QX})^{-1}, \tag{14}$$

for some positive semidefinite matrix $\boldsymbol{\Xi}$, and all other quantities are as above. The slower rate of convergence results from the middle matrix. For any permutation matrix Π^*, we have

$$(\mathbf{X}^\top\Pi^*\mathbf{V}(\beta^*)\Pi^{*\top}\mathbf{X}) = \sum_{i=1}^{n} \mathbf{x}_{\pi^*(i)}\psi''(\mathbf{x}_i^\top\beta^*)\mathbf{x}_{\pi^*(i)}^\top = O_{\mathbf{P}}(n)$$

for typical random designs. The two outer matrices in (14) are the same as in the expression for the asymptotic covariance of the LL method as provided in (11) and (12). Using the results for the LL method above then yields that altogether (14) scales as $O_{\mathbf{P}}(n/K^2)$, which is generally slower than the rate $O_{\mathbf{P}}(K^{-1})$ of the LL method.

Finally, Chambers [9] (cf. also [26]) discusses an improvement of the LL method to improve its statistical efficiency based on the observation that

$$\begin{aligned}\mathrm{Cov}(\mathbf{y}) &= \mathrm{Cov}(\Pi^*\mathbf{y}^*) \\ &= \mathbf{E}_{\Pi^*}[\mathrm{Cov}(\Pi^*\mathbf{y}^*|\Pi^*)] + \mathrm{Cov}_{\Pi^*}(\mathbf{E}[\Pi^*\mathbf{y}^*|\Pi^*]) \\ &= \mathrm{Cov}(\mathbf{y}^*) + \mathrm{Cov}_{\Pi^*}(\Pi^*\boldsymbol{\mu}^*(\beta^*)),\end{aligned}$$

which suggests that the second term on the right hand side, which varies across different blocks, should be integrated into the estimating equations to weigh blocks according to their variance contribution (in the case of linear regression, this amounts to a specific generalized least squares fit). However, even if $\mathrm{Cov}(\mathbf{y})$ were known, the above improvement can apparently not achieve a faster rate than $O_{\mathbf{P}}(K^{-1})$.

5. CASE STUDY

We consider the bike sharing data available on the UCI machine learning repository [29]. The data set contains seasonal and weather information along with daily counts of rental bikes used between 2011 and 2012 in Washington, DC. The objective is to predict the number of ride-sharing bikes used during any given day (variable `count`) based on the categorical predictor variables `season` (spring, summer, fall, winter), `year` (2011 or 2012), `weathersit` (describing the weather situation on a given day in four categories from good to highly inclement weather), `weekday` (Monday through Sunday, numbered 0 to 6) and `workingday` (binary

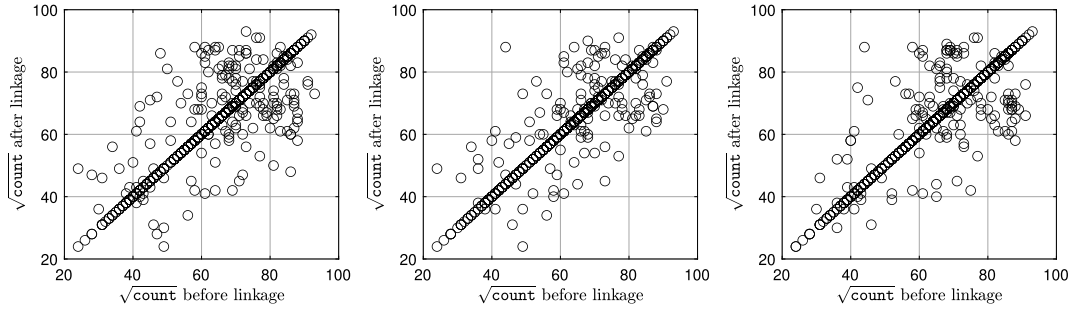

Figure 6. Scatterplots of the response variable before and after linkage for selected permutations.

variable indicating whether a given day is a working day as opposed to a holiday or Saturday/Sunday), as well as the three continuous predictor variables `atemp` (feeled temperature), `hum` (a humidity index) and `wind` (windspeed).

Overall, the data set consists of 731 instances (days). We apply a square root transformation to the response variable `count`. The transformed response variable is treated as if it followed a Poisson GLM with log-link, which can formally be regarded as a quasi-likelihood approach. The use of the transformation yields substantial improvement in terms of model fit compared to a Poisson model in which the raw counts are used as the response variable.

In order to further improve the fit of the model, we delete observations satisfying one of the following criteria: (1) days affected by an extreme weather condition (e.g., blizzard, hurricane or windstorm), (2) national holidays (including Thanksgiving and Christmas), (3) especially hot days with temperatures exceeding 31.8 degrees Celsius. The resulting thinned data set includes 692 instances that are used to fit the following (Poisson-like) regression model:

(15)
$$\begin{aligned}\log(\sqrt{\texttt{count}}) &= \beta_0^* + \textstyle\sum_{j=2}^4 \beta_{\texttt{s}j}^* \cdot \mathbb{I}(\texttt{season} = j) + \beta_{\texttt{y}}^* \cdot \texttt{year} \\ &+ \beta_{\texttt{wor}}^* \cdot \texttt{workingday} + \beta_{\texttt{a}}^* \cdot \texttt{atemp} + \beta_{\texttt{hum}}^* \cdot \texttt{hum} + \beta_{\texttt{wi}}^* \cdot \texttt{wind} \\ &+ \beta_{\texttt{y*a}}^* \cdot (\texttt{yr} * \texttt{atemp}) + \beta_{\texttt{we}}^* \cdot \mathbb{I}(\texttt{weekday} \in \{4,5,6\}) \\ &+ \textstyle\sum_{j=2}^4 \beta_{\texttt{wea}j}^* \cdot \mathbb{I}(\texttt{weathersit} = j) \\ &+ \textstyle\sum_{j=2}^4 \beta_{\texttt{s}j\texttt{*a}}^* \cdot \big(\mathbb{I}(\texttt{season} = j) * \texttt{atemp}\big)\end{aligned}$$

In total, the linear predictor consists of 17 terms apart from the intercept including interaction term between `season` and `atemp` and between `year` and `atemp`.

In order to mimic mismatch error introduced by record linkage, the response variable (`count`) is put into a separate file that additionally contains varying combinations of variables that are used for record linkage (see below for a list of those combinations). One-to-one linkage is enforced by breaking ties between potentially matching records given the variables used for matching uniformly at random; in order to account for that randomness, we consider 100 independent replications, and the results reported subsequently represent averages over those replications unless noted otherwise. Put differently, the underlying random permutations Π^* are of the block form $\Pi^* = \mathrm{bdiag}(\Pi_1^*, \ldots, \Pi_K^*)$, where K denotes the number of unique combinations of values assumed for the matching variables. The average fraction of mismatches k/n obtained in this way approximately equals 0.23. Figure 6 illustrates the discrepancy between actual response and response after linkage based on the following combination of matching variables: `month`, `holiday`, `weekday`, `workingday`, `temp`[7]. In practice, the combination of variables used for matching may not be (fully) disclosed to the data analyst that operates on the merged data. Therefore, in addition to the case in which the matching variables are fully known, we also consider cases in which only one or two of the matching variables out of five overall are known (cf. Table 1). Note that here, none of the matching variables contain any sensitive information; the omission of subsets is intended to demonstrate the impact of less knowledge about the linkage process at a conceptual level.

Table 1. List of matching variables provided to the data analyst (the combination of five being the full list), the resulting number of blocks K and the symbols for the associated constraint matrices $\mathbf{C}$ used for the proposed estimator "Proposed$^{\mathbf{C}}$" in its constrained form (6) as well as the symbols for the associated matrices $\mathbf{Q}$ used for the methods of Lahiri-Larsen and Chambers.

matching variables known	K	(6)	(10)
`month`, `holiday`, `weekday` `workingday`, `temp`	535	$\mathbf{C}^1$	$\mathbf{Q}^1$
`month`, `temp`	167	$\mathbf{C}^2$	$\mathbf{Q}^2$
`temp`	34	$\mathbf{C}^3$	$\mathbf{Q}^3$

Results. In order to assess the performance of the proposed approach and the baseline competitors discussed in §4, estimation of the regression parameters of model (15) is performed based on the merged file contaminated by mismatch error. The resulting parameter estimates are then used to

[7] `temp` denotes the temperature in degree Celsius on a given day (integer-valued).

evaluate the deviance on the file $\{(\mathbf{x}_i, y_i^*)\}_{i=1}^n$ containing predictors and response in their correct correspondence, i.e.,

$$2\sum_{i=1}^{n}\left[y_i^*\log\left(\frac{y_i^*}{\mu_i^{\text{est}}}\right)-(y_i^*-\mu_i^{\text{est}})\right], \quad \text{where } \mu_i^{\text{test}}=\exp(\beta_0^{\text{est}}+\mathbf{x}_i^\top\beta^{\text{est}}), \tag{16}$$

where β^{est} is a placeholder for the estimates based on the merged file $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ delivered by i) the proposed method, ii) the methods of Lahiri-Larsen (LL) and Chambers, and (iii) the naive estimator without any adjustment for mismatches. Both i) and ii) are evaluated for full and only partial knowledge of the matching variables. The ultimate reference for the quantity (16) is obtained by substituting $\beta^{\text{est}}=\widehat{\beta}^{\text{oracle}}$, where the oracle refers to the estimator based on the file $\{(\mathbf{x}_i, y_i^*)\}_{i=1}^n$ without any mismatched pairs. The quantity (16) is hence intended to measure the drop in model fit that is induced by parameter estimates in the presence of mismatch error.

Specific figures regarding (16) are presented in Table 2. Note that in case that all matching variables are known to the data analyst, the performance of the LL estimator is rather close to the oracle and better than that of the proposed estimator (6) in its constrained form (265.46 vs. 269.92 with a standard error of 0.21). However, as less information about matching variables is available, the proposed estimator performs on par (2nd column) respectively dramatically better (3rd column) than the LL estimator; in the latter case, the performance of the LL estimator is even considerably worse than that of the naive estimator. This is somewhat in alignment with what is predicted in §4: as the number of blocks K drops, so does the performance of the LL estimator. By contrast, the proposed estimator achieves a solid improvement over the naive estimator even in the complete absence of information about the matching variables used. The Chambers estimator performs weaker than the proposed and the LL estimator if the full set of matching variables is provided; in the two other cases with less information, we experience numerical difficulties (hence the value NA): Newton iterations used to obtain a root of the estimating equations (13) converge to far suboptimal points leading to a deviance exceeding that of the intercept-only model.

Regarding the selection of the tuning parameter λ for the proposed approach, we create a separate validation set free of mismatches whose size is 20% of the total number of samples, and select λ so as to minimize the counterpart to (16) on the validation set. Note that the existence of such a validation set is reasonably realistic, at least if sufficient information about matching variables is provided and at least some of the resulting combinations are unique, i.e., they yield singleton blocks for which mismatches can be ruled out. The corresponding results based on this data-driven selection of λ are labelled "Proposed(λ)" in Table 2, to be distinguished from "Proposed(oracle)" in which λ is optimized to minimize the performance measure (16) directly. We note that the performance of "Proposed(λ)" is only slightly inferior to that of "Proposed(oracle)".

Table 2. Deviances (16) for several competitors as described in the text, averaged over 100 random block-structured permutations (the corresponding standard errors are given in parentheses[8]). "Proposed(oracle)" and "Proposed(oracle)$^{\mathbf{C}}$" refer to the choice of λ that directly minimizes (16), while "Proposed(λ)" and "Proposed(λ)$^{\mathbf{C}}$" refer to the choice of λ based on a validation set; the superscript $\mathbf{C}$ indicates the use of constraints (6) with $\mathbf{C}\in\{\mathbf{C}^j\}_{j=1}^3$.

	$(\mathbf{Q}^1,\mathbf{C}^1)$	$(\mathbf{Q}^2,\mathbf{C}^2)$	$(\mathbf{Q}^3,\mathbf{C}^3)$
Oracle		263.40	
Lahiri-Larsen	265.46	274.88	436.68
Chambers	272.70(0.40)	NA	NA
Proposed(Oracle)		282.63(0.39)	
Proposed$^{\mathbf{C}}$(Oracle)	269.92(0.21)	275.00(0.29)	278.34(0.32)
Proposed(λ)		285.04(0.48)	
Proposed$^{\mathbf{C}}$(λ)	270.98(0.36)	276.29(0.36)	281.74(0.52)
Naive		316.86(1.07)	
Intercept only		2540.44	

Permutation Recovery. In this paragraph, we describe how the proposed approach can be leveraged to reduce mismatch error in the response variable contained in the merged file. We henceforth suppose that the data analyst is equipped with full knowledge of the matching variables `month`, `holiday`, `weekday`, `workingday`, `temp` used during the creation of the merged file. For ease of presentation, we here confine ourselves to two specific random permutations $\Pi_{\min}$ and $\Pi_{\max}$ out of the ensemble $\{\Pi^{(1)},\ldots,\Pi^{(100)}\}$ that were generated, defined by

$$\Pi_{\min}=\min_{1\le i\le N}\|\Pi^{(i)}\mathbf{y}^*-\mathbf{y}^*\|_2,\ \Pi_{\max}=\max_{1\le i\le N}\|\Pi^{(i)}\mathbf{y}^*-\mathbf{y}^*\|_2,$$

representing a best and a worst case scenario, respectively. Given $(\mathbf{X},\Pi_{\min}\mathbf{y}^*)$ and $(\mathbf{X},\Pi_{\max}\mathbf{y}^*)$, we compute the proposed estimator with constraint matrix $\mathbf{C}=\mathbf{C}^1$ (cf. Table 2 top) and use the resulting solution $\widehat{\beta}$ in place of β^* in the optimization problem defining the maximum likelihood estimator for the unknown permutation (8). The resulting optimization problem is given by

$$\min_{\Pi\in\mathcal{P}(\mathcal{G})}-\langle\Pi\mathbf{y},\mathbf{X}\widehat{\beta}\rangle, \tag{17}$$

where $\mathbf{y}=\Pi_{\min}\mathbf{y}^*$ and $\mathbf{y}=\Pi_{\max}\mathbf{y}^*$, respectively, and $\mathcal{P}(\mathcal{G})$ denotes the set of all block-wise permutation matrices induced by the resulting index subsets $\mathcal{G}=\{G_j\}_{j=1}^K$, $K=535$, corresponding to identical values for the matching variables. Note that the minimizer of (17) can be obtained by pairing the order statistics of the linear predictor and the response

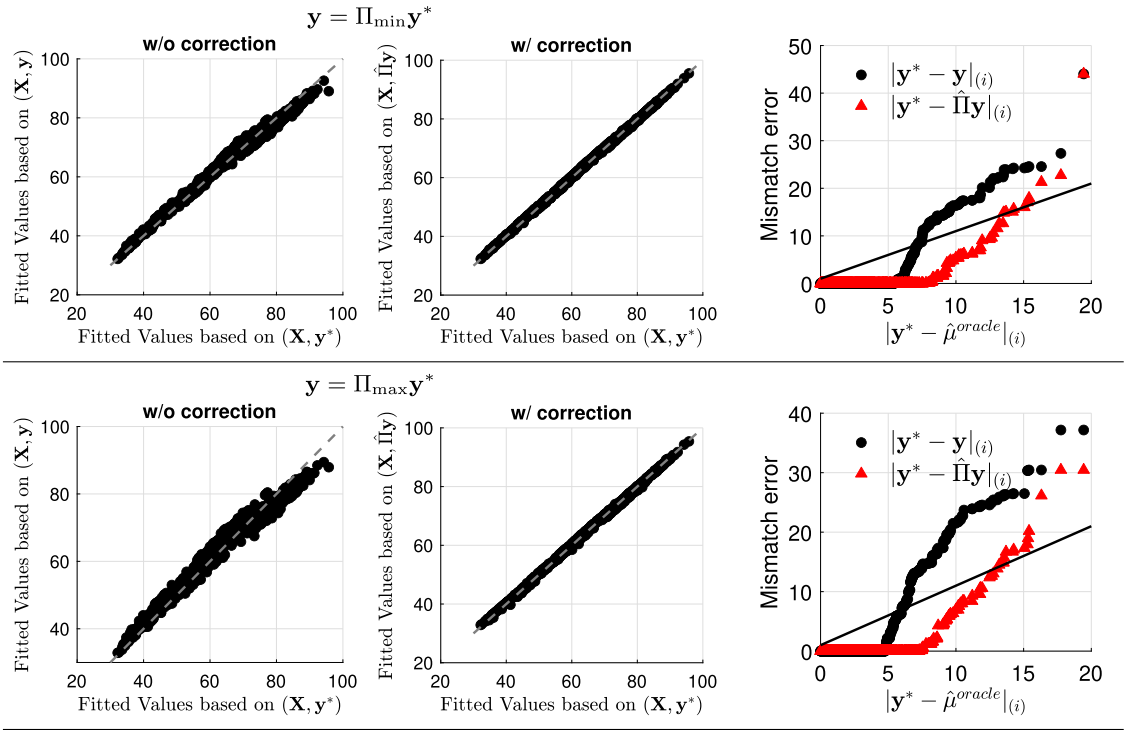

Figure 7. Left/Middle: Fitted values based on mismatch-free data $(\mathbf{X}, \mathbf{y}^*)$ vs. fitted values based on the merged data $(\mathbf{X}, \mathbf{y})$ [left] and corrected data $(\mathbf{X}, \widehat{\Pi}\mathbf{y})$ [middle], with $\widehat{\Pi}$ denoting the minimizer of (17); "fitted values" here refer to ordinary (Quasi-) GLM estimation based on the data given in parentheses. Right: Q-Q plots of the absolute differences between the true responses and their fitted values based on the oracle estimator vs. the absolute mismatch errors in the merged file $\{|y_i^* - y_i|\}_{i=1}^n$ [dots] and their counterparts after correction based on (17) [triangles].

within each of the sets $\{G_j\}_{j=1}^K$. While permutation recovery, i.e., $\{\widehat{\Pi} = \Pi_{\min}\}$ and $\{\widehat{\Pi} = \Pi_{\max}\}$, respectively, where $\widehat{\Pi}$ denotes the minimizer of (17) turns out to be out of reach here, a substantial reduction in mismatch error is achieved, i.e., $\widehat{\Pi}\mathbf{y}$ is visibly closer to $\mathbf{y}^*$ than $\mathbf{y}$ as shown in Figure 7. The corrected response $\widehat{\Pi}\mathbf{y}$ can be used to refit the regression model. Figure 7 indicates that the resulting fitted values exhibit a much better agreement with the fitted values obtained from a mismatch-free data set.

6. CONCLUSION

In this paper, we have presented a method based on ℓ_1-penalization to account for mismatch error in the response in linked files, and have highlighted its benefits compared to established methods for this scenario. We have also explored how to directly reduce mismatch error by estimating the underlying permutation associated with the true correspondence between predictor-response pairs. The proposed approach is computationally appealing, supported by theoretical guarantees, and bears considerable potential regarding the adjustment for mismatch error in post-linkage analysis. At the same time, the approach presented herein prompts several directions of future research. Concerning the estimation of the regression parameter, procedures for a data-driven choice of the tuning parameter, such as with the help of a hold-out set (possibly also contaminated by mismatch error) as considered in the case study herein, merit further investigation. In a next step, it also appears worthwhile exploring the use of observation-specific penalization factors $\{\lambda_i\}_{i=1}^n$ instead of a single global value λ that could prove particularly beneficial in Poisson and Gamma regression in light of heteroscedasticity. Moreover, it is of great practical relevance to be able to conduct statistical inference for the regression parameter (confidence intervals and tests of linear hypotheses). A promising approach with regard to this aspect is the use of techniques developed for constructing confidence intervals in lasso regression, e.g., [62, 63, 64].

Concerning estimation of the permutation, we have focused on exact permutation recovery and on approximate recovery with small Hamming distance. The results are coupled to stringent conditions, which are often not met in practice. Nevertheless, the results presented in the case study of the previous section indicate that approximate permutation

[8]The oracle estimators and the Lahiri-Larsen method do not differ across permutations by construction, hence no standard error is reported.

recovery can be achieved with respect to alternative metrics like the ℓ_2-distance between the true response $\mathbf{y}^*$ and the estimator $\widehat{\Pi}\mathbf{y}$, and elaborating the corresponding theory constitutes a further promising direction.

SUPPLEMENT

Online supplement (http://intlpress.com/site/pub/files/_supp/sii/2023/0016/0003/SII-2023-0016-0003-s001.pdf) is organized as follows. Sections A through D contain complete proofs of the results stated in the paper. Sections E through G contain complementary technical discussions. Section H contains additional simulation results and figures complementing those in §2.5.

ACKNOWLEDGEMENTS

The first author and the last author was partially supported by the NSF Grant CCF-1849876 and NSF-SES 2120-318. The authors would like to thank the reviewers for their thoughtful and encouraging comments that have led to numerous improvements over an earlier draft.

Received 19 February 2021

REFERENCES

[1] Newcombe,H. and Kennedy,J. Record linkage: making maximum use of the discriminating power of identifying information. *Communications of the ACM*, 5(11):563–566, 1962.

[2] Fellegi,I. P. and Sunter,A. B. A theory for record linkage. *Journal of the American Statistical Association*, 64:1183–1210, 1969.

[3] Christen,P. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer, 2012.

[4] Herzog,T., Scheuren,F. and Winkler,W. *Data quality and record linkage techniques*. Springer, 2007. MR2654644

[5] Neter,J., Maynes,S. and Ramanathan,R. The effect of mismatching on the measurement of response error. *Journal of the American Statistical Association*, 60:1005–1027, 1965. MR0193729

[6] Scheuren,F. and Winkler,W. Regression analysis of data files that are computer matched I. *Survey Methodology*, 19:39–58, 1993.

[7] Scheuren,F. and Winkler,W. Regression analysis of data files that are computer matched II. *Survey Methodology*, 23:157–165, 12 1997.

[8] Lahiri,P. and Larsen,M. D. Regression analysis with linked data. *Journal of the American Statistical Association*, 100(469): 222–230, 2005. MR2156832

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

[10] Han,Y. and Lahiri,P. Statistical analysis with linked data. *International Statistical Review*, 87:139–157, 2019. MR3957348

[11] Kim,G. and Chambers,R. Regression Analysis under incomplete linkage. *Computational Statistics and Data Analysis*, 56:2756–2770, 2012. MR2915160

[12] DeGroot,M. and Goel,P. Estimation of the correlation coefficient from a broken random sample. *The Annals of Statistics*, 8: 264–278, 1980. MR0560728

[13] DeGroot,M. and Goel,P. The matching problem for multivariate normal data. *Sankhya, Series B*, 38:14–29, 1976. MR0652955

[14] Unnikrishnan,J., Haghighatshoar,S. and Vetterli,M. Unlabeled sensing with random linear measurements. *IEEE Transactions on Information Theory*, 64:3237–3253, 2018. MR3798374

[15] Wainwright,M. J. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. doi: 10.1017/9781108627771. MR3967104

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

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

[18] Slawski,M., Rahmani,M. and Li,P. 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.

[19] Zhang,H., Slawski,M., and Li,P. Permutation recovery from multiple measurement vectors in unlabeled sensing. In *IEEE International Symposium on Information Theory (ISIT)*, 2019.

[20] Tsakiris,M., Peng,L., Conca,A., Kneip,L., Shi,Y., and Choi,H. An algebraic-geometric approach to shuffled linear regression. to appear in IEEE Transactions on Information Theory, 2020. MR4130666

[21] Tsakiris,M. Eigenspace conditions for homomorphic sensing. arXiv:1812.07966, December 2018.

[22] McCullagh,P. and Nelder,J. *Generalized Linear Models*. Chapman and Hall, London, 1989. MR3223057

[23] Slawski,M. and Ben-David,E. Linear regression with sparsely permuted data. *Electronic Journal of Statistics*, 13:1–36, 2019. MR3896144

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

[25] Enamorado,T. T., Fifield,B. and Imai,K. Using a probabilistic model to assist merging of large-scale administrative records. *American Political Science Review*, 113:353–371, 2019.

[26] Chambers,R. and Diniz da Silva,A. 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. MR4049654

[27] Pananjady,A., Wainwright,M. J. and Courtade,T. A. Denoising linear models with permuted data. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 446–450. IEEE, 2017.

[28] Negahban,S. N., Ravikumar,P. Wainwright,M. J., Yu,B., et al. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27 (4):538–557, 2012. MR3025133

[29] Fanaee-T,H. and Gama,J. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, 2(2-3):113–127, 2014.

[30] Pananjady,A., Wainwright,M., and Cortade,T. Linear regression with shuffled data: Statistical and computational limits of permutation recovery. *IEEE Transactions on Information Theory*, 3826–3300, 2018. MR3798377

[31] Emiya,V., Bonnefoy,A., Daudet,L. and Gribonval,R. Compressed sensing with unknown sensor permutation. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1040–1044. IEEE, 2014.

[32] Peng,L. and Tsakiris,M. C. Linear regression without correspondences via concave minimization. *IEEE Signal Processing Letters*, 27:1580–1584, 2020.

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

[34] Gutman,R., Afendulis,C. and Zaslavsky,A. A Bayesian Procedure for File Linking to Analyze End-of-Life Medical Costs. *Journal of the American Statistical Association*, 108:34–47, 2013. MR3174601

[35] Slawski,M., Ben-David,E. and Li,P. Two-stage approach to

multivariate linear regression with sparsely mismatched data. *J. Mach. Learn. Res.*, 21(204):1–42, 2020. MR4209490
[36] Slawski,M., Diao,G. and Ben-David,E. A pseudo-likelihood approach to linear regression with partially shuffled data. *Journal of Computational and Graphical Statistics*, 30(4):991–1003, 2021. MR4356600
[37] Shi,X., Li,X. and Cai,T. Spherical regression under mismatch corruption with application to automated knowledge translation. *Journal of the American Statistical Association*, 116(536):1953–1964, 2021. MR4353725
[38] Carpentier,A. and Schlüter,T. Learning relationships between data obtained independently. In *Proceedings of the International Conference on Artifical Intelligence and Statistics (AISTATS)*, pages 658–666, 2016.
[39] Rigollet,P. and Weed,J. Uncoupled isotonic regression via minimum Wasserstein deconvolution. *Information and Inference*, 8: 691–717, 2019. MR4045481
[40] Flammarion,N., Mao,C. and Rigollet,P. Optimal Rates of Statistical Seriation. *Bernoulli*, 25:623–653, 2019. MR3892331
[41] Ma,R., Cai,T. T. and Li,H. Optimal permutation recovery in permuted monotone matrix model. *Journal of the American Statistical Association*, 116(535):1358–1372, 2021. MR4309278
[42] Balabdaoui,F., Doss,C. R. and Durot,C. Unlinked monotone regression. *Journal of Machine Learning Research*, 22:172, 2021. MR4318528
[43] Wang,G., Zhu,J., Blum,R. S., Willett,P., Marano,S., Matta,V. and Braca,P. Signal amplitude estimation and detection from unlabeled binary quantized samples. *IEEE Transactions on Signal Processing*, 66(16):4291–4303, 2018. MR3859099
[44] She,Y. and Owen,A. Outlier Detection Using Nonconvex Penalized Regression. *Journal of the American Statistical Association*, 106:626–639, 2012. MR2847975
[45] Tibshirani,J. and Manning,C. D. Robust logistic regression using shift parameters (long version). *arXiv preprint arXiv:1305.4987*, 2013. MR2634010
[46] Yang,E., Tewari,A. and Ravikumar,P. On robust estimation of high dimensional generalized linear models. In *International Joint Conference on Artificial Intelligence (IJCAI)*, volume 13, 2013.
[47] Zhang,L. C. *Analysis of Integrated Data*, chapter 2. CRC Press, 2019.
[48] Abowd,J., Abramowitz,J., Levenstein,M., McCue,K., Patki,D., Raghunathan,T., Rodgers,A., Shapiro,M. and Wasi,N.. Optimal Probabilistic Record Linkage: Best Practice for Linking Employers in Survey and Administrative Data. Technical report, Center for Economic Studies, U.S. Census Bureau, 2019.
[49] Burkard,R., Dell'Amico,M. and Martello,S. *Assignment Problems: Revised Reprint*. SIAM, 2009. MR2488749
[50] Laska,J. N., Davenport,M. A. and Baraniuk,R. G. Exact Signal Recovery from Sparsely Corrupted Measurements through the Pursuit of Justice. In *Asilomar Conference on Signals, Systems and Computers*, pages 1556–1560, 2009.
[51] Nguyen,N. and Tran,T. Robust Lasso with Missing and Grossly Corrupted Observations. *IEEE Transactions on Information Theory*, 59:2036–2058, 2013. MR3043781
[52] Bhatia,K., Jain,P., Kamalaruban,P. and Kar,P. Consistent robust regression. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2110–2119, 2017.
[53] Liu,T. Y. and Jiang,H. Minimizing sum of truncated convex functions and its applications. *Journal of Computational and Graphical Statistics*, 28(1):1–10, 2019. MR3939367
[54] Donoho,D. and Johnstone,I. Ideal spatial adaption by Wavelet shrinkage. *Biometrika*, 81:425–455, 1994. MR1311089
[55] Bertsekas,D. *Nonlinear Programming*. Athena Scientific, 2nd edition edition, 1999. MR3444832
[56] Vershynin,R. *High-Dimensional Probability. An Introduction with Applications in Data Science*. Cambridge University Press, 2018. MR3837109
[57] Koltchinskii,V. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*. Springer, 2011. MR2829871
[58] Van de Geer,S. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36:614–645, 2008. MR2396809
[59] Wang,Z., Ben-David,E. and Slawski,M. Supplement to "Estimation in exponential family regression based on linked data contaminated by mismatch error". October 2020. http://intlpress.com/site/pub/files/_supp/sii/2023/0016/0003/SII-2023-0016-0003-s001.pdf
[60] Frénay,B. and Verleysen,M. Classification in the presence of label noise: a survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25:845–869, 2013. MR4378310
[61] Dalzell,N. and Reiter,J. Regression Modeling and File Matching Using Possibly Erroneous Matching Variables. *Journal of Computational and Graphical Statistics*, 27:728–738, 2018. MR3890865
[62] Zhang,C. H. and Zhang,S. Confidence intervals for low dimensional parameters in high-dimensional linear models. *Journal of the Royal Statistical Society Series B*, 76:217–242, 2014. MR3153940
[63] Javanmard,A. and Montanari,A. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15:2869–2909, 2014. MR3277152
[64] van de Geer,S., Bühlmann,P., Ritov,Y. and Dezeure,R. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42:1166–1202, 2014. MR3224285

Zhenbang Wang
Department of Statistics, George Mason University
Fairfax, VA, USA
E-mail address: zwang39@gmu.edu

Emanuel Ben-David
U.S. Census, CSRM
Suitland, MD, USA
E-mail address: emanuel.ben.david@census.gov

Martin Slawski
Department of Statistics, George Mason University
Fairfax, VA, USA
E-mail address: mslawsk3@gmu.edu