JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2021, VOL. 00, NO. 0, 1–13
https://doi.org/10.1080/10618600.2020.1870482

A Pseudo-Likelihood Approach to Linear Regression With Partially Shuffled Data

Martin Slawski[a], Guoqing Diao*[b], and Emanuel Ben-David[c]

[a]Department of Statistics, George Mason University, Fairfax, VA; [b]Department of Biostatistics & Bioinformatics, George Washington University, Washington, DC; [c]U.S. Census Center for Statistical Research & Methodology (CSRM), Washington, DC

ABSTRACT

Recently, there has been significant interest in linear regression in the situation where predictors and responses are not observed in matching pairs corresponding to the same statistical unit as a consequence of separate data collection and uncertainty in data integration. Mismatched pairs can considerably impact the model fit and disrupt the estimation of regression parameters. In this article, we present a method to adjust for such mismatches under "partial shuffling" in which a sufficiently large fraction of (predictors, response)-pairs are observed in their correct correspondence. The proposed approach is based on a pseudo-likelihood in which each term takes the form of a two-component mixture density. expectation-maximization schemes are proposed for optimization, which (i) scale favorably in the number of samples, and (ii) achieve excellent statistical performance relative to an oracle that has access to the correct pairings as certified by simulations and case studies. In particular, the proposed approach can tolerate considerably larger fraction of mismatches than existing approaches, and enables estimation of the noise level as well as the fraction of mismatches. Inference for the resulting estimator (standard errors, confidence intervals) can be based on established theory for composite likelihood estimation. Along the way, we also propose a statistical test for the presence of mismatches and establish its consistency under suitable conditions. Supplemental files for this article are available online.

ARTICLE HISTORY
Received October 2019
Revised December 2020

KEYWORDS
Broken sample problem; Expectation-maximization algorithm; Mixture models; Pseudo-likelihood; Record linkage

1. Introduction

A tacit assumption in linear regression is that each of the (predictors, response)-pairs $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ is associated with the same underlying statistical unit. However, there are scenarios in which the $\{\mathbf{x}_i\}_{i=1}^n$ and the $\{y_i\}_{i=1}^n$ were collected separately, and there is uncertainty about which of the pairs $\{(\mathbf{x}_i, y_j)\}_{i\le j}$ are in correspondence to each other. Pioneering work by DeGroot, Feder, and Goel (1971), DeGroot and Goel (1976, 1980), Goel (1975), and Goel and Ramalingam (1987) has formalized this setting under the notion of "broken sample": it is assumed that $\{(\mathbf{x}_{\pi^*(i)}, y_i)\}_{i=1}^n$ are iid observations from some joint distribution $P_{\mathbf{x},y;\theta^*}$, where θ^* is an unknown parameter and π^* is an unknown permutation of $\{1,\dots,n\}$. To give an example, $P_{\mathbf{x},y;\theta^*}$ might be a Gaussian distribution with θ^* representing the mean and covariance matrix. Depending on the problem, inference for both θ^* and π^* can be of interest.

More recently, there has been a surge of interest in the above setup in the context of linear regression, driven by applications in engineering and promising new developments in the mathematical signal processing and machine learning literature. Specifically, the following model has been studied under the terms "unlabeled sensing" (Unnikrishnan, Haghighatshoar, and Vetterli 2018), "regression with unknown permutation" (Pananjady, Wainwright, and Cortade 2018; Emiya et al. 2014), and "regression with shuffled data" (Abid, Poon, and Zou 2017; Hsu, Shi, and Sun 2017):

$$\begin{aligned} & y_i = \mathbf{x}_{\pi^*(i)}^\top \beta^* + \sigma_* \varepsilon_i, \quad i = 1,\dots,n, \\ \iff\ & \mathbf{y} = \Pi^* \mathbf{X} \beta^* + \sigma_* \boldsymbol{\epsilon}, \ \ \mathbf{y} = (y_i)_{i=1}^n, \\ & \Pi^* = \big(I(\pi^*(i) = j)\big)_{i,j=1}^n, \ \ \mathbf{X}^\top = [\mathbf{x}_1 \dots \mathbf{x}_n], \\ & \boldsymbol{\epsilon} = (\varepsilon_i)_{i=1}^n. \end{aligned} \tag{1}$$

In (1), $\beta^* \in \mathbb{R}^d$ denotes the regression parameter, the $\{\varepsilon_i\}_{i=1}^n$ represent iid zero-mean and unit-variance errors, and σ_* is referred to as "noise level." Model (1) has been considered from the point of view of signal recovery (with β^* representing an unknown signal of interest) based on (noisy) linear sensing at the $\{\mathbf{x}_i\}_{i=1}^n$, with the caveat that those linear measurements are received in an unknown order. For example, each measurement may come with an inaccurate time stamp, and as result, measurements are received in a shuffled order (Balakhrisnan 1962). Specific applications of (1) in signal processing and sensors networks are reviewed in Unnikrishnan and Vetterli (2013), Unnikrishnan, Haghighatshoar, and Vetterli (2018), Pananjady, Wainwright, and Cortade (2017, 2018), and Haghighatshoar and Caire (2017).

Another important domain in which model (1) is of interest is data integration. Specifically, consider two data files A and B, with A containing the response variable y and B containing predictor variables $\mathbf{x}$ for a set of statistical units common to A

CONTACT Martin Slawski mslawsk3@gmu.edu Department of Statistics, George Mason University, Washington, DC 20006.
*Work was conducted while the author was affiliated with George Mason University.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JCGS.

and B. For example, A may contain the annual income of a set of individuals, while B may contain a collection of demographic variables, and the goal is to fit a linear regression model for income based on those variables. To perform this task, file A needs to be merged with file B, which is straightforward only if both files are equipped with unique identifiers. However, it is common that the data analyst does not have access to such identifiers, for example, because of privacy concerns. In this case, linkage of A and B needs to be based on a combination of variables that are contained in both files (so-called matching variables), with the possibility of ambiguities and the potential for *linkage error*, that is, a record in A is not matched to the correct counterpart in B. Therefore, model (1) can be used to account for errors (mismatches) in post-linkage regression analysis, a long-standing problem in statistics dating back to the work in Neter, Maynes, and Ramanathan (1965) that is particularly relevant in the area of official statistics and the work of government agencies like the U.S. Census Bureau (Scheuren and Winkler 1993, 1997). The latter regularly combines data from a variety of sources such as administrative data, sample survey, and census data. The main purpose of combining data is to reuse existing data, reduce the cost of data collection, research, and the burden on responders. In spite of its relevance to this application, model (1) has hardly been considered directly. Instead, the common approach is to use information about the linkage process, for example, the probability of a mismatch given a certain configuration for the matching variables, to construct suitable estimators that curb the impact of linkage error on the regression fit (e.g., Lahiri and Larsen 2005; Chambers 2009; Hof and Zwinderman 2012, 2015; Gutman, Afendulis, and Zaslavsky 2013; Dalzell and Reiter 2018; Han and Lahiri 2019). However, information about the linkage process may be scarce or unavailable, in which case it can be useful to resort to (1).

1.1. Related Work

Several recent articles have studied estimation of β^* and/or π^* under model (1), predominantly under random Gaussian design with $\{\mathbf{x}_i\}_{i=1}^n \overset{\text{iid}}{\sim} N(0, I_d)$ and Gaussian errors. Unnikrishnan, Haghighatshoar, and Vetterli (2018) showed that in the noiseless case ($\sigma_* = 0$), β^* can be uniquely recovered by exhaustive search over permutations if $n > 2d$. Regarding the noisy case, a series of properties have been established for the least squares problem

$$\min_{\Pi \in \mathcal{P}_n,\, \beta \in \mathbb{R}^d} \left\| \mathbf{y} - \Pi \mathbf{X} \beta \right\|_2^2, \tag{2}$$

where $\mathcal{P}_n$ denotes the set of n-by-n permutation matrices. Problem (2) is a specific quadratic assignment problem (Burkard, Dell'Amico, and Martello 2009). A result in Pananjady, Wainwright, and Cortade (2018) shows that (2) is NP-hard. The article Pananjady, Wainwright, and Cortade (2018) also derives necessary and sufficient conditions for exact and approximate recovery of Π^* based on (2), and elaborates on the significance of the signal-to-noise ratio (SNR) $\|\beta^*\|_2^2/\sigma^2$ in this context. An excessively large SNR of the order n^2 is proved to be a necessary condition for approximate permutation recovery for *any* estimator. In a similar spirit, Hsu, Shi, and Sun (2017) showed that the SNR needs to be at least of the order $d/\log\log n$ to ensure approximate recovery of β^*. In fact, problem (2) can be shown to suffer from overfitting due to the extra freedom in optimizing Π (Abid, Poon, and Zou 2017; Slawski and Ben-David 2019).

Tractable algorithms for (2) with provable guarantees are scarce at this point: the scheme in Hsu, Shi, and Sun (2017) has polynomial time complexity, but is "not meant for practical deployment" as the authors state themselves. The convex relaxation of (2) in which $\mathcal{P}_n$ is replaced by its convex hull, the set of doubly stochastic matrices, was observed to yield poor performance (Emiya et al. 2014). Wu (1998) and Abid and Zou (2018) discussed alternating minimization as well as the use of the expectation-maximization (EM) algorithm (combined with MCMC sampling) in which Π^* constitutes missing data. A recent article by Tsakiris (2018) discusses a branch-and-bound scheme with promising empirical performance on small datasets; the theoretical properties of the approaches in Wu (1998), Abid and Zou (2018), and Tsakiris (2018) remain to be investigated.

In view of the aforementioned computational and statistical barriers, Slawski and Ben-David (2019) discussed a simplified setting of (1) in which $\pi^*(i) = i$ except for at most $k \ll n$ elements of $\{1, \ldots, n\}$; π^* is called *k-sparse* in this case. Slawski and Ben-David showed that under this restriction on π^*, the constrained least squares estimator corresponding to (2) has desirable statistical properties if the fraction k/n is not too large. Moreover, a convex relaxation of that constrained least squares problem yields an estimator of β^* that is consistent under suitable conditions on k/n; the permutation can be estimated subsequently by sorting (cf. Equation (9)). The articles (Slawski, Rahmani, and Li 2019; Slawski, Ben-David, and Li 2020) consider extensions to a multivariate regression setup in which the $\{y_i\}_{i=1}^n$ have dimension $m \geq 1$ each. It is shown that permutation recovery, that is, the event $\{\widehat{\Pi} = \Pi^*\}$ for a suitable estimator $\widehat{\Pi}$ of Π^*, can succeed without stringent conditions on the SNR once m is at least of the order $\log n$. Zhang, Slawski, and Li (2019) complemented this result with matching information-theoretic lower bounds. Motivated by applications in automatic term translation, Shi, Lu, and Cai (2020) considered a closely related setup which the authors term "spherical regression with mismatch corruption" with responses and predictors being contained in the unit sphere of $\mathbb{R}^m$ (in Shi, Lu, and Cai (2020), $m = d$). In addition to sparsity, Shi, Lu, and Cai additionally assumed Π^* to have a block structure. On the other hand, in Shi, Lu, and Cai, Π^* is not required to be a permutation to allow a slightly more general class of mismatches.

Finally, it is worth mentioning that instead of linear regression, Carpentier and Schlüter (2016) and Rigollet and Weed (2019) studied isotonic regression, that is, $y_i = f^*(x_{\pi^*(i)}) + \sigma_* \varepsilon_i$, $1 \leq i \leq n$, with $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$ being one-dimensional, and a nondecreasing regression function f^*.

1.2. Contributions

In this article, we build on the setup of sparse permutations as put forth in Slawski and Ben-David (2019). The approach proposed herein improves over the approach in Slawski and

Ben-David (2019) with regard to two important aspects. One of the downsides therein is that mismatches are treated as generic data contamination, which leads to a substantial loss of information. As a result, performance degrades severely as the fraction of mismatches k/n increases. A second drawback of the approach is its dependence on a tuning parameter involving the noise level σ_*, which is generally not known nor easy to estimate. By contrast, the approach proposed herein is in principle tuning-free (apart from the choice of a suitable initial solution), and produces estimates of the noise level σ_* and the fraction of mismatches $\alpha_* = k/n$ in addition to an estimate of the regression parameter. Moreover, the approach is far less affected as α_* increases and empirically performs rather closely to the oracle least squares estimator equipped with knowledge of π^* (cf. Section 4); while in theory, an arbitrary constant fraction of mismatches can be tolerated, $\alpha_* \approx 0.7$ typically constitutes the limit in practice. In addition, the proposed approach also avoids a quadratic runtime in n that is incurred for alternatives such as the Lahiri–Larsen estimator (Lahiri and Larsen 2005). Optimization is based on a pseudo-likelihood having the form of a likelihood for fitting a two-component mixture model, with one component corresponding to a regular linear regression model without mismatches and the second component accounting for extra variation due to mismatches. Despite the nonconvexity of the resulting optimization problem, reasonable approximate solutions can be obtained via the EM algorithm and one of its variants (Titterington 1984; Lange 1995) whose initialization is discussed in detail in Section 3.4. The EM schemes are easy to implement and exhibit only a linear dependence in n. By leveraging well-developed theory on composite likelihood estimation, asymptotic standard errors and confidence intervals for $(\beta^*, \sigma_*, \alpha_*)$ can be obtained (cf. Theorem 1). Along the way, we also propose a test for the null hypothesis $H_0 : \Pi^* = I_n$, that is, a test for the presence of mismatches, and show its consistency under suitable conditions (Section 2.2).

2. Approach

We start by fixing notation. Densities are denoted by f and the corresponding random variables appear as subscript. Moreover, we write $U \sim f$ to express that the random variable U follows the distribution specified by the density f. We also use uppercase letter notation to indicate distributions, for example, $U \sim N(m, s^2)$ for the Normal distribution with mean m and variance s^2.

To formally introduce the approach proposed herein, we make the following assumptions.

(A1) The permutation π^* is assumed to be chosen uniformly at random from the set of permutations $\{\pi : \sum_{i=1}^n I(\pi(i) \neq i) = k\}$ moving exactly k indices of $\{1, \dots, n\}$.

(A2) Conditional on π^*, the pairs $\{(\mathbf{x}_{\pi^*(i)}, y_i)\}_{i=1}^n$ are iid zero-mean random variables, drawn from a joint distribution with density $f_{\mathbf{x},y}(\mathbf{x}, y) = f_{y|\mathbf{x}}(y) \cdot f_{\mathbf{x}}(\mathbf{x})$ with $f_{y|\mathbf{x}} \sim N(\mathbf{x}^\top \beta^*, \sigma_*^2)$.

Define indicator variables $z_i = I(\pi^*(i) \neq i)$, $i = 1, \dots, n$, and fix an arbitrary index $i \in \{1, \dots, n\}$. Under (A2), it then holds that

$$y_i|\{\mathbf{x}_i, z_i = 0\} \sim N(\mathbf{x}_i^\top \beta^*, \sigma_*^2), \qquad y_i|\{\mathbf{x}_i, z_i = 1\} \sim f_y, \quad (3)$$

where $f_y(y) = \int f_{y|\mathbf{x}}(y) f_{\mathbf{x}}(\mathbf{x})\, d\mathbf{x}$. In fact, note that conditional on $\{z_i = 1\}$, y_i is independent of $\mathbf{x}_i$, and as a result the conditional distribution coincides with the marginal distribution of y. In conclusion, (A1) and (3) imply that $y_i|\mathbf{x}_i$ follows a two-component mixture with proportions $1 - \alpha_*$ and $\alpha_* = k/n$, that is, with some slight abuse of notation

$$y_i|\mathbf{x}_i \sim (1 - \alpha_*) N(\mathbf{x}_i^\top \beta^*, \sigma_*^2) + \alpha_* f_y. \quad (4)$$

Remarks.

(i) Assumption (A1) can be considerably relaxed without affecting (4). Specifically, it suffices to assume that the indicators $\{z_i\}_{i=1}^n$ are independent of $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ and satisfy $\mathbf{P}(z_i = 1) = \alpha_*$, $1 \leq i \leq n$. In fact, the latter does not even require π^* to be a permutation of $\{1, \dots, n\}$.

(ii) The zero-mean assumption in (A2) is merely made for convenience as it eliminates the need for an intercept.

(iii) Other regression settings such as logistic regression can be covered by appropriately modifying the model for $y_i|\{\mathbf{x}_i, z_i = 0\}$ in (3).

Since estimation of the marginal density in f_y (4) can be performed based on the $\{y_i\}_{i=1}^n$ only and is thus not affected by the presence of π^*, f_y can be assumed to be effectively known given that n is sufficiently large and can consequently be estimated with small error, be it in a parametric (e.g., by assuming $f_y \sim N(0, \tau_*^2)$) or in nonparametric fashion (by density estimation). Observe that f_y implicitly depends on β^* via the linear predictor $\mathbf{x}^\top \beta^*$. Knowledge of the distribution of the latter as in the setting discussed subsequently can thus benefit statistical efficiency in estimation, but specific assumptions regarding the distribution of $\mathbf{x}$ or $\mathbf{x}^\top \beta^*$ are not critical to our approach.

In the sequel, we focus on isotropic Gaussian design as considered in Pananjady, Wainwright, and Cortade (2018) and Slawski and Ben-David (2019), that is, $f_{\mathbf{x}} \sim N(0, I_d)$. In this case, note that $f_y \sim N(0, \|\beta^*\|_2^2 + \sigma_*^2)$ since y is the sum of two independent Gaussian random variables $\mathbf{x}^\top \beta^* \sim N(0, \|\beta^*\|_2^2)$ and $\sigma_* \varepsilon \sim N(0, \sigma_*^2)$. Accordingly, we have

$$y_i|\mathbf{x}_i \sim (1 - \alpha_*) N(\mathbf{x}_i^\top \beta^*, \sigma_*^2) + \alpha_* N(0, \|\beta^*\|_2^2 + \sigma_*^2). \quad (5)$$

The above considerations suggest the following "likelihood" approach

$$\max_{\beta \in \mathbb{R}^d, \sigma^2 > 0, \alpha \in [0,1]} \prod_{i=1}^n f(y_i|\mathbf{x}_i), \quad (6)$$

where $f(y_i|\mathbf{x}_i)$ refers to the density of the above Gaussian mixture distribution, that is,

$$f(y_i|\mathbf{x}_i) = \frac{1 - \alpha}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mathbf{x}_i^\top \beta)^2}{2\sigma^2}\right) + \frac{\alpha}{\sqrt{2\pi(\sigma^2 + \|\beta\|_2^2)}} \exp\left(-\frac{y_i^2}{2(\sigma^2 + \|\beta\|_2^2)}\right), \quad (7)$$

$$i = 1, \dots, n.$$

The quotation marks above indicate that the objective in (6) is not a genuine likelihood function since $\{y_i|\mathbf{x}_i\}_{i=1}^n$ are not

independent observations from the Gaussian mixtures given in (7). However, we can still treat (6) within the framework of *pseudo likelihood*, or more specifically *composite likelihood*. The pseudo-likelihood (6) is composed of likelihoods of individual observations, which constitutes the most basic variant of a composite likelihood. Nevertheless, the approach enjoys several attractive properties including asymptotic normality at the standard rate and a closed form expression for the asymptotic covariance matrix, while avoiding the computational barrier that is associated with the unknown permutation as elaborated in the introduction.

The asymptotic normality result is stated in Theorem 1. Denote by $\theta^* = (\beta^*, \sigma_*^2, \alpha_*)$ the unknown parameter, which is supposed to be an interior point of $\mathbb{R}^d \times [0,\infty) \times [0,1]$. Let further $\ell_{\mathrm{p}}(\theta) = \sum_{i=1}^n \ell_{i,\mathrm{p}}$, $\ell_{i,\mathrm{p}} := -\log(f(y_i|\mathbf{x}_i;\theta))$ denote the negative pseudo log-likelihood with $f(y_i|\mathbf{x}_i;\theta)$ as in (7), $\theta = (\beta, \sigma^2, \alpha)$. The global minimizer $\operatorname{argmin}_\theta \ell_{\mathrm{p}}(\theta)$ is denoted by $\widehat{\theta}_n$. Equipped with those definitions and the following assumption (A3), we can state the following result.

(A3) The $\{\ell_{i,\mathrm{p}}\}_{i=1}^n$ satisfy the regularity conditions specified in Theorem 5.23 or in Theorem 5.41 in van der Vaart (1998).

Theorem 1. Under (A1), (A2), and (A3), $n^{1/2}(\widehat{\theta}_n - \theta^*) \to N(0, H_*^{-1} G^* H_*^{-1})$ in distribution as $n \to \infty$, where $H_* = \mathbf{E}\left[-\nabla_\theta^2 \log f(y|\mathbf{x};\theta^*)\right]$ and $G^* = \mathbf{E}\left[\nabla_\theta \log f(y|\mathbf{x};\theta^*)\, \nabla_\theta \log f(y|\mathbf{x};\theta^*)^\top\right]$. Here, ∇_θ and ∇_θ^2 denote the gradient and Hessian with respect to θ, respectively, $f(y|\mathbf{x};\theta)$ denotes the density of a generic pair $(\mathbf{x}, y)$ according to (4), and $\mathbf{E}$ denotes the expectation with respect to that density. Moreover, H_* and G^* can be consistently estimated by

$$\widehat{H} = n^{-1}\nabla_\theta^2 \ell_p(\widehat{\theta}_n), \qquad \widehat{G} = n^{-1}\sum_{i=1}^n \nabla_\theta \ell_{i,p}(\widehat{\theta}_n)\, \nabla_\theta \ell_{i,p}(\widehat{\theta}_n)^\top. \tag{8}$$

Theorem 1 can be proved by invoking well-established theory on composite likelihood theory (see, e.g., Lindsay 1988; Varin, Reid, and Firth 2011). Hence, we omit the details of the proof. Explicit expressions for the estimators $\widehat{H}$ and $\widehat{G}$, which are relevant in practice to estimate standard errors and to construct asymptotic confidence intervals, are provided in the supplement.

Remark. In this article, we do not develop any novel approach for estimating the permutation π^*. If the latter is of interest, the plug-in approach in Slawski and Ben-David (2019) can be applied. The latter is based on the optimization problem

$$\begin{aligned}\min_{\Pi\in\mathcal{P}_n} \left\|\mathbf{y} - \Pi\mathbf{X}\widehat{\beta}\right\|_2^2 &= -2\max_{\Pi\in\mathcal{P}_n}\langle \mathbf{y}, \Pi\mathbf{X}\widehat{\beta}\rangle + c \\ &= -2\sum_{i=1}^n y_{(i)}(\mathbf{X}\widehat{\beta})_{(i)} + c, \end{aligned} \tag{9}$$

where $c = \|\mathbf{y}\|_2^2 + \|\mathbf{X}\widehat{\beta}\|_2^2$ does not depend on Π, and the subscript (i) denotes the ith order statistic, that is, $y_{(1)} < \cdots < y_{(n)}$ (assuming the absence of ties). The relations in (9) imply that for fixed β, the optimal permutation can be found by sorting $\{y_i\}_{i=1}^n$ and $\{\mathbf{x}_i^\top\beta\}_{i=1}^n$. Statistical properties of the plug-in approach (9) are studied in Slawski and Ben-David (2019) independent of specific properties of $\widehat{\beta}$.

To account for partial shuffling, approach (9) can be refined by identifying the set of mismatches $S_* = \{i : z_i = 1\}$ first, and then solving (9) only with respect to observations in that set. Note that the mixture model representation (3) naturally lends itself to an estimator of S_* by estimating the corresponding posterior probabilities $\mathbf{P}(z_i = 1|\mathbf{x}_i, y_i)$, $1 \le i \le n$, given $(\widehat{\beta}, \widehat{\sigma}^2, \widehat{\alpha})$.

2.1. Connection to Robust Regression

The above pseudo-likelihood approach can be related to robust M-estimation as we elaborate below. This connection puts the present work in perspective with prior work (Slawski and Ben-David 2019). Consider the negative pseudo log-likelihood that follows from (6) and (7), up to additive constants:

$$\sum_{i=1}^n -\log\left\{\frac{1-\alpha}{\sigma}\exp\left(-\frac{(y_i - x_i^\top\beta)^2}{2\sigma^2}\right) + \frac{\alpha}{\sqrt{\sigma^2 + \|\beta\|_2^2}}\exp\left(-\frac{y_i^2}{2(\sigma^2+\|\beta\|_2^2)}\right)\right\}. \tag{10}$$

With α, σ^2 and $\tau = \sqrt{\sigma^2 + \|\beta\|_2^2}$ considered as fixed, the above expression can be written as the following loss function $L(\beta)$, up to additive constants:

$$\begin{aligned} L(\beta) &= \sum_{i=1}^n \ell\left(\left|\frac{r_i(\beta)}{\sigma}\right|; \gamma, \left|\frac{y_i}{\tau}\right|\right), \\ r_i(\beta) &:= y_i - \mathbf{x}_i^\top\beta,\ 1 \le i \le n, \\ \ell(z; a, b) &:= -\log\left(\exp\left(-\frac{z^2}{2}\right) + a\exp\left(-\frac{b^2}{2}\right)\right), \\ \gamma &:= \frac{\alpha}{1-\alpha}\cdot\frac{\sigma}{\tau}. \end{aligned} \tag{11}$$

Figure 1 visualizes $\ell(\cdot; a, b)$ for selected values of a and b; the function scales have been rescaled to the range $[0,1]$. The shape of ℓ resembles a "capped loss" such as Tukey's bisquare (e.g., Maronna, Martin, and Yohai 2006) commonly employed in robust regression. In fact, ℓ is uniformly bounded by $-\log(a) + b^2/2$. For $\alpha = 0$, ℓ reduces to ordinary squared loss. As α increases, ℓ levels off more quickly, and behaves like an indicator loss in the limit $\alpha \to 1$. The above connection also highlights the advantages of the pseudo-likelihood approach compared to a plain robust M-estimation approach. The pseudo-likelihood can be interpreted as the combination of observation-specific and self-calibrated robust losses, where "calibration" refers to tuning parameters that control the robustness versus efficiency trade-off (more informally speaking, parameters that control the range in which the loss function levels off). Moreover, the pseudo-likelihood integrates estimation of the parameters α_* and σ_*^2 of potential interest.

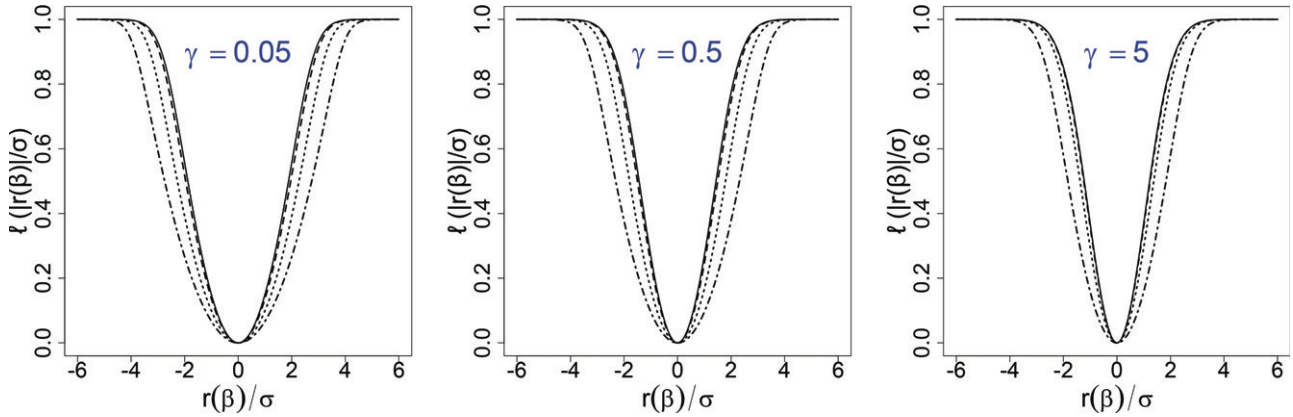

Figure 1. Visualization of the loss function (11) for $\gamma \in \{0.05, 0.5, 5\}$ (from left to right) and $y/\tau \in \{0.5, 1, 2, 3\}$ (solid, dashed, dotted, dashed-dotted).

2.2. Testing for the Presence of Mismatches

Before applying the approach developed at the beginning of this section, it is appropriate to test for the presence of mismatches. We here consider a statistical test for the hypothesis $H_0 : \Pi^* = I_n$, or equivalently $H_0 : \alpha_* = 0$. While one possible direction is the formulation of this test within the setting of mixture models (Chen and Li 2009; Zhu and Zhang 2004), a much more straightforward test can be performed based on the residuals $\widehat{\boldsymbol{\epsilon}} = \mathrm{P}_{\mathbf{X}}^{\perp}\mathbf{y}$ of the ordinary least squares fit, where $\mathrm{P}_{\mathbf{X}}^{\perp}$ denotes the orthoprojector on the orthogonal complement $\mathcal{U}$ of the column space of $\mathbf{X}$. Letting $\mathbf{U}$ denote the n-by-$(n-d)$ matrix whose columns form an orthonormal basis of $\mathcal{U}$, we have $\widehat{\boldsymbol{\epsilon}} = \mathrm{P}_{\mathbf{X}}^{\perp}\mathbf{y} = \mathbf{U}\mathbf{U}^{\top}\mathbf{y}$ and thus $\xi := \mathbf{U}^{\top}\widehat{\boldsymbol{\epsilon}} = \mathbf{U}^{\top}\mathbf{y}$. This yields

$$\begin{aligned} \xi &= \mathbf{U}^{\top}\mathbf{y} = \mathbf{U}^{\top}\Pi^*\mathbf{X}\beta^* + \sigma_*\mathbf{U}^{\top}\boldsymbol{\epsilon} = \mathbf{U}^{\top}\Pi^*\mathbf{X}\beta^* + \zeta, \\ &\overset{H_0}{=} \zeta \quad \text{where } \zeta \sim N(0, \sigma_*^2 I_{n-d}) \quad \Rightarrow \frac{\|\zeta\|_2^2}{\sigma_*^2} \overset{H_0}{\sim} \chi^2(n-d). \end{aligned} \tag{12}$$

This suggests the use of a chi-squared test based on the residual sums of squares. For model (1) with Gaussian design, that is,

$$y_i = \mathbf{x}_{\pi^*(i)}^{\top}\beta^* + \sigma_*\varepsilon_i, \quad \mathbf{x}_i \sim N(0, I_d), \quad i = 1, \ldots, n, \tag{13}$$

results in Pananjady, Wainwright, and Cortade (2018) imply that the power of the chi-square test tends to one under the alternative hypothesis as $n \to \infty$ and $\|\beta^*\|_2 - \sigma_*\sqrt{\log(n-d)} \to \infty$. Note that at least for small k, this condition appears inevitable since $\max_{1 \le i \le n-d} |\zeta_i|/\sigma_* = O_{\mathbf{P}}\left(\sqrt{\log(n-d)}\right)$. More specifically, we have the following statement (cf. supplement for a proof).

Proposition 1. Suppose model (13) holds. Then for any $t \in (0, k)$, we have

$$\left\|\mathbf{U}^{\top}\Pi^*\mathbf{X}\beta^*\right\|_2^2 \le \frac{n-d}{n} \cdot \frac{t}{2}\left\|\beta^*\right\|_2^2$$

with probability at most

$$6\exp\left(-\frac{k}{10}\left[\log\frac{k}{t} + \frac{t}{k} - 1\right]\right) + \exp(-(n-d)/24).$$

Observe that the above proposition implies a high probability lower bound on the quantity $\left\|\mathbf{U}^{\top}\Pi^*\mathbf{X}\beta^*\right\|_2$ appearing in (12), which affirms the claim preceding the proposition.

The effect size associated with the above test grows with the ratio $\|\beta^*\|_2/\sigma_*$ and the fraction of mismatches α_*. Their significance with regard to the power of the test is corroborated by the empirical results depicted in Figure 2. For the latter, $\|\beta^*\|_2$ is fixed to one while varying values of α_* of σ_* are considered in the left and right panel, respectively.

Unknown variance. The test considered above relies on σ_*^2 to be known, which is often not the case in practice. However, consistent estimation of σ_*^2 in the situation of mismatches generally appears to be at least as challenging as the given testing problem, and the fact that the proposed test statistic involves the usual variance estimator indicates a close entanglement of both aspects.

A common scenario in the analysis of linked files is that a subset $Q \subset \{1 \le i \le n : z_i = 0\}$ of observations are known to be correctly matched. For instance, certain combinations of variables used during record linkage turn out to be unique, hence there is no danger of mismatching the corresponding observations. In this case, σ_*^2 can be estimated based on Q, and a test statistic given by the ratio of independent residual sums of squares can be employed. Specifically, we consider

$$\frac{|Q|-d}{n-|Q|-d} \cdot \frac{\left\|\mathrm{P}_{\mathbf{X}_{Q^c}}^{\perp}\mathbf{y}_{Q^c}\right\|_2^2}{\left\|\mathrm{P}_{\mathbf{X}_Q}^{\perp}\mathbf{y}_Q\right\|_2^2} \overset{H_0}{\sim} F(n-|Q|-d, |Q|-d), \tag{14}$$

where $F(a, b)$ denotes the F-distribution with a and b degrees of freedom, $Q^c = \{1, \ldots, n\} \setminus Q$, and the superscripts in $\mathbf{X}$ and $\mathbf{y}$ refer to the corresponding row sub-matrices and -vectors.

A comparison of the empirical power of the test statistics (12) and (14) given $|Q|/n = 0.1$ can be found in Figure 2.

3. EM Algorithm

The pseudo-likelihood (6) corresponds to the "regular" likelihood of a mixture model, and inherits the computational properties of the latter. In particular, likelihood maximization in mixture models is nonconvex, and thus one cannot hope to find the global optimum of (6) in practice. At the same time, established

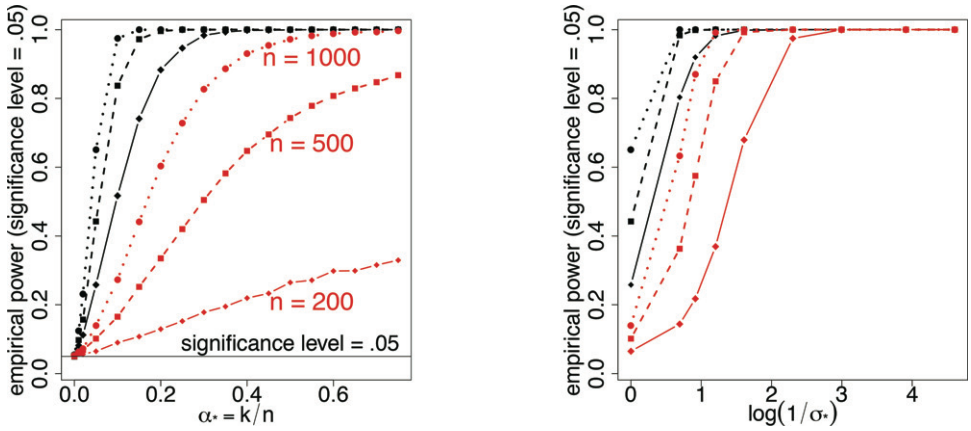

Figure 2. Empirical power of the test statistics in (12) and (14) with $|Q|/n = 0.1$ (light color) for a significance level of 0.05. The results are based on 10,000 replications from model (13) under assumption (A1) with β^* drawn uniformly at random from the unit sphere ($d = 10$) in dependence of $n \in \{200, 500, 1000\}$ (shown in different line types/symbols), $\alpha_* = k/n$ and σ_*. Left: Empirical power for varying α_* and fixed $\sigma_* = 1$. Right: Empirical power for varying σ_* and fixed $\alpha_* = 0.05$.

computational approaches for fitting mixture models like the EM algorithm can be employed for finding approximate maximizers of (6) in practice. The usual convergence properties of the EM algorithm continue to hold regardless of the fact that (6) is a misspecified likelihood. In fact, for the derivation of the E-step we can simply pretend that the $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ are independent observations following the mixture distributions (7): the key property of the EM algorithm to increase the likelihood at each iteration does not require the likelihood to be correctly specified. In the following paragraphs, we provide the specifics of the proposed computational scheme. We first note that given the indicator variables $\{z_i\}_{i=1}^n$, the complete data negative (pseudo) log-likelihood in $(\beta, \sigma^2, \alpha)$ is given by

$$\sum_{i=1}^n \Big\{ (1-z_i)\Big(-\log(1-\alpha) + \frac{\log(\sigma^2)}{2} + \frac{(y_i - \mathbf{x}_i^\top \beta)^2}{2\sigma^2} \Big) + z_i \Big(-\log(\alpha) + \frac{\log(\sigma^2 + \|\beta\|_2^2)}{2} + \frac{y_i^2}{2(\sigma^2 + \|\beta\|_2^2)} \Big) \Big\}. \tag{15}$$

3.1. Both σ_*^2 and $\|\beta^*\|_2$ Known

Denote $\tau_*^2 = \sigma_*^2 + \|\beta^*\|_2^2$. Given current iterates $\widehat{\beta}^{(k)}, \widehat{\alpha}^{(k)}$, the E-step is given by

$$\pi_i^{(k)} := \mathbf{E}^{(k)}[z_i] = \frac{\frac{\widehat{\alpha}^{(k)}}{\tau_*} \exp\Big(-\frac{y_i^2}{2\tau_*^2}\Big)}{\frac{\widehat{\alpha}^{(k)}}{\tau_*} \exp\Big(-\frac{y_i^2}{2\tau_*^2}\Big) + \frac{(1-\widehat{\alpha}^{(k)})}{\sigma_*} \exp\Big(-\frac{(y_i - \mathbf{x}_i^\top \widehat{\beta}^{(k)})^2}{2\sigma_*^2}\Big)}, \quad i = 1, \ldots, n, \tag{16}$$

where $\mathbf{E}^{(k)}$ denotes the expectation if the unknown parameters of the underlying distribution were given by $(\widehat{\alpha}^{(k)}, \widehat{\beta}^{(k)})$. Accordingly, in light of (15) the M-step is given by the optimization problem

$$\min_{\alpha} \sum_{i=1}^n \Big(-(1-\pi_i^{(k)}) \log(1-\alpha) - \pi_i^{(k)} \log(\alpha) \Big) + \min_{\beta} \sum_{i=1}^n (1-\pi_i^{(k)})(y_i - \mathbf{x}_i^\top \beta)^2. \tag{17}$$

Both optimization problems have a closed form solution. The optimization problem in β amounts to a weighted least squares fit of (predictors, response)-pairs $(\mathbf{x}_i, y_i)_{i=1}^n$ and weights $\omega_i^{(k)} = 1 - \pi_i^{(k)}$, $i = 1, \ldots, n$. This yields the updates

$$\widehat{\alpha}^{(k+1)} \leftarrow \frac{1}{n} \sum_{i=1}^n \pi_i^{(k)}, \quad \widehat{\beta}^{(k+1)} \leftarrow (\mathbf{X}^\top \mathbf{W}^{(k)} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W}^{(k)} \mathbf{y},$$
$$\mathbf{W}^{(k)} = \mathrm{diag}(\omega_1^{(k)}, \ldots, \omega_n^{(k)}),$$

which is well in line with the robust regression viewpoint in Section 2.1. Alternatively, the M-step (17) can be performed subject to the additional constraint $\|\beta\|_2^2 \le \|\beta^*\|_2^2$. The latter is straightforward to accommodate.

3.2. Plug-In Approach

It is straightforward to estimate τ_*^2 via

$$\widehat{\tau}^2 = \frac{1}{n} \sum_{i=1}^n y_i^2$$

since $\mathbf{E}[y_i^2] = \tau_*^2$, $i = 1, \ldots, n$. After substituting τ_*^2 by the above estimator and σ_*^2 by an iterate $\widehat{\sigma}^{2(k)}$, the scheme of the previous subsection can still be applied. The counterpart to the E-step (16) is given by

$$\pi_i^{(k)} \leftarrow \frac{\frac{\widehat{\alpha}^{(k)}}{\widehat{\tau}} \exp\Big(-\frac{y_i^2}{2\widehat{\tau}^2}\Big)}{\frac{\widehat{\alpha}^{(k)}}{\widehat{\tau}} \exp\Big(-\frac{y_i^2}{2\widehat{\tau}^2}\Big) + \frac{(1-\widehat{\alpha}^{(k)})}{\widehat{\sigma}^{(k)}} \exp\Big(-\frac{(y_i - \mathbf{x}_i^\top \widehat{\beta}^{(k)})^2}{2\widehat{\sigma}^{2(k)}}\Big)}, \quad i = 1, \ldots, n, \tag{18}$$

and the iterate for σ^2 is updated as

$$\widehat{\sigma}^{2(k+1)} \leftarrow \frac{1}{\sum_{i=1}^n (1-\pi_i^{(k)})} \sum_{i=1}^n (1-\pi_i^{(k)})(y_i - \mathbf{x}_i^\top \widehat{\beta}^{(k)})^2. \quad (19)$$

Upon termination, to account for the usual bias of the ML estimator of the error variance in linear regression with Gaussian errors, $\widehat{\sigma}^2$ may be replaced by a bias-corrected counterpart in which the number of predictor variables d is subtracted in the denominator of (19), cf. Aitkin (1981) for a related discussion in the context of regression with censored response.

The constraint $\sigma^2 + \|\beta\|_2^2 \le \widehat{\tau}^2$ can optionally be imposed; however, with the presence of this constraint, the simple closed form updates for $\widehat{\sigma}^{2(k)}$ and $\widehat{\beta}^{(k)}$ in the M-step are no longer valid. A compromise is the relaxed constraint $\|\beta\|_2^2 \le \widehat{\tau}^2$ that still gives rise to closed form updates.

It is worth noting that the approach discussed in this subsection does not strongly hinge on the assumption $\mathbf{x} \sim N(0, I_d)$, and remains applicable outside this setting as demonstrated in Section 4. That assumption only enters in the specification $f_y \sim N(0, \widehat{\tau}^2)$, which can either simply be adopted as a reasonable approximation or replaced by alternative models (parametric or nonparametric).

3.3. Simultaneous Estimation of All Parameters

The plug-in approach of the previous section is convenient due to its closed form updates by means of a reduction to weighted least squares estimation. In addition, it does not require assumptions on the distribution of the $\{\mathbf{x}_i\}_{i=1}^n$. However, for isotropic Gaussian design, the plug-in approach essentially disregards the part of the complete data log-likelihood that is associated with the $\{z_i\}_{i=1}^n$. It is a cleaner, but also computationally significantly more involved approach to avoid the use of the auxiliary (and eventually redundant) parameter τ_*^2, and to integrate all terms of the complete data log-likelihood (15). While the E-step (18) remains unchanged with the only modification that $\widehat{\tau}^2$ gets replaced by $\|\widehat{\beta}^{(k)}\|_2^2 + \widehat{\sigma}^{2(k)}$, the M-step has no longer a closed form solution for $\widehat{\beta}^{(k+1)}, \sigma^{(k+1)}$. Instead, the latter result as the minimizers of the optimization problem

$$\min_{\substack{\beta\in\mathbb{R}^d \\ \sigma^2>0}} \sum_{i=1}^n (1-\pi_i^{(k)}) \frac{\log(\sigma^2)}{2} + \sum_{i=1}^n (1-\pi_i^{(k)}) \frac{(y_i - \mathbf{x}_i^\top \beta)^2}{2\sigma^2} \quad (20)$$

$$+ \sum_{i=1}^n \pi_i^{(k)} \frac{\log(\sigma^2 + \|\beta\|_2^2)}{2} + \sum_{i=1}^n \pi_i^{(k)} \frac{y_i^2}{2(\sigma^2 + \|\beta\|_2^2)}.$$

This optimization problem fails to be convex. Rather than solving this problem, we suggest to update the parameters via one iteration of Fisher scoring, which is also known as Titterington's algorithm in the literature on the EM algorithm (Titterington 1984). Specifically, we consider the following update:

$$\begin{bmatrix} \widehat{\beta}^{(k+1)} \\ \widehat{\sigma}^{(k+1)} \end{bmatrix} = \begin{bmatrix} \widehat{\beta}^{(k)} \\ \widehat{\sigma}^{(k)} \end{bmatrix} + \gamma^{(k)} d^{(k)}, \qquad d^{(k)} := -F^{(k)} g^{(k)}, \quad (21)$$

where the step-size $\gamma^{(k)} \in [0, 1]$ is chosen by backtracking line search (e.g., Bertsekas 1999, sec. 1.2), and the update direction $d^{(k)}$ depends on the gradient $g^{(k)}$ of the expected complete data negative log-likelihood as well as $F^{(k)}$, the expected Fisher information (with respect to the current parameters $\widehat{\beta}^{(k)}$ and $\widehat{\sigma}^{(k)}$). Since the gradient of the *expected complete* data log-likelihood is known to coincide with the gradient of the *incomplete* data log-likelihood (Lange 1995, p. 426), the above update scheme reduces the latter at each iteration. Expressions for $g^{(k)}$ and $F^{(k)}$ are provided in the supplement.

3.4. Initialization

While the EM iterations above can be shown to yield descent at each iteration, they are not guaranteed to produce the global minimizer of the incomplete data negative log-likelihood (10). As a result, careful initialization, that is, choice of the initial iterates $\widehat{\beta}^{(0)}, \widehat{\sigma}^{2(0)}$, and $\widehat{\alpha}^{(0)}$ can greatly benefit the performance. As a starting point, one might consider $\widehat{\beta}^{(0)} = \widehat{\beta}^{\mathrm{LS}}$, where $\widehat{\beta}^{\mathrm{LS}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ denotes the ordinary least squares estimator, that is, the naive approach that ignores the presence of mismatches. The result below indicates that under a uniform-at-random model for Π^*, this naive approach is still useful to the extent that $\frac{\widehat{\beta}^{\mathrm{LS}}}{\|\widehat{\beta}^{\mathrm{LS}}\|_2}$ provides an essentially unbiased estimator of $\frac{\beta^*}{\|\beta^*\|_2}$.

Proposition 2. Consider model (13) with $n \ge d+1$ and suppose that Π^* is chosen uniformly at random according to assumption (A1), and let $\widehat{\beta}^{LS} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ denote the ordinary least squares estimator. We then have

$$\mathbf{E}_{\mathbf{X},\varepsilon,\pi^*}[\widehat{\beta}^{LS}] = (1-\alpha_*)\beta^*,$$
$$\mathrm{cov}_{\mathbf{X},\varepsilon,\pi^*}[\widehat{\beta}^{LS}] = \frac{c_*^2}{n-d} I_d + O(\|\beta^*\|_2^2/n^2),$$

where $c_*^2 = (2\alpha_* - \alpha_*^2)\|\beta^*\|_2^2 + \sigma_*^2$.

Proposition 2 (proved in the supplement) suggests that $\widehat{\beta} = \frac{1}{1-\alpha_*}\widehat{\beta}^{\mathrm{LS}}$ as an unbiased estimator. Since α_* is typically unknown and generally not easy to estimate, an alternative is

$$\widehat{\beta} = \frac{\widehat{\beta}^{\mathrm{LS}}}{\|\widehat{\beta}^{\mathrm{LS}}\|_2} \cdot \widehat{\|\beta^*\|_2}, \qquad \widehat{\|\beta^*\|_2} = \left(\frac{1}{n}\sum_{i=1}^n y_i^2 - \sigma_*^2\right)^{1/2}, \quad (22)$$

which requires knowledge of σ_*^2 (if $\|\beta^*\|_2^2 \gg \sigma_*^2$, the variance of the errors σ_*^2 can be disregarded). While potentially giving rise to an unbiased estimator, Proposition 2 also asserts that the variance of the components of $\widehat{\beta}^{\mathrm{LS}}$ (and in turn the MSE) is rather substantial, growing with $\|\beta^*\|_2^2$ and α_*. In particular, this implies that $\widehat{\beta}^{\mathrm{LS}}$ and its rescaled counterparts discussed above exhibit a poor statistical efficiency relative to the oracle estimators based on knowledge of Π^* or the set of correct matches $\{1 \le i \le n : z_i = 0\}$. In light of this, another option is to employ robust regression methods like Huber's estimator as considered in Slawski and Ben-David (2019) even though the latter is limited to the regime of small to moderate α_*.

Connection to the Lahiri–Larsen estimator. In their seminal work on regression with linked data, Lahiri and Larsen (2005) proposed the following estimator

$$\widehat{\beta}^{\mathrm{LL}} = (\mathbf{X}^\top\mathbf{Q}^\top\mathbf{Q}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Q}^\top\mathbf{y}, \quad \text{where } \mathbf{Q} = \mathbf{E}[\Pi^*]. \quad (23)$$

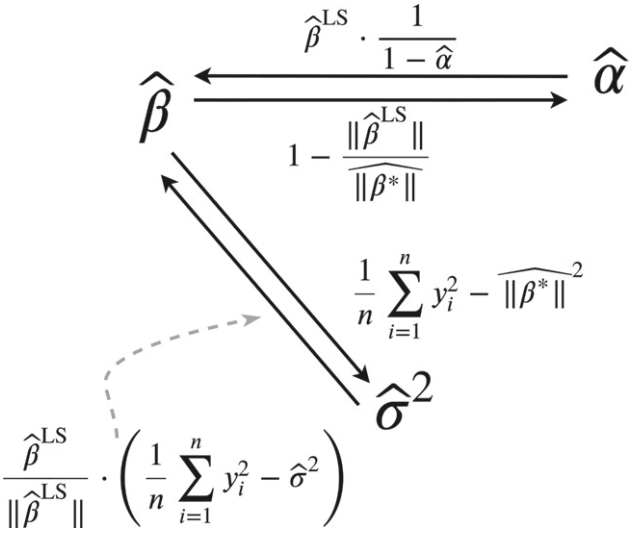

Figure 3. Diagram visualizing the interdependence for initial estimators $(\widehat{\beta}, \widehat{\sigma}, \widehat{\alpha})$ in light of Proposition 2 and relation (22). Note that given an estimator $\widehat{\beta}$, one can use $\widehat{\|\beta^*\|_2} = \|\widehat{\beta}\|_2$.

It is easy to see that the above estimator is unbiased, that is, $\mathbf{E}_{\pi^*,\varepsilon}[\widehat{\beta}^{\mathrm{LL}}] = \beta^*$ uniformly in $\mathbf{X}$. Assuming that Π^* is drawn uniformly at random from the set of k-sparse permutations of n elements, the matrix $\mathbf{Q}$ is given by

$$\mathbf{Q} = \left(1 - \alpha_* - \frac{\alpha_*}{n-1}\right) I_n + \frac{\alpha_*}{n-1}\mathbf{1}\mathbf{1}^\top.$$

Discarding all terms in $\mathbf{Q}$ involving $\alpha_*/(n-1)$, the estimator (23) reduces to the estimator in Proposition 2. It is not hard to establish asymptotic equivalence of the two estimators; the formal derivation is omitted for the sake of brevity.

Equipped with an estimator of β^*, the quantities α_* and σ_* can be estimated according to Figure 3. Estimation of the three quantities is generally interdependent in the sense that one of the three parameters is supposed to be known or accurately estimable. The latter requirement becomes significantly more difficult to meet as the fraction of mismatches α_* increases.

4. Empirical Results

4.1. Gaussian Design

Data is generated according to model (13) with $n = 200$, $d = 10$, and β^* drawn uniformly at random from the corresponding sphere. We vary $\sigma_* \in \{0.1, 0.2, 0.5, 1\}$ and $\alpha_* \in \{0.1, 0.2, \ldots, 0.7\}$, and the permutation π^* is drawn uniformly at random according to (A1). For each configuration of (σ_*, α_*), 100 independent replications are considered. To estimate the parameters $(\beta_*, \sigma_*, \alpha_*)$, we consider the approaches in Sections 3.2 and 3.3. The former is computationally simpler as it reduces to solving a sequence of weighted least squares problems. Both approaches were initialized with $\widehat{\beta}^{(0)} = \widehat{\beta}^{\mathrm{LS}}$, $\widehat{\sigma}^{(0)} = n^{-1/2}\left\|\mathbf{y} - \mathbf{X}\widehat{\beta}^{(0)}\right\|_2$, and $\widehat{\alpha}^{(0)} = 0.5$. More sophisticated initialization schemes as discussed in Section 3.4 did not yield substantial gains in performance. The approaches in Sections 3.2 and 3.3 are compared to the oracle estimator

$$\widehat{\beta}^{\mathrm{o}} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\Pi^{*\top}\mathbf{y}, \tag{24}$$

and the robust regression method in Slawski and Ben-David (2019) (abbreviated SB-D19); the latter is given an additional advantage by equipping it with knowledge of the noise level σ_*, which is required for the optimal choice of the regularization parameter. The performance of $\widehat{\beta}^{\mathrm{LS}}$ turns out to be rather far from competitive and is thus not reported.

Results. Figure 4 displays (i) the ℓ_2-estimation errors $\left\|\widehat{\beta} - \beta^*\right\|_2 / \left\|\widehat{\beta}^{\mathrm{o}} - \beta^*\right\|_2$ relative to that of the oracle estimator (left column), (ii) the relative error for the noise level $|\widehat{\sigma}/\sigma_* - 1|$ (middle column), and the absolute error for the fraction of mismatches $|\widehat{\alpha} - \alpha_*|$ (right column). The plots show medians of these measures of error over 100 independent replications and bootstrap standard error bars. A table representation of the same results can be found in the supplement, which also provides a collection of complementary results including separate investigations of bias/standard error/coverage of confidence intervals, as well as results regarding the identification of mismatches (mismatch recovery rate) and departures from isotropic Gaussian design in model (13).

Figure 4 shows that for both variants of the proposed approach, the estimation error for the regression coefficients in ℓ_2-norm are largely within factors of three or less of the oracle estimator. Note that the error for the latter roughly scales as $\sigma_*\sqrt{d/n}$, while the error of a less powerful oracle eliminating all mismatches $\{i : \pi^*(i) \neq i\}$ and performing a least squares fit with the remaining observation roughly scales as $\sigma_*\sqrt{d/((1-\alpha_*)n)}$; the performance observed for $\widehat{\beta}$ is thus not far from this second oracle, and goes along with massive improvements over $\widehat{\beta}^{\mathrm{LS}}$ if $\|\beta^*\|_2/\sigma_* \gg 1$. The errors can be seen to vary more strongly with the fraction of mismatches α_* than with the noise level σ_*. For example, for $\alpha_* = 0.1$, the errors are within a factor of 1.3 of the oracle, and increase to a factor of 3 and higher as α_* reaches 0.6; in particular, note the visible change of slope for α_* between 0.5 and 0.7. Figure 4 also indicates that the computationally more complex approach in Section 3.3 performs slightly better than the plug-in approach in Section 3.2, with visible differences for the smallest value of σ_* under consideration ($\sigma_* = 0.1$). As σ_* increases, these differences vanish. Both approaches significantly outperform the robust regression method SB-D19 whose performance degrades much more severely with the fraction of mismatches α_*. The differences are most pronounced for $\sigma_* \in \{0.1, 0.2\}$, and partially disappear for $\sigma_* = 1$. The latter observation can be explained by the fact that in this setting, $\|\beta^*\|_2/\sigma_* = 1$ in which case the error induced by mismatches is of the same order as that induced by additive noise.

The middle and right columns of Figure 4 show that the proposed approach also enables estimation of the parameters σ_* and α_* with small error in most settings. For $\alpha_* \in \{0.6, 0.7\}$ and/or $\sigma_* = 1$, estimation becomes a serious challenge and as result, the estimators $\widehat{\beta}$, $\widehat{\alpha}$, and $\widehat{\sigma}$ become less reliable. We hypothesize that the reason is rather of a computational than of a statistical nature, given the nonconvexity of the negative pseudo-likelihood on the one hand (whose impact becomes more pronounced as α_* increases) and the assertions of Theorem 1 on the other hand.

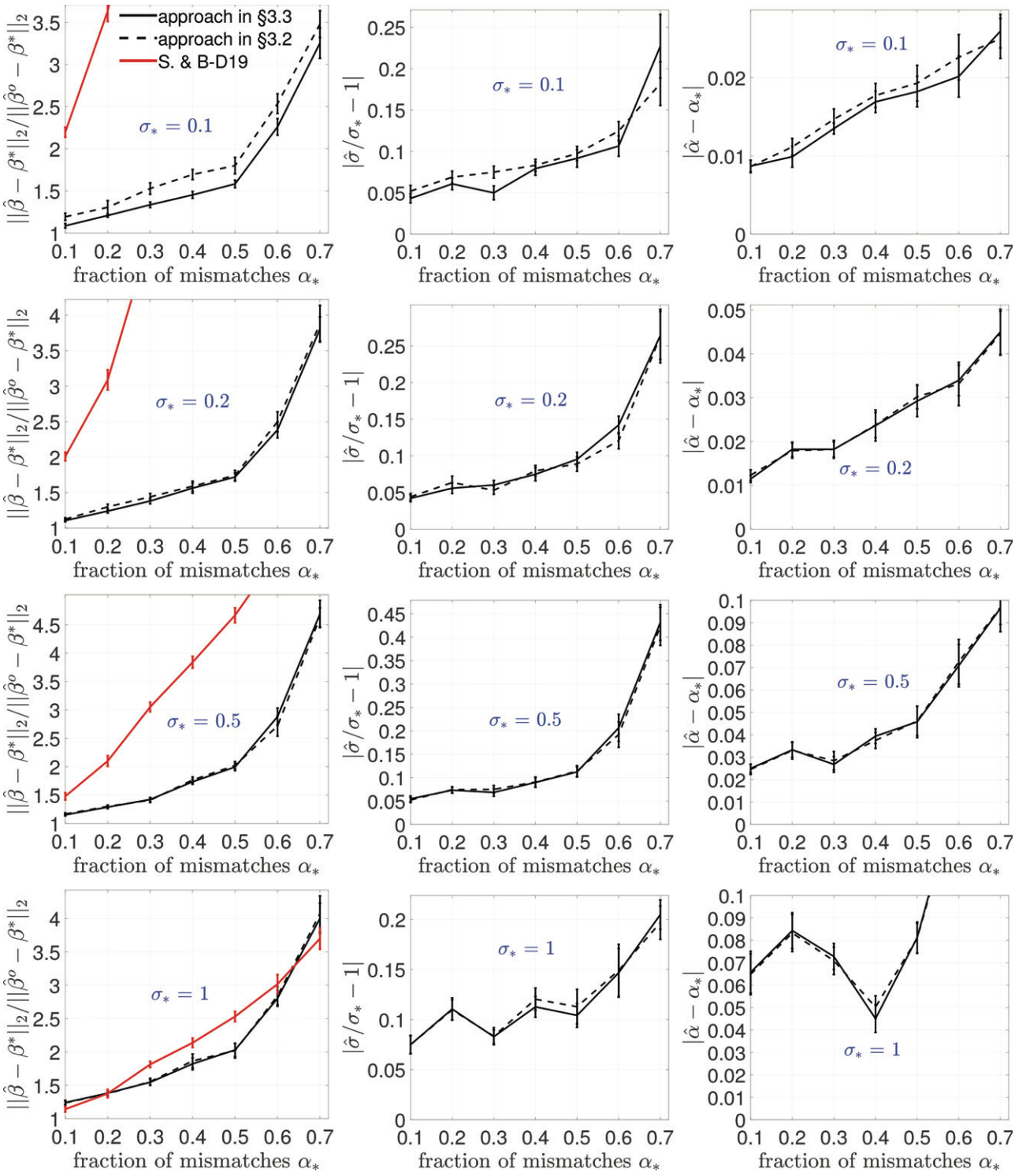

Figure 4. Visualization of the simulation results according to the metrics described in the text for varying noise level σ_* (top to bottom), divided by $\widehat{\beta}$, $\widehat{\sigma}$, and $\widehat{\alpha}$ (left to right).

4.2. CPS Wage Data

We use the CPS wage dataset available from STATLIB (*http://lib.stat.cmu.edu/datasets/*) containing information on wages and other characteristics of $n = 534$ workers, including sex, number of years of education, years of work experience, type of occupation, and union membership. To mimic the situation in record linkage, we complement this dataset with synthetic demographic information (first name, last name, zip code etc.) generated by the R package `generator` (Hendricks 2015) matching the information on sex and age in the original dataset. We recreate the response variable log(`wage`) (logarithm of the hourly wage) according to

$$\begin{aligned}\log(\texttt{wage}) = \beta_0^* + \beta_1^* \cdot I(\texttt{sex} = \text{“}F\text{”}) + \beta_2^* \cdot \texttt{experience} \\ + \beta_3^* \cdot \texttt{experience}^2 + \beta_4^* \cdot \texttt{education} + \\ + \beta_5^* \cdot I(\texttt{occupation} = \text{“Sales”})\end{aligned}$$

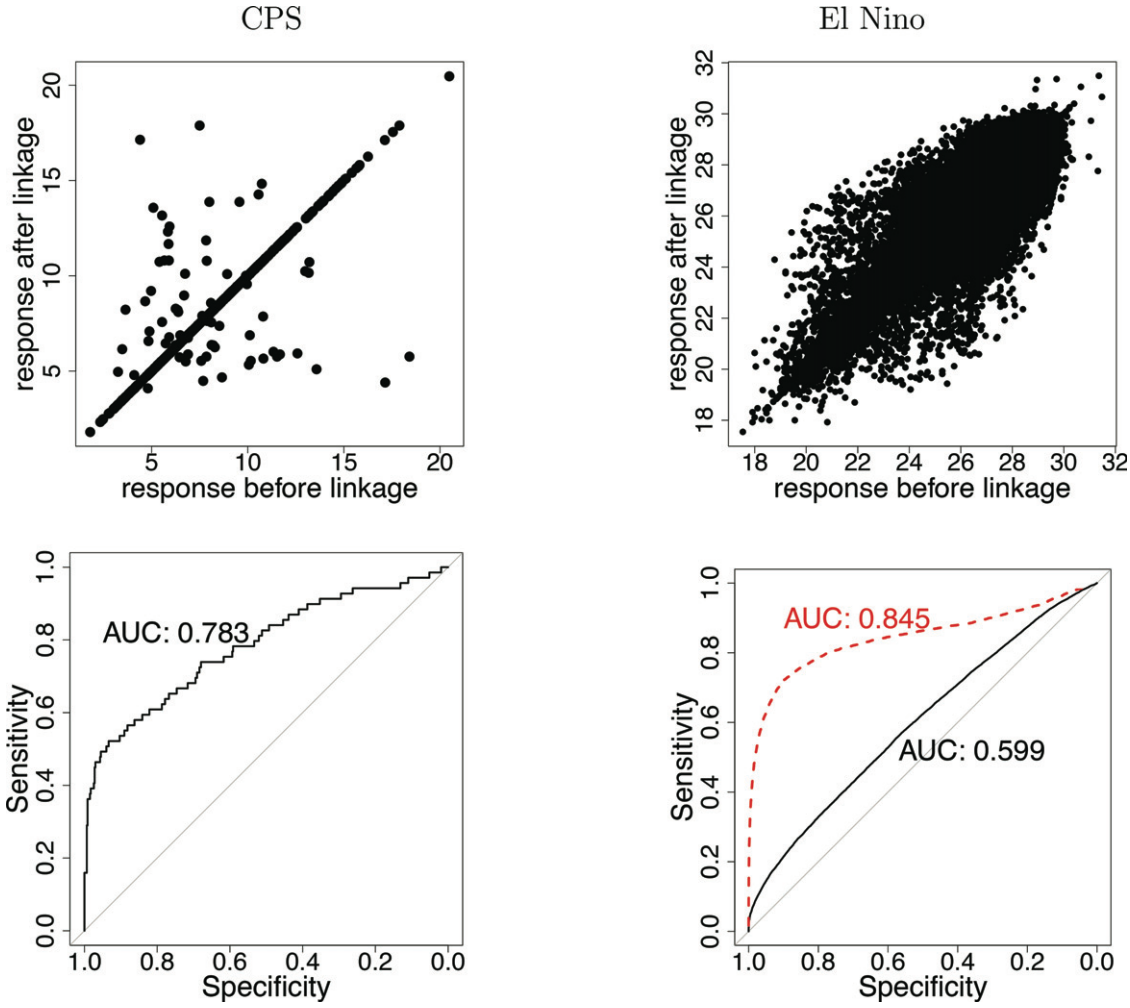

Figure 5. Top: Response variable before and after linkage for the CPS wage data (left) and the El Nino data (right); the angle bisector corresponds to the situation without mismatch. Bottom: ROC curves regarding the discrimination between mismatches and correct matches for the CPS wage data (left) and the El Nino data (right) based on the estimated posterior probabilities (18); the dashed ROC curve (right) is obtained when observations are considered mismatched only if the difference in the response before/after linkage exceeds three times the residual standard error.

$$
\begin{aligned}
&+ \beta_6^* \cdot I(\texttt{occupation} = \text{“Clerical”}) \\
&+ \beta_7^* \cdot I(\texttt{occupation} = \text{“Service”}) \\
&+ \beta_8^* \cdot I(\texttt{occupation} = \text{“Professional”}) \\
&+ \beta_9^* \cdot I(\texttt{occupation} = \text{“Other”}) \\
&+ \beta_{10}^* \cdot I(\texttt{union} = \text{“}Y\text{”}) \\
&+ \sigma_* \varepsilon, \quad \varepsilon \sim N(0,1).
\end{aligned}
$$

Here, $I(\ldots)$ represents indicator variables: “F” is short for female, “union = Y[es]” indicates membership in a union, and the variable `occupation` represents one of six occupational categories (reference category is “Management”). The variables `experience` and `education` represent work experience and education (in #years), respectively, and are both treated as numerical variables. The regression coefficients $\beta_0^*, \ldots, \beta_{10}^*$ were chosen as the coefficients from the least squares fit of the same model with the original wages. By recreating the response, we fully maintain the correlation structure of the predictors while achieving a better model fit (the choice $\sigma_* = 1.5$ leads to an R^2 close to 0.7), which helps to demonstrate the impact of linkage error and the ability of the proposed approach to provide remedy. Linkage error is generated by splitting the entire file into two files, one of which only contains the response variable and the zip code of the individuals while the second file contains all variables except for the response. The thus obtained two files were linked based on the variable zip code using the R package `fastLink` (Enamorado, Eifield, and Imai 2018). Since zip code does not represent a unique identifier, a fraction of $\alpha_* \approx 0.13$ of the records are incorrectly matched. Figure 5 displays the discrepancy between the response before and after file linkage. We compare the following approaches (i) oracle least squares based on the original undivided file, (ii) naive least squares ignoring linkage error, (iii) the robust regression method (SB) in Slawski and Ben-David (2019), (iv) the Lahiri–Larsen (LL) estimator (23), where the matrix $\mathbf{Q}$ is constructed by assuming that matching among observations with the same zip code is done uniformly at random, and (v) the proposed approach in the variant of Section 3.2 in which the solution of (iii) along with a robust estimator of the noise level (properly rescaled median absolute deviation of the residuals) is used for initialization. These five approaches are compared in terms of $\|\widehat{\beta} - \widehat{\beta}^{\text{o}}\|_2$ and $\sum_{i=1}^n (y_i - x_{\pi^*(i)}^\top \widehat{\beta})^2$ (mean squared error on the original data). For a more detailed comparison, regression coefficients and standard errors of (i), (ii), and (iii) are reported in Table 1.

Table 1. Summary of results for the CPS wage data.

	Oracle	Proposed	LL	SB	Naive
$\|\widehat{\beta} - \widehat{\beta}^{o}\|_2$	0	0.03	0.07	0.17	0.20
$\sum_{i=1}^{n}(y_i - x_{\pi^*(i)}^{\top}\widehat{\beta})^2$	23.46	23.57	23.68	24.22	24.82

	Oracle		Proposed		Naive	
$\widehat{\beta}_0$	1.01	(0.083)	1.02	(0.098)	1.2	(0.1)
$\widehat{\beta}_1$	−0.21	(0.02)	−0.22	(0.024)	−0.19	(0.026)
$\widehat{\beta}_2$	0.03	(0.0026)	0.03	(0.0035)	−0.03	(0.0033)
$\widehat{\beta}_3$	−0.0004	(0.000058)	−0.0005	(0.00008)	−0.0004	(0.000073)
$\widehat{\beta}_4$	0.07	(0.0049)	0.07	(0.006)	0.06	(0.0062)
$\widehat{\beta}_5$	−0.32	(0.045)	−0.30	(0.063)	−0.26	(0.058)
$\widehat{\beta}_6$	−0.23	(0.038)	−0.23	(0.035)	−0.23	(0.048)
$\widehat{\beta}_7$	−0.37	(0.04)	−0.38	(0.046)	−0.35	(0.051)
$\widehat{\beta}_8$	−0.05	(0.036)	−0.06	(0.031)	−0.08	(0.045)
$\widehat{\beta}_9$	−0.20	(0.037)	−0.19	(0.034)	−0.18	(0.048)
$\widehat{\beta}_{10}$	0.20	(0.025)	0.21	(0.023)	0.18	(0.031)
$\widehat{\sigma}^2$	0.045	(0.003)	0.043	(0.0044)	0.072	(0.004)
$\widehat{\alpha}$	NA		0.14	(0.027)	NA	

NOTE: Top: ℓ_2-estimation error for the regression coefficients and mean squared error. Bottom: Parameter estimates and their standard error estimates (in parentheses) of the proposed method in comparison to oracle and naive least squares.

Results. The figures in Table 1 show that the proposed approach exhibits similar performance as the oracle estimator. The estimated regression coefficients and their standard errors are rather close, and the fraction of mismatches is also estimated accurately ($\widehat{\alpha} = 0.14$ compared to $\alpha_* = 0.13$). For comparison, the changes in the regression coefficients are more noticeable for the naive least squares solution, which also yields a considerably reduced fit as is indicated by an inflation of the estimated residual variance (0.072 compared to 0.045). The robust regression method in Slawski and Ben-David (2019) yields improvements relative to the naive approach, but they are less substantial relative to the proposed approach. The latter also outperforms the Lahiri–Larsen estimator, which is equipped with additional information in terms of the matrix $\mathbf{Q}$.

4.3. El Nino Data

We here build on the case study presented in §3.2 in Slawski and Ben-David (2019) that is based on the El Nino dataset (Dheeru and Taniskidou 2017). The latter contains meteorological measurements recorded by a sensor network known as the Tropical Atmosphere Ocean Array consisting of ∼70 buoys placed across the equatorial Pacific. Sensors positioned at those buoys record zonal and meridional wind speeds (abbreviated `zon` and `mer`), relative humidity (`humidity`), air temperature (`air.temp`), sea surface temperature, and subsurface temperatures down to a depth of 500 m (`s.s.temp`). The regression model considered in Slawski and Ben-David (2019) is given by

$$\texttt{air.temp} = \beta_0^* + \beta_{\texttt{z}}^* \cdot \texttt{zon.winds} + \beta_{\texttt{m}}^* \cdot \texttt{mer.winds} + \beta_{\texttt{h}}^* \cdot \texttt{humidity} + \beta_{\texttt{s}}^* \cdot \texttt{s.s.temp} + \varepsilon.$$

Each set of measurements is uniquely identified by the buoy identifier and the day of its recording. In Slawski and Ben-David (2019), this information is discarded, and the response variable (`air.temp`) is put into a separate file that additionally contains the longitude and latitude of the measurement as an inexact identifier. The latter is used subsequently as matching variable by `fastLink` to merge the response variable with the predictor variables. The right panel of Figure 5 shows that the error induced by mismatches is rather substantial, with a fraction of 0.82 of the $n = 93{,}935$ observations being mismatched. However, not all mismatches lead to substantial changes in the response: for example, only a fraction of 0.16 of the observations is associated with an error in the response larger than twice the residual standard error from the oracle least squares fit. The fact that the majority of mismatches does not lead to major errors is ultimately a consequence of the fact that meteorological measurements sharing the same (latitude, longitude)-pair exhibit spatial correlation even though they may not correspond to the same observational unit.

Results. According to Table 2, the proposed method is not far from the oracle estimator. The estimates for the regression coefficients are noticeably closer than those of the naive approach. The latter yields a poor fit, with the residual variance inflated by more than a factor of two (0.594 vs. 0.259). The proposed approach also outperforms the method in Slawski and Ben-David (2019) as well as the Lahiri–Larsen estimator (with $\mathbf{Q}$ constructed analogously to the previous subsection, with zip code replaced by (longitude, latitude)) in terms of the ℓ_2-estimation error and mean squared error. The performance of the Lahiri–Larsen is suboptimal here due to a small number of distinct (latitude, longitude)-pairs relative to the sample size (about 5k vs. 94k); by construction of the matrix $\mathbf{Q}$, the Lahiri–Larsen estimator here amounts to averaging predictors and response with the same (latitude, longitude), and a subsequent weighted least squares fit with the thus obtained averages. The effective sample size is hence reduced to 5k. Moreover, it is worth noting that the proposed approach estimates the fraction of mismatches as approximately 0.073, whereas the nominal fraction of mismatches is around 0.82. This gap results from the fact that the majority of mismatches do not substantially change the response compared to the error of the regression model as explained above; cf. also the ROC curves in Figure 5. The estimate $\widehat{\alpha} = 0.073$ turns out to be close to the fraction of mismatches that change the response by three times the

Table 2. Summary of results for the El Nino data.

	Oracle	Proposed	LL	SB	Naive
$\|\widehat{\beta} - \widehat{\beta}^0\|_2$	0	0.04	1.56	0.59	1.57
$\sum_{i=1}^n (y_i - x_{\pi^*(i)}^\top \widehat{\beta})^2$	24.4k	24.7k	25.2k	25.1k	25.9k

	Oracle		Proposed		Naive	
$\widehat{\beta}_0$	5.15	(0.049)	5.12	(0.073)	6.72	(0.075)
$\widehat{\beta}_z$	−0.056	(0.00055)	−0.044	(0.00081)	−0.037	(0.00083)
$\widehat{\beta}_m$	−0.031	(0.00058)	−0.038	(0.00078)	−0.045	(0.00089)
$\widehat{\beta}_h$	−0.022	(0.00034)	−0.016	(0.00048)	−0.017	(0.00051)
$\widehat{\beta}_s$	0.844	(0.0011)	0.827	(0.0017)	0.774	(0.0017)
$\widehat{\sigma}^2$	0.259	(0.0012)	0.358	(0.0026)	.594	(0.0027)
$\widehat{\alpha}$	NA		0.073	(0.0014)	NA	

NOTE: Top: ℓ_2-estimation error for the regression coefficients and mean squared error. Bottom: Parameter estimates and their standard error estimates (in parentheses) of the proposed method in comparison to oracle and naive least squares.

residual standard error or more. Even though the proposed estimator is effective in curbing the impact of mismatch error, the underlying assumption in Section 2 that the response is independent of the predictors in the case of a mismatch appears to be violated; this is a consequence of noticeable correlations between the predictors and the variables used for linking as mentioned above. The mixture model (4) is thus at least mildly misspecified. In light of this, it is plausible that $\widehat{\alpha}$ and $\widehat{\sigma}^2$ result as underestimate and (slight) overestimate, respectively, while the estimates for the regression coefficients are subject to a minor attenuation effect.

5. Conclusion

In this article, we have presented a pseudo-likelihood method to account for mismatches in the response variables in linear regression, an important problem in the analysis of linked files. The proposed method is computationally scalable, requires at most minimum tuning, provides estimators of all parameters of interest, and achieves promising empirical performance according to the results in the preceding section. In light of these appealing properties, we hope that the method will be widely adopted to deal with the scenarios discussed herein. Owing to its simple modular structure, the method considered herein can be generalized to a variety of other regression models including multiple response variables, generalized linear models and nonparametric regression, which will be investigated in future work. Another interesting direction of research concerns the adjustment for mismatches in the situation where a subset of the predictor variables is contained in the same file as the response, while the remaining predictors are contained in a separate file.

Supplementary Materials

Supplement: The supplementary file provides proofs of the statements herein and further technical result and derivations, as well as additional simulation results. (.pdf file)

Code.zip: R and MATLAB code that reproduces the results of the simulation studies and the real data analysis. (.zip file)

Acknowledgments

We sincerely thank two reviewers and an associate editor for their thorough reading of our manuscript and many excellent comments and suggestions that led to substantial improvements of this work.

Funding

The first author was partially supported by the NSF grant CCF-1849876.

References

Abid, A., Poon, A., and Zou, J. (2017), "Linear Regression With Shuffled Labels," arXiv no. 1705.01342. [1,2]

Abid, A., and Zou, J. (2018), "Stochastic EM for Shuffled Linear Regression," in *Allerton Conference on Communication, Control, and Computing*, pp. 470–477. [2]

Aitkin, M. (1981), "A Note on the Regression Analysis of Censored Data," *Technometrics*, 23, 161–163. [7]

Balakhrisnan, A. (1962), "On the Problem of Time Jitter in Sampling," *IRE Transactions on Information Theory*, 8, 226–236. [1]

Bertsekas, D. (1999), *Nonlinear Programming* (2nd ed.), Boston: Athena Scientific. [7]

Burkard, R., Dell'Amico, M., and Martello, S. (2009), *Assignment Problems: Revised Reprint*, Philadelphia, PA: SIAM. [2]

Carpentier, A., and Schlüter, T. (2016), "Learning Relationships Between Data Obtained Independently," in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 658–666. [2]

Chambers, R. (2009), "Regression Analysis of Probability-Linked Data," Technical Report, Statistics New Zealand. [2]

Chen, J., and Li, P. (2009), "Hypothesis Test for Normal Mixture Models: The EM Approach," *The Annals of Statistics*, 37, 2523–2542. [5]

Dalzell, N., and Reiter, J. (2018), "Regression Modeling and File Matching Using Possibly Erroneous Matching Variables," *Journal of Computational and Graphical Statistics*, 27, 728–738. [2]

DasGupta, S., and Gupta, A. (2003), "An Elementary Proof of a Theorem of Johnson and Lindenstrauss," *Random Structures and Algorithms*, 22, 60–65.

DeGroot, M., Feder, P., and Goel, P. (1971), "Matchmaking," *The Annals of Mathematical Statistics*, 42, 578–593. [1]

DeGroot, M., and Goel, P. (1976), "The Matching Problem for Multivariate Normal Data," *Sankhyā: The Indian Journal of Statistics*, Series B, 38, 14–29. [1]

——— (1980), "Estimation of the Correlation Coefficient From a Broken Random Sample," *The Annals of Statistics*, 8, 264–278. [1]

Dheeru, D., and Taniskidou, E. K. (2017), "UCI Machine Learning Repository," available at *http://archive.ics.uci.edu/ml*. [11]

Emiya, V., Bonnefoy, A., Daudet, L., and Gribonval, R. (2014), "Compressed Sensing With Unknown Sensor Permutation," in *Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1040–1044. [1,2]

Enamorado, E., Eifield, B., and Imai, K. (2018), "Fast Probabilistic Record Linkage With Missing Data," R Package Version 0.5.0. [10]

Goel, P. (1975), "On Re-Pairing Observations in a Broken Sample," *The Annals of Statistics*, 3, 1364–1369. [1]

Goel, P., and Ramalingam, T. (1987), "Some Properties of the Maximum Likelihood Strategy for Re-Pairing a Broken Random Sample," *Journal of Statistical Planning and Inference*, 16, 237–248. [1]

Gutman, R., Afendulis, C., and Zaslavsky, A. (2013), "A Bayesian Procedure for File Linking to Analyze End-of-Life Medical Costs," *Journal of the American Statistical Association*, 108, 34–47. [2]

Haghighatshoar, S., and Caire, G. (2017), "Signal Recovery From Unlabeled Samples," in *International Symposium on Information Theory (ISIT)*, pp. 451–455. [1]

Han, Y., and Lahiri, P. (2019), "Statistical Analysis With Linked Data," *International Statistical Review*, 87, S139–S157. [2]

Hendricks, P. (2015), "`generator`: Generate Data Containing Fake Personally Identifiable Information," R Package Version 0.1.0, available at *https://CRAN.R-project.org/package=generator*. [9]

Hof, M. H. P., and Zwinderman, A. H. (2012), "Methods for Analyzing Data From Probabilistic Linkage Strategies Based on Partially Identifying Variables," *Statistics in Medicine*, 31, 4231–4242. [2]

——— (2015), "A Mixture Model for the Analysis of Data Derived From Record Linkage," *Statistics in Medicine*, 34, 74–92. [2]

Hsu, D., Shi, K., and Sun, X. (2017), "Linear Regression Without Correspondence," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 1531–1540. [1,2]

Lahiri, P., and Larsen, M. D. (2005), "Regression Analysis With Linked Data," *Journal of the American Statistical Association*, 100, 222–230. [2,3,7]

Lange, K. (1995), "A Gradient Algorithm Locally Equivalent to the EM Algorithm," *Journal of the Royal Statistical Society*, Series B, 57, 425–437. [3,7]

Lindsay, B. (1988), "Composite Likelihood Methods," *Contemporary Mathematics*, 80, 221–239. [4]

Maronna, R., Martin, R., and Yohai, V. (2006), *Robust Statistics: Theory and Methods*, New York: Wiley. [4]

Neter, J., Maynes, S., and Ramanathan, R. (1965), "The Effect of Mismatching on the Measurement of Response Error," *Journal of the American Statistical Association*, 60, 1005–1027. [2]

Pananjady, A., Wainwright, M., and Cortade, T. (2017), "Denoising Linear Models With Permuted Data," arXiv no. 1704.07461. [1]

——— (2018), "Linear Regression With Shuffled Data: Statistical and Computational Limits of Permutation Recovery," *IEEE Transactions on Information Theory*, 64, 3826–3300. [1,2,3,5]

Rigollet, P., and Weed, J. (2019), "Uncoupled Isotonic Regression via Minimum Wasserstein Deconvolution," *Information & Inference*, 8, 691–717. [2]

Scheuren, F., and Winkler, W. (1993), "Regression Analysis of Data Files That Are Computer Matched I," *Survey Methodology*, 19, 39–58. [2]

——— (1997), "Regression Analysis of Data Files That Are Computer Matched II," *Survey Methodology*, 23, 157–165. [2]

Shi, X., Lu, X., and Cai, T. (2020), "Spherical Regresion Under Mismatch Corruption With Application to Automated Knowledge Translation," *Journal of the American Statistical Association* (to appear). [2]

Slawski, M., and Ben-David, E. (2019), "Linear Regression With Sparsely Permuted Data," *Electronic Journal of Statistics*, 13, 1–36. [2,3,4,7,8,10,11]

Slawski, M., Ben-David, E., and Li, P. (2020), "A Two-Stage Approach to Multivariate Linear Regression With Sparsely Mismatched Data," *Journal of Machine Learning Research*, 21, 1–42. [2]

Slawski, M., Rahmani, M., and Li, P. (2019), "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)*. [2]

Titterington, M. (1984), "Recursive Parameter Estimation Using Incomplete Data," *Journal of the Royal Statistical Society*, Series B, 46, 257–267. [3,7]

Tsakiris, M. (2018), "Eigenspace Conditions for Homomorphic Sensing," arXiv no. 1812.07966. [2]

Tukey, J. (1977), *Exploratory Data Analysis* (3rd ed.), Reading, MA: Addison-Wesley.

Unnikrishnan, J., Haghighatshoar, S., and Vetterli, M. (2018), "Unlabeled Sensing With Random Linear Measurements," *IEEE Transactions on Information Theory*, 64, 3237–3253. [1,2]

Unnikrishnan, J., and Vetterli, M. (2013), "Sampling and Reconstruction of Spatial Fields Using Mobile Sensors," *IEEE Transactions on Signal Processing*, 61, 2328–2340. [1]

van der Vaart, A. (1998), *Asymptotic Statistics*, Cambridge: Cambridge University Press. [4]

Varin, C., Reid, N., and Firth, D. (2011), "An Overview of Composite Likelihood Estimation," *Statistica Sinica*, 21, 5–42. [4]

Wu, Y. N. (1998), "A Note on Broken Sample Problem," Technical Report, Department of Statistics, University of Michigan. [2]

Zhang, H., Slawski, M., and Li, P. (2019), "Permutation Recovery From Multiple Measurement Vectors in Unlabeled Sensing," in *2019 IEEE International Symposium on Information Theory (ISIT)*, pp. 1857–1861. [2]

Zhu, H.-T., and Zhang, H. (2004), "Hypothesis Testing in Mixture Regression Models," *Journal of the Royal Statistical Society*, Series B, 66, 3–16. [5]