Regularized optimal transport of covariates and outcomes in data recoding

Valérie Garès, Jérémy Omer

Regularized optimal transport of covariates and outcomes in data recoding

Valérie Garès[1] & Jérémy Omer[2]

[1] *Univ Rennes, INSA, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France, valerie.gares@insa-rennes.fr*

[2] *Univ Rennes, INSA, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France, jeremy.omer@insa-rennes.fr*

Abstract

When databases are constructed from heterogeneous sources, it is not unusual that different encodings are used for the same outcome. In such case, it is necessary to recode the outcome variable before merging two databases. The method proposed for the recoding is an application of optimal transportation where we search for a bijective mapping between the distributions of such variable in two databases. In this article, we build upon the work by Garès et al. [9], where they transport the distributions of categorical outcomes assuming that they are distributed equally in the two databases. Here, we extend the scope of the model to treat all the situations where the covariates explain the outcomes similarly in the two databases. In particular, we do not require that the outcomes be distributed equally. For this, we propose a model where joint distributions of outcomes and covariates are transported. We also propose to enrich the model by relaxing the constraints on marginal distributions and adding an L^1 regularization term. The performances of the models are evaluated in a simulation study, and they are applied to a real dataset.

Keywords. Merging databases; Variable recoding; Linkage; Optimal Transportation; Statistical matching; Heterogeneous sources; Domain adaptation; Epidemiology.

1 Introduction

Sharing and producing information from heterogeneous sources becomes a major issue. An objective when merging databases is to associate, mix and include databases from different sources in order to provide a strong knowledge database. This allows us to extract more information from merged data than we would obtain from using the databases separately [3, 10]. Here we focus on a specific issue related to data merging, the recoding problem: an issue that occurs when merging two databases where a variable is not coded in the same scale in both databases. For instance, recoding variables may be necessary before merging databases collected during the same study. Indeed, the survey questionnaire

may be modified when collecting the same information in two waves of recruitment with different subjects or in two waves with the same subjects at different ages. It can also be useful to merge two databases collected during different studies that focus on the same information. As an illustration, a previous work on data recoding [9] is applied on a French cohort study, the ELFE study, where the variable of interest is the answer to the question :"how would you rate your overall health?". During the first baseline data collection wave (January to April 2011), the different possible answers were proposed in a five points ordinal scale: "excellent", "very well", "well", "fair", "bad" and during the second baseline data collection wave (May to December 2011), another five points ordinal scale was used: "very well", "well", "medium", "bad" and "very bad".

As illustrated in Figure 1, the problem can be formalized in terms of two databases A and B: A contains the observations of a vector X of P covariates and of one outcome Y measured on n_A units; B contains the observations of X and of one outcome Z for n_B other subjects. In particular, Z is not observed for the subjects of A, and neither is Y for the subjects of B. More generally, there is no subject for whom Y and Z are simultaneaously observed. What is more, we assume that Y and Z are categorical variables that refer to the same latent variable. Hence Y and Z can be seen as two different encodings of the same variable that can have different numbers of modalities (or categories). The problem can then be enunciated as a recoding problem where we would like to predict Z in database A (which, symmetrically, is equivalent to the prediction of Y in database B). More precisely, since the subjects of A are only characterized by the values of X and Y, we wish to estimate the distribution of Z given the values of X and Y in base A.

Variable recoding can be seen as a missing data problem. In this context, the missing value depends only on the database to which the subject belongs and not on the variable itself. The missingness mechanism can then be considered as *missing at random*. This problem has been widely studied in the literature ([15]) and many existing methods for treating missing data could be used. Moreover, Y and Z refer to the same information, which can be interpreted as a latent variable. Methods of prediction of this latent class could also be applied (e.g. *class latent analysis* or *trait latent analysis* [1, 17]). Finally, methods for classification learning first estimate the distribution of Z from covariates using database B and predict Z in A in a second step ([19]). The major drawback of such approaches is that the resulting methods use the information contained in each database in two independent steps that cannot capture the interrelations between them.

Recently, Garès et al. [9] have proposed a method based on optimal transportation (OT). Assuming that the distributions of Y and Z are the same in the two databases, the OT theory ([20]) provides a map that pushes the distribution of Y forward to the distribution of Z. This approach has shown better performance when compared to classical methods such as multiple imputation [9] and machine learning methods [21], but it still exhibits two limits. First, there are several contexts where it will not be true that Y has the same distributions in the two databases. For instance, this has already been observed when comparing North American NHANES study and the French National Health

Survey. The "self-rated overall health" outcome is not distributed identically in the two databases, where the rates of functional limitations and education level are different [6]. The second limit is that their OT model does not actually solve the recoding problem. The authors use OT to derive only the joint distribution of Y and Z. To actually predict Z in A, they have to execute a greedy nearest neighbor algorithm, which makes some arbitrary decisions.

Our aim is to build upon the work of Garès et al. [9] to develop a recoding method that requires less restrictive assumptions and directly targets the solution of the recoding problem. In particular, we consider that Y and Z may be distributed differently in A and B, but we still focus on databases where covariates explain the outcomes Y and Z similarly in the two databases. This restriction remains necessary, because the only information we have to characterize the subjects is the set of common covariates. In particular, if outcomes are independent from covariates, recoding is doomed to failure unless additional information is provided.

Our main contribution is in the development of an OT model where the joint distribution of X and Y is pushed forward to that of X and Z. This allows to derive the distribution of Y given the values of X and Z directly from the optimal solution of the model. Moreover, since the observations in the two databases can only give us access to estimators of the marginal distributions and of a particular cost function, the solution of the OT model is itself only an estimator of the optimal distribution of (X, Y, X, Z). From the observation that all the estimators used in the OT model are strongly consistent, we show that the estimator of the distribution of (X, Y, X, Z) is also strongly consistent. As an extension of the OT model, we then propose to relax the constraints on marginals, because they may be too restrictive in the presence of errors in the estimations of their right-hand side. We also add a regularization term to the objective function to smooth the variations of outcomes with respect to covariates. Such regularization has already been employed by [8] with success in similar models.

The remainder of the article is organized as follows. The data recoding problem we focus on is formalized in Section 2. We then recall some elements about OT theory and describe the OT model introduced by Garès et al. [9] in Section 3. In Section 4, we introduce the OT model where the joint distribution of X and Y is transported to that of X and Z. Since the model we solve is defined from empirical estimators, we then study the convergence of its solutions. In Section 5, we develop an extension of the OT model where constraints on marginals are relaxed and a regularization term is considered. Finally, in Section 6, the performances of the models are assessed on a benchmark of simulated databases, and they are applied on a real example extracted from the National Child Development (NCDS) data collection [18] in Section 7.

Database A

	$X \in \mathbb{R}^P$	$Y \in \mathbb{R}$	$Z \in \mathbb{R}$
1	Observed	Observed	Unobserved
...			
...			
n_A			

Database B

	$X \in \mathbb{R}^P$	$Y \in \mathbb{R}$	$Z \in \mathbb{R}$
1	Observed	Unobserved	Observed
...			
...			
n_B			

Table 1: Problem formulation

2 Definition of the problem and notations

2.1 Formal statement of the data recoding problem

Let A and B be two independent databases corresponding to two sets of subjects. For more concise notations, we assume without loss of generality that the two databases have equal sizes, so that they can be written as $A = \{i_1, \ldots, i_n\}$ and $B = \{j_1, \ldots, j_n\}$. Let $((X_i, Y_i, Z_i))_{i \in A}$ and $((X_j, Y_j, Z_j))_{j \in B}$ be two sequences of i.i.d. discrete random variables with outcomes in $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$, where $\mathcal{X}$ is a finite subset of $\mathbb{R}^P$, and $\mathcal{Y}$ and $\mathcal{Z}$ are finite subsets of $\mathbb{R}$. Variables $(X_i, Y_i, Z_i), i \in A$, are i.i.d copies of (X^A, Y^A, Z^A) and $(X_j, Y_j, Z_j), j \in B$, are i.i.d copies of (X^B, Y^B, Z^B). By independence of the databases, we moreover assume that (X^A, Y^A, Z^A) and (X^B, Y^B, Z^B) are independent. Every random variable is defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Finally, we assume that for all $x \in \mathcal{X}$ the probability distributions of Y^A and Z^A given that $X^A = x$ are respectively equal to those of Y^B and Z^B given that $X^B = x$, i.e.,

$$\begin{aligned} &\mathbb{P}(Y^A = y \mid X^A = x) = \mathbb{P}(Y^B = y \mid X^B = x),\ \forall x \in \mathcal{X}, \forall y \in \mathcal{Y},\ \text{and} \\ &\mathbb{P}(Z^A = z \mid X^A = x) = \mathbb{P}(Z^B = z \mid X^B = x),\ \forall x \in \mathcal{X}, \forall z \in Z. \end{aligned} \tag{1}$$

In particular, assumption (1) implicitly states that $\mathbb{P}(Y^A = y \mid X^A = x)$ is defined if and only if $\mathbb{P}(Y^B = y \mid X^B = x)$ is defined, i.e., $\mathbb{P}(X_A = x) = 0$ if and only if $\mathbb{P}(X_B = x) = 0$. Without loss of generality, we thus restrict the domain of the covariates $\mathcal{X}$ to the values, x, such that $\mathbb{P}(X_A = x) \neq 0$ and $\mathbb{P}(X_B = x) \neq 0$.

The data recoding problem consists in the prediction of $(Z_i)_{i \in A}$ from independent observations of $((X_i, Y_i))_{i \in A}$ and $((X_j, Z_j))_{j \in B}$. To be more specific, the problem is to propose an estimator of the distribution of Z^A conditional to $X^A = x$ and $Y^A = y$, $\left\{\mathbb{P}(Z^A = z | X^A = x, Y^A = y), z \in \mathcal{Z}\right\}$, $x \in \mathcal{X}, y \in \mathcal{Y}$, from the observations of X and Y in A, $\{(x_i, y_i)\}_{i \in A}$, and of X and Z in B, $\{(x_j, z_j)\}_{j \in B}$.

Observe that since Y and Z are never jointly observed for the same subject, the assumption on conditional distributions (1) cannot be tested from the data.

2.2 Distributions and estimators

In the remainder of the article, the discrete probability distribution of a discrete random variable V with possible outcomes in $\mathcal{V}$ is given by

$$\mu^V = \sum_{v \in \mathcal{V}} \mu_v^V \delta_v,$$

where δ_v is the Dirac delta measure centered at v. If $\mathcal{V}$ is finite with cardinal $|V|$, μ^V will also refer to the vector of weights $(\mu_v^V)_{v \in V}$.

Vectors $((x_i, y_i))_{i \in A}$ and $((x_j, z_j))_{j \in B}$ are realizations of two n samples of random variables $((X_i, Y_i))_{i \in A}$ and $((X_j, Y_j))_{j \in B}$ with unknown joint distribution $\mu^{(X^A, Y^A)}$ and $\mu^{(X^B, Z^B)}$. As a consequence, we will consider their unbiased empirical estimators given by

$$\hat{\mu}_{n,x,y}^{(X^A,Y^A)} = \frac{1}{n} \sum_{i \in A} \mathbb{1}_{\{X_i = x, Y_i = y\}}, \ \forall x \in \mathcal{X}, y \in \mathcal{Y}, \tag{2}$$

$$\hat{\mu}_{n,x,z}^{(X^B,Z^B)} = \frac{1}{n} \sum_{j \in B} \mathbb{1}_{\{X_j = x, Z_j = z\}}, \ \forall x \in \mathcal{X}, z \in \mathcal{Z}. \tag{3}$$

The strong law of large numbers applied to the sequences of i.i.d. random variables $((X_i, Y_i))_{i \in A}$ and $((X_j, Z_j))_{j \in B}$ directly yields that

$$\hat{\mu}_{n,x,y}^{(X^A,Y^A)} \xrightarrow[n \to +\infty]{a.s.} \mu_{x,y}^{(X^A,Y^A)}, \forall x \in \mathcal{X}, \forall y \in \mathcal{Y}$$
$$\hat{\mu}_{n,x,z}^{(X^B,Z^B)} \xrightarrow[n \to +\infty]{a.s.} \mu_{x,z}^{(X^B,Z^B)}, \forall x \in \mathcal{X}, \forall z \in \mathcal{Z}$$

In what follows, we will also need an estimator of $\mu^{(X^A,Z^A)}$. Let $x \in \mathcal{X}$ ($\mathbb{P}(X^B = x) \neq 0$) and $z \in \mathcal{Z}$, then

$$\begin{aligned} \mathbb{P}(X^A = x, Z^A = z) &= \mathbb{P}(Z^A = z \mid X^A = x)\mathbb{P}(X^A = x) \\ &= \mathbb{P}(Z^B = z \mid X^B = x)\mathbb{P}(X^A = x) \\ &= \mathbb{P}(X^B = x, Z^B = z)\frac{\mathbb{P}(X^A = x)}{\mathbb{P}(X^B = x)}, \end{aligned}$$

where the second equality is a direct application of the assumption (1). Denoting as $\hat{\mu}_n^{X^A}$ and $\hat{\mu}_n^{X^B}$ the unbiased empirical estimators of μ^{X^A} and μ^{X^B}, we then consider the estimator of $\mu^{(X^A,Z^A)}$ given by

$$\hat{\mu}_{n,x,z}^{(X^A,Z^A)} \hat{\mu}_{n,x}^{X^B} = \hat{\mu}_{n,x,z}^{(X^B,Z^B)} \hat{\mu}_{n,x}^{X^A}, \ \forall x \in \mathcal{X}, \ \forall z \in Z. \tag{4}$$

The almost sure convergence of all the estimators in the expression of $\hat{\mu}_{n,x,z}^{(X^A,Z^A)}$ yields

$$\hat{\mu}_{n,x,z}^{(X^A,Z^A)} \xrightarrow[n \to +\infty]{a.s.} \mu_{x,z}^{(X^A,Z^A)}, \forall x \in \mathcal{X}, \forall z \in \mathcal{Z}.$$

Remark 1 (Estimators for comparable populations)**.** *Observe that if we add the assumption that X^A and X^B are copies of the same random variable X, we fall back in the framework of [9] where Z is distributed equally in bases A and B. We immediately see that in such case, we can use the unbiased estimator $\hat{\mu}_n^{(X^A,Z^A)} = \hat{\mu}_n^{(X^B,Z^B)}$.*

3 State of the art on optimal transportation models for data recoding

3.1 Abstract statement of the optimal transportation problem

Consider a pile of sand distributed with density f, that has to be moved to fill a hole (with the same volume) according to a new distribution, with a prescribed density g. Consider a map T describing this movement: $T(x)$ represents the destination of the sand originally located at x. The OT problem consists in finding a map T such that the average displacement is minimal, where the displacement between two points x and y is measured by a given cost function c. This is the original statement of the transportation problem due to Gaspard Monge [16].

Consider $\mathbb{X}$ and $\mathbb{Y}$ two Radon spaces. Let μ^X be a probability measure on $\mathbb{X}$, μ^Y a probability measure on $\mathbb{Y}$ and $c : \mathbb{X} \times \mathbb{Y} \longrightarrow [0, \infty]$ a Borel-measurable function. Let two random variables, X and Y, which respectively follow distributions μ^X and μ^Y. Kantorovich's formulation of the OT problem consists in finding a measure $\gamma \in \Gamma(\mu^X, \mu^Y)$ that realizes the infimum:

$$\inf \left\{ \mathbb{E}[c(X,Y)] = \int_{\mathbb{X}\times\mathbb{Y}} c(x,y)\, \mathrm{d}\gamma(x,y) \middle| \gamma \in \Gamma(\mu,\nu) \right\}, \tag{5}$$

where $\Gamma(\mu^X, \mu^Y)$ is the set of measures on $\mathbb{X} \times \mathbb{Y}$ with marginals μ^X on $\mathbb{X}$ and μ^Y on $\mathbb{Y}$ [14]. Kantorovich's formulation plugs the problem in a linear setting and the solution is achievable thanks to compacity argument. It can be shown [20] that a minimizer for this problem always exists as soon as the cost function c is lower semi-continuous.
In this work, we will consider the Kantovorich's formulation adapted to the discrete case, known as Hitchcock's problem [12].

3.2 Optimal transportation of outcomes

The first OT approach for data recoding is described in [9]. The authors of [9] consider a problem similar to that stated in Section 2. The only difference is that they make the additional assumption that Y^A and Z^A respectively follow the same distributions as Y^B and Z^B.

In this setting, they aim at solving the OT problem (5) that pushes μ^{Y^A} forward to μ^{Z^A}. As a consequence, variable γ of (5) is a discrete measure with marginals μ^{Y^A} and

μ^{Z^A}, represented by a $|\mathcal{Y}| \times |\mathcal{Z}|$ matrix. The cost, denoted as c is a $|\mathcal{Y}| \times |\mathcal{Z}|$ matrix, $(c_{y,z})_{y\in\mathcal{Y},z\in\mathcal{Z}}$. To be specific, their goal is in the identification of

$$\gamma^* \in \text{argmin}_{\gamma\in\mathbb{R}_+^{|\mathcal{Y}|\times|\mathcal{Z}|}} \left\{ \langle \gamma | c \rangle : \gamma \mathbf{1}_{|\mathcal{Z}|} = \mu^{Y^A}, \gamma^T \mathbf{1}_{|\mathcal{Y}|} = \mu^{Z^A} \right\}, \tag{6}$$

where $\langle \cdot | \cdot \rangle$ is the dot product, $\mathbf{1}$ is a vector of ones with appropriate dimension and M^T is the transpose of matrix M. In the cost function considered by Garès et al., $c_{y,z}$ measures the average distance between the covariates of subjects of A satisfying $Y = y$ and subjects of B satisfying $Z = z$, that is

$$c_{y,z} = \mathbb{E}\left[d(X^A, X^B) \mid Y^A = y, Z^B = z \right], \tag{7}$$

where d can be any distance function defined on $\mathcal{X} \times \mathcal{X}$. The significance of their choice is that it allows for a clear connection between the structures observed in databases A and B.

Remark 2. *$d(X^A, X^B)$ is the distance between vectors of categorical covariates. Here a Hamming distance from the associated complete disjunctive tables is used but other distances are adapted to ordinal categorical variables: Spearman, Chebyshev, Kendall or Cayley distances.*

The above model cannot be solved in reality, since the distributions of X^A, X^B, Y^A and Z^A are not known. As a consequence, they use the unbiased empirical estimators $\hat{\mu}_n^{X^A}$ of $\mu_n^{X^A}$ and $\hat{\mu}_n^{X^B}$ of $\mu_n^{X^B}$. Observations Y and Z are only available in A and B (respectively), so they define the two empirical estimators

$$\begin{aligned} \hat{\mu}_{n,y}^{Y^A} &= \frac{1}{n} \sum_{i\in A} \mathbb{1}_{\{Y_i = y\}}, \ \forall y \in \mathcal{Y} \\ \hat{\mu}_{n,z}^{Z^A} &= \frac{1}{n} \sum_{j\in B} \mathbb{1}_{\{Z_j = z\}}, \ \forall z \in \mathcal{Z}, \end{aligned} \tag{8}$$

where they use their assumption that $\mu^{Z^A} = \mu^{Z^B}$. Finally, denoting as

$$\kappa_{n,y,z} \equiv \sum_{i\in A} \sum_{j\in B} \mathbb{1}_{\{Y_i = y, Z_j = z\}}$$

the number of pairs $(i, j) \in A \times B$ such that $Y_i = y$ and $Z_j = z$, the cost matrix c is estimated by

$$\hat{c}_{n,y,z} = \begin{cases} \frac{1}{\kappa_{n,y,z}} \sum_{i\in A} \sum_{j\in B} \mathbb{1}_{\{Y_i = y, Z_j = z\}} \times d(X_i, X_j), & \forall y \in \mathcal{Y}, z \in \mathcal{Z} : \kappa_{n,y,z} \neq 0, \\ 0, & \forall y \in \mathcal{Y}, z \in \mathcal{Z} : \kappa_{n,y,z} = 0. \end{cases} \tag{9}$$

Plugging the values observed for these estimators in (6) yield a linear programming model denoted as $\hat{\mathcal{P}}_n^0$. The solution $\hat{\gamma}_n$ can then be interpreted as an estimator $\hat{\mu}_n^{(Y^A,Z^A)}$ of the joint distribution of Y^A and Z^A, $\mu^{(Y^A,Z^A)}$.

One specificity of this approach is that it does not directly identify an estimator of the distributions that are searched for, $\mu^{Z^A|X^A=x,Y^A=y}, \forall x \in \mathcal{X}, \forall y \in \mathcal{Y}$. As a consequence, the authors of [9] predict Z^A in a second step where they execute a nearest neighbor algorithm. For this, they partition A and B according to the observations by defining

$$\mathcal{O}_y = \{i \in A \mid y_i = y\}, \forall y \in \mathcal{Y} \text{ and } \mathcal{O}_z = \{j \in B \mid z_j = z\}, \forall z \in \mathcal{Z}.$$

They also define $N_{y,z}$ as the (rounded) expected number of subjects with modalities $(Y^A, Z^A) = (y, z)$ if (Y^A, Z^A) follows $\hat{\mu}_n^{(Y^A,Z^A)}$. Algorithm 1 details how the prediction is then made. For a subject $i \in A$, the nearest modality $z \in \mathcal{Z}$ is then computed according to the average Euclidean distance between its covariate x_i and the covariates of the subjects of B with modality $z_j = z$. One issue worth noticing is that the predictions $\hat{z}$ returned by the algorithm depend on the order in which the pairs $(\tilde{y}, \tilde{z})$ are picked. Here, the arbitrary choice is to take them by descending values of $N_{y,z}$, but other orders might be just as worthy.

for $y \in \mathcal{Y}, z \in \mathcal{Z}$ **do**
  $N_{y,z} \leftarrow \text{round}(n \times \hat{\mu}_{n,y,z}^{(Y^A,Z^A)})$;
while $\max_{y,z} N_{y,z} \geq 0$ **do**
  Let $(\tilde{y}, \tilde{z}) \in \text{argmax}_{y,z}\{N_{y,z}\}$;
  for $i \in \mathcal{O}_{\tilde{y}}$ **do**
    $d_i \leftarrow \frac{1}{|\mathcal{O}_{\tilde{z}}|} \sum_{j \in \mathcal{O}_{\tilde{z}}} d(x_i, x_j)$// `distance used to get nearest neighbors`
  for $k = 1, \ldots, N_{\tilde{y},\tilde{z}}$ **do**
    Let $i_{\min} \in \text{argmin}_{i \in \mathcal{O}_{\tilde{y}}}\{d_i\}$;
    $\hat{z}_{i_{\min}} \leftarrow \tilde{z}$;
    $\mathcal{O}_{\tilde{y}} \leftarrow \mathcal{O}_{\tilde{y}} \setminus \{i_{\min}\}$;
  $N_{\tilde{y},\tilde{z}} \leftarrow 0$;
Return $\hat{z}_i, \forall i \in A$;

Algorithm 1: Nearest neighbor algorithm for the prediction of $Z_i, i \in A$

Algorithm 1 returns individual predictions for each subject. This is only relevant if there is a real need to distinguish subjects even though their observations of X and Y are equal. To compare with the solutions of our models, we can recover an estimation of the probability of Z^A given the values of X^A and Y^A by

$$\hat{\mu}_{n,x,y}^{Z^A=z|X^A=x,Y^A=y} = \frac{1}{|\mathcal{O}_{x,y}|} \sum_{i \in \mathcal{O}_{x,y}} \mathbb{1}_{\hat{z}_i=z}, \ \forall x \in \mathcal{X}, y \in \mathcal{Y}, z \in \mathcal{Z}, \tag{10}$$

where, similarly to $\mathcal{O}_y$, $\mathcal{O}_{x,y}$ is defined as $\mathcal{O}_{x,y} = \{i \in A \mid x_i = x, y_i = y\}$.

In the remainder, the overall method described in this section will be referred to as OUTCOME.

4 Optimal transportation of the joint distribution of outcomes and covariates

The method developed in [9], OUTCOME, has two main drawbacks. First, it relies on the strong assumption that Y^A and Z^A follow the same distribution as Y^B and Z^B. Second, it requires an independent post-treatment step where the predictions are computed with a nearest neighbor algorithm. In particular, this means that the choice made in the second step are not explicitly taken into account in the OT model. Moreover, the nearest neighbor algorithm has a greedy behavior where the quality of the predictions may depend on arbitrary choices in its execution.

In this section, we describe a second OT approach that tackles the above two issues.

4.1 Kantovorich's formulation

In the framework presented in Section 2, we propose to search for an optimal transportation between the two joint distributions of (X^A, Y^A) and (X^A, Z^A) with marginals $\mu^{(X^A,Y^A)}$ and $\mu^{(X^A,Z^A)}$ repectively. Under Kantorovich's formulation in a discrete setting, we will then search for

$$\gamma^* \in \text{argmin}_{\gamma \in \mathcal{D}} < c, \gamma >,$$

where c is a given cost matrix and $\mathcal{D}$ is the set of joint distributions with marginals $\mu^{(X^A,Y^A)}$ and $\mu^{(X^A,Z^A)}$. It is natural to see any element $\gamma \in \mathcal{D}$ as the vector of joint probabilities $\mathbb{P}((X^A = x, Y^A = y), (X^A = x', Z^A = z))$ for all $x, x' \in \mathcal{X}$, $y \in \mathcal{Y}$ and $z \in \mathcal{Z}$. Since this probability nullifies for all $x \neq x'$, we will define $\gamma \in \mathcal{D}$ as a vector of $\mathbb{R}^{|\mathcal{X}|\times|\mathcal{Y}|\times|\mathcal{Z}|}$, where $\gamma_{x,y,z}$ stands for an estimation of the joint probability $\mathbb{P}(X^A = x, Y^A = y, Z^A = z)$. These notations lead to the more detailed OT model

$$\mathcal{P} : \begin{cases} v^* = \min \ < c, \gamma > \\ \qquad \text{s.t.} \ \sum_{z \in Z} \gamma_{x,y,z} = \mu^{(X^A,Y^A)}_{x,y}, \ \forall x \in \mathcal{X}, \forall y \in \mathcal{Y} \\ \qquad \quad \sum_{y \in Y} \gamma_{x,y,z} = \mu^{(X^A,Z^A)}_{x,z}, \ \forall x \in \mathcal{X}, \forall z \in \mathcal{Z} \\ \qquad \quad \gamma_{x,y,z} \geq 0, \ \forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \forall z \in \mathcal{Z} \end{cases} \tag{11}$$

The above model can be solved only if the marginals $\mu^{(X^A,Y^A)}$ and $\mu^{(X^A,Z^A)}$ are known. As discussed in Section 2, this is not the case, but we can build unbiased estimators $\hat{\mu}_n^{X^A,Y^A}$

and $\hat{\mu}_n^{X^A,Z^A}$ as in (2) and (4). As for the cost matrix, we keep the one used in OUTCOME, meaning that it does not depend on the value of x. More formally,

$$c_{x,y,z} = \mathbb{E}\left[d(X^A, X^B) \mid Y^A = y, Z^B = z\right], \forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \forall z \in \mathcal{Z}. \tag{12}$$

In the remainder, we drop x in the list of indices of c. As a consequence, we also use the estimator $\hat{c}_n$ as defined in (9).

Proposition 1. *For all $y \in \mathcal{Y}$, $z \in \mathcal{Z}$, $\hat{c}_{n,y,z} \xrightarrow[n\to+\infty]{a.s.} c_{y,z}$, or, stated otherwise:*

$$\|\hat{c}_n - c\|_\infty \xrightarrow[n\to+\infty]{a.s.} 0$$

Proof. Let $y \in \mathcal{Y}, z \in \mathcal{Z}$. Given that X^A and X^B have finite probability distributions, the expression of c can be formulated as

$$\begin{aligned}
c_{y,z} &= \sum_{(x^A,x^B)\in\mathcal{X}^2} d(x^A, x^B)\mu_{(x^A,x^B)}^{X^A,X^B|Y^A=y,Z^B=z}, \\
&= \sum_{(x^A,x^B)\in\mathcal{X}^2} d(x^A, x^B)\mu_{x^A}^{X^A|Y^A=y}\mu_{x^B}^{X^B|Z^B=z}, \\
&= \sum_{(x^A,x^B)\in\mathcal{X}^2} d(x^A, x^B)\frac{\mu_{x^A,y}^{(X^A,Y^A)}}{\mu_y^{Y^A}}\frac{\mu_{x^B,z}^{(X^B,Z^B)}}{\mu_z^{Z^B}},
\end{aligned}$$

where the second equality is by independence of (X^A, Y^A, Z^A) and (X^B, Y^B, Z^B).

The estimator given in (9) can be rewritten as

$$\begin{aligned}
\hat{c}_{n,y,z} &= \frac{1}{\sum_{i\in A}\mathbb{1}_{\{Y_i=y\}}\sum_{j\in B}\mathbb{1}_{\{Z_j=z\}}} \times \sum_{(x^A,x^B)\in\mathcal{X}^2}\sum_{i\in A}\sum_{j\in B}\mathbb{1}_{\{Y_i=y,X_i=x^A\}}\mathbb{1}_{\{Z_j=z,X_j=x^B\}} \times d(x^A, x^B), \\
&= \sum_{(x^A,x^B)\in\mathcal{X}^2}\frac{\sum_{i\in A}\mathbb{1}_{\{Y_i=y,X_i=x^A\}}}{\sum_{i\in A}\mathbb{1}_{\{Y_i=y\}}}\frac{\sum_{j\in B}\mathbb{1}_{\{Z_j=z,X_j=x^B\}}}{\sum_{j\in B}\mathbb{1}_{\{Z_j=z\}}} \times d(x^A, x^B), \\
&= \sum_{(x^A,x^B)\in\mathcal{X}^2}\frac{\hat{\mu}_{n,x^A,y}^{(X^A,Y^A)}}{\hat{\mu}_{n,y}^{Y^A}} \times \frac{\hat{\mu}_{n,x^B,z}^{(X^B,Z^B)}}{\hat{\mu}_{n,z}^{Z^B}} \times d(x^A, x^B),
\end{aligned}$$

where the third equality is by (2).

To conclude, observe that we expressed $c_{y,z}$ and $\hat{c}_{n,y,z}$ as two finite sums such that the terms of $\hat{c}_{n,y,z}$ are strongly consistent estimators of those of $c_{y,z}$. □

The resulting model is

$$\hat{\mathcal{P}}_n : \begin{cases}
\hat{v}_n = \min \ < \hat{c}_n, \gamma > \\
\qquad \text{s.t.} \sum_{z\in Z}\gamma_{x,y,z} = \hat{\mu}_{n,x,y}^{(X^A,Y^A)}, \ \forall x \in \mathcal{X}, \forall y \in \mathcal{Y} \\
\qquad\quad \sum_{y\in Y}\gamma_{x,y,z} = \hat{\mu}_{n,x,z}^{(X^A,Z^A)}, \ \forall x \in \mathcal{X}, \forall z \in \mathcal{Z} \\
\qquad\quad \gamma_{x,y,z} \geq 0, \ \forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \forall z \in \mathcal{Z}
\end{cases} \tag{13}$$

The optimal solution, $\hat{\gamma}_n$, of $\hat{\mathcal{P}}_n$ then corresponds to an estimation of the distribution of (X^A, Y^A, Z^A). We thus deduce an estimation of the distribution of Z^A given the values of X^A and Y^A as

$$\hat{\mu}_{n,z}^{Z^A|X^A=x,Y^A=y} = \frac{\hat{\gamma}_{n,x,y,z}}{\hat{\mu}_{n,x,y}^{(X^A,Y^A)}},\ \forall x \in \mathcal{X}, y \in \mathcal{Y}, z \in \mathcal{Z}. \tag{14}$$

In contrast to OUTCOME, the method that consists in solving $\hat{\mathcal{P}}_n$ to solve the recoding problem is referred to as JOINT in what follows.

4.2 Consistency of the optimal transport estimator

In this section we study the asymptotic behavior of the estimator, $\hat{v}_n$, of v^*, and of the optimal solutions of $\hat{\mathcal{P}}_n$ with respect to those of $\mathcal{P}$. To do this we rewrite $\mathcal{P}$ and $\hat{\mathcal{P}}_n$ as generic linear programs in standard form

$$\mathcal{P} : v^* = \min\left\{\langle c|\gamma\rangle : A\gamma = b, \gamma \geq 0\right\} \quad \text{and} \quad \hat{\mathcal{P}}_n : \hat{v}_n = \min\left\{\langle \hat{c}_n|\gamma\rangle : A\gamma = \hat{b}_n, \gamma \geq 0\right\},$$

where $c, \hat{c}_n \in \mathbb{R}^p$, $b, \hat{b}_n \in \mathbb{R}^m$ and $A \in \mathbb{R}^{m\times p}$. For $\gamma \in \mathbb{R}^p$ and $S \subset \mathbb{R}^p$, we also define the point-to-set distance as

$$d(\gamma, S) = \inf_{\gamma' \in S} \|\gamma - \gamma'\|,$$

where $\|\cdot\|$ is some norm of $\mathbb{R}^p$. We then introduce the deviation measure of set $S \subset \mathbb{R}^p$ from set $S' \subset \mathbb{R}^p$ as

$$\mathbb{D}(S, S') = \sup_{\gamma \in S} \inf_{\gamma' \in S'} \|\gamma - \gamma'\|.$$

Theorem 1. *Let S^* be the set of optimal solutions of $\mathcal{P}$ and $\hat{S}_n$ the set of optimal solutions of $\hat{\mathcal{P}}_n$. Then*

$$\hat{v}_n \xrightarrow[n\to+\infty]{a.s.} v^*, \quad \textit{and} \quad \mathbb{D}(\hat{S}_n, S^*) \xrightarrow[n\to+\infty]{a.s.} 0.$$

Proof. It is known (see e.g. [22]) that the value function of a linear program

$$(\hat{b}_n, \hat{c}_n) \mapsto \min\{\langle \hat{c}_n|x\rangle : Ax = \hat{b}_n, x \geq 0\}$$

is continuous for any constraint matrix A, hence the a.s. convergence of $\hat{c}_n$ to c and of $\hat{b}_n$ to b yields $\hat{v}_n \xrightarrow[n\to+\infty]{a.s.} v^*$.

From Hoffman error bound lemma [13], there is a constant $K_1 = K_1(c) > 0$ such that $\forall \gamma \in \hat{S}_n$,

$$\begin{aligned} d(\gamma, S^*) &\leq K_1(|\langle c|\gamma\rangle - v^*| + \|A\gamma - b\|) \\ &\leq K_1(|\langle c|\gamma\rangle - \langle \hat{c}_n(\omega)|\gamma\rangle + \langle \hat{c}_n(\omega)|\gamma\rangle - v^*| + \left\|\widehat{b}_n(\omega) - b\right\|) \\ &\leq K_1(K_2\left\|\widehat{c}_n(\omega) - c\right\| + |\hat{v}_n(\omega) - v^*| + \left\|\widehat{b}_n(\omega) - b\right\|) \end{aligned}$$

where K_2 is such that $\|\gamma\| < K_2$. Using the a.s. convergence of $\hat{c}_n$, $\hat{b}_n$, and $\hat{v}_n$, we show that, $\forall \epsilon > 0$, $\exists \Omega_1 \in \mathcal{F}$ with $\mathbb{P}(\Omega_1) = 1$, $\forall \omega \in \Omega_1$, $\exists N \in \mathbb{N}, \forall n \geq N$, such that

$$\forall \gamma \in \hat{S}_n(\omega),\ d(\gamma, S^*) \leq \epsilon,$$

which yields the result.

□

The convergence of $\mathbb{D}(\hat{S}_n, S^*)$ justifies that we estimate a solution of $\mathcal{P}$ with one of $\hat{\mathcal{P}}_n$. However it cannot justify the overall approach that consists in searching for a solution of $\mathcal{P}$ to derive the conditional distributions $\mu^{Z^A|X^A=x,Y^A=Y}, x \in \mathcal{X}, y \in \mathcal{Y}$. The quality of our estimation of these distributions will depend on how the cost function reflects some unobserved properties of the distributions. In the choice of our cost function, we follow the intuition that subjects with outomes $Y_i = y$ and $Z_i = z$ should be frequent when y and z are close to each other in the space of the covariates. If for instance, there is some prior distribution $\bar{\mu}^{(X^A,Y^A,Z^A)}$, the result would certainly be improved by minimizing $\left\|\gamma - \bar{\mu}^{X^A,Y^A,Z^A}\right\|_1$ instead.

5 Improving the models with relaxation and regularization

In this section, we study how $\hat{\mathcal{P}}_n$ can be enriched by including a term of error in the constraints on the marginals of γ. We then add a regularization term expressing that the transportation map should not vary too quickly with respect to X.

5.1 Relaxation of the constraints on marginals

Due to the possible errors in the estimations of the terms of $\mathcal{P}$, the constraints of $\hat{\mathcal{P}}_n$ may drive its optimal solution away from the true values of the marginals of $\mu^{(X^A,Y^A,Z^A)}$. As a consequence, it might be more meaningful to allow for small violations of the constraints of $\hat{\mathcal{P}}_n$. This is done by adding slack variables in the constraints such that they sum to zero and the norm 1 of the vector of slack variables is bounded by some value α_n. The

equality constraints of $\hat{\mathcal{P}}_n$ are then relaxed as follows.

$$\sum_{z\in Z}\gamma_{x,y,z} = \hat{\mu}_{n,x,y}^{(X^A,Y^A)} + e_{x,y}^{X,Y},\ \forall x\in\mathcal{X},\forall y\in\mathcal{Y} \tag{15}$$

$$\sum_{y\in Y}\gamma_{x,y,z} = \hat{\mu}_{n,x,z}^{(X^A,Z^A)} + e_{x,z}^{X,Z},\ \forall x\in\mathcal{X},\forall z\in\mathcal{Z} \tag{16}$$

$$\sum_{x\in\mathcal{X},y\in\mathcal{Y}} e_{x,y}^{X,Y} = 0,\ \sum_{x\in\mathcal{X},z\in\mathcal{Z}} e_{x,z}^{X,Z} = 0 \tag{17}$$

$$-e_{x,y}^{X,Y,+} \le e_{x,y}^{X,Y} \le e_{x,y}^{X,Y,+},\ \forall x\in\mathcal{X},\forall y\in\mathcal{Y} \tag{18}$$

$$-e_{x,z}^{X,Z,+} \le e_{x,z}^{X,Z} \le e_{x,z}^{X,Z,+},\ \forall x\in\mathcal{X},\forall z\in\mathcal{Z} \tag{19}$$

$$\sum_{x\in\mathcal{X},y\in\mathcal{Y}} e_{x,y}^{X,Y,+} \le \alpha_n,\ \sum_{x\in\mathcal{X},z\in\mathcal{Z}} e_{x,z}^{X,Z,+} \le \alpha_n \tag{20}$$

Constraints (17) guarantee that γ remains a distribution of probability despite the relaxation, and constraints (18)–(20) are the linearization of the constraints that bound the norm 1 of the two vectors of error $e^{X,Y}$ and $e^{X,Z}$. The linearization requires the introduction of extra variables $e^{X,Y,+}$ and $e^{X,Z,+}$ that are constrained to be larger than the absolute value of $e^{X,Y}$ and $e^{X,Z}$ by (18)–(19). Observe that the application of the central limit theorem to the right member shows that the standard deviation of the estimation error is in $\mathcal{O}(\frac{1}{\sqrt{n}})$. As a consequence, we will set $\alpha_n := \frac{\alpha}{\sqrt{n}}$, where α is a parameter to be calibrated by simulations.

5.2 Regularization of the objective function

The introduction of regularization terms is not unusual in the applications of OT. For instance, Cuturi [5] considers an entropy term that encourages sparser joint distributions. This allows for faster computation of the solution and improves the results of classic transport on classification problems. In domain adaptation, Courty et al. [4] also argue that sparsity should be promoted in the OT map. In addition to an entropy term, they minimize some $\ell_p - \ell_q$ mixed norm term that encourages the affectation of each individual to only one class. Another class of regularization techniques aim at minimizing the variations of the transportation map. In discrete OT, one issue is the lack of definition of the gradient. Ferradans et al. [8] impose a graph structure in the discrete distributions they wish to transport to use a classic graph gradient operator. Transposed to our setting, their approach comes down to building an undirected graph, $G = (\mathcal{X}, E_{\mathcal{X}}, w)$, where $w \in \mathbb{R}^{E_{\mathcal{X}}}$ is a vector of weights on the edges of G. The edges of G link the pairs of elements of $\mathcal{X}$ defined as neighbors. Typically, an edge $\{x_i, x_j\}$ is created if x_j is among the k nearest neighbors of x_i for some distance, d, defined on $\mathcal{X}$ and some parameter $k \ge 1$. It is then classic to define $w_{i,j} := d(x_i, x_j)^{-1}$ for all $\{i,j\} \in E_{\mathcal{X}}$. Denoting $\Gamma \in \mathbb{R}^{|\mathcal{X}|\times D}$ ($D \in \mathcal{Z}^+$) the term they wish to regularize, the gradient of G at Γ is defined as

$$\Delta\Gamma = \left(w_{i,j}(\Gamma_{i,\cdot} - \Gamma_{j,\cdot})\right)_{\{i,j\}\in E_{\mathcal{X}}}.$$

Finally, the regularity of a transport map is defined by some norm, $J_{p,q}$, of the graph gradient:

$$J_{p,q}(\Delta\Gamma) = \sum_{\{i,j\}\in E_{\mathcal{X}}} \left(\left\| w_{i,j} \left(\Gamma_{i,\cdot} - \Gamma_{j,\cdot} \right) \right\|_q \right)^p,$$

where $\|\cdot\|_p$ is the ℓ^p norm in $\mathbb{R}^D$.

In data recoding, it is not clear that sparsity should be sought. However, we expect some regularity in the variations of the conditional distribution $\mu^{Y^A,Z^A|X^A=x}$ with respect to x if the covariates are correlated to the outcomes. As a consequence, we add a regularization term similar to that considered by Ferradans et al. [8], where the term we wish to regularize is directly $(\frac{\gamma_{x,y,z}}{\hat{\mu}^{X^A}_{n,x}})_{x\in\mathcal{X},y\in\mathcal{Y},z\in\mathcal{Z}}$. In this work, we favor the norm that will have the smallest impact on the model. In particular, with the graph anisotropic total variation ($p = 1$, $q = 1$), the regularization term can be linearized so that the optimization model remains a linear program. Other typical values have been tested during preliminary tests ($p = 1, 2$ and $q = 1, 2$) without significant impact in the numerical results. Consequently, the regularization term is given by:

$$\sum_{\{x_i,x_j\}\in E_{\mathcal{X}}} \left(w_{i,j} \sum_{y\in Y,z\in Z} \left| \frac{\gamma_{x_i,y,z}}{\hat{\mu}^{X^A}_{n,x_i}} - \frac{\gamma_{x_j,y,z}}{\hat{\mu}^{X^A}_{n,x_j}} \right| \right).$$

After linearization of this term and relaxation of the constraints on marginals, we obtain the following regularized linear program.

$$\hat{\mathcal{P}}^R_n : \begin{cases} \hat{v}_n = \min & < \hat{c}_n, \gamma > + \lambda \sum_{(x_i,x_j)\in E_{\mathcal{X}}} w_{i,j} \sum_{y\in\mathcal{Y},z\in\mathcal{Z}} r^+_{i,j,y,z} \\ \text{s.t.} & \text{constraints (15)–(20)} \\ & \gamma_{x_i,y,z}/\hat{\mu}^{X^A}_{n,x_i} - \gamma_{x_j,y,z}/\hat{\mu}^{X^A}_{n,x_j} \leq r^+_{i,j,y,z},\ \forall\{x_i,x_j\}\in E_{\mathcal{X}}, y\in\mathcal{Y}, z\in\mathcal{Z} \\ & \gamma_{x_i,y,z}/\hat{\mu}^{X^A}_{n,x_i} - \gamma_{x_j,y,z}/\hat{\mu}^{X^A}_{n,x_j} \geq -r^+_{i,j,y,z},\ \forall\{x_i,x_j\}\in E_{\mathcal{X}}, y\in\mathcal{Y}, z\in\mathcal{Z} \\ & \gamma_{x,y,z} \geq 0,\ \forall x\in\mathcal{X}, \forall y\in\mathcal{Y}, \forall z\in\mathcal{Z} \end{cases} \tag{21}$$

The constant $\lambda \in \mathbb{R}^+$ is a regularization parameter to be calibrated numerically.

Remark 3. *Observing that the OT model remains a linear program, there is no real difficulty in extending the consistency results of Theorem 1 to the regularized model. The only significant difference is that the constraint matrix of $\hat{\mathcal{P}}^R_n$ is an estimation, A_n, of the true constraint matrix A. The extension of Theorem 1 is a consequence of the almost sure convergence of A_n to A.*

The method that computes a solution of the recoding problem from that of $\hat{\mathcal{P}}^R_n$ is called R-JOINT in the remainder of the article.

5.3 Improvement of the transport of outcomes

The validation of JOINT and R-JOINT will be mostly based on comparisons with OUTCOME. Since several improvements have been proposed in the above sections, we propose to adapt OUTCOME in order to include as many of these features as possible.

First and foremost, it can be incorrect to use the estimator $\hat{\mu}^{Z^A}$ given in (8) if Y and Z do not follow the same distributions in the two bases. Instead, we use similar arguments as in Section 2.2 to derive the strongly consistent estimators

$$\hat{\mu}^{Z^A}_{n,z} = \sum_{x \in \mathcal{X}} \frac{\hat{\mu}^{(X^B,Z^B)}_{n,x,z} \hat{\mu}^{X^A}_{n,x}}{\hat{\mu}^{X^B}_{n,x}}, \ \forall z \in Z. \tag{22}$$

The second improvement is in the relaxation of the constraints on the marginals. Following the same steps as in Section 5.1, we obtain the following relaxed linear program.

$$\hat{\mathcal{P}}^{0-R}_n : \begin{cases} \hat{v}_n = \min \ < \hat{c}_n, \gamma > \\ \qquad \sum_{z \in \mathcal{Z}} \gamma_{y,z} = \hat{\mu}^{Y^A}_{n,y} + e^Y_y, \ \forall y \in \mathcal{Y} \\ \qquad \sum_{y \in \mathcal{Y}} \gamma_{y,z} = \hat{\mu}^{Z^A}_{n,z} + e^Z_z, \ \forall z \in \mathcal{Z} \\ \qquad \sum_{y \in \mathcal{Y}} e^Y_y = 0, \ \sum_{z \in \mathcal{Z}} e^Z_z = 0 \\ \qquad -e^{Y,+}_y \leq e^Y_y \leq e^{Y,+}_y, \ \forall y \in \mathcal{Y} \\ \qquad -e^{Z,+}_z \leq e^Z_z \leq e^{Z,+}_z, \ \forall z \in \mathcal{Z} \\ \qquad \gamma_{y,z} \geq 0, \ \forall y \in \mathcal{Y}, \forall z \in \mathcal{Z} \end{cases} \tag{23}$$

However, we do not include any regularization term in $\hat{\mathcal{P}}^0_n$, because, the one that we used in Section 5.2 can only be defined if the decision variables are also indexed by the elements of $\mathcal{X}$.

In OUTCOME, the solution of the linear program is then completed with the execution of the nearest neighbor procedure described in Algorithm 1. For each $i \in A$, this algorithm tries to affect to subject i the outcome $\hat{z}_i$ that minimizes the distance $d(i,z) := \frac{1}{|\mathcal{O}_z|} \sum_{j \in \mathcal{O}_z} d(x_i, x_j)$ among all $z \in Z$, while satisfying the marginal distributions given by the solution of $\hat{\mathcal{P}}^0_n$. The limit of this approach is that there cannot be any general characterization of the predictions returned by Algorithm 1. Instead, it seems natural to search for the predictions $(\hat{z}_i)_{i \in A}$ that minimize the total distance $\sum_{i \in A} d(i,z)$. This can be formalized with the following linear program, where $\hat{\gamma}_n$ is an optimal solution of $\hat{\mathcal{P}}^{0-R}_n$

and $\{n\hat{\gamma}_{n,y,z}\}_{y\in\mathcal{Y},z\in\mathcal{Z}}$ are rounded so that they sum to n.

$$\begin{cases} \min\limits_{\delta} \sum\limits_{i\in A} d(i,z)\delta_{i,z} \\ \quad \sum\limits_{z\in\mathcal{Z}} \delta_{i,z} = 1,\ \forall i \in A \\ \quad \sum\limits_{i\in\mathcal{O}_y} \delta_{i,z} = \text{round}(n \times \hat{\gamma}_{n,y,z}),\ \forall y \in \mathcal{Y}, \forall z \in \mathcal{Z} \\ \quad \delta_{i,z} \geq 0,\ \forall i \in A, \forall z \in \mathcal{Z} \end{cases} \tag{24}$$

From an optimal solution $\hat{\delta}$ of the above model, we then predict outcome $z \in \mathcal{Z}$ for $i \in A$ if $\hat{\delta}_{i,z} = 1$. Since there is no integrality constraint in the model, $\hat{\delta}$ can be fractional. But one can notice that the constraint matrix is totally unimodular with integer right-hand side, hence every extreme solution is integer. As a consequence, we can get an integer optimal solution $\hat{\delta}$ by solving the linear program with, e.g., the simplex algorithm. We finally deduce the solution to the recoding problem by adapting (10)

$$\hat{\mu}_{n,z}^{Z|X^A=x,Y=y} = \frac{1}{|\mathcal{O}_{x,y}|} \sum_{i\in\mathcal{O}_{x,y}} \hat{\delta}_{i,z},\ \forall x \in \mathcal{X}, y \in \mathcal{Y}, z \in \mathcal{Z}. \tag{25}$$

The overall method described in this section is called R-OUTCOME in the remainder of the article.

6 Experimental validation

In this section, the relevance of the models we developed is assessed by means of simulation studies. Each database is constructed by generating n independent samples of (X, Y, Z) according to predefined distributions that may vary between A and B. In Section 6.1, we first describe a default simulation scenario and we introduce variations from this scenario to study the marginal impacts of the parameters of the simulations. We then solve every simulation of each scenario with the four methods described above: OUTCOME, R-OUTCOME, JOINT and R-JOINT. The results are discussed in Section 6.3.

6.1 Simulation design

In all our simulations, $(X_i, Y_i, Z_i)_{i\in A}$ and $(X_j, Y_j, Z_j)_{j\in B}$ are obtained by discretization of continuous random variables as follows. Let $\{U_i\}_{i\in A}$ be a family of i.i.d. 3-dimensional random vectors with multi-variate normal distribution $\mathcal{N}(m^A, \mathbf{\Sigma}^A)$. Likewise, $\{U_j\}_{j\in B}$ is a family of i.i.d. random vectors with distribution $\mathcal{N}(m^B, \mathbf{\Sigma}^B)$. For simplicity, we take

$\mathbf{\Sigma}^A = \mathbf{\Sigma}^B = \mathbf{\Sigma}$, where

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & 0.2 & 0.2 \\ 0.2 & 1 & 0.2 \\ 0.2 & 0.2 & 1 \end{pmatrix}.$$

In contrast, we may have $m^A \neq m^B$ when the distributions of X^A and X^B are different. For the discretization, for some $i \in A$, we denote as t_1 the median of $U_{i,1}$, $t_{2,1}$ and $t_{2,2}$ the tertiles of $U_{i,2}$, and $t_{3,1}$, $t_{3,2}$ and $t_{3,3}$ the quartiles of $U_{i,3}$. For all $i \in A \cup B$, we then discretize $U_{i,1}$ into two modalities by setting

$$X_{i,1} = \mathbb{1}_{\{U_{i,1} > t_1\}}.$$

Covariate $X_{i,2}$ is the discretization of $U_{i,2}$ into three modalities defined by

$$X_{i,2} = \mathbb{1}_{\{t_{21} < U_{i,2} \leq t_{22}\}} + 2 \times \mathbb{1}_{\{U_{i,2} > t_{22}\}}.$$

Finally, we set

$$X_{i,3} = \mathbb{1}_{\{t_{31} < U_{i,3} \leq t_{32}\}} + 2 \times \mathbb{1}_{\{t_{32} < U_{i,3} \leq t_{33}\}} + 3 \times \mathbb{1}_{\{U_{i,3} > t_{33}\}}.$$

Observe that the values of $t_1, \ldots, t_{33}$ are defined once from the quantiles of U in base A, so that if $U_i, i \in A$, and $U_j, j \in B$, have different means, X^A and X^B will have different distributions.

For all $i \in A \cup B$, we then construct Y_i and Z_i by two different discretizations of a single latent variable V_i. In the default scenario, V_i depends linearly on U_i as follows.

$$V_i = a_1 U_{i,1} + a_2 U_{i,2} + a_3 U_{i,3} + \sigma W_i, \tag{26}$$

where $a \in \mathbb{R}^3$ is a given parameter of the scenario and W_i follows a standard normal distribution (with $\{W_i\}_{i \in A \cup B}$ i.i.d. random variables). As above, we build Y_i by discretization of V_i into three modalities using the tertiles of V_j for some $j \in A$. In contrast, Z_i is obtained by discretization of V_i into four modalities using the quartiles of V_k for some $k \in B$.

A scenario following the above definition is completely defined by the values of m^A, m^B, a, σ and n. In the remainder, σ will be set so that, R^2, the coefficient of determination of V from U reaches a given value. The default scenario, denoted as $\mathrm{S}_{\mathrm{ref}}$, is characterized by $m^A = (0,0,0)$, $m^B = (1,0,0)$, $a = (1,1,1)$, $R^2 = 0.5$ and $n = 1000$. Taking the results obtained for $\mathrm{S}_{\mathrm{ref}}$ as reference, the impact of the elements characterizing the simulations will be studied through the following scenarios.

Sample size. Keeping m^A, m^B, a and σ unchanged, we allow n to vary in

$$\{50, 100, 250, 500, 1000, 2500, 5000, 10000\}.$$

The resulting scenarios are denoted as Sn-50,..., Sn-10000, where Sn-1000$\equiv \mathrm{S}_{\mathrm{ref}}$. In every other scenario, we set $n = 1000$.

Measure of association between outcomes and covariates. We investigate the impact of R^2 by letting the parameter vary in $\{0.01, 0.05, 0.1, 0.9\}$ while keeping m^A, m^B and a unchanged. The corresponding scenarios are denoted as SR-0.01, ..., SR-0.9 ($\mathrm{S_{ref}} \equiv$SR-0.5).

Distribution of the covariates. Keeping R^2 and a as in $\mathrm{S_{ref}}$, we investigate the impact of differences in the distributions of X^A and X^B by considering the following four scenarios:

- SX-1: $m^A = m^B = (0, 0, 0)$,
- SX-2: $m^A = (0, 0, 0)$, $m^B = (0.5, 0, 0)$,
- SX-3: $m^A = (0, 0, 0)$, $m^B = (1, 1, 0)$,
- SX-4: $m^A = (0, 0, 0)$, $m^B = (1, 2, 0)$.

Non-linearity in the association between V and U. Here, we investigate the impact of the association between V and U as expressed in (26). Keeping the values of σ, a, m^A and m^B as in $\mathrm{S_{ref}}$, we modify the expression of V with nonlinear expressions as follows.

- SNL-1: $V_i = a_1(U_{i,1})^2 + a_2(U_{i,2})^2 + a_3(U_{i,3})^2 + \sigma W_i, \forall i \in A \cup B$,
- SNL-2: $V_i = \exp(a_1 U_{i,1} + a_2 U_{i,2} + a_3 U_{i,3}) + \sigma W_i, \forall i \in A \cup B$.

Heterogeneous groups. The expression of V as a continuous function of U involves that the groups of subjects with same outcomes will be grossly homogeneous in the space of the covariates. We investigate, the impact of a more heterogeneous structure by modifying only the discretization of V. Scenario SHG is then obtained as $\mathrm{S_{ref}}$ until the discretization of V. At this stage, we also use the tertiles of V in base A, t_1^A and t_2^A, and the quartiles of V in base B, t_1^B, t_2^B and t_3^B. In contrast to the default scenario, we keep only two modalities for Y and three modalities for Z by merging the extreme two groups as follows.

- $Y_i = \mathbb{1}_{\left\{t_1^A \leq V_i < t_2^A\right\}}$.
- $Z_i = \mathbb{1}_{\left\{t_1^B \leq V_i < t_2^B\right\}} + 2 \times \mathbb{1}_{\left\{t_2^B \leq V_i < t_3^B\right\}}$.

Robustness to different conditional distributions. Finally, we wish to evaluate the importance of satisfying the assumption that the distributions of Y and Z given X are the same in the two databases. For this, we allow vector a to be different in the two databases when computing V (see (26)). More formally,

$$\begin{cases} V_i = a_1^A U_{i,1} + a_2^A U_{i,2} + a_3^A U_{i,3} + \sigma W_i, \forall i \in A, \\ V_i = a_1^B U_{j,1} + a_2^B U_{j,2} + a_3^B U_{j,3} + \sigma W_j, \forall j \in B, \end{cases}$$

with $a^A, a^B \in \mathbb{R}^3$. Keeping R^2, m^A and m^B as in $\mathrm{S}_{\mathrm{ref}}$, we consider the following three scenarios.

- Sa-1: $a^A = (1, 1, 1)$ and $a^B = (1, 1, 2)$
- Sa-2: $a^A = (1, 1, 1)$ and $a^B = (1, 1.5, 2)$
- Sa-3: $a^A = (1, 1, 1)$ and $a^B = (3, 1.5, 2)$

6.2 Experimental setup

Every method has been implemented using the Julia language [2], and we used the JuMP library [7] to model linear programs. Every model is then solved with the simplex algorithm of CLP solver[1] on a single thread of an Intel(R) Core(TM)i7-3770 CPU @ 3.40GHz processor.

We wish to compare the performance of the four methods, OUTCOME, R-OUTCOME, JOINT and R-JOINT, on all the scenarios defined in the previous section. To evaluate the performance of the methods, we compute the rate of error, ρ, made in the distribution $\hat{\mu}_n^{X^A|Y^A,Z^A}$ returned by each method as

$$\begin{aligned} &\frac{1}{2}\mathbb{E}_{X^A,Y^A}\left(\left\|\hat{\mu}_n^{Z^A|X^A,Y^A} - \mu^{Z^A|X^A,Y^A}\right\|_1\right) \\ &= \frac{1}{2}\sum_{x\in\mathcal{X},y\in\mathcal{Y}}\left\|\hat{\mu}_n^{Z^A|X^A=x,Y^A=y} - \mu^{Z^A|X^A=x,Y^A=y}\right\|_1 \times \mu_{x,y}^{X^A,Y^A}. \end{aligned} \tag{27}$$

where the $\frac{1}{2}$ factor ensures that the error lies in $[0, 1]$. Due to the discretization performed in the simulations, there is no simple analytical expression of the true distributions $\mu^{X^A|Y^A,Z^A}$ and μ^{X^A,Y^A}. We replace them with the empirical distributions observed in a simulation with $n = 10^5$ subjects.

As already observed, the relaxation and the regularization both involve one parameter, respectively denoted as α and λ. Since the models should be robust to differences in the simulations, we wish to keep the same parameter values in every scenario. To calibrate the parameters, we run each method on ten simulations of each scenario using a wide range of parameter values. We set the parameters to values that consistently produced small rates of error for every scenario. More precisely, $\alpha = 0.4$ in R-OUTCOME and $\alpha = 0.4$, $\lambda = 0.1$ in R-JOINT.

6.3 Simulation results

Since OUTCOME has been shown to produce better results than multiple imputation and several machine learning methods [21], we compare our OT methods only to OUTCOME.

[1] See the COIN-OR webpage of CLP for more details: `https://projects.coin-or.org/Clp/wiki`

For each scenario, we execute the four OT methods to solve 100 simulations. For each simulation, we compute an estimation of the distribution of Z^A given X^A and Y^A, and we also solve the symmetric problem where we search for an estimation of the distribution of Y^B given X^B and Z^B. The results we display for each simulation correspond to the average of the errors made in these two distributions. For a more synthetic presentation of the results, we draw boxplots for each scenario and each method where similar scenarios are grouped in the same figure. Those boxplots are given in Figures 1–5. In each one of these boxplots, the rate of error (27) appears on the ordinate axis and the scenario appears in abscissa.

First and foremost, a global look at the results shows that the regularized transport of joint distributions, R-JOINT, brings an impressive improvement when compared to the previously developed transport of outcomes, OUTCOME. In particular, it allows to reduce the average error by a factor 4.5 for $\mathrm{S}_{\mathrm{ref}}$, and for most scenarios, R-JOINT estimates $\mu^{X^A|Y^A,Z^A}$ with an average rate of error below 10%. The only scenarios where the improvement factor is less than 2 are the most nonlinear scenario, SNL-2, and that farthest from the hypotheses of Section 2, Sa-3. The comparison of R-JOINT with JOINT and R-OUTCOME indicates that this improvement is due to both the initial choice of transporting joint distributions and to the relaxation and regularization of the transport model.

For a more precise analysis, we draw our attention to the impact of the size of the databases, n. In Figure 1, we clearly observe that n must reach a critical value before the mean and variance of the rate of error stabilize. With the choice we made in our simulations, the graphs indicate that there is no much hope for good estimations if $n \leq 250$. We also observe that the performances of R-JOINT and R-OUTCOME stabilize faster when n increases than those of JOINT and OUTCOME. Such behavior was expected given that the relaxation of the constraints on marginal distributions (see Section 5.1) is designed to reduce the impact of errors in their empirical estimators.

As already stated in the previous sections, our method is motivated by the assumption that covariates somehow explain the outcomes. It is thus expected that small values of R^2 will yield poor results. The boxplots in Figure 2 provide more insight with this respect. In particular, we see that the mean and variance of the rate of error are both very high with every method when $R^2 = 0.01$. We still observe a negative impact on the results of R-OUTCOME, JOINT and R-JOINT when $R^2 = 0.05$, but they do not appear to be highly sensitive to the coefficient of determination when it is larger than 0.1. This is a valuable finding, since it extends the range of databases where OT methods could be relevant for data recoding. Another observation is that the rate of error of OUTCOME displays rather small variations with the value of R^2 as soon as $R^2 \geq 0.05$. Our interpretation is that the constraint on marginal distributions of the other methods rely more on the association between covariates and outcomes. Nonetheless, the average rate of error remains higher with OUTCOME than with R-JOINT for every value of R^2.

Figure 3 allows to measure the impact of differences in the distributions of X^A and X^B. More specifically, μ^{X^A} and μ^{X^B} are equal in SX-1 and they are more and more

different from SX-2 to SX-4. Contrary to the other three methods, OUTCOME relies on the hypothesis that these two distributions are equal. The corresponding boxplots clearly indicate that the violation of this hypothesis has a dramatic impact on the results of this method. In particular, the average rate of error of OUTCOME reaches 62% in scenario SX-4. The other methods are much less sensitive to differences between μ^{X^A} and μ^{X^B} even though errors tend to be larger in SX-4.

In contrast, every method relies on the hypothesis that the distributions of Y and Z given X are the same in both databases. Scenarios Sa-1 to Sa-3 measure how deviations from this assumption impact the results. From the boxplots displayed in Figure 4, we get that rather small deviations do not severely impact the rate of errors (see Sa-1 and Sa-2). On the other hand, the errors observed for Sa-3 show that the methods developed in this article may not be appropriate for large deviations from this assumption.

Finally, we investigate how the four methods behave when the simulation of V differs from the procedure described for the default scenario. In Figure 5, we draw the boxplots of the errors encountered for scenarios SNL-1, SNL-2 and SHG. It is interesting that OUTCOME seems to be less sensitive to nonlinearity in the association between U and V than the other methods – even though it still makes more error. In particular, when solving simulations of SNL-2, R-JOINT has the largest variance in the rate of error, and it only reduces the average rate of error of OUTCOME by less than 10%. Also, it appears that relaxation and regularization do not bring a significant improvement when solving simulations of SNL-2 and SHG. Indeed, the average rates of errors are very similar with JOINT and R-JOINT for these scenarios, but the variance is smaller with JOINT. Overall, the results observed for SNL-1 and SHG indicate that small deviations from the linear association between U and V will not have a dramatic impact on the errors made by the methods we develop in this article. In contrast, Scenario SNL-2 has been built as an extreme case to challenge the OT methods. It is our opinion that specific methods should be developed for such highly nonlinear association between outcomes and covariates.

7 Application on a real dataset: NCDS study

Method OUTCOME is applied to the ELFE study in [9] to recode a self-rated overall health outcome that is coded in different scales in two databases. As the two scales are never simultaneously observed on the same individual, it is not possible to compare the different methods on this study. Instead, we choose to apply the four methods on the National Child Development Study (NCDS) study where one identical outcome has been observed in two different scales for all individuals.

The NCDS project is a continuing survey which follows the lives of over 17,000 people born in England, Scotland and Wales in the same week of the year 1958. It is a well-known study, because its results have greatly contributed to the improvement of maternity services in the United Kingdom. This survey collects specific information in many distinct

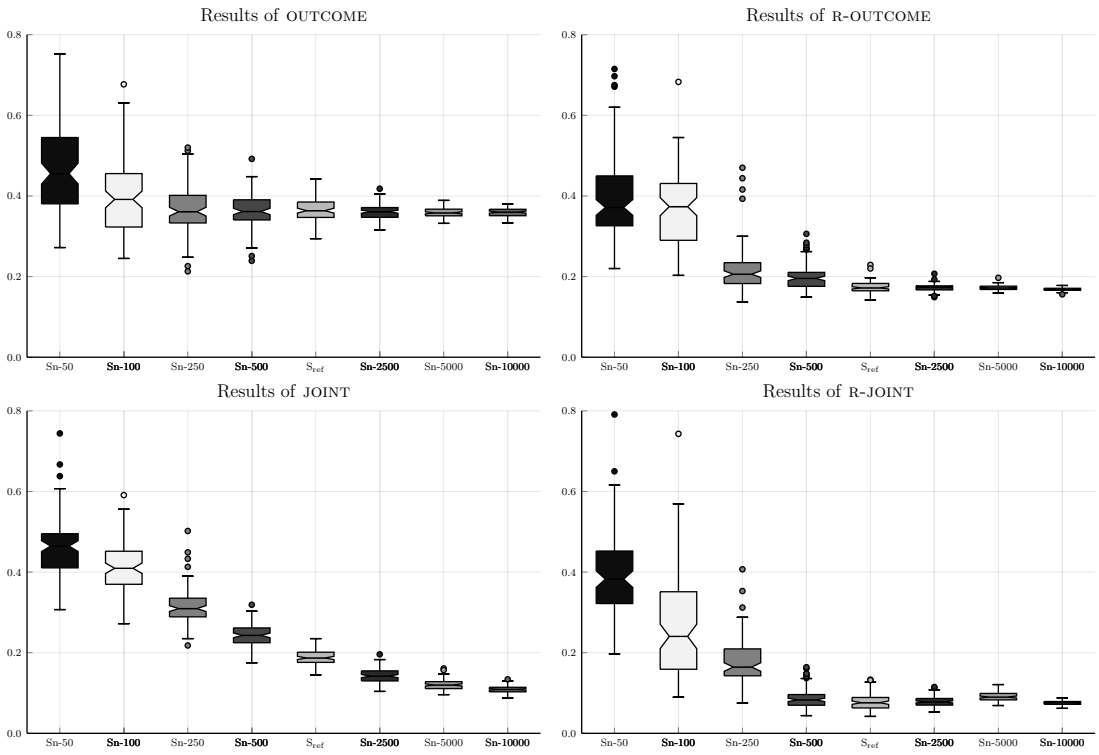

Figure 1: Evolution of the error with the size of the databases

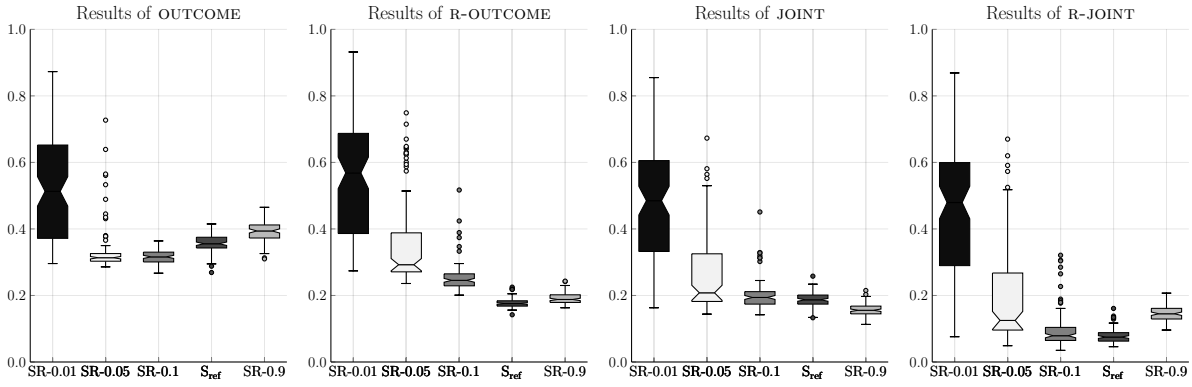

Figure 2: Error in the distribution for SR scenarios

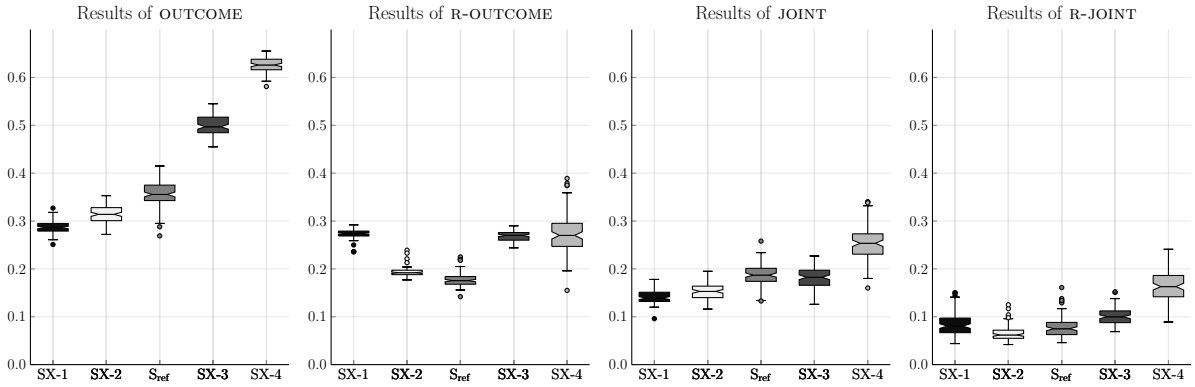

Figure 3: Error in the distribution for SX scenarios

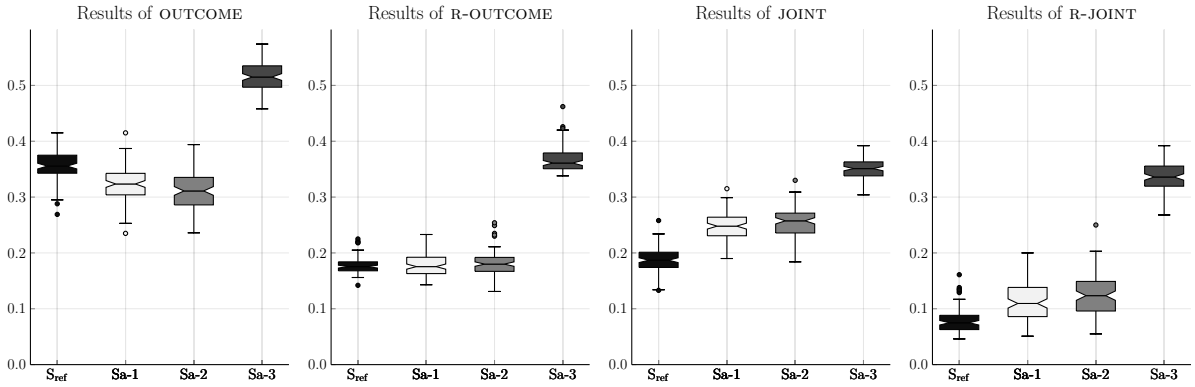

Figure 4: Error in the distribution for Sa scenarios

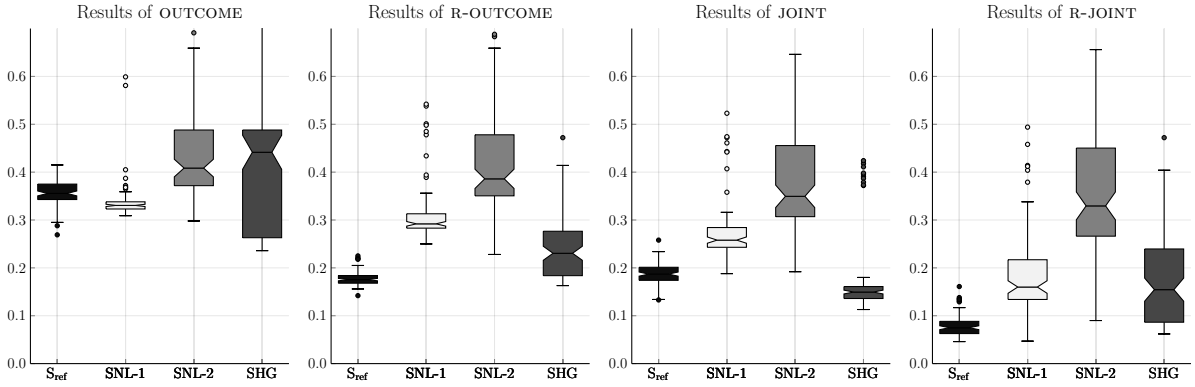

Figure 5: Error in the distribution for SNL scenarios

fields like physical and educational development, economic circumstances, employment, family life, health behaviour, well-being, social participation and attitudes. In particular, the NCDS Activity Histories 1974-2013 [18] merges all data on work and non-work activities in successive waves into one longitudinal dataset. In this dataset, two measurements scales of the social class of the participants built from profession were collected: the Goldthorp social class 1990 scale (GSS90) and the RGs social Class 1991 scale (RGS91). GSS90 is an ordered scale with 11 categories, whereas RGS91 is an ordered scale with only 6 categories (see [11] for a detailed description of these two scales).

In our numerical tests, we split the dataset in two databases A and B with equal lengths, $n = 4015$. We assume that the social class outcome is known only in the GSS90 scale in base A, whereas it is known only in the RGS91 scale in base B. Moreover, we select the following four covariates for their ability to predict the outcome coded in each of the two respective scales:

- the gender of the participant coded in 2 obviously not ordered modalities,
- the health status (4 ordered modalities),
- the employment status at wave 5 (7 not ordered modalities),
- the study level at wave 4 (assessed in 2 not ordered modalities).

In order to study the benefit of our models compared to OUTCOME, we consider two scenarios. In the first scenario, we arbitrarily store the first half of the NCDS subjects in base A, and the second half in base B. As a consequence, we can assume that the outcomes are distributed identically in the two databases. In the second scenario, we artificially create two unbalanced databases by introducing differences in the distribution of covariates between the two database. For this, we sample the 4015 subjects of database A so that 80% subjects have made long studies against 20% who have made short studies. The remaining subjects are then stored in database B.

In the fist scenario, where the two database are balanced, the average error with OUTCOME and R-OUTCOME is 15.3% against 10.6% with JOINT and R-JOINT. We observe that relaxation and regularization do not improve the results. Actually, relaxation alone even deteriorates the results, but regularization allows to compensate this deterioration. This shows that using R-JOINT remains a good compromise in the more restricted framework explored by [9].

In the second scenario where databases are unbalanced, the average error is 30.5% with OUTCOME, 18.1% with R-OUTCOME while it is 16.2% with JOINT and 14.9% with R-JOINT. These results highlight the importance of the choice of the estimators of marginal distributions (see (10) and (25)) to reduce the negative impact of differences in the distributions of covariates. The transport of the joint distribution of covariates and outcomes provides an additional reduction in the recoding error. R-JOINT then stands out as the best performing method.

8 Conclusions and perspectives

This research consists in the development of a new OT method for data recoding, R-JOINT. The originality is that the transport considers the joint distribution of covariates and outcomes whereas a previous method [9], OUTCOME, was based on the transport of the distribution of outcomes. In R-JOINT, we also improve the OT model by relaxing the constraints on marginal distributions and by adding a regularization term that smooths the variations of the transportation map in the space of the covariates. Based on this insight, we propose several improvements in OUTCOME. Experimental tests on simulated databases validate the relevance of the method by exhibiting a reduction of the data recoding error by a factor 3 to 4 in most scenarios. What is more, R-JOINT is by design less sensitive to differences in the distributions of the covariates between the two databases. Several tests indicate that our method should be able to recode outcomes with small rates of errors for a large range of databases.

In future research, there would be value in extending this OT method to support covariates with continuous distributions. We have focused on discrete random variables for simplicity, but the motivation of our method is unchanged with continuous or mixed covariates. It is important that the OT model takes the joint distribution of outcomes and covariates into consideration. The key challenge in such extension will be to focus on well chosen subsets of covariates' values so that the OT model remains tractable.

Another lead would be to design methods that are more adapted to large differences in the distributions of outcomes given the covariates values. One could also look into data recoding when the association between outcomes and covariates is highly nonlinear.

This work could also be extended to record linkage when the objective is to link de-identified research datasets at the patient level and no common individual identifier is available. Indeed, Y and Z do not need to refer to the same information in our theoretical framework. However, in this application, the two databases A and B will have common individuals, so the assumption that databases are independent databases is not respected.

Acknowledgement

We would like to thank Mounir Haddou, Loïc Hervé and Jean-François Dupuy for their support and for sharing their insights into the theoretical aspects of the methods we developed.

This research has received the help from "Région Occitanie" Grant RBIO-2015-14054319 and Mastodons-CNRS Grant.

The authors are grateful to the Centre for Longitudinal Studies and to the Institute of Education for the use of the National Child Development Study data and to the Economic and Social Data Service for making them available.

References

[1] D.J. Bartholomew, M. Knott, and I. Moustaki. *Latent Variable Models and Factor Analysis: A Unified Approach*. (3rd ed). Wiley, 2011.

[2] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: A Fresh Approach to Numerical Computing. *Society for Industrial and Applied Mathematics Review*, 59(1):65–98, 2017.

[3] I. Bloch. Fusion d'informations en traitement du signal et des images. *Hermes Science Publication*, 2003.

[4] N. Courty, R. Flamary, and D. Tuia. Domain adaptation with regularized optimal transport. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases 2014*, LNCS, pages 1–16, Nancy, France, 2014.

[5] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 2292–2300, USA, 2013. Curran Associates Inc.

[6] C. Delpierre, G. D. Datta, M. Kelly-Irving, V. Lauwers-Cances, L. Berkman, and T. Lang. What role does socio-economic position play in the link between functional limitations and self-rated health: France vs. USA? *European Journal of Public Health*, 22(3):317–321, 2012.

[7] Iain Dunning, Joey Huchette, and Miles Lubin. JuMP: A Modeling Language for Mathematical Optimization. *SIAM Review*, 59(2):295–320, 2017.

[8] S. Ferradans, N. Papadakis, G. Peyré, and J.-F. Aujol. Regularized Discrete Optimal Transport. *Lecture Notes in Computer Science*, 7893 LNCS:428–439, jul 2013.

[9] V. Garès, C. Dimeglio, G. Guernec, R. Fantin, B. Lepage, M. R. Kosorok, and N. Savy. On the use of optimal transportation theory to recode variables and application to database merging. Preprint, `https://hal.archives-ouvertes.fr/hal-01905857`, October 2018.

[10] D.L. Hall and J. Llinas. An introduction to multisensor data fusion. *Proceedings of the IEEE, International Conference on Intelligent Robots and Systems*, 85:6–23, 1997.

[11] Maggie Hancock. National Child Development Study, Activity Histories (1974-2013). Technical report, Centre for Longitudinal Studies, March 2016.

[12] F.L. Hitchcock. The distribution of a product from several sources to numerous localities. *Journal of mathematics and physics / Massachusetts Institute of Technology.*, 20:224–230, 1941.

[13] Alan J Hoffman. On approximate solutions of systems of linear inequalities. In *Selected Papers Of Alan J Hoffman: With Commentary*, pages 174–176. World Scientific, 2003.

[14] L. Kantorovich. On the transfer of masses. *Doklady Akademii Nauk SSSR*, 37:7–8, 1942.

[15] R.J. Little and D. Rubin. *Statistical Analysis with Missing Data*. Wiley, NY, 1987.

[16] G. Monge. Mémoire sur la Théorie des Déblais et des Remblais. *Histoire de l'Académie royale des sciences de Paris*, pages 666–704, 1781.

[17] A. Skrondal and S. Rabe-Hesketh. *Generalized Latent Variable Modeling*. New York: Chapman and Hall/CRC, 2004.

[18] Centre for Longitudinal Studies University of London, Institute of Education. National Child Development Study: Activity Histories, 1974-2013 [data collection], *2nd Edition*, UK Data Service, 2016. `http://doi.org/10.5255/UKDA-SN-6942-3`.

[19] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, second edition, New York, NY, USA, 2000.

[20] C. Villani. Optimal transport, old and new. *Grundlehren des mathematischen Wissenschaften, Springer-Verlag*, 338, 2009.

[21] D. Vuillemenot, V. Gares, C. Dimeglio, G. Guernec, B. Lepage, Kosorok M.R., N. Savy, and P. Saint-Pierre. Comparison of OT-algorithm and machine learning approach to merge databases. Technical report, IRMAR, 2019.

[22] R. J.-B. Wets. On the continuity of the value of a linear program and of related polyhedral-valued multifunctions. In R. W. Cottle, editor, *Mathematical Programming Essays in Honor of George B. Dantzig Part I*, volume 24, pages 14–29. Springer Berlin Heidelberg, 1985.