Optimal transport for data integration

Valérie Garès, Chloé Friguet, Nicolas Courty

Optimal transport for data integration

Valérie Garès[1] & Nicolas Courty [2] & Chloé Friguet[3]

[1] *IRMAR, INSA Rennes, Rennes, France, valerie.gares@insa-rennes.fr*
[2] *IRISA, Univ. Bretagne Sud, Vannes, France, nicolas.courty@univ-ubs.fr*
[3] *IRISA, Univ. Bretagne Sud, Vannes, France, chloe.friguet@univ-ubs.fr*

Résumé. Les méthodes d'appariement statistique consistent à intégrer deux ou plusieurs sources de données, relatives à une même population cible. Ces sources partagent un sous-ensemble de covariables tout en disposant d'autres sous-ensembles de variables distincts. Le but est de construire un ensemble unique de données synthétique dans lequel toutes les variables des différentes sources sont disponibles conjointement. Une méthode basée sur une application du transport optimal a été proposée dans Garès and Omer (2020), dans le cas où les variables distinctes des différentes sources de données sont catégorielles. La distribution jointe des variables partagées et distinctes est transportée dans un jeu de données. L'approche proposée ici utilise également le transport optimal pour la distribution des variables partagées et distinctes, mais intègre de plus l'estimation d'une fonction pour prédire les variables distinctes dans l'autre source. Les performances de la méthode sont évaluées via une étude par simulation de Monte Carlo.

Mots-clés. Intégration des données, Appariement statistique, Recodage des variables, Adaptation de domaines, Transport optimal.

Abstract. Statistical matching methods consist in integrating two or more data sources, related to the same target population, which share a subset of covariates while each data source has its own distinct subset of variables. The aim is to derive a unique synthetic data set in which all the variables, coming from the different sources, are jointly available. A method based on an application of optimal transport theory has been proposed by Garès and Omer (2020), in the case where the distinct variables in the different data sources are categorical. Joint distribution of shared and distinct variables is transported within the data sources. Although the method demonstrated good performance, the proposed approach also transports the distribution of shared and distinct variables and estimate a function to predict the missing variables. The performances are assessed through a Monte Carlo simulation study.

Keywords. Data integration, Statistical matching, Variable recoding, Domain adaptation, Optimal transport.

1 Introduction

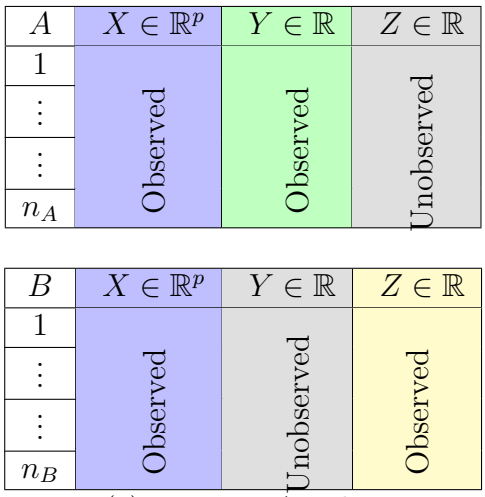
This work addresses the challenge of integrating different data sources considering a Statistical Matching strategy based on Optimal Transport theory.

It is motivated by an application in variable recoding. For example, such issues are encountered when a variable is not coded in the same scale in two datasets. As an illustration, a previous work on data recoding Gares et al. (2019) 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".

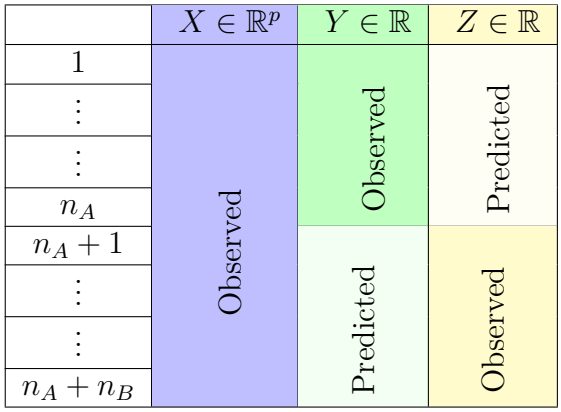
Data recoding is a particular case of Statistical Matching (SM) which consists in determining a model allowing to aggregate information contained in two or more data sources, coming from the same target population. The aim is to derive a unique synthetic dataset in which all the variables, coming from the different sources, are jointly available. The quality and accuracy of statistical analysis carried out retrospectively can therefore be optimised.

As illustrated in Figure 1a, the problem can be formalized in terms of two data sources A and B, with respectively n_A and n_B observations. It is assumed that A and B have disjoint samples and share a subset of p variables X (called *covariates*). Besides, A and B have another distinct subset of variables, denoted respectively Y in A and Z in B (called *outcomes*). As a consequence, there is no unit for which Y and Z are simultaneously observed.

Figure 1: Problem formulation. Statistical matching provides joint information on variables collected through sources A and B, with both common (covariates X) and distinct (outcomes Y and Z respectively) variables



(a) Datasets A and B



(b) Synthetic dataset: joint information

Two main SM approaches are usually considered D'Orazio et al. (2006): (1) the *macro* approach which aims to identify associations between the variables Y and Z, such as joint distributions or correlations and (2) the *micro* approach which consists in generating a complete database in which the data of all the variables are available for each unit, as in Figure 1b

.

However, the association measures between Y and Z conditionally on X can not be estimated and they are therefore generally assumed to be zero (known as the Conditional Independence Assumption (CIA)). CIA is required for non-parametric hot-deck approaches, the methods based on likelihood or propensity score. Multiple imputation methods are also proposed and don't require the CIA. It is worth noting that CIA can not be tested from the datasets and remains a strong assumption.

Besides, Optimal Transport (OT) has been proposed recently as an efficient tool to deal with SM issues in Gares et al. (2019). Outcome Z, observed in B but not in A, is viewed as an additional information that can be specified in the model to improve the estimation of Y in B. However, inequalities between the joint distributions of (X, Y, Z) in A and B may be possible. The distribution of Y is transported forward to the distribution of Z within a data source and a cost function is proposed as an average distance between covariates X from each data sources. This approach has shown better performances when compared to missing data imputation methods such as multiple imputation, non-parametric hot-deck approaches or a statistical learning method. It is assumed that distributions μ^Y, μ^Z, $\mu^{Y|X}$ and $\mu^{Z|X}$ remain unchanged across data sources. This algorithm is extended in Garès and Omer (2020), considering the transport of the joint distribution between (X, Y) and (X, Z) within a data source. The constraints on marginals of OT problem are also relaxed, because they may be too restrictive in the presence of errors in the estimations. A regularization term is added to the objective function to smooth the variations of outcomes with respect to covariates. Only $\mu^{Y|X}$ and $\mu^{Z|X}$ are assumed to remain unchanged across data sources.

In machine learning literature, transfer learning refers to the training of a model on a (largely) labelled dataset, called source domain, and applied to a new unlabeled data set, called target domain. It often occurs in various application that source and target domain differ regarding their conditional or/and marginal distributions, so models trained on source data can not be directly applied to target domain. Domain Adaptation (DA) is a set of proposed techniques to overcome this problem. Recently, OT has been proposed to solve DA issues, under target shift or covariate shift assumptions Courty et al. (2016, 2017). To handle these assumptions, the algorithm jointly optimises a function f to predict a (categorical or continuous) output Y given a input X, and minimizes the OT loss between the joint source distribution (X, Y) and an estimated target joint distribution $(X, f(X))$ Courty et al. (2017).

In this article, we propose a new formulation of the solution proposed in Garès and Omer (2020) for statistical matching as an extension of the model developed in Courty et al. (2017) for domain adaptation. Two classifiers f and g are introduced to predict the missing values for Z in A and Y in B. The proposed algorithm minimizes the OT loss between the joint distribution of $(X, Y, g(X, Y))$ in A and the joint distribution of $(X, f(X, Z), Z)$ in B.

The remainder of the article is organized as follows. In Section 2, the OT model defined in Garès and Omer (2020) to transport the distribution of covariates and outcomes within data source is firstly reviewed. Then, the new formulation is introduced in Section 3, considering OT between the joint distribution of covariates and estimated outcomes. In Section 4, the proposed method is evaluated through simulation studies.

2 Background on optimal transport for statistical matching

Notations and assumptions. Let A and B be two data sources containing distinct units as in 1a. It is further assumed, without loss of generality, that $n_A = n_B = n$. Let's denote the set of indices in both datasets by $A = \{i_1, \ldots, i_n\}$ and $B = \{j_1, \ldots, j_n\}$. $(X_i, Y_i, Z_i)_{i\in A}$ and $(X_j, Y_j, Z_j)_{j\in B}$ are two sequences of *i.i.d.* random variables with values in $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$, where $\mathcal{X} \subset \mathbb{R}^P$ 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* replications 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).

The following assumption is required in Garès and Omer (2020).

Assumption 1. *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.,*

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

All random variables are defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The probability distribution of a random variable V with possible values in $\mathcal{V}$ is given by μ^V. In the case of a discrete variable, $\mu^V = \sum_{v\in\mathcal{V}} \mu_v^V \delta_v$, where δ_v is the Dirac delta measure centered at v. If $\mathcal{V}$ is finite with cardinality $|V|$, μ^V will also refer to the vector of probabilities $(\mu_v^V)_{v\in\mathcal{V}}$.

Optimal transport theory. Consider $\mathcal{X}$ and $\mathcal{Y}$ two Radon spaces. Let μ^X be a probability measure on $\mathcal{X}$, μ^Y a probability measure on $\mathcal{Y}$ and $c : \mathcal{X} \times \mathcal{Y} \longrightarrow [0, \infty]$ 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 $\gamma \in \Gamma(\mu^X, \mu^Y)$ such that:

$$\inf \left\{ \mathbb{E}[c(X, Y)] = \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\, \mathrm{d}\gamma(x, y) \,\middle|\, \gamma \in \Gamma(\mu, \nu) \right\}, \tag{2}$$

where $\Gamma(\mu^X, \mu^Y)$ is the set of measures on $\mathcal{X} \times \mathcal{Y}$ with marginals μ^X on $\mathcal{X}$ and μ^Y on $\mathcal{Y}$. Kantorovich's formulation plugs the problem in a linear setting and the solution is achievable thanks to compacity argument.

Review of regularized optimal transport of covariates and outcomes in data recoding Garès and Omer (2020) The aim is to search for an OT between the two joint distributions of (X^A, Y^A) and (X^A, Z^A) with marginals $\mu^{(X^A,Y^A)}$ and $\mu^{(X^A,Z^A)}$ respectively. Under Kantorovich's formulation in a discrete setting, it is restated as:

$$\gamma^* \in \underset{\gamma\in\Gamma\left(\mu^{(X^A,Y^A)},\mu^{(X^A,Z^A)}\right)}{\operatorname{argmin}} \sum_{\mathcal{X}\times\mathcal{Y}\times\mathcal{X}\times\mathcal{Z}} c_{x,y,x',z}\, \gamma_{x,y,x',z},$$

where c is a given cost function. Any element $\gamma \in \Gamma\left(\mu^{(X^A,Y^A)}, \mu^{(X^A,Z^A)}\right)$ corresponds to the vector of joint probabilities $\mathbb{P}\Big((X^A = x, Y^A = y), (X^A = x', Z^A = z)\Big)$ for all $x, x' \in \mathcal{X}$,

$y \in \mathcal{Y}$ and $z \in \mathcal{Z}$. Since this probability is null for all $x \neq x'$, γ is defined as a vector of $[0,1]^{|\mathcal{X}|\times|\mathcal{Y}|\times|\mathcal{Z}|}$, where $\gamma_{x,y,z}$ stands for the joint probability $\mathbb{P}(X^A = x, Y^A = y, Z^A = z)$. The following OT model is therefore introduced:

$$\mathcal{P}_1 : \begin{cases} \min\limits_{\gamma} \sum\limits_{\mathcal{X}\times\mathcal{Y}\times\mathcal{X}\times\mathcal{Z}} c_{x,y,z}\,\gamma_{x,y,z} \\ \text{s.t.} \sum\limits_{z\in Z} \gamma_{x,y,z} = \mu_{x,y}^{(X^A,Y^A)},\ \forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \\ \quad \sum\limits_{y\in Y} \gamma_{x,y,z} = \mu_{x,z}^{(X^A,Z^A)},\ \forall x \in \mathcal{X}, \forall z \in \mathcal{Z}, \\ \quad \gamma_{x,y,z} \geq 0,\ \forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \forall z \in \mathcal{Z}. \end{cases} \tag{3}$$

The above model can be solved only if the cost c and the marginals $\mu^{(X^A,Y^A)}$ and $\mu^{(X^A,Z^A)}$ are known. Thanks to assumption 1, consistent estimators $\hat{\mu}_n^{(X^A,Y^A)}$ and $\hat{\mu}_n^{(X^A,Z^A)}$ of $\mu^{(X^A,Y^A)}$ and $\mu^{(X^A,Z^A)}$ can be derived.

Let's define the unbiased estimators of $\mu^{(X^A,Y^A)}$ and $\mu^{(X^B,Z^B)}$ by:

$$\hat{\mu}_{n,x,y}^{(X^A,Y^A)} = \frac{1}{n}\sum_{i\in A} \mathbb{1}_{\{X_i=x,Y_i=y\}},\ \forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \tag{4}$$

$$\hat{\mu}_{n,x,z}^{(X^B,Z^B)} = \frac{1}{n}\sum_{j\in B} \mathbb{1}_{\{X_j=x,Z_j=z\}},\ \forall x \in \mathcal{X}, \forall z \in \mathcal{Z}. \tag{5}$$

Similarly, $\hat{\mu}_n^{X^A}$ and $\hat{\mu}_n^{X^B}$ denote the unbiased empirical estimators of μ^{X^A} and μ^{X^B}.

Under assumption 1, the estimator of $\mu^{(X^A,Z^A)}$ is, $\forall x \in \mathcal{X}$, $\forall z \in Z$:

$$\tilde{\mu}_{n,x,z}^{(X^A,Z^A)} = \begin{cases} \frac{\hat{\mu}_{n,x,z}^{\left(X^B,Z^B\right)} \hat{\mu}_{n,x}^{X^A}}{\hat{\mu}_{n,x}^{X^B}}, & \text{if } \hat{\mu}_{n,x}^{X^B} \neq 0, \\ 0, & \text{if } \hat{\mu}_{n,x}^{X^B} = 0. \end{cases} \tag{6}$$

Moreover, the cost measure is defined by:

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

with d a distance on the metric space $\mathcal{X}$, and is estimated by:

$$\hat{c}_{n,y,z} = \begin{cases} \frac{1}{\kappa_{n,y,z}} \sum_{i\in A}\sum_{j\in B}\ \mathbb{1}_{\{Y_i=y,Z_j=z\}} \times d(X_i, X_j), & \forall y \in \mathcal{Y}, \forall z \in \mathcal{Z} : \kappa_{n,y,z} \neq 0, \\ 0, & \forall y \in \mathcal{Y}, \forall z \in \mathcal{Z} : \kappa_{n,y,z} = 0, \end{cases} \tag{8}$$

with

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

As datasets A and B are independent, $\hat{c}_{n,y,z}$ is a consistent estimator of $c_{x,y,z}$, $\forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \forall z \in \mathcal{Z}$. Plugging the observed values for these estimators in (3) yields a linear programming model denoted as $\hat{\mathcal{P}}_{1,n}$. The solution $\hat{\gamma}_n$ can then be interpreted as an estimator $\hat{\mu}_n^{(X^A,Y^A,Z^A)}$ of the joint distribution of X^A, Y^A and Z^A, $\mu^{(X^A,Y^A,Z^A)}$.

An estimation of the distribution of Z^A given the values of X^A and Y^A is then deduced:

$$\tilde{\mu}_{n,z}^{Z^A|X^A=x,Y^A=y} = \begin{cases} \frac{\hat{\gamma}_{n,x,y,z}}{\hat{\mu}_{n,x,y}^{(X^A,Y^A)}}, & \forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \forall z \in \mathcal{Z} : \hat{\mu}_{n,x,y}^{(X^A,Y^A)} \neq 0, \\ 0, & \forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \forall z \in \mathcal{Z} : \hat{\mu}_{n,x,y}^{(X^A,Y^A)} = 0. \end{cases} \tag{9}$$

Due to the possible errors in the estimations of the terms of $\hat{\mathcal{P}}_{1,n}$, the equality constraints is relaxed by adding slack variables in the constraints, such that they sum to zero and the ℓ_l-norm of the vector of slack variables is bounded.

A regularization term is also added, considering $\left(\frac{\gamma_{x,y,z}}{\hat{\mu}_{n,x}^{X^A}}\right)_{x\in\mathcal{X},y\in\mathcal{Y},z\in\mathcal{Z}}$. Some regularity in the variations of the conditional distribution $\mu^{Y^A,Z^A|X^A=x}$ with respect to x is then expected.

3 Optimal transport between the joint distribution of covariates and estimated outcomes between two datasets

In this section, we extend the model developed by Courty et al. (2017) and propose to search for an optimal transport between the two joint distributions of (X^A, Y^A, Z^A) and (X^B, Y^B, Z^B) leading to the following OT model:

$$\gamma^* \in \underset{\gamma\in\Gamma\left(\mu^{(X^A,Y^A,Z^A)},\mu^{(X^B,Y^B,Z^B)}\right)}{\operatorname{argmin}} \sum_{(X\times\mathcal{Y}\times\mathcal{Z})^2} c\Big((x,y,z),(x',y',z')\Big)\,\gamma_{(x,y,z),(x',y',z')}, \tag{10}$$

where $c((x,y,z),(x',y',z')) = d(x,x') + \alpha_1\mathcal{L}_1(y,y') + \alpha_2\mathcal{L}_2(z,z')$ is a cost measure combining both the distances between the modalities of X and loss functions $\mathcal{L}_1$ and $\mathcal{L}_2$ measuring the discrepancy between y and y' and z and z' respectively. Since z and y' are not observed, we replace them by $f(x,y)$ and $g(x',z')$ respectively where $f : \mathcal{X}\times\mathcal{Y} \to \mathcal{Z}$ and $g : \mathcal{X}\times\mathcal{Z} \to \mathcal{Y}$ are prevision rules. We thus consider the following joint distributions $(X^A, Y^A, f(X^A, Y^A))$ and $(X^B, g(X^B, Z^B), Z^B)$. α_1 and α_2 are crucial hyper-parameters which need to be calibrated balancing the alignment of covariates and outcomes. A regularization term, such that entropic, can also be added, such as in Courty et al. (2014).

Transport of individuals. The empirical version of (10) is given by

$$\hat{\mathcal{P}}_{2,n} : \begin{cases} \min\limits_{f,g,\gamma} \sum\limits_{i\in A, j\in B} c\Big((x_i,y_i,f(x_i,y_i)),(x_j,g(x_j,z_j),z_j)\Big)\gamma_{i,j}, \\ \text{s.t.} \sum\limits_{i\in A}\gamma_{i,j} = \dfrac{1}{n},\ \forall j \in B, \\ \quad \sum\limits_{j\in B}\gamma_{i,j} = \dfrac{1}{n},\ \forall i \in A, \\ \quad \gamma_{i,j} \geq 0,\ \forall i \in A, \forall j \in B. \end{cases} \tag{11}$$

The problem can be written as

$$\min_{f,g} \mathcal{W}_1\Big(\widehat{\mu}^{A,f}, \widehat{\mu}^{B,g}\Big),$$

where $\mathcal{W}_1$ is the 1-Wasserstein distance for the cost c,

$$\mu^{A,f} = \Big(x, y, f(x,y)\Big)_{(x,y)\sim\mu^{(X^A,Y^A)}} \text{ and } \mu^{B,g} = \Big(x, z, g(x,z)\Big)_{(x,z)\sim\mu^{(X^B,Z^B)}},$$

and

$$\widehat{\mu}^{A,f} = \frac{1}{n}\sum_{i\in A} \delta_{(x_i,y_i,f(x_i,y_i))} \text{ and } \widehat{\mu}^{B,g} = \frac{1}{n}\sum_{j\in B} \delta_{(x_j,z_j,g(x_j,z_j))}.$$

We will assume that f and g belongs to the function space $\mathcal{H}$ which is a Reproducing Kernel Hilbert Space (RKHS) or a function space parameterized by some parameters $w \in \mathbb{R}^p$. For example, linear models, neural networks and kernel methods belong to such a space.

Problem $\widehat{\mathcal{P}}_{2,n}$ is smooth and the constraints are separable according to γ, f and g. Hence, a natural way to solve the problem in Equation (11) is to alternate optimization on parameters γ, f and g. This algorithm is known as Block Coordinate Descent (BCD) or Gauss-Seidel. Solving $\widehat{\mathcal{P}}_{2,n}$ when f and g are fixed is a classic linear programming problem. The optimization problem with fixed γ and f leads to a new learning problem given by:

$$\min_{g\in\mathcal{H}} \sum_{i\in A, j\in B} \mathcal{L}_1\Big(g(x_j,z_j), y_i\Big)\gamma_{i,j}. \tag{12}$$

The optimization problem is similar with fixed γ and g.

For estimating a multiclass classifier with a one-against-all strategy, in Courty et al. (2017), they suggest to use the Hinge function for classification problem equal to $\mathcal{L}_1(y, f(x,z)) = \max(0, 1 - yf(x,z))^2$. Let define P such that $P^B_{j,y} = 1$ if the unit j is of class $y \in \mathcal{Y}$ and $P^B_{j,y} = 0$ otherwise. Denote as g_y the decision function related to the y-vs-all problem. They replace the problem defined in 12 by:

$$\min_{g_y\in\mathcal{H}, y\in\mathcal{Y}} \sum_{j\in B,} \sum_{\mathcal{Y}} \hat{P}^B_{j,y}\mathcal{L}_1(1, g_y(x_j,z_j)) + (1-\hat{P}^B_{j,y})\mathcal{L}_1(-1, g_y(x_j,z_j)), \tag{13}$$

where $\widehat{P}^B$ is the transported class proportion matrix given by $\hat{P}^B = n\gamma^T P^A$.

4 Simulations

Simulation scenarios. Simulation scenarios are chosen to be the same as in Garès and Omer (2020). Each dataset is constructed by generating n independent samples of (X, Y, Z) according to predefined distributions that may vary between A and B. In all our simulations, $(Y_i, Z_i)_{i\in A}$ and $(Y_j, Z_j)_{j\in B}$ are obtained by discretization of continuous multivariate random variables. Let $\{U_i\}_{i\in A}$ be a family of *i.i.d.* 3-dimensional random vectors with multivariate 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 consider $\mathbf{\Sigma}^A = \mathbf{\Sigma}^B = \mathbf{\Sigma}$, where $\mathbf{\Sigma}_{i,i} = 1$, $\forall i =$

$1, 2, 3$ and $\mathbf{\Sigma}_{i,j} = 0.2$ for $i \neq j$. 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.

$$V_i = a_1 U_{i,1} + a_2 U_{i,2} + a_3 U_{i,3} + \sigma W_i, \tag{14}$$

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 $m^A = (0, 0, 0)$, $m^B = (1, 0, 0)$, $a = (1, 1, 1), R^2 = 0.6$ and $n = 1000$.

Methods. In this work, we compare the different algorithms without any relaxation of the constraints ($\alpha = 0$) or regularization ($\lambda = 0$) to focus on the objective function performance. We compare the different OT algorithms:

OT Transport of the joint distribution of covariates and outcomes within a data source ($\widehat{\mathcal{P}}_{1,n}$).

OTE Transport of covariates and estimated outcomes between data sources ($\widehat{\mathcal{P}}_{3,n}$). For classifiers f and g, we used neural networks with one hidden layer with relu activation function and 160 neurons and softmax activation function for the last layer, adam optimizer, 100 epochs and a batch size equal to 10. We used 10 iterations for the BCD algorithm.

OTE-boost Transport of covariates and estimated outcomes between data sources ($\widehat{\mathcal{P}}_{3,n}$) using eXtreme Gradient Boosting algorithms with depth equal to 5 and 1000 estimators for classifiers f and g. We used 10 iterations for the BCD algorithm.

We choose $\alpha_1 = 0.25$ and $\alpha_2 = 0.33$ to balance the alignment of covariates and outcome.

Evaluation of the methods. To evaluate the performance of the methods, we compute the accuracy of prediction of Z in A and Y in B given by: $\frac{1}{n} \sum_{i \in A} \mathbb{1}_{\{\hat{z}_i = z_i\}} + \sum_{j \in B} \mathbb{1}_{\{\hat{y}_j = y_j\}}$ where, for the first algorithms, we use the usual prediction provided by the maximum a posteriori decision rule $\widehat{z}_i^A = \mathrm{argmax}_{z \in \mathcal{Z}} \left\{ \tilde{\mu}_{n,z}^{Z^A | X^A = x_i, Y^A = y_i} \right\}, \forall i \in A$.

Effectiveness of different methods to the different data assumptions 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.

Effect of the R^2 Keeping m^B and a as in $\mathrm{S}_{\mathrm{ref}}$, we investigate the impact of the R^2 choosing $R^2 \in \{0.2, 0.4, 0.6, 0.8\}$.

Covariate shift assumption Keeping R^2 and a as in $\mathrm{S}_{\mathrm{ref}}$, we investigate the impact of differences in the distributions of X^A and X^B by considering the following four scenarios: 1: $m^A = m^B = (0, 0, 0)$, $m^A = (0, 0, 0)$, 2: $m^B = (1, 0, 0)$, 3: $m^A = (0, 0, 0)$, $m^B = (1, 1, 0)$, and 4: $m^A = (0, 0, 0)$, $m^B = (1, 2, 0)$.

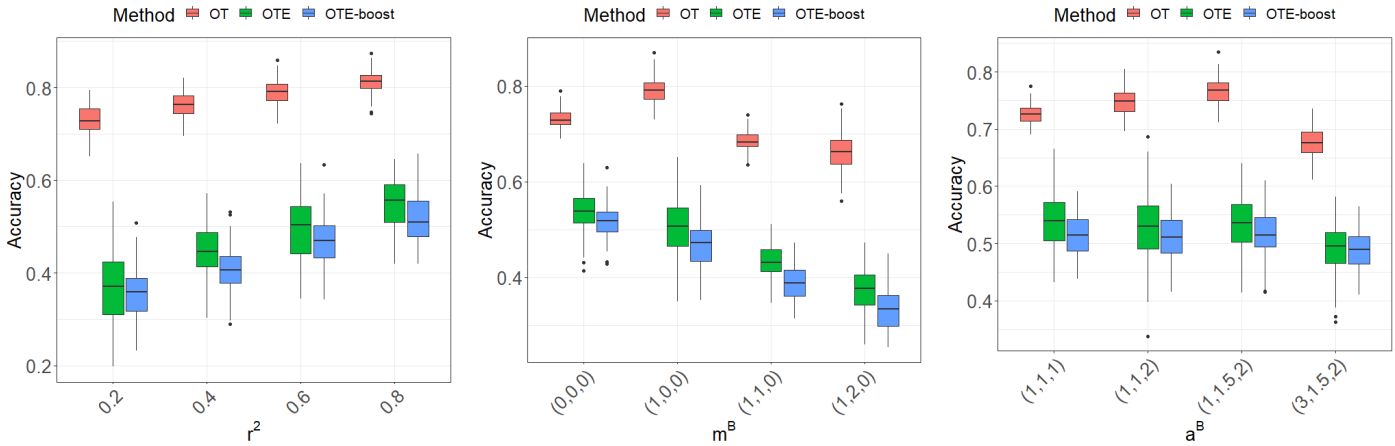
Changes in conditional distribution $Y|X$ Finally, we wish to evaluate the importance of satisfying the assumption that the distributions of Y and Z given X are the same in the two databases. For this, we allow the vector a to be different in the two databases when computing V. More formally,

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

with $a^A, a^B \in \mathbb{R}^3$. Keeping R^2, m^A as in $\mathrm{S}_{\mathrm{ref}}$ and $m^B = (0, 0, 0)$, we consider the following four scenarios: 1: $a^A = (a^B 1, 1, 1)$, 2: $a^A = (1, 1, 1)$ and $a^B = (1, 1, 2)$, 3: $a^A = (1, 1, 1)$ and $a^B = (1, 1.5, 2)$, 4: $a^A = (1, 1, 1)$ and $a^B = (3, 1.5, 2)$.

Results The performances of (OT) are larger than the ones of (OTE) and (OTE-boost) for all scenarios. The performances of the algorithms increase as R^2 increases (2a) and decrease as the difference between the covariate distribution increases (2b). The change in conditional distribution has a small negative effect on the performances (2c) .

Figure 2: Effect of different scenario parameters on the methods performances (Accuracy)



(a) R^2 (b) Covariate shift (c) Changes in $Y|X$

References

Courty, N., Flamary, R., Habrard, A., and Rakotomamonjy, A. (2017). Joint distribution optimal transportation for domain adaptation. In *Advances in Neural Information Processing Systems*, pages 3730–3739.

Courty, N., Flamary, R., and Tuia, D. (2014). 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.

Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. (2016). Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865.

D'Orazio, M., Di Zio, M., and Scanu, M. (2006). *Statistical matching: Theory and practice*. John Wiley & Sons.

Gares, V., Dimeglio, C., Guernec, G., Fantin, R., Lepage, B., Kosorok, M. R., and Savy, N. (2019). On the use of optimal transportation theory to recode variables and application to database merging. *The international journal of biostatistics*, 16(1).

Garès, V. and Omer, J. (2020). Regularized optimal transport of covariates and outcomes in data recoding. *Journal of the American Statistical Association*, pages 1–14.