Learning relationships between data obtained independently

Alexandra Carpentier
Universität Potsdam
carpentier@math.uni-potsdam.de

Teresa Schlüter
London School of Economics and WBGU
teresaschlueter@googlemail.com

Abstract

The aim of this paper is to provide a new method for learning the relationships between data that have been obtained independently. Unlike existing methods like matching, the proposed technique does not require any contextual information, provided that the dependency between the variables of interest is monotone. It can therefore be easily combined with matching in order to exploit the advantages of both methods. This technique can be described as a mix between quantile matching, and deconvolution. We provide for it a theoretical and an empirical validation.

1 Introduction

The Big Data phenomenon is made possible by the parallelisation of data acquisition. The data are not collected by a centralised organism, but by several contributors. This is a strength since it allows a huge variety and quantity of data to be made available. However, a necessary step for studying data from different contributors is *merging* these datasets. For instance, consider the classical problem of knowing which part Y of its income X an individual spends on housing. Formally, we model the relation between X and Y by

$$Y = h(X, Z) + \epsilon,$$

where ϵ is a noise, Z is (in case they are provided) some additional *contextual variables* (like e.g. age, sex, job, etc) and $h_Z := h(., Z)$ is the dependence function between Y and X given the contextual variables Z. The objective is to estimate the dependence relation h_Z. An obstacle for answering this question is that in most countries, the data on wages $\mathbf{X}$ are collected by one kind of agent (e.g. the office for national statistics), and the data on housing transactions $\mathbf{Y}$ are collected by another kind of agent (e.g. financial institutions specialized in mortgage lending). The dependency h_Z between these variables cannot be established immediately, since the two data sets have been collected independently : $\mathbf{X}$ and $\mathbf{Y}$ are independent. Standard results on regression don't apply in this context. Estimating h_Z, or *merging the variables* $\mathbf{X}$ *and* $\mathbf{Y}$, is then challenging. This kind of problem becomes more and more common with the growing importance of social networks such as twitter, facebook, etc. For instance, it is desirable to combine the data collected by these social networks by merging the user profiles, in the interest of a fuller analysis of their content.

A very popular method for overcoming this problem is called *matching*, see [18, 15, 6]. Suppose that in addition to collecting data on incomes and house prices, the agents also collect *contextual data* (corresponding to the contextual variables Z) such as the age, sex, job, etc. In other words, the two independent datasets are of the form $(\mathbf{X}, \mathbf{Z}(1))$ and $(\mathbf{Y}, \mathbf{Z}(2))$ where $\mathbf{Z}(1)$ and $\mathbf{Z}(2)$ are the contextual data of respectively $\mathbf{X}$ and $\mathbf{Y}$. The matching procedure consists in associating the data in $\mathbf{X}$ with the data in $\mathbf{Y}$ that are most similar in terms of the contextual data, i.e. in associating the $(X_i, Z(1)_i)$ with the $(Y_j, Z(2)_j)$ such that a given distance between $Z(1)_i$ and $Z(2)_j$ is minimised. Once the matched dataset is available, h_Z can be inferred using a suitable regression method. Also using this idea of cleverly taking advantage of contextual variables, many other methods have been considered for dealing with the problems of fusing data coming from independent sources, as *data fusion, linkage, data integration, etc* [1, 10, 3, 11].

These approaches are very reasonable, but they rely heavily on the contextual data $\mathbf{Z}(1)$ and $\mathbf{Z}(2)$. If these contextual data are not very detailed, these methods can perform poorly. For instance, the most complete UK database on housing transactions which is collected by the Land Registry contains the house market transaction prices, some structural characteristics of the house and the geographical location, but no in-

Appearing in Proceedings of the 19^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 41.

formation on the buyer. On the other hand, the micro databases that contain information on the wage of individuals have few if no information on the type of lodging that the individuals occupy, apart from sometimes an approximate geographic location. In this baseline problem, the approximate geographic location is the only matching variable available. In some highly populated area, there will be many matches. In this situation, an additional noise, or even a model misspecification will be introduced by the imprecision of the matching. In the case of the social network example, the collected data are often partially anonymized, which makes the merging process difficult.

In this paper, we develop an alternative method for learning the relationships between two variables that have been collected independently. Our approach *does not rely on contextual variables* and can therefore be used to improve on any given matching method - it is particularly interesting in the situation where many possible matches are available. It can even be used when *no contextual variables are available*, i.e. when the only available data are $\mathbf{X}$ and $\mathbf{Y}$ and when they have been collected independently.
The necessary assumptions in order for our method to work is that the dependence function h_Z given the contextual variables Z is monotone (increasing or decreasing) and that the noise distribution is known (exactly or by an estimate). In the income and house price example, it is clear that h_Z is an increasing function given standard contextual variables Z such as sex, age, job, etc - the larger the income of an individual, the more it will pay for its house *on average* given the standard contextual variables. In the social network example, a high utilisation of a given social mainstream network is *on average* positively connected with a high utilisation of another mainstream social media. This situation - an increasing relation h_Z given the contextual variables Z - is in fact fairly common, since we only need *averaged* monotonicity (through h).
The fact that our method can even be used when no contextual variables are available can seem very surprising, since it is counter intuitive that the relation h (when there are no contextual variables, $h_Z = h$) between $\mathbf{X}$ and $\mathbf{Y}$ can be deduced from data $\mathbf{X}$ and $\mathbf{Y}$ that are independent. The reason why this is possible is the monotonicity of h. In the simple case when the noise ϵ is null, if it is known that h is e.g. increasing, then the t percentile of $\mathbf{X}$ corresponds to the t percentile of $\mathbf{Y}$. The relation h between $\mathbf{X}$ and $\mathbf{Y}$ can then be inferred by matching the quantiles of the distributions of $\mathbf{X}$ and $\mathbf{Y}$. When there is noise, the problem is slightly more involved, and this is the situation that we develop in this paper. Our approach mixes elements of *quantile regression and quantile comparison*, and of *deconvolution*, applied in a non-standard way to the problem of merging data sets. Quantile regression and quantile comparison (see e.g. [14, 19, 12, 9]) consists in fitting a distribution with other distributions, or comparing a distribution with other distributions, by using the quantiles of the distributions. Distribution deconvolution (see e.g. [8, 5, 13, 7, 4, 2, 5, 16, 17]) consists in estimating the distribution of a random variable using noisy samples.

This paper is structured as follows. In Section 2, we describe the setting of the paper. In Section 3, we provide the estimator of h_Z, and associated bounds on its performance. Finally, we present numerical experiments on world bank data, and on census and land registry data in Section 4 for assessing the practical impact of our method. The Appendix contains the proofs of the main theorems and additional experiments.

2 Setting

In this section, we present in a formal way the setting and the objective.

2.1 Presentation of the model

Let $d \in \mathbb{N}$ be the dimension of the contextual variables Z. In our approach, we allow $d = 0$, i.e. no contextual variables are available. We do not make assumptions on the way the contextual variables are generated.

We assume the following modelling for the explicative variables X:

$$X|Z \sim f_Z,$$

where f_Z is a density on $\mathbb{R}$ (and F_Z is the associated distribution function). We now assume the following modelling for the explained variable Y:

$$Y := h(X, Z) + \epsilon,$$

where $\epsilon|Z \sim \xi_Z$ is a noise that is independent of X, where ξ_Z is a density on $\mathbb{R}$ of mean 0, and where for any Z, $h(., Z)$ is a function from $\mathbb{R}$ to $\mathbb{R}$ which is *monotone*. We assume that we know the distribution ξ_Z. We write $f_{h,Z}$ for the density of $h(X, Z)|Z$ (and $F_{h,Z}$ for the associated distribution function), and g_Z for the density of $Y|Z$. Given the model, we have

$$g_Z = f_{h,Z} * \xi_Z,$$

i.e. g_Z is the convolution of $f_{h,Z}$ and ξ_Z.

Remark on the monotonicity assumption: Assuming that for any Z, $h(., Z)$ is monotone, e.g. nondecreasing, means that given the contextual variables Z and *on average*, a larger X corresponds to a larger Y. In the rent and wage example, it is a very reasonable assumption that richer people, given standard contextual variables as their sex, age, socio professional category, location, etc spend *on average* more

on lodging. Another example concerns the number of followers on two social media such as e.g. twitter and facebook. It is reasonable to assume that someone who is very active on facebook is more likely to be active on twitter. This assumption is not very restrictive since it is made *on average*, and we do not make the assumption that there is a strict order relationship that holds for every individual, which is a much stronger assumption.

2.2 Data

We assume that we are given two databases:

$$(\mathbf{X}, \mathbf{Z}(1)) \quad \text{and} \quad (\mathbf{Y}, \mathbf{Z}(2))$$

with respectively m and n individuals, where $X_i|Z(1)_i \sim f_{Z(1)_i}$ and $Y_i|Z(2)_i \sim g_{Z(2)_i}$ and are *totally independent*, as e.g. in the case where they are collected from two independent sources. For instance, $\mathbf{X}$ can be a dataset containing a sample of wages of individual in a geographical unit, and $\mathbf{Y}$ a dataset containing a sample of house prices in the same geographical unit, and $\mathbf{Z}(1), \mathbf{Z}(2)$ can be standard categorical variables such age age and sex. The problem here is that these datasets have been collected independently of each other, and one does not know which individual, of a given wage, buys which house.

Objective: Infer the function h from the data $(\mathbf{X}, \mathbf{Z}(1))$ and $(\mathbf{Y}, \mathbf{Z}(2))$.

2.3 Preliminary matching procedure

Matching procedures (see e.g. [18, 15, 6]) aim at merging $(\mathbf{X}, \mathbf{Z}(1))$ and $(\mathbf{Y}, \mathbf{Z}(2))$ by finding good matches between the contextual variables $\mathbf{Z}(1)$ and $\mathbf{Z}(2)$. The basic procedure can be summarized as follows. Let d be a distance function between the data points in $\mathbf{Z}(1)$ and the data points in $\mathbf{Z}(2)$. A non-robust matching procedure aims at associating any $(X_i, Z(1)_i)$ of the first dataset with the point $(Y_j, Z(2)_j)$ of the second dataset that minimises $d(Z(1)_i, Z(2)_j)$. A more robust generalisation of this method, related to nearest neighbours methods, is to match, for any z, the points $\mathcal{D}_1(z)$ of the first dataset whose contextual variables $\mathbf{Z}(1)$ are $\upsilon-$close to z, with points $\mathcal{D}_2(z)$ of the second dataset whose contextual variables $\mathbf{Z}(2)$ are $\upsilon-$close to z, for $\upsilon > 0$.

Particularly in the case of not very detailed contextual variables (age, sex, etc), a typical matching approach would then not return a one to one match from $\mathbf{X}$ to $\mathbf{Y}$, but would map subsets of $\mathbf{X}$ to subsets of $\mathbf{Y}$, in function of Z. So a matching procedure outputs, for values of Z that exist in the dataset, the following subsets of $\mathbf{X}$ and $\mathbf{Y}$ that correspond to Z :

$$\mathbf{X}^Z \quad \text{and} \quad \mathbf{Y}^Z.$$

This provides a first merging of the variables, but in the case where for given Z, the function $h(., Z)$ is not constant (which is often the case for not very detailed contextual variables), this does not allow for a reconstruction of h, and therefore for the determination of the relation between X and Y.

In this paper is to refine a such procedure (or in the case where there are no contextual variables and therefore where no matching is possible, our aim is to link as well as we can X and Y), and is thus to estimate, for any Z in the set of contextual variables, the relation between X and Y given Z, i.e. the function

$$h_Z(.) := h(., Z),$$

given the data $\mathbf{X}^Z \quad \text{and} \quad \mathbf{Y}^Z$.

Revisited objective: Infer the function h_Z from the data $\mathbf{X}^Z \quad \text{and} \quad \mathbf{Y}^Z$.

Remark on the model The post-matching model is $y = h_Z(x^Z) + \epsilon$ - so that if for a given context Z we observed a dataset of the form (X^Z, Y^Z) (therefore with with cross-information), X^Z and Y^Z would not be independent - and one could use standard techniques as e.g. regression. But in our setting we observe the data X^Z and Y^Z from different, independent sources - making de facto X^Z and Y^Z independent (and fully independent, not just knowing the order statistics). Regression techniques are not applicable there since we do not have the information of which x^Z in dataset X^Z corresponds to which y^Z in dataset Y^Z.

3 Methods and results

We now present our main procedure and results.

3.1 Main procedure and results

We assume that we dispose of a matching procedure based on the contextual variables Z as described in Subsection 2.3. We restrict to the case of discrete contextual variables for the theoretical results, and we consider exact matching according to these variables (the subsets $\mathbf{X}^Z \quad \text{and} \quad \mathbf{Y}^Z$ correspond to exactly the same Z) but this can be easily generalised in practice. If no contextual variables Z are available ($d = 0$), then we use as convention in the rest of this section

$$\mathbf{X}^Z := \mathbf{X} \quad \text{and} \quad \mathbf{Y}^Z := \mathbf{Y}.$$

Let Z be a given value of the contextual variables such that $(\mathbf{X}^Z, \mathbf{Y}^Z)$ are non-empty and let n_Z be the number of data in the smallest of these two sets. Let $\hat{F}_Z$ be the empirical distribution estimator of F_Z defined over the samples $\mathbf{X}^Z$. We assume that there is a *deconvolution estimator* $\hat{F}_{h,Z}$ that estimates the distribution $F_{h,Z}$, based on $\mathbf{Y}^Z$ (separating its density $g_Z = f_{h,Z} * \xi_Z$ form the noise ξ_Z), and satisfies the following assumption.

Algorithm 1 The procedure MatchMerge

Input:
$(\mathbf{X}, \mathbf{Z}(1))$ and $(\mathbf{Y}, \mathbf{Z}(2))$
A matching method with respect to Z
A deconvolution method from the noise ξ_z

Main procedure :
Apply the matching method to the data and obtain for all Z $\mathbf{X}^Z$ and $\mathbf{Y}^Z$
for Z s.t. $n_Z > 1$ **do**
Compute the empirical estimator $\hat{F}_Z$ of F_Z on $\mathbf{X}^Z$
Compute the deconvolution estimator $\hat{F}_{h,Z}$ of $F_{h,Z}$ using $\mathbf{Y}^Z$ and ξ_Z
Set $\hat{h}_Z = \hat{F}_{h,Z}^{-1} \circ \hat{F}_Z$.
end for

Output :
Return $\hat{h}(.,.) = \hat{h}_.(.)$

Assumption 3.1 (Deconvolution estimator available for the explained residuals)**.** *Let $\delta > 0$. Let $x \in \mathbb{R}$. There exists an estimator $\hat{F}_{h,Z}$ of $F_{h,Z}$ computed using the $\mathbf{Y}^Z$ and the knowledge of ξ and that is such that with probability larger than $1 - \delta$*

$$|F_{h,Z}(x) - \hat{F}_{h,Z}(x)| \leq \psi(\delta, n_z) := \psi. \tag{1}$$

The existence of a such *deconvolution estimator*, satisfying Assumption 3.1 is standard under some regularity conditions. A discussion on the existence of it is provided in the Subsection A.2.

Theorems 3.1 and 3.2 below give some properties on the efficiency of the estimate $\hat{h}_Z$ of h_Z defined as

$$\hat{h}_Z = \hat{F}_{h,Z}^{-1} \circ \hat{F}_Z,$$

where $\hat{F}_{h,Z}^{-1}$ is the pseudo inverse of $\hat{F}_{h,Z}$. The entire procedure for computing $\hat{h}_Z$, which we call MatchMerge, for constructing this estimator of h is summarized in Algorithm 1.

The following theorem provides a first theoretical guarantees for $\hat{h}_Z(.)$, as well as a confidence statement in any point u.

Theorem 3.1. *Let $u \in \mathbb{R}$. Let Assumption 3.1 be satisfied for $\delta > 0$. We have with probability $1 - 2\delta$*

$$h_Z \circ F_Z^{-1}(F_Z(u) - \psi - \phi) \leq \hat{h}_Z(u) \leq h_Z \circ F_Z^{-1}(F_Z(u) + \psi + \phi).$$

The proof of this result is in the Appendix, Subsection A.3. This theorem provides a bound on the accuracy of the estimate of $\hat{h}_Z(u)$ if a rather mild condition (Assumption 3.1 is verified). It is important to note here that apart from the assumption that h_Z is a monotone function (without loss of generality, let us say that h_Z is non-decreasing), no additional assumptions are made on h_Z. And even though the datasets $\mathbf{X}^Z$, $\mathbf{Y}^Z$ are independent, it is possible to recover the link function h in this non-parametric model. The bound in this theorem is not explicit, it depends on $h_Z, F_Z, F_{h,Z}$. The next theorem gives an explicit bound (depending on ϕ, ψ), provided that an additional assumption is made on F_Z, h_Z.

Assumption 3.2 (Hölder assumption)**.** *Let $(\alpha, L) > 0$. A function G is Hölder continuous on $\mathcal{V}$ if for any $(x, y) \in \mathcal{V}^2$, we have*

$$|G(x) - G(y)| \leq L|x - y|^\alpha.$$

The Hölder assumption is mild if α is small and L is large. In particular, functions that are differentiable on a compact (or Lipschitz) verify it with $L = \sup |G'|$ and $\alpha = 1$, but it is more general than this (functions of the form x^α verify it in 0 for any $\alpha > 0$ with parameters α and $L = 1$).

This theorem provides a more specific theoretical guarantee in the case where the distributions are Hölder smooth.

Theorem 3.2. *Let $u \in \mathbb{R}$. Let Assumption 3.1 be satisfied for $\delta > 0$ and assume that for any Z, h_Z is monotone (without loss of generality, non-decreasing). Let $L, M, \alpha, \beta > 0$. Assume that h_Z is $(\alpha, L)-$Hölder on $[u - M(\psi + \phi)^\beta, u + M(\psi + \phi)^\beta]$, and that F_Z^{-1} is $(\beta, M)-$Hölder on $[F_Z(u) - \psi - \phi, F_Z(u) + \psi + \phi]$ (where F_Z^{-1} is the pseudo-inverse of F_Z) as in Assumption 3.2.*

Then with probability larger than $1 - 2\delta$

$$|\hat{h}_Z(u) - h_Z(u)| \leq LM^\alpha(\psi + \phi)^{\alpha\beta}.$$

The proof of this result is in the Appendix, Subsection A.4.

The bound in Theorem 3.2 implies that it is possible to recover the link function between the data X, Y, Z,

$$h(x, z) = h_z(x),$$

with a good precision. Our approach enables us to complement matching in this way. Again, if no contextual variables Z are available ($d = 0$), our approach is still applicable by just skipping the matching step.

Remark on the noise assumption In order for our theoretical results to holds, we need to know the distribution of the noise to perform the deconvolution. Let f^* be the Fourier transform of the noise, and f be the Fourier transform of a distribution. If $|1 - 1/f^*| \geq |1/f - 1/f^*|$, i.e. if the distance defined as $d(g, g') = |1/g - 1/g'|$, between the Dirac Fourier transform and

f^* is larger than the distance between f and f^*, then deconvolution with f is better than quantile matching. The distance relation stated before is quite sensitive to parameters such as variance or range, so if f^* is closer to f in terms of these quantities than to a Dirac mass in 0, it is likely that our method applied using f in the deconvolution step will be more efficient than simple quantile matching.

In the case qhere $\tilde{h}$ is separable, we introduce a specific and more efficient method in the Supplementary Material, Appendix A.1.

4 Experiments

In this section, we apply our methods on two real datasets : the first dataset contains macro-economic data from the world bank on development in 45 African countries, and the second contains micro-economic data from the UK land registry on house transaction prices in London. In order to assess the performance of our method, we have deliberately chosen datasets which contain simultaneously the two variables we want to merge and therefore their relationship - but we split in these two applications these datasets when we apply our method and consider the first variable on the first part and the second variable on the second part, so that the two variables are independent and so that we are in real condition for applying our method. The datasets we use are publicly available and more informations on them can be found in the description of the experiments. We also performed simulations and applied our method on synthetic data, this is to be found in the Appendix B due to space constraint.

4.1 Experiment on real data I: Urbanisation and life expectancy

In the first example we look at the impact of urbanisation on life expectancy in 45 African countries, in 2006. We use world bank macro-economic data (the Africa Development Indicators (ADI) dataset). We consider the two variables $\mathbf{Y}$ =“average life expectancy” and $\mathbf{X}$ =“urbanisation percentage”. The life expectancy is clearly an increasing function of urbanisation since a more urbanised country implies that more infrastructure (electricity, hospitals, etc) is available, and that accessing these public services is much easier.

In this dataset, the cross-information is available (the data-set provides, for each country, the average life expectancy and the urbanisation percentage). This cross information will be used for performance assessment only, and of course, not by our method. We do not provide our algorithm with the dependence structure $(\mathbf{X}, \mathbf{Y})$, but we provide it with independent subsamples of size 30 of $\mathbf{X}$ and $\mathbf{Y}$. We plot the estimate $\hat{h}$ obtained with our method using different deconvolution distributions in the deconvolution process (the “true” noise distribution ξ is not available), as well as the points (X_i, Y_i) for illustrating how the estimator of the link function captures the dependence structure from the monotonicity constraint. The results are plotted in Figure 1, each curve corresponds to an estimate $\hat{h}$ for which the deconvolution is made using a different deconvolution distribution. We provide results for normal deconvolution distribution with various variances (left plot), and uniform deconvolution distribution with various ranges (right plot). As expected, the smaller the variance of the deconvolution distribution, the closer the estimator $\hat{h}$ is to the points from which it is constructed. However, as in standard regression, this comes together with the problem of over-fitting, as illustrated in Table 1 where the risk (the square root of the MSE) is evaluated on an independent sample of 20 countries. This evaluation shows clearly that taking a deconvolution distribution of small variance does not provide an estimator that has good generalization properties, and it is therefore better to take this into account instead of doing a simple quantile matching.

From the curves in Figure 1, it seems that urbanisation has a multiplier effect on the life expectancy: indeed, the function h is convex until a threshold. However, there is clearly endogeneity in this model. In particular, the richer the country, the more urbanised it is (in Africa). And the richer the country, the higher the life expectancy. If we want to measure the true effect of urbanisation on life expectancy, we should get rid of this side effect. We should thus control for this effect using a control variate Z =“GDP per head”, as explained in Subsection A.1. We assume here that there is a linear underlying model, and that we have

$$X_i = \tilde{X}_i + \alpha_1 + \beta_1 Z_i \quad \text{and} \quad Y_i = \tilde{Y}_i + \alpha_2 + \beta_2 Z_i.$$

We estimate the $\tilde{X}_i$ and the $\tilde{Y}_i$ using the control Z and doing a linear regression (as in Algorithm 2). As explained before, urbanisation, even after the control should have a positive effect on the controlled life expectancy, because urbanisation enlarges the access to important public services. The results in this controlled setting are plotted in Figure 2. The points are the estimates of the controlled $(\tilde{X}_i, \tilde{Y}_i)$, and each curve corresponds to a deconvolution with a different noise deconvolution. We observe here that the curve $\hat{h}$ is now concave - controlling by the GDP per head has cancelled the multiplier effect, although the impact of urbanisation on life expectancy is still positive. It means that some urbanisation has a very positive effect on life expectancy (because it implies better access to vital infrastructure), but that this positive effect is sub linear (the multiplier effect, coming from the fact that the GDP is positively correlated with urbanisation and life expectancy, is suppressed). The MSE

(again, evaluated on an independent sample) is displayed in Table 2

4.2 Experiment on real data II : Property prices and percentage of educated residents

In the second example we look at the impact of the the neighborhood share of high skilled residents on local property prices. We use data from the 2011 UK census on the share of residents holding a university degree in electoral wards and prices of 2011 housing transactions available for download at the land registry website[1]. We consider the two variables $\mathbf{Y}$ ="average price" and $\mathbf{X}$ ="percentage of high skilled residents". House prices in an area are clearly increasing in the local concentration of well educated workers as workers holding university degrees are paid the highest wages and subsequently spend more on housing. We are interested in reconstructing the map of house prices in London, using the percentages of degrees (with geographical co-variate), and the house prices (but without using the geographical information) : we want to merge the percentages of degrees (plus geographical location) with the house prices. We believe that this example is interesting, because unlike the UK, most countries do not provide refined geographical data for house transaction, whereas geographical census data are usually available in developed countries.

Figure 3 a) shows the distribution of average house prices in 2011 for 733 London wards. House prices are highest in the west of Inner London and in the southwest of Outer London. With our method we try to reconstruct the local price pattern on the basis of the degree variable without making use of the geographical information available for the house prices. We divide at random our sample in two datasets, one that will serve the purpose of constructing the estimate and the other the one of evaluating it through the MSE (300 for construction of the estimates and 433 for evaluating their performances). Figure 3 c) and d) show reconstructions of the original price map using respectively the distributions $U([-0.5, 0.5])$, and $U([-2.5, 2.5])$ for the deconvolution (the "true" noise distribution of the noise). The choice of a uniform deconvolution distribution ξ is reasonable in this case (since the variables are bounded and rescaled between 0 and 100), and we considered two ranges for ξ (1 and 5). As a comparison method we also reconstruct the map based on quantile matching, which is equivalent to using a Dirac mass in 0 in the deconvolution step (Figure 3 b)). All three methods are able to generate a price pattern very similar to the original spatial distributing shown in a). The qualitative difference is that the larger the noise with which one deconvolves, the more contrasted the picture gets - removing the smoothing that is due to the noise - and d) is the most contrasted picture.

Figure 4 shows the estimate $\hat{h}$ obtained with our method using the different deconvolution distributions, plotted with the second part of the sample for evaluation. We evaluate the performance under the different deconvolution distributions by calculating the mean squared error (MSE), using the distance of each point to the function $\hat{h}$. Quantile matching results in a MSE of **8.62**, whereas the uniform distributions yield an MSE of **8.54** for $\xi = U([-0.5, 0.5])$, respectively **8.23** for $\xi = U([-2.5, 2.5])$. Our method outperforms a simple quantile match, when using a reasonable deconvolution distribution (here uniform, since the noise is bounded).

Remark regarding the deconvolution noise in the empirical results Empirically, if the noise distribution is unknown, we found out that in all studied cases, deconvolving with some "reasonable" distribution - even if it is not the correct one - outperforms quantile matching. The "reasonable" distribution does not need to be the true noise distribution, which is unknown, but it can be a distribution deduced from some prior knowledge on the noise. This knowledge does not need to be precise. In the examples of Subsection 4.1, deconvolving with a distribution of very small variance is giving bad results with respect to the MSE while bigger noises are performing significantly better, see Table 1 and 2 (improvement ranges between 30% to more than 50% in Table 2). We did not show results for pure quantile matching because they are slightly worse than deconvolution with $\mathcal{N}(0, 0.1)$. Also, deconvolution with the uniform distribution or the Gaussian distribution of variance ranging from 1 to 2×5^2 provides comparable results - although these distributions are significantly different and largely span the set of "reasonable" distributions. Similar results can be seen in the MSE in Figure 4 for the second example and in the Appendix on synthetic data. This highlights the fact that one does not need to have a precise knowledge of the noise distribution to gain something significant with respect to quantile matching.

Conclusion

We developed in this paper a new method for merging variables. It can be used as a complement to matching - or also on its own when no contextual variables are available. It is easy to implement and provides good results, both in theory and in practice, provided that the dependence function h is monotone.

Acknowledgements AC's work is supported since 2015 by the DFG's Emmy Noether grant MuSyAD (CA 1488/1-1).

[1] http://www.landregistry.gov.uk/market-trend-data/public-data/transaction-data

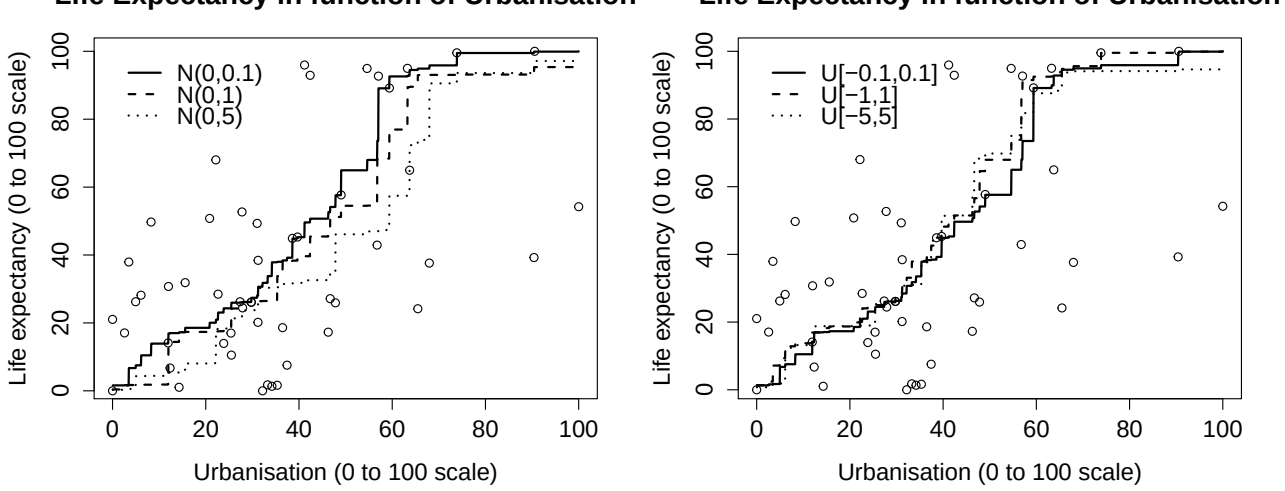

Figure 1: Life expectancy in function of urbanisation percentage (both rescaled on $[0, 100]$) in 45 African countries. Each point corresponds to a country. The lines correspond to $\hat{h}$ estimated using different deconvolution distributions i.e. $\mathcal{N}(0, 0.1)$, $\mathcal{N}(0, 1)$ and $U([-5, 5])$.

Table 1: Risk of the estimates obtained in Figure 1.

	Normal distribution			Uniform distribution		
Distribution	$\mathcal{N}(0, 0.1)$	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 5)$	$U([-0.1, 0.1])$	$U([-1, 1])$	$U([-5, 5])$
MSE	40.4	37.5	38.2	39.4	31.2	33.4

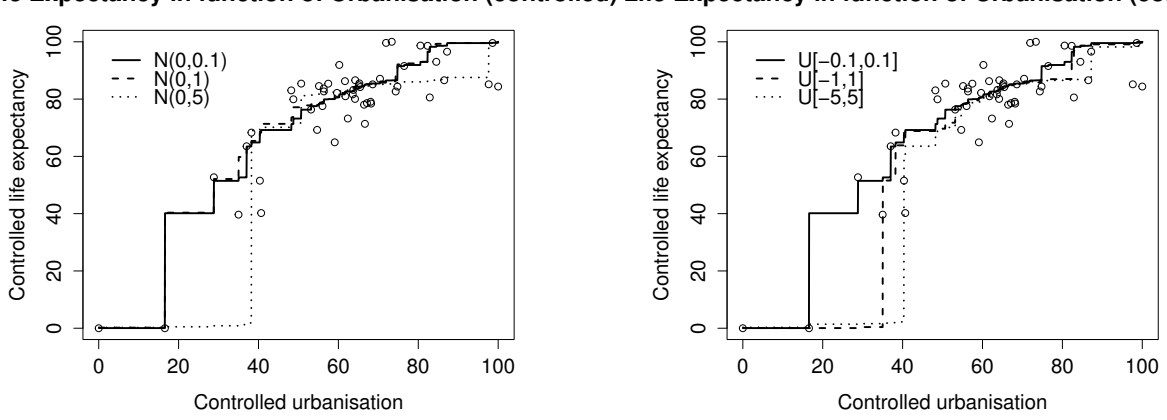

Figure 2: Controlled life expectancy in function of controlled urbanisation percentage (both rescaled on $[0, 100]$) in 45 African countries. Each point corresponds to a country. The lines correspond to $\hat{h}$ estimated using different deconvolution distributions i.e. $\mathcal{N}(0, 0.1)$, $\mathcal{N}(0, 1)$ and $U([-0.5, 0.5])$.

Table 2: Risk of the estimates obtained in Figure 2.

	Normal distribution			Uniform distribution		
Distribution	$\mathcal{N}(0, 0.1)$	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 5)$	$U([-0.1, 0.1])$	$U([-1, 1])$	$U([-5, 5])$
MSE	17.4	7.9	8.6	15.5	7.3	8.2

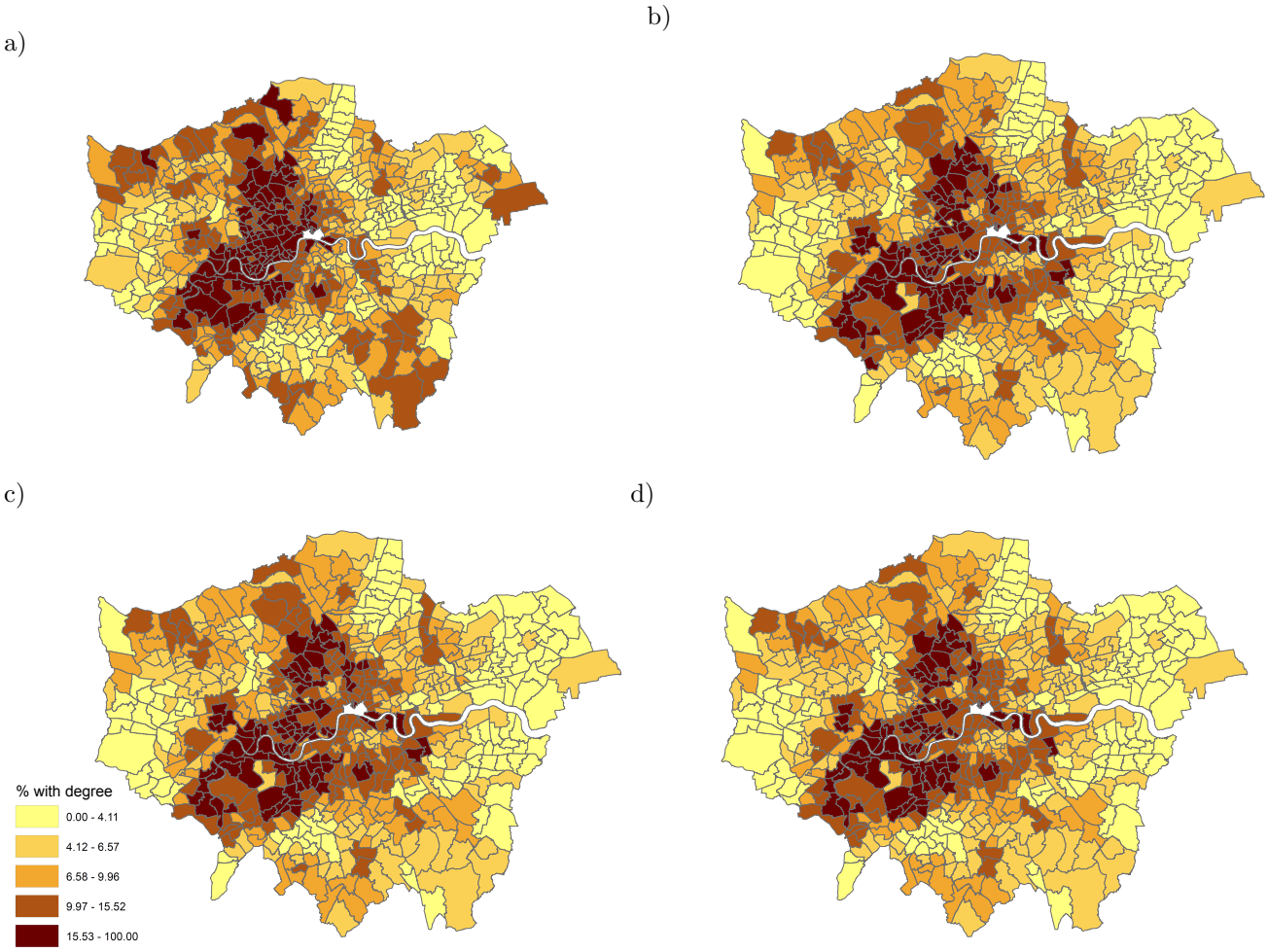

Figure 3: Reconstruction of the house price map using the percentage of residents holding a degree in each ward. Maps from Up to Down and Left to Right: (UL) True map of average house prices in 733 London wards, (UR) Reconstruction of the map based on quantile matching $\xi = 0$, (DL) Reconstruction of the map based on our method with $\xi = U([-0.5, 0.5])$ and (DR) Reconstruction of the map based on our method with $\xi = U([-2.5, 2.5])$

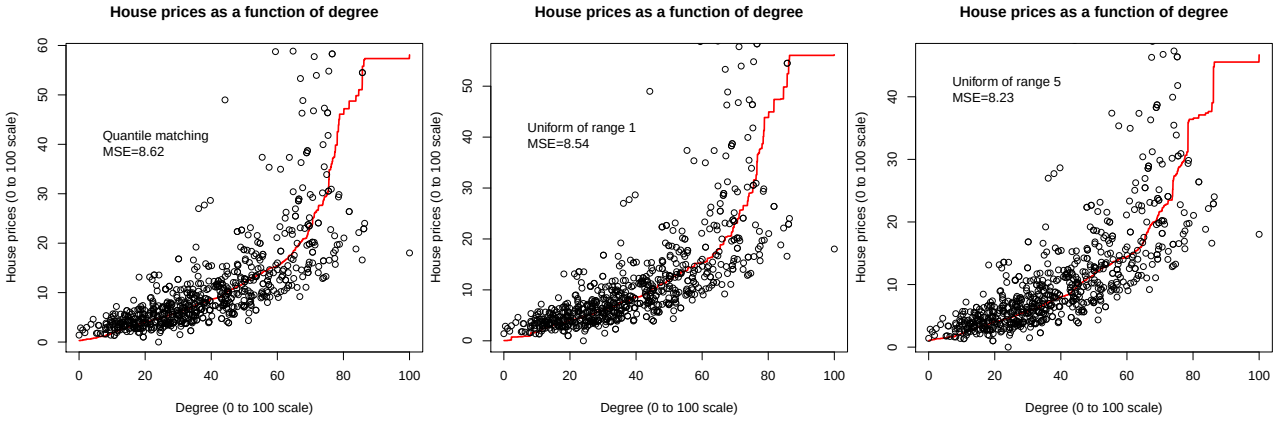

Figure 4: House prices in a geographical unit in function of degree percentage in the same geographical unit (both rescaled on $[0, 100]$) in 733 London wards. Each point corresponds to a ward. The red lines correspond to $\hat{h}$ estimated using different distributions ξ. From left to right: quantile matching ($\xi = 0$), $U([-0.5, 0.5])$ and $U([-2.5, 2.5])$. The MSE (mean squared error) is the mean squared error between the line and the points.

References

[1] Sören Auer, Jens Lehmann, Axel-Cyrille Ngonga Ngomo, and Amrapali Zaveri. Introduction to linked data and its lifecycle on the web. In *Reasoning Web. Semantic Technologies for Intelligent Data Access*, pages 1–90. Springer, 2013.

[2] Anthony J Bell and Terrence J Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995.

[3] Jens Bleiholder and Felix Naumann. Data fusion. *ACM Computing Surveys (CSUR)*, 41(1):1, 2008.

[4] Raymond J Carroll and Peter Hall. Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association*, 83(404):1184–1186, 1988.

[5] Tony F Chan and Chiu-Kwong Wong. Total variation blind deconvolution. *Image Processing, IEEE Transactions on*, 7(3):370–375, 1998.

[6] William W Cohen and Jacob Richman. Learning to match and cluster large high-dimensional data sets for data integration. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 475–480. ACM, 2002.

[7] Clifford B Cordy and David R Thomas. Deconvolution of a distribution function. *Journal of the American Statistical Association*, 92(440):1459–1465, 1997.

[8] Itai Dattner, A Goldenshluger, A Juditsky, et al. On deconvolution of distribution functions. *The Annals of Statistics*, 39(5):2477–2501, 2011.

[9] Kjell A Doksum and Gerald L Sievers. Plotting with confidence: Graphical comparisons of two populations. *Biometrika*, 63(3):421–434, 1976.

[10] Xin Luna Dong and Felix Naumann. Data fusion: resolving data conflicts for integration. *Proceedings of the VLDB Endowment*, 2(2):1654–1655, 2009.

[11] Xin Luna Dong and Divesh Srivastava. Big data integration. In *Data Engineering (ICDE), 2013 IEEE 29th International Conference on*, pages 1245–1248. IEEE, 2013.

[12] John HJ Einmahl, Ian W McKeague, et al. Confidence tubes for multiple quantile plots via empirical likelihood. *The Annals of Statistics*, 27(4):1348–1367, 1999.

[13] Jianqing Fan et al. On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, 19(3):1257–1272, 1991.

[14] F Lombard. Nonparametric confidence bands for a quantile comparison function. *Technometrics*, 47(3), 2005.

[15] Alvaro E Monge, Charles Elkan, et al. The field matching problem: Algorithms and applications. In *KDD*, pages 267–270, 1996.

[16] Eric Moulines, J Cardoso, and Elisabeth Gassiat. Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 5, pages 3617–3620. IEEE, 1997.

[17] Jean-Luc Starck, Mai K Nguyen, and Fionn Murtagh. Wavelets and curvelets for image deconvolution: a combined approach. *Signal Processing*, 83(10):2279–2283, 2003.

[18] Volker Walter and Dieter Fritsch. Matching spatial data sets: a statistical approach. *International Journal of Geographical Information Science*, 13(5):445–473, 1999.

[19] Rand R Wilcox and David M Erceg-Hurn. Comparing two dependent groups via quantiles. *Journal of Applied Statistics*, 39(12):2655–2664, 2012.