Édith DARIN
Leïla FARDEAU
Loïc MACHEREL

ENSAE Third Year
First Semester
2017-2018

The Method of Auxiliary-to-Study Tilting:
A tale about Semiparametric Data Combination Problems

Review, Comments, Criticisms and Application of
"Efficient Estimation of Data Combination Models
by the Method of Auxiliary-to-Study Tilting (AST)"
by Bryan. S. Graham, Cristine Campos de Xavier Pinto, Daniel Egel

Course:
Semi and Non-Parametric Econometrics

Professors:
Xavier d'Haultfoeille
Laurent Davezies

Contents

Introduction

Broad context: State of the Art With the increasing use of machine learning techniques, recent research in econometrics have applied more and more of these methods. Thus, research on a number of semiparametric and nonparametric problems have seen significant improvements, among which semiparametric missing data and data combination models through a renewal of techniques based on propensity scores. In particular, propensity score weighting aims at constructing weights that directly balance covariates or functions of covariates between two populations so that reweighted data will mimic some characteristics of a randomized experiment: reweighting allows obtaining similar distribution of observed baseline covariates between treated and untreated subjects.

One of these approaches, developed by Graham, Pinto and Egel in a 2016 article [Graham et al., 2016], deals with semiparametric data combination models. It aims at developping a locally efficient estimator, for some semiparametric data combination issues. One of it's major improvements is that it has a double robustness property (i.e. it remains consistent even if the propensity score is misspecified), as we will see.

Tilting parameters: From Inverse Probability Tilting to Auxiliary-to-Study Tilting (AST) This method is an extension of a novel reweighting procedure for moment condition models with missing data developed by the same authors in a 2012 article [Graham et al., 2012], and named inverse probability tilting (IPT). IPT is an inverse probability weighting procedure in which the propensity score is not estimated through the conditional maximum likelihood estimate (CMLE) but with a method of moments. Inverse probability weighting is broadly a weighting mentod that relies on computing weights based on some variables that are "good" predictors of selection in order to reweight observations in a second step.

It is a semiparametric missing data model with data missing at random. The first assumption is that it must be identified, among other key assumptions, authors define a parametric propensity score model in order to derive their results on efficiency : indeed they find that the estimator is locally efficent (unlike the standard inverse probability weights with propensity scores estimated by CMLE, in most cases). But, they also do study the properties of IPT when the assumption on the propensity score model fails: in fact, one of the IPT's most interesting properties is it's double robustness. Note that it is also verified in the case of the model of Auxiliary Tilting. This property broadly corresponds to the fact that on the one hand, if properly specified, the AST estimator will have low samplng variation ; but, on the other hand, even when it is misspecified, it will remain consistent under some conditions.

1 A synthetic review of the article

1.1 Understanding the theory

1.1.1 Explaining the model

Inverse Probability Weighting Methods applied to Data combination Data combination models are econometric models that aim at recovering (or artificially constructing) information on the *study population* of interest on a *study sample*, by using information from a sample drawn from an auxiliary population, i.e. the *auxiliary sample*. Of course both study and auxiliary samples are random samples.

More formally, let $Z = (W', X', Y')'$ denote a random vector drawn from the study population with distribution function F_s. The study sample, with N_s observations only contains measurements of (Y, W). The combination of the study and the auxiliary population is referred to as the *merged population*. On the other hand the auxiliary sample contains information on (X, W), and has N_a observations and has distribution function F_a. Let $E_s[.]$ (respectively $E_a[.]$) denote expectation with respect to the study (auxiliary) population. For a known function $\psi(z, \gamma)$:

$$\exists!\ \gamma_0 \mid E_s[\psi(Z, \gamma_0)] = 0 \tag{1}$$

One of the key features of these models is that the distribution of the variables of interest must be independent of the sample they are in, conditional on a set of proxy variables that are used to estimate the propensity score. Note that these variables must be observed in both samples.

In this context, the propensity score can be defined as the probability that one observation belongs to the study sample, conditional on the proxy variables. In fact that is how Chen, Hong and Tarozzi (2008) define the one they use in order to construct an Augmented Inverse Probability Weighting for auxiliary data.

Semiparametric data combination model In this case the data combination model is defined as such (Assumption 1) :

i. (Identification)

For some $\psi(z, \gamma_0) = \psi_s(Y, W, \gamma_0) - \psi_a(X, W, \gamma_0)$ *so that equation (1) holds with:*

$$E_s[\psi(z, \gamma)] \neq 0\ ,\ \forall \gamma \neq \gamma_0, \gamma \in \mathcal{G}, z \in \mathcal{Z}$$

This assumption implies that the model is globally identifiable. If we consider the merged sample, it can be translated as $E[\psi(Z, \gamma_0)|D] = 0$

ii. (Conditional distributional equality)

$$F_s(x|w) = F_a(y|w)\ \forall\ w \in \mathcal{W},\ x \in \mathcal{X} \text{ and } y \in \mathcal{Y}$$

If we consider the merged population, this amounts to assuming conditional independance on the distribution function:

$$F(y|w, d = 1) = F(y|w, d = 0) \text{ and } F(x|w, d = 1) = F(x|w, d = 0)$$

iii. (Weak overlap)

$$S_s \subset S_a \text{ with } S_j = \{w \mid f_j(w) > 0\}$$

iv. (Multinomial Sampling)

With probability $Q_0 \in (\xi, 1 - \xi)$ *with* $0 < \xi < 1$*, we draw a unit at random from* F_s *and record its realizations*

of Y and W, otherwise we draw a unit from F_a and record its realizations of X and W. Let D_i denote the indicator of "the i-th draw corresponds to a study population unit" ($i \in \{1, ..., N\}$).

Thus, the sampling distribution induced by this scheme has density :

$$f(z,d)Q_0^d(1-Q_0)^{1-d}f_s(z)^d f_a(z)^{1-d} \tag{2}$$

This assumption allows us to consider the merged sample $\{()D_i, W_i, (1-D_i)X_i', D_iY_i')'\}_{i=1}^N$as a random sample drawn from a *merged population*. Note that this is a pesudo-population as in many cases does not correspond to a real population.

v. **(Propensity Score Model)**
There is a unique $\delta_0 \in \mathcal{D}$, $r(W)$ a known vector of linearly independent functions of W with a constant in the first row, and known function of $G(.)$ such that :

(a) $G(.)$ is strictly increasing, differentiable and maps into the unit interval with $\lim_{v\to+\infty} G(v) = 0$ and $\lim_{v\to-\infty} G(v) = 1$

(b) $\frac{f_s(w)}{f_a(w)} = \frac{1-Q_0}{Q_0}\frac{G(r(w)'\delta_0)}{1-G(r(w)'\delta_0)}$ $\forall w \in \mathcal{W}$

Therefore, we write $p_0 = G(r(w)'\delta_0)$ the propensity score. Assumptions (iii) and (iv) imply that it is bounded away from one. In the merged sample, we also have : $p_0(w) = E[D|W = w]$.

It is important to note that the Propensity Score Model defined by the last assumption is parametric.

Equation (2) immediately implies that: $f(z|d = 1) = f_s(z)$ and $f(z|d = 0) = f_a(z)$. By Bayes' Law, he therefore have :

$$f_s(w) = f(w|d=1) = \frac{p_0(w)f(w)}{Q_0} \text{ and } f_a(w) = f(w|d=0) = \frac{(1-p_0(w))f(w)}{1-Q_0}$$

Which implies :

$$f_s(w) = f_a(w)\left[\frac{1-Q_0}{Q_0}\frac{p_0(w)}{(1-p_0(w))}\right] \tag{3}$$

The last equation along with assumptions (ii) and (iii) allows identification of γ_0 :

$$E_s[\psi(Z,\gamma)] = E\left[\frac{D}{Q_0}\psi_s(Y,W,\gamma)\right] - E\left[\frac{1-D}{Q_0}\frac{p_0(W)}{1-p_0(W)}\psi_a(X,W,\gamma)\right] \tag{4}$$

The Auxiliary-to-Study Tilting procedure The AST estimator is an estimator of γ_0 based on the method of moments.

Let's assume that the merged sample is arranged such that it's first N_s observations correspond to the study sample and the other N_a to the auxiliary sample. The AST estimator is computed in three steps.

1. The propensity score parameter δ is estimated by conditional maximum likelihood. The optimim is therefore given by the following expression :

$$\frac{1}{N}\sum_{i=1}^{N}\frac{D_i - G(r(W_i)'\hat{\delta}_{ML})}{G(r(W_i)'\hat{\delta}_{ML})[1-G(r(W_i)'\hat{\delta}_{ML}]} \times \frac{\partial G(r(W_i)'\hat{\delta}_{ML})}{\partial(r(W_i)'\hat{\delta}_{ML})} \times r(W_i) = 0 \tag{5}$$

2. Then, we compute a reweighting of both study and auxiliary samples. We thus estimate our *tilting parameters*, denoted by λ_a and λ_s. Let $t(W)$ denote a vector of known linearly independent functions of W

with a constant 1 in it's first row. δ_0 has been estimated by conditional maximum likelihood, giving $\hat{\delta}_{ML}$ and Q_0 is estimated through $\hat{Q}_{ML}$ corresponding to the one we get setting $\delta_0 = \hat{\delta}_{ML}$. We now estimate the tilt of the auxiliary sample by solving:

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1-D_i}{1-G(r(W_i)'\hat{\delta}_{ML})+t(W_i)'\hat{\lambda}_a}-1\right)\times\frac{G(r(W_i)'\hat{\delta}_{ML})}{\hat{Q}_{ML}}t(W_i)=0 \tag{6}$$

Denoting :

$$\hat{F}_s^{\text{eff}}(w)=\sum_{i=1}^{N}\hat{\pi}^{\text{eff}}\mathbf{1}_{(W_i\leq w)}$$

$$\text{with } \hat{\pi}^{\text{eff}}=\frac{G(r(W_i)'\hat{\delta}_{ML})}{\sum_{i=1}^{N}G(r(W_i)'\hat{\delta}_{ML})}$$

Some rearrangements allow showing that the solution to (6) is chosen to form a reweighting so that $\sum_{i=1}^{N}\hat{\pi}^{\text{eff}}t(W_i)$ is numerically identical to the efficient estimate of $E_s[t(W_i)]$ based on $\hat{F}_s^{\text{eff}}(w)$, i.e. :

$$\sum_{i=N_s+1}^{N}\hat{\pi}_i^a t(W_i)=\sum_{i=1}^{N}\hat{\pi}^{\text{eff}}t(W_i)$$

With :

$$\hat{\pi}_i^a=\frac{G(r(W_i)'\hat{\delta}_{ML})}{\sum_{i=1}^{N}G(r(W_i)'\hat{\delta}_{ML})}\times\frac{1}{G(r(W_i)'\hat{\delta}_{ML}+t(W_i)'\hat{\lambda}_a)}\ ,\ \forall i=N_s+1,...,N$$

And we compute the tilt of the study sample, by solving :

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1-D_i}{1-G(r(W_i)'\hat{\delta}_{ML}+t(W_i)'\hat{\lambda}_s}-1\right)\times\frac{G(r(W_i)'\hat{\delta}_{ML})}{\hat{Q}_{ML}}t(W_i)=0 \tag{7}$$

so that :

$$\sum_{i=1}^{N_s}\hat{\pi}_i^s t(W_i)=\sum_{i=1}^{N}\hat{\pi}^{\text{eff}}t(W_i)$$

for :

$$\hat{\pi}_i^s=\frac{G(r(W_i)'\hat{\delta}_{ML})}{\sum_{i=1}^{N}G(r(W_i)'\hat{\delta}_{ML})}\times\frac{1}{G(r(W_i)'\hat{\delta}_{ML}+t(W_i)'\hat{\lambda}_s)}\ ,\ \forall i=1,...,N_s$$

3. Once we have these two tilts, we choose $\hat{\gamma}^{AST}$ to solve:

$$\sum_{i=1}^{N_s}\hat{\pi}_i^s\psi_s(Y_i,W_i,\hat{\gamma}_{AST})=\sum_{i=N_s+1}^{N}\hat{\pi}_i^a\psi_a(X_i,W_i,\hat{\gamma}_{AST}) \tag{8}$$

Note that the AST estimate of γ_0 is based on two distinct estimates of the distribution function of our study population :

- the study tilt: $\{\hat{\pi}_i^s\}_{i=1}^{N_s}$
- the auxiliary tilt: $\{\hat{\pi}_i^a\}_{i=N_s+1}^{N}$

1.1.2 Inputs: how the authors contribute to the research state

Our aim in this section is to show what the authors are bringing to the former state-of-the-art. Their inputs cover from change in the methodological frame by setting the analytical frame at the cross-road of different econometrical issues thus providing a solution to a wider range of challenges, to a new tool – the AST technique itself – and a new understanding of widely-used object: the propensity score.

The theoretical framework: data combination The aspect that first dragged our attention and that influenced our choice for discussing this article was the fact that the framework the authors are definying, encompasses a broad range of econometrical challenges. Indeed they mention potential applications from Average Treatment effect on the Treated estimand to poverty mapping and to counterfactual distributions. This really rich panorama is only made possible by the setting level they are considering: translating all previous mentionned challenges as data combination problems, i.e. a situation where you want to get information out of two samples that differ on their interest variable. In our opinion this is a clever synthetisis that describes data missing problem – the two samples are drawn out of the same population of interest which is the focus of your study –, policy evaluation issue – one sample (the control group) helps you to understand the focus sample (the treated) or even discrimination analysis, – the quantity of interest is the remaning discrepancy between your two samples – the one with the discriminated population and the one with the dominating- when you control out for background characteristics.

However despite how logical it seems to sum up all these issue into a framework of *data combination*, it has apparently not a dedicated research field, although Ridder and Moffitt were already in 2007 "[hoping], to stimulate further research on data combination" with their opus that "brought together research that until now was rather disjoint" in "Econometrics of data combination" [Ridder and Moffitt, 2007].

Purposely setting their articles in that reseach branch makes them gain a lot of explanation power. For example although the setting is very similar to the one of Chen et al. [Chen et al., 2008], Graham *et alii* twist their epistemological background by considering that the "verify-out-of-sample" case of the former is actually a definition of data combination problem which is a step towards building bridges between different fields of econometrics.

Improved properties of the estimator To have a fine understanding of the novelty the estimator they are proposing is bringing, we have mostly to compare it with previous ones and among them especially the one of [Chen et al., 2008] since it is the closest. Graham et al. demonstrate in this article that the semiparametric AST estimator is locally efficient, double robust and can be computed with a high-dimensional W.

First comment to make is that they are providing a fully-described estimator with a computed efficiency bound. This is a perfect way to introduce the performance of an estimator since the information bound will be used as a benchmark of sampling variation.

Secondly it is interesting to notice that they decided to depart from a working hypothesis of a non-parametrical form of the propensity score such as in [Chen et al., 2008], leading to a semi-parametric problem, in order to get the double robustness thanks to Assumption 2. Thus we can enhance the dilemma that they were facing: adopting a non-parametrical solution which is robust to missspecification but which is very computer-intensive and might fail in practice with an high dimensional W. Their solution that breaks this dilemma is then to accept a semiparametrisation of the estimator whith a logit form that, in this setting, leads to double robustness (robustness to propensity score missspecification and robustness to missspecification of conditionnal expectation functions).

A new way of conceptualizing the propensity score Let's recall what are the basics use and understanding of the propensity score [Rosenbaum and Rubin, 1983]. It was first defined as the *probability to be assigned to a treatment given a vector of observed covariates*. We can generalize out of the context of observationnal study and policy evaluation: it is the probability of being in one group given background characteristics. From that, standard methods are inverse probability weighting of the focus group, matching between the two group of individus having the same propensity score or stratification based on the score.

That said, we can infer another characteristic of the propensity score which to be a *balacing score*, which means that it used to make similar two samples based on these background covariates such that both sample are after reweighting balanced along these covariates.

In the AST setting the authors add another interpretation of the propensity score which we found striking. First we have to twist the definition of the auxiliary sample as a "biased sample of the study population" and then to see the propensity score as the knowledge about the biasing function which we know up to a finite-dimensional parameter. This version unleashes the full power of the propensity score which reveals the latent variable which based on the covariate has split in a non-random way the *pseudo* master generating population into the study and the auxiliary sample.

1.1.3 In-depth review: some clarifications

A "flexible parametric estimator" First we want to come back to the estimation itself and understand in which sense it is semiparametric or more exaclty as they have written "flexible parametric". As already stated, Graham et allii are departing from usual lessons about propensity score and the effect of missspecification [Drake, 1993] that motivates the use of non-parametric techniques (to machine learning ones [Lee et al., 2010] since they purposely declare the propensity score under a paramatric form such that a priori it seems that the extraction of information from the auxiliary tilt might be a little bit gross and not finely tuned.

This is just because they transfer the flexibility from the propensity score estimation step to the tilting parameter estimation step, in other words from $r(W)$ to $t(W)$. Let's recall that equations 6 and 7 can be seen as an optimization program under constraint, the constraint being an adequation of the weighted $t(W)$ in both sample, which is actually a predefined set of moment conditions in selected background covariates. Thus there is no need to specify the entire distribution of W.

We obtain therefore that the flexible control for differences in background characteristisc, in other words the semi-parametric side of the estimator, comes from the calibration of the study population distributions of X and Y with estimation of the distribution reweighted W sharing a finite number of moment with W in the master population.

Assumptions implication We wanted to review and understand the practical consequences of some crucial part of Assumption 1.

First of all the *conditional distributional equality* assumption. They define it as conditionned on the background variable, the two potential outcomes, X and Y would have behave the same. To begin, it has for consequence that W has to entail all covariates that describe the discrepancy between the two samples, or otherly put W needs to be exhaustive to balance the two samples. Moreover we can rephrase this assumption as *independance of outcome from belonging to one of the sample* given the recorded covariates (or unconfoundedness assumption in the literature devoted to econometrics of treatment assignment), to understand that actually this assumption is quite strong and it is crucial since without it, we could'nt treat the merged sample as a random sample from the joint distribution of X, Y and W and thus study the

inference. For example if we are in a setting with two independant samples that were surveyed at different time, we have to assume furthermore that X and Y are constant over time.

The *Weak overlap* hypothesis states that the auxiliary distribution of W is similar or at least envelops the study distribution of W. We can link this condition to the concept of *interpolation* (vs extrapolation). Indeed since we want to extract information from the auxiliary sample, it should not be too far away from the one we already have to build bridges. The consequence is that the propensity score is then bounded away from 0 and 1 since a propensity score of 1 means that the entities in the study sample do not share any characteristics with the one in the auxiliary sample such that the identification is fragile. And it derives from that $\Sigma_{RR}(\gamma_0)$, the error term, the degree of ineffeciency depends on $\frac{p_0(W)}{1-p_0(W)}$, greater when that takes extreme values i.e when overlap is poor. The notion of convex hull here is of prime importance and it's a good tool for the practician, and an easy routine to implement.

Difference between $r(W)$ and $t(W)$ We wanted to focus on these two objects since they are both a vector of known linearly independent functions of the same covariates W, put another way a synthetic representation of some background characteristics. Let's recall first that $r(W)$ appears for the computation of the propensity score, whereas $t(W)$ contains the moment constraints for exact balance of the two samples. Still one can wonder why we will choose $t(W) \neq r(W)$ since a priori we want a perfect balance for all the covariates (a part from computation cost point of view – to limit the constraint dimension –). Nevertheless in that case – and with the assumption 2 in mind – we can notice that if they coincide, then, only the auxiliary/control units will need to be re-weighted to impose exact balance. This is because the efficient estimate $\bar{t}_{\text{eff}} = \frac{\sum_{i=1}^{n} G(r(W_i)'\hat{\delta}_{ML})t(W_i)}{\sum_{i=1}^{n} G(r(W_i)'\hat{\delta}_{ML})}$ is numerically identical to the sub-sample mean of $t(W)$ across treated units alone (cf. G is the logit function).

1.1.4 Blurred areas and critics

We want to gather in this section points that remain for us either not fully convincing or quite opaque concerning equally the theory or its implementation.

Multinomial sampling and independancy Let's use the topology defined by Chen *et alii* [Chen et al., 2008] about the difference between missing-data context – the *verify-in-sample* case when both the study and the auxiliary sample are random ones from the population of interest – and data combination context – the *verify-out-sample* case, where auxiliary sample is collected independantly from the study sample, which is the population of interest. Graham *et alii* purposely set their article in the second case such that apparently the auxiliary population is supposed to be a random independant sample of the same pseudo *merged population* as the study one. This is perfectly fine when we're actually combining two datasets. But in the case where we are considering as study a split of a real dataset, the remaining part being the auxiliary (as in counterfactual analysis – c.f. the example of the article we will discuss later) are we allow to consider that the two samples are still independent? Indeed we might consider that the two parts of the dataset are independent from each other but only if the survey followed a complex survey design stratified on the variable we're splitting on. The importance of the independancy of the dataset remains still to be proven in the AST framework but the difference in the settings dragged our attention and we have been wondering if it was correct to bind them under the same model.

Data separation and Maximum Likelihood Estimation The AST method unleashes its full potential when considering the propensity score under a logit parametric form (for the double robustness) estimated with maximum likelihood fit (for the efficiency). But we might encounter complete separation in the dataset,

meaning that one background covariates of W (or a linear combination of some subset of W) when it was greater than some threshold is associated with only one outcome value of the D – the dummy that assigns individual i either to study sample or to auxiliary sample. In other words D_i has a loss in discrimination power for some characteristics. This kind of setting arise quite frequently in observationnal data because of the finite size of the data sample and because of social-related mechanisms (cf. in the Black/White example of Study/Auxiliary sample maybe there are only Whites above a certain income threshold)[1]. And the complete separation issue, as studied by Albert and Anderson [Albert and Anderson, 1984] leads to problem in estimating the maximum of likelihood in log-linear model. Therefore Maximum Likelihood Estimation might fail in practice.

Variable selection and Maximum Likelihood Estimation Here we want to briefly highlight the literature about variable selection for propensity score estimation, ie in W that would be relevant both for $r(W)$ and $t(W)$ in our settings. It's not completely obvious among academicians which variables should be included as control covariates. The one correlated with group assignment or the one correlated with outcome variable or both. We will refer here to the article of Brookhart *et alii* on variable selection [Brookhart et al., 2006]. Furthemore we might encounter in our datasets the curse of dimensionnality, not mentionned by Graham *et alii*, ie a large set of background covariates which is one case where it is interesting to have non-parametric estimation.

Assumption 2: grandeur and decadence The whole AST method is based on Assumption 2 that posits a working model for the conditional expectation function of ψ_a and ψ_s given W. It's crucial since it ensures the efficiency of the estimator but also the double robustness when the propensity score is missspecified (cf. Monte-Carlo example).

The problem remains that the form of this assumption (or equivalently the form of $\psi(X, Y, W)$) varies with the kind of issue studied – ATT, counterfactual analysis, etc.– such that a fine understanding of the methodology is required for each setting, without a routine check easily implementable.

Furthermore the way it is introduced in the paper cast some doubt about the transparency of this assumption: "[the model] attains the bound provided by Theorem 1 if Assumption 2 *happens to be true* in the population being sampled from, but *is not part of the prior restriction* used to calculate the bound." Prior restriction being all the assumption related to the setting of the data combination model. It was out of our reach to understand clearly what they mean by that since according to them "Our estimator is not efficient with respect to this augmented model" which is the one when Assumption 2 is added to the prior.

Eventually we will conclude on the fact that on one hand for sure the Auxiliary-To-Tilting methodolgy allows to a flexible control of difference in covariates but on the other hand it needs quite restrictive assumption that might hinder to the full range of application that are provided as potential example.

1.2 Comments on the applications

1.2.1 Application : the role of premarket differences in wage inequalities

In an illustrative empirical application, the authors extend a previous work od Neal and Johnson [Neal and Johnson, 1996] in which they aim at explaining the wages difference between Blacks and Whites in the United States, emphasizing the importance of individuals' cognitive achievement. In fact these disparities,

[1]Note that this issue is not completely tackled with a pre-analysis of the overlap assumption since overlap means auxiliary distribution covering study distribution but not necessarly the inverse.

that are quite large might reflect differences in social background that might have affected their human capital accumularion. To do so, the authors of this study used the National Longitudinal Survey of Youth 1979 (NLSY79) in order compute the least-square fit of the logarithm og hourly wages on a constant, the dummy of "the individual is black", age and an Armed Forces Qualification Test (AFQT) percentile score measured at age 16 to 18. They found that including AFQT score in the regression makes the coefficient of black dummy drop by 2/3.

Graham, Pinto and Egel explore the role of cognitive skills by comparing the differences between blacks and a hypothetical population of whites whose distribution of ages and AFQT scores coincides with the Black distribution by using the method of Auxiliary-to-Study Tilting.

Relevant populations and variables of interest In this case, the study population of interest is the employed Black men residing in the United States and aged between 16 and 18 in 1979. The auxiliary population is the White men of the same age in the US. The NLSY79 takes a random draw of these populations and the study sample is the blacks in this sample and the auxiliary sample, the whites of the same sample.

The Y vector contains real average wages averaged between 1990 and 1993 of the study sample and X contains average wages of the white men. The vector W contains both samples' year of birth and AFQT scores.

Authors first note that if the ages are, as expected, similar between the two samples, distribution of the AFQT differs significantly. Reweighting with the AST method will allow comparing the features of the observed distribution of black young men's wages with those of a hypothetical white population which has AFQT similar. scores and age ditribution

Choice of ψ In order to decompose the wage distributions into quantiles of the counterfactual distribution, we write the α-quantile of it which corresponds to q^α such that :

$$E_s[1(X \leq q^\alpha) - \alpha] = 0$$

Which amounts to setting : $\psi_s(Y, W, \gamma_0) = 0$ and $\psi_a(X, W, \gamma_0) = \alpha - \mathbf{1}_{(X \leq q^\alpha)}$

Moment Conditional Expectation Functions The vector $t(W)$ contains a constant, two year of birth dummies, a quadratic polynomial in AFQT scores and 12 dummies of "AFQT score lies below ζ" with $\zeta = -2, -1.75, ..., 0.25, 0.5$.

1.2.2 Limits

In this article, the authors argue that this method is a good one for solving data combination problems by reweighting the auxiliary sample when W is high dimensional. But, as we see it in this case, but also in all the applications of Auxiliary-to-Study tilting that have been done so far (at least to our knowledge), we always have 2 variables in W.

Moreover, we see that there is no check of Assumption 2. Yet, as we saw previously, some crucial results rely on it.

1.2.3 Assessing the relevance of double robustness and efficiency results: Monte Carlo experiments

In order to verify their main results on the AST estimator, the authors lead a number of Monte Carlo experiments. They simulate a stylized evaluation program on which the researcher wants to estimate the Average Treatment Effect and make the assumptions vary so that they can observe their effect on how the AST estimator performs.

We will not detail the experiments here, but we found worthwhile to report their main findings as the results they aim at assesing here are of utmost importance. However note that $G(.)$ is estimated with a logit and that $Y|W, D$ and $X|W, D$ are normally distributed.

$p_0(w)$	$q_\alpha(w)$	AST consistent	AST median unbiased
linear	linear	yes	yes
quadratic	linear	yes	yes
linear	quadratic	yes	yes
quadratic	quadratic	no	no

Moreover, doing these simulations allows comparing finite sample distributions to their asymptotic counterparts. We thus see that AST's asymptotic bias and it's median bias on the one hand and AST's standard deviation and it's asypmtotic standard error on the other hand are quite similar for each of these simulations.

1.3 Implementation

In this part, we will focus more on the algorithmic side of the problem thanks to the code provided in open source, explaining what techniques and peculiarities in coding were used in two different cases : the first one being the difference in premarket conditions between Black and White young men as described in the article, the second only mentioned as the evaluation of ATT in job training. Both of these implementations used propensity score under the logit form without previous incorporated step that check the assumption 2.

1.3.1 "Premarket" cognitive achievements differences role in earnings between Black and White men

Choice of $t(W)$ and $r(W)$ The algorithm, implemented in Matlab, uses as $t(W)$ a vector containing a constant, two dummies on year of birth (in 1963 and 1964), the transformed AFQT score (from percentile to a real) and its square, and dummies on the transformed AFQT score (under -1.75, -1.5,..., 0.5). Then the $r(W)$ is chosen as $t(W)$, meaning that the propensity score is computed with the same function of covariate as the constraints of the tilting optimization problem such that only the Black men sample will need to be reweighted.

Bootstraping Though not reported on the paper, the algorithmic part used bootstraping to try to evaluate confidence intervals in quantile differences (with both raw and adjusted data). Even though not directly present in the code, we assume the proportions of people under different bounds -their Table 1 - is computed this way, explaining why they give standard deviation along their estimate.

Each bootstraping step draw would be explained by the need to make sure that the tilting stays roughly the same without outliers or with more of them, which is linked to the convex hull checking we will discuss in the next implementation.

1.3.2 ATT evaluation in job training

This implementation was made using Python, and a step-by-step approach is described in their attached notebook.

Rescaling of variables Numerical variables are scaled down with a factor from 10 to 1000. It simplifies the information given without losing too much information, and thus the computation.

Computation of λ_s and λ_a In the logit case, a smarter computation is proposed in the additional material and implemented in the algorithm. We will not go into details as the supplementary material covers the proof, but one important thing to require is that it aims at computing a simpler (quadratic) function than the one needed for λ_s or λ_a, whose optimization is simpler.

Checking the convex hull As we mentioned earlier, checking the convex hull on the control variables is generally a good practice. Here, the convex hull is drawn on the earnings in 1974 and 1975 of the control data. As the treatment data average is very close to the edge of the hull, this means that a reweighing aiming at making the two averages (treatment and control) coincide will put a large weight on a reduced number of points in the control tilt, which might be hard to compute or even impossible

Regularizer To go around the previous problem, a "regularization" parameter (between 0 and 1, 1 meaning no regularization) is implemented in the code. To be short, it corresponds to a relaxing on the optimization condition in the computation of λ_s and λ_a, at the cost of longer computation, used when overlap is poor between the two populations.

Choice of $t(W)$ and $r(W)$ As in the previous implementation, $t(W) = r(W)$ at first. However, a the author then chooses to affect simply a constant to $r(W)$. It is possible because as the data is taken in a context of randomized controlled trial, the propensity score is constant because of the random assignment

2 Proposition of application: a study of attrition in remote data collection of food security indicators

2.1 Framework of the implementation

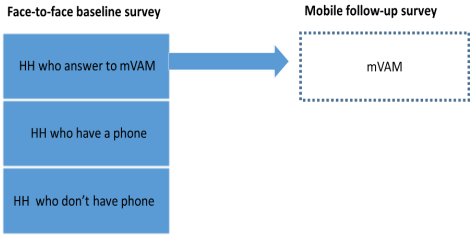
We offer to apply the Auxiliary-To-Tilting method to a data setting we encounter during one member's internship. Let's first describe the framework to understand how AST can be useful.

Challenge mobile Vulnerability Assessment and Mapping (mVAM) surveys are about collecting food security data by mobile phone for the UN World Food Programme (WFP). Compared to face-to-face survey, it's a faster, cheaper and more flexible data collection tool. However, concerns about representativeness of sample casts doubts on the validity of the conclusion derived from it.

Are mVAM respondents similar to the population for which we want to assess food security situation?

Selection bias Reaching the perfect sampling design which makes the sample fully representative of targeted population is an hard task in the context of remote survey because several bias specific to this data collection tool arise and above all: the selection bias. We can only reach people that have a phone working, open and whose battery is loaded, who know how to use it, who chose to share their phone numbers and finally whose network coverage is good enough to contact them. In the context where WFP works that leads to select into our sample more well-off people, leaving aside household potentially less food secure than the ones which own a phone. Consequently, we have some trouble finding out if the mVAM sample is similar to the targeted population.

Data Collection Framework One setting that allows us to understand better what kind of household are reached through mobile is to get phone numbers through a first traditional comprehensive face-to-face baseline survey representative of the targetd population.



Then we conduct a mobile follow-up on this phone sample pool. Some will answer, other not but at least thanks to the baseline survey we can know a wide range of social, demographical and economic characteristics of the non-respondents.

Making mVAM data representative Based on these covariates, we can predict the probability of participation to the follow-up survey, this probability being called the *propensity score*. Example: you are more likely to participate to mVAM – i.e. to answer the phone call – if you are a young male head of your household whose assets show that you are wealthy.

The second step as in traditional complex survey design, the idea is to weight our mVAM observations with inverse of the propensity score, in order to give more importance to people that were less likely to be reached.

Implementation of AST Our idea, which is to be honest a little twist of the main challenge in that setting made to apply AST, is to use the version of AST for counterfactual distribution. Indeed we want to know that if the mVAM people had the same background covariates they would have the same food security indicator as the non-respondents. In other words, even though we wouldn't be able to interpret as easily the difference between tilted distributions of respondents and non respondents as in the case of wage discrimination between Black and White people, reducing it to 0 would hint that there exist a reweighing procedure where previously collected data on mVAM would be "sufficient" to estimate their food security even without follow-up phone interviews.

Challenges We also use this implementation as a way to answer questions that were raised during our study of the paper :

- Is it possible to use categorical variables ? After studying the code, we concluded that even though no categorical data can be used as is, we can work around this problem by adding dummies on every modality of the data.
- Does the implementation of the AST holds with a high dimension of W ? We tend to answer with the affirmative with our implementation, even though simply adding more variables does not mean that the tilting will get better.

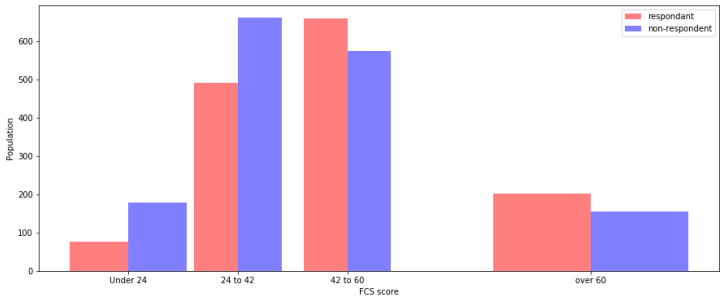
2.2 Results of the implementation

Presentation of the data The detailed implementation is present in the attached notebook, we are only going to resume here our procedure and results.

Our data was obtained through one of our members internship and concerns a sample of population of 1429 respondents and 1569 non respondents, for a total population of 2998.

Our covariate W data includes many variables concerning demography, education, wealth, education and mobile phone use. More details will be given along our different experiments.

Our X and Y in the data are the food security indicator distribution one for the respondents and the other for the non-respondents, which repartitions can be seen on the following histogram :



Bins are divided following the official definition:

- under 24 food security is considered "insufficient".
- between 24 and 42 is the "borderline" category.
- above 42 is an "acceptable" level.
- We further divide the "acceptable" population between under and above 60 to have more categories to compare.

We notice that the repartition of the respondents seem more positioned on the right on the histogram than the non-respondents. Finding a tilting that would correct our non-respondent repartition "to the right" of our histogram would be a good start: it would be the same idea as the original paper when they compare differences of repartition between Black and White population at different hourly wage thresholds.

First try: only a few covariates Our first tilting experiment is a modest attempt at simply trying to reduce the difference there is between respondents and non-respondents to belong to the categories listed above concerning food security score.

We choose $t(W)$ as a constant and linear vectors of the size of the household, the amount of money spent on food, dummy on whether the head of the household is a woman, and another on whether or not they are literate. We set $r(W) = t(W)$

We compute for each population what is the probability of a random draw to be under a certain FCS threshold, then tilt our auxiliary (non-respondents) distribution and do the same. We proceed by bootstrap (draw a sample with replacement from the original sample with same size as the latter) . Results are reported in the table below, tilted non respondents probabilities are the mean of these probabilities on bootstraped samples and empirical standard error between brackets.

	Respondent	Non-respondent	Tilted non-respondent
Pr(FCS_score < 24)	0.053184	0.113448	0.098644 (0.007786)
Pr(FCS_score < 42)	0.396781	0.534736	0.484393 (0.014964)
Pr(FCS_score < 60)	0.857943	0.900574	0.870136 (0.011118)

As we can see, even tough the tilt is not perfect, it already yields a distribution closer to what we hope to have.

On a side note, our tilting works fine with the default regularizer parameter, and as expected the results are roughly the same as with using a parameter of 1/2 (not reported here) - the tilt is not "difficult" to find, as the overlap is good, which means that relaxing the overlap hypothesis is not necessary.

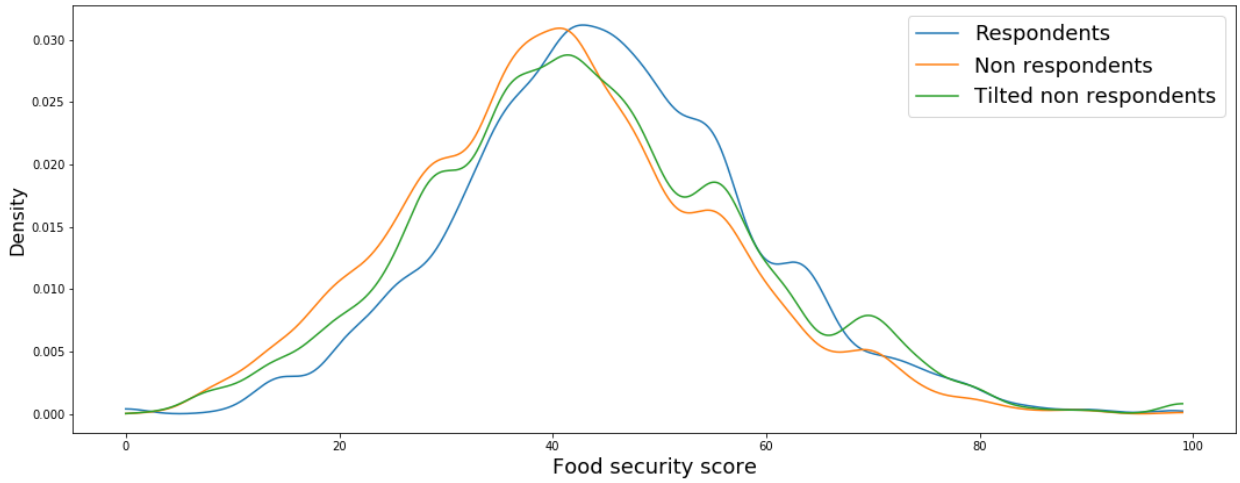
Second tilt: Adding more covariates On our second try, we want to test if the AST works fine with a higher-dimensional W, i.e adding more covariates. In addition to the previous tilt covariates, we add variables concerning age and gender composition of the household, education, marital and job situations, cattle ownership, access to drinking water and use of a cell phone. Categorical variables are converted to dummies variables on each modality as explained previously.

The first difficulty encountered is that we cannot use the default parameter on the regularizer anymore - no tilt is found and thus an error is returned on most bootstraped samples. Adopting the same regularizer parameter as the authors - i.e 1/2 - seems to work fine. It makes sense that overlap becomes poorer as we add more variables, even more so with many binary (the dummies) variables.

However, as we can see below, our tilting, if a bit better, is still not as close as the study distribution as we would wish.

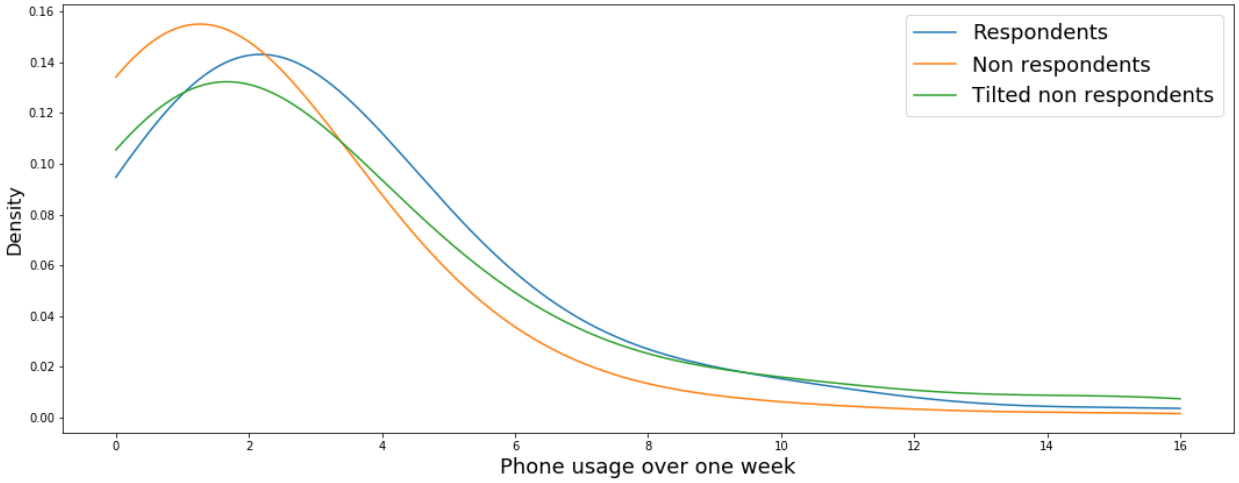
	Respondents	Non Respondents	Tilted non Respondents
Pr(FCS_score < 24)	0.053184	0.113448	0.083963 (0.007270)
Pr(FCS_score < 42)	0.396781	0.534736	0.466950 (0.016078)
Pr(FCS_score < 60)	0.857943	0.900574	0.857733 (0.011796)

Having a look at the estimated densities may help us understand more on how the tilting behaves:



Our first figure represents the densities of the distribution of the FCS score on the study and auxiliary population, as well as the tilted auxiliary distribution. Each of these is estimated through a Gaussian Kernel.

As expected, the tilted density is "between" the study and untilted auxiliary distribution most of the time. However, there are some values that seem to be "overreweighted", as the tilted density around 70 seems to show. This hints that in our case, an outlier in regards to the covariates lies in the auxiliary population with



a food security score around 70, but this outlier might not be one in the study distribution, explaining that is has a bigger reweighting.

We can also estimate the densities of one of the covariates for each sample, below on phone usage.

As expected, phone usage is more widespread among respondents, and the tilting tends to "correct" this in part.

Conclusion on the study We hope to have demonstrated that using and AST in our context is relevant, even though our dataset already had relatively close distributions between auxiliary and study sample.

As the tilting works to reduce the difference observed between study and auxiliary population, we can assume that there exist a reweighting procedure that would help us better evaluate the distribution of non-respondents, even though it might be necessary to gather more data on the initial face-to-face interview to find a perfect tilt. Our implementation is still really basic, and developing further our $t(W)$ and $r(W)$, for example by adding more interactions between the variables, could strengthen our model. We however ran into the same questioning as the author on how to choose such interactions, as there is yet no lead on how to make those decisions from the data.

Concerning the affirmation that the AST works well on a high dimensional W, the difficulties we encountered do not completely contradict this. We used many dummy variables, and overlap becomes mechanically poorer as the number of possible combinations grows, however simply gathering a bigger sample study could solve these difficulties.

References

A. Albert and J. Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10, 1984.

M. A. Brookhart, S. Schneeweiss, K. J. Rothman, R. J. Glynn, J. Avorn, and T. Stürmer. Variable selection for propensity score models. *American journal of epidemiology*, 163(12):1149–1156, 2006.

X. Chen, H. Hong, and A. Tarozzi. Semiparametric efficiency in gmm models of nonclassical measurement errors, missing data and treatment effects. 2008.

C. Drake. Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*, pages 1231–1236, 1993.

B. S. Graham, C. C. d. X. Pinto, and D. Egel. Inverse probability tilting for moment condition models with missing data. *Review of Economic Studies*, 79(3):1053–1079, 2012.

B. S. Graham, C. C. d. X. Pinto, and D. Egel. Efficient Estimation of Data Combination Models by the Method of Auxiliary-to-Study Tilting (AST). *Journal of Business and Economic Statistics*, 34(2):288–301, 2016.

B. K. Lee, J. Lessler, and E. A. Stuart. Improving propensity score weighting using machine learning. *Statistics in medicine*, 29(3):337–346, 2010.

D. A. Neal and W. R. Johnson. The Role of Premarket Factors in Black-White Wage Differences. *Journal of Political Economy*, 104(5):869–895, 1996.

G. Ridder and R. Moffitt. The econometrics of data combination. *Handbook of econometrics*, 6:5469–5547, 2007.

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.