8 Supplement

8.1 Proof of Lemma 3.1

Since the result is exactly symmetric when non-instrumented units are matched we prove it only for the case when instrumented units are matched. Assume $\mathbf{w} \in \mathbb{R}^p$. For a given unit i with $z_i = 1$, suppose we could find a $\boldsymbol{\theta}^{i*}$ as defined in the AME-IV problem. Let us define another unit k with $z_k = 0$, and $\mathbf{x}_k \circ \boldsymbol{\theta}^{i*} = \mathbf{x}_i \circ \boldsymbol{\theta}^{i*}$, by definition of $\mathtt{MG}(\boldsymbol{\theta}^{i*}, \mathbf{x}_i)$ it must be that $\mathbf{x}_k \in \mathtt{MG}(\boldsymbol{\theta}^{i*}, \mathbf{x}_i)$. So $\mathbf{1}_{[\mathbf{x}_i \neq \mathbf{x}_k]} = J - \boldsymbol{\theta}^{i*}$, where J is a vector of length p that has all entries equals to 1.

Assume there is another unit j with $z_j = 0$, and $j \neq k$.

If $j \in \mathtt{MG}(\boldsymbol{\theta}^{i*}, \mathbf{x}_i)$, then $\mathbf{1}_{[\mathbf{x}_i \neq \mathbf{x}_j]} = J - \boldsymbol{\theta}^{i*}$. So

$$\mathbf{w}^T \mathbf{1}_{[\mathbf{x}_i \neq \mathbf{x}_k]} = \mathbf{w}^T(J - \boldsymbol{\theta}^{i*}) = \mathbf{w}^T \mathbf{1}_{[\mathbf{x}_i \neq \mathbf{x}_j]}$$

If $j \notin \mathtt{MG}(\boldsymbol{\theta}^{i*}, \mathbf{x}_i)$, let us define $\boldsymbol{\theta}^{ij} = J - \mathbf{1}_{[\mathbf{x}_i \neq \mathbf{x}_j]}$, obviously $\boldsymbol{\theta}^{ij} \neq \boldsymbol{\theta}^{i*}$. Since $\boldsymbol{\theta}^{i*} \in \underset{\boldsymbol{\theta} \in \{0,1\}^p}{\operatorname{argmax}} \boldsymbol{\theta}^T \mathbf{w}$, we have:

$$\begin{aligned}
\mathbf{w}^T \mathbf{1}_{[\mathbf{x}_i \neq \mathbf{x}_k]} &= \mathbf{w}^T(J - \boldsymbol{\theta}^{i*}) \\
&= \mathbf{w}^T - \mathbf{w}^T \boldsymbol{\theta}^{i*} \\
&< \mathbf{w}^T - \mathbf{w}^T \boldsymbol{\theta}^{ij} \\
&= \mathbf{w}^T(J - \boldsymbol{\theta}^{ij}) \\
&= \mathbf{w}^T \mathbf{1}_{[\mathbf{x}_i \neq \mathbf{x}_j]}.
\end{aligned}$$

Therefore,

$$k \in \underset{\substack{j=1,\dots,n \\ Z_j = 0}}{\operatorname{argmin}} \mathbf{w}^T \mathbf{1}_{[\mathbf{x}_i \neq \mathbf{x}_j]}.$$

This concludes the proof.

8.2 Asymptotic Variance and Confidence Intervals for LATE Estimates

To construct estimators for the variance of $\hat{\lambda}$ we use an asymptotic approximation, that is, we will try to estimate the asymptotic variance of $\hat{\lambda}$, rather than its small sample variance. The strategy we use to do this is the same as Imbens and Rubin (2015), with the difference that our data is grouped: we adapt their estimators to grouped data using canonical methods for stratified sampling. In order to define asymptotic quantities for our estimators, we must marginally expand the definitions of potential outcomes introduced in our paper. In practice, while our framework has been presented under the assumption that the potential outcomes and treatments are fixed, we now relax that assumption and instead treat $y_i(1), y_i(0), t_i(1), t_i(0)$ as realizations of random variables $Y_i(1), Y_i(0), T_i(1), T_i(0)$, which are drawn from some unknown distribution $f(Y_i(1), Y_i(0), T_i(1), T_i(0))$. In this case the SUTVA assumption requires that each set of potential outcomes and treatments is independently drawn from the same distribution for all units. As usual, lowercase versions of the symbols above denote observed realizations of the respective random variables.

The asymptotic behaviour of our method is straightforward. Since the covariates we consider are discrete (say binary for convenience) there are only a finite number of possible covariate combinations one can observe. If the sample size n increases and the probability of observing all combinations of covariates is positive then asymptotically all possible combinations of covariates will be observed. In fact, most units will be matched exactly when $n \gg p$. This means that our matched groups will only contain exactly matched units, and therefore be exactly equal to a stratified fully randomized experiment in which the strata are the matched groups, by our Assumption 3 of our paper. By this principle, asymptotic results for IV estimation in stratified experiments, such as those in Imbens and Rubin (2015), apply asymptotically.

Recall as well that in this scenario we have a set of m matched groups $\mathtt{MG}_1, \dots \mathtt{MG}_m$ indexed by ℓ, such that each unit is only in one matched group. We denote the number of units in matched group ℓ that have $z_i = 1$ with n_ℓ^1 and the number of units in matched group ℓ with $z_i = 0$ with n_ℓ^0. Finally the total number of units in matched group ℓ is $n_\ell = n_\ell^0 + n_\ell^1$.

We make all the assumptions listed in Section3 but we must require a variant of (A3), to be used instead of it. This assumption is:
(A3') $\Pr(Z_i = 1 | i \in \mathtt{MG}_\ell) = \Pr(Z_k = 1 | k \in \mathtt{MG}_\ell) = \frac{n_\ell^1}{n_\ell}, \forall i, k.$
That is, if two units are in the same matched group, then they have the same probability of receiving the instrument. This probability will be equal to the ratio of instrument 1 units to all units in the matched group because we hold these quantities fixed. Note that this more stringent assumption holds when matches are made exactly, and is common in variance computation for matching estimators (see, for example, Kang et al. (2016)).

We keep our exposition concise and we do not give explicit definitions for our variance estimands. These are all standard and can be found in Imbens and Rubin (2015).

We have to start from estimating variances of observed potential outcomes and treatments within each matched group. We do so with the canonical approach:

$$\hat{s}^2_{\ell 0} = \frac{1}{n^0_\ell - 1} \sum_{i \in \mathtt{MG}_\ell} \left(y_i(1-z_i) - \frac{1}{n^0_\ell} \sum_{i \in \mathtt{MG}_\ell} y_i(1-z_i) \right)^2$$

$$\hat{s}^2_{\ell 1} = \frac{1}{n^1_\ell - 1} \sum_{i \in \mathtt{MG}_\ell} \left(y_i z_i - \frac{1}{n^1_\ell} \sum_{i \in \mathtt{MG}_\ell} y_i z_i \right)^2$$

$$\hat{r}^2_{\ell 0} = \frac{1}{n^0_\ell - 1} \sum_{i \in \mathtt{MG}_\ell} \left(t_i(1-z_i) - \frac{1}{n^0_\ell} \sum_{i \in \mathtt{MG}_\ell} t_i(1-z_i) \right)^2 = 0$$

$$\hat{r}^2_{\ell 1} = \frac{1}{n^1_\ell - 1} \sum_{i \in \mathtt{MG}_\ell} \left(t_i z_i - \frac{1}{n^1_\ell} \sum_{i \in \mathtt{MG}_\ell} t_i z_i \right)^2 ,$$

where: $\hat{s}^2_{\ell 0}$ is an estimator for the variance of potential responses for the units with instrument value 0 in matched group ℓ, $\hat{s}^2_{\ell 1}$ for the variance of potential responses for the units with instrument value 1 in matched group ℓ, $\hat{r}^2_{\ell 0}$ for the variance of potential treatments the units with instrument value 0 in matched group ℓ, and $\hat{r}^2_{\ell 1}$ is an estimator for the variance of potential treatments for the units with instrument value 1 in matched group ℓ. The fact that $\hat{r}^2_{\ell 0} = 0$ follows from Assumption A4.

We now move to variance estimation for the two ITTs. Conservatively biased estimators for these quantities are given in Imbens and Rubin (2015). These estimators are commonly used in practice and simple to compute, hence why they are often preferred to unbiased but more complex alternative. We repeat them below:

$$\widehat{Var}(\widehat{\mathtt{ITT}}_y) = \sum_{\ell=1}^{m} \left(\frac{n_\ell}{n}\right)^2 \left(\frac{\hat{s}^2_{\ell 1}}{n^1_\ell} + \frac{\hat{s}^2_{\ell 0}}{n^0_\ell} \right)$$

$$\widehat{Var}(\widehat{\mathtt{ITT}}_t) = \sum_{\ell=1}^{m} \left(\frac{n_\ell}{n}\right)^2 \frac{\hat{r}^2_{\ell 1}}{n^1_\ell} .$$

To estimate the asymptotic variance of $\hat{\lambda}$ we also need estimators for the covariance of the two ITTs both within each matched group, and in the whole sample. Starting with the former, we can use the standard sample covariance estimator for $Cov(\widehat{\mathtt{ITT}}_{y\ell}, \widehat{\mathtt{ITT}}_{t\ell})$:

$$\widehat{Cov}(\widehat{\mathtt{ITT}}_{y\ell}, \widehat{\mathtt{ITT}}_{t\ell}) = \frac{1}{n^1_\ell(n^1_\ell - 1)} \times \sum_{i \in \mathtt{MG}_\ell} \left(y_i z_i - \frac{1}{n^1_\ell} \sum_{i \in \mathtt{MG}_\ell} y_i z_i \right) \times \left(t_i z_i - \frac{1}{n^1_\ell} \sum_{i \in \mathtt{MG}_\ell} t_i z_i \right) .$$

The reasoning behind why we use only units with instrument value 1 to estimate this covariance is given in Imbens and Rubin (2015), and follows from A4. We can use standard techniques for covariance estimation in grouped data to combine the estimators above into an overall estimator for $Cov(\widehat{\mathtt{ITT}}_y, \widehat{\mathtt{ITT}}_t)$ as follows:

$$\widehat{Cov}(\widehat{\mathtt{ITT}}_y, \widehat{\mathtt{ITT}}_t) = \sum_{\ell=1}^{m} \left(\frac{n_\ell}{n}\right)^2 \widehat{Cov}(\widehat{\mathtt{ITT}}_{y\ell}, \widehat{\mathtt{ITT}}_{t\ell}).$$

Once all these estimators are defined, we can use them to get an estimate of the asymptotic variance of $\hat{\lambda}$. This quantity is obtained in Imbens and Rubin (2015) with an application of the delta method to convergence of the two ITTs. The final estimator for the asymptotic variance of $\hat{\lambda}$, which we denote by σ^2, is given by:

$$\hat{\sigma}^2 = \frac{1}{\widehat{\mathtt{ITT}}_t^2} \widehat{Var}(\widehat{\mathtt{ITT}}_y) + \frac{\widehat{\mathtt{ITT}}_y^2}{\widehat{\mathtt{ITT}}_t^4} \widehat{Var}(\widehat{\mathtt{ITT}}_t) - 2 \frac{\widehat{\mathtt{ITT}}_y}{\widehat{\mathtt{ITT}}_t^3} \widehat{Cov}(\widehat{\mathtt{ITT}}_y, \widehat{\mathtt{ITT}}_t).$$

Using this variance, $1 - \alpha\%$ asymptotic confidence intervals can be computed in the standard way.

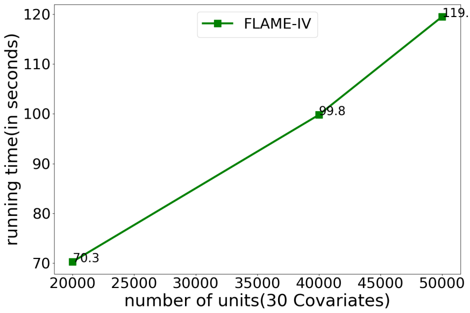

Figure 6: Running Time for FLAME-IV on large dataset.

8.3 The FLAME-IV Algorithm

We adapt the Algorithm described in Wang et al. (2019) to the IV setting. The algorithm is ran as described in that paper, except instrument indicator is used instead of the treatment indicator as input to the algorithm. Here we give a short summary of how the algorithm works and refer to Wang et al. (2019) for an in-depth description.

FLAME-IV takes as inputs a training dataset $D = \{(x_i, t_i, z_i, y_i)\}_{i=1}^n$, consisting of covariates, instrument indicator, treatment indicator, and outcome for every unit that we wish to match on, as well as a holdout set D^H consisting of the same variables for a different set of units that aren't used for matching but to evaluate prediction error and match quality. The algorithm then first checks if any units can be matched exactly with at least one unit with the opposite instrument indicator. If yes, all the units that match exactly are put into their own matched group and removed from the pool of units to be matched. After this initial check, the algorithm starts iterating through the matching covariates: at each iteration, Match Quality is evaluated on the holdout set after removing each covariate from the set of matching covariates. The covariate whose removal leads to the smallest reduction in MQ, is discarded and the algorithm proceeds to look for exact matches on all the remaining covariates. Units that can be matched exactly on the remaining covariates are put into matched groups, and removed from the set of units to be matched. Note that MQis recomputed after removing each remaining covariate at each iteration because the subset of covariates that it is evaluated on always is always smaller after each iteration (it does not include the covariate removed prior to this iteration). The algorithm will proceed in this way, removing covariates one by one, until either: a) MQ goes below a pre-defined threshold, b) all remaining units are matched, or c) all covariates are removed. Experimental evidence presented in Wang et al. (2019) suggests a threshold of 5% of the prediction error with all of the covariates. The matched groups produced by the algorithm can then be used with the estimators described in the paper to estimate desired treatment effects. Units left unmatched after the algorithm stops are not used for estimation. The algorithm ensures that at least one instrumented and one non-instrumented unit are present in each matched groups, but has no guarantees on treatment and control units: matched groups that do not contain either treated or control units are not used for estimation.

One of the strengths of FLAME-IV is that it can be implemented in several ways that guarantee performance on large datasets. An implementation leveraging bit vectors is described in Wang et al. (2019), optimizing speed when datasets are not too large. A native implementation of the algorithm on any database management system that uses SQL as a query language is also given in the same paper: this implementation is ideal for large relational databases as it does not require data to be exported from the database for matching.

While FLAME-IV is a greedy solution to the AME-IV problem, an optimal solution could be obtained by adapting the DAME (Dynamic Almost Matching Exactly) procedure described in Dieng et al. (2019) to the IV setting by using instrument indicators as treatment indicators in the input to the algorithm. Resulting matched groups with no treated or control units should be discarded as we do here. The same estimators we employ in this paper can also be employed with the same properties for matched groups constructed with this methodology.

8.4 More Running Time Results on Large Dataset

Figure 6 shows the results of running time for FLAME-IV on a larger dataset. The running time is still very short($<$ 2 min) on the large dataset for FLAME-IV. Full matching can not handle a dataset of this size.

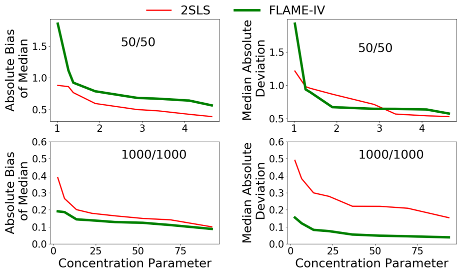

Figure 7: Performance for linear generation model with confounded instrument at various sample sizes. Here, 2SLS has an advantage because the data are generated according to a 2SLS model. FLAME-IV (either early-stopping or run-until-no-more-matches) performs similarly to 2SLS on large datasets, with smaller absolute bias of the median and median absolute deviation. On the smaller datasets, FLAME-IV has a slightly larger bias than 2SLS.

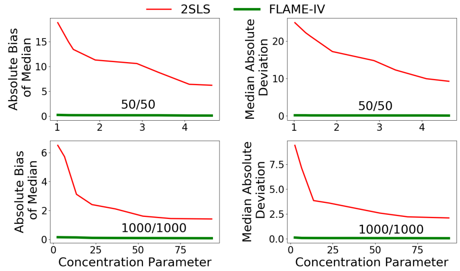

Figure 8: Performance for nonlinear generation model with confounded instrument assignment and different sample sizes. Here, the 2SLS model is misspecified. FLAME-IV (either early-stop or run-until-no-more-matches) outperforms 2SLS on both datasets, having smaller absolute bias of median and median absolute deviation.

8.5 Additional Simulations with Confounded Instrument Assignment

Here we present results from simulations similar to those in Section 5.1, but where, in addition to treatment assignment, instrument assignment is also confounded. Instrument is assigned as follows:

$$Z_i' = \rho X_i' \tag{13}$$

$$Z^* = \texttt{Median}\{Z_i'\}\forall i \tag{14}$$

$$Z_i = \mathbf{I}_{[Z_i' \geq Z^*]} \tag{15}$$

where X_i' contains last two variables for unit i, and $\rho \sim N(0.1, 0.01)$.

Results for a linear outcome model, same as in Equation (9) are displayed in Figure 7, and results for a nonlinear outcome model as in Eq: (10) are displayed in Figure 8. Results are largely similar to those obtained when instrument assignment is unconfounded. This suggests that our method performs equally well when instrument assignment is confounded.

8.6 Sample Matched Groups

Sample matched groups are given in Table 3. These groups were produced by FLAME-IVon the data from Pons (2018), introduced in Section 6. The algorithm was ran on all of the covariates collected in the original study except for territory. Here we report some selected covariates for the groups. The first group is comprised of electoral districts in which previous turnout was relatively good but PS vote share was low. This suggest that existing partisan splits are being taken into account by FLAME-IVfor matching. Municipalities in the second group have slightly lower turnout at the previous election but a much larger vote share for PS. Note also that treatment adoption is very high in the second group, while low in the first: this suggest that the instrument is weak in Group 1 and strong in Group 2.

Territory	Last Election PS Vote Share	Last Election Turnout	Population (in thousands)	Share Male	Share Unemployed	Treated	Instrumented
Matched Group 1							
Plouguenast et environs	(0.01, 0.05]	(0.77, 0.88]	(0, 450]	(0.47, 0.57]	(0, 0.1]	0	1
Lorrez-le-Bocage-Préaux et environs	(0.01, 0.05]	(0.77, 0.88]	(0, 450]	(0.47, 0.57]	(0, 0.1]	0	1
La Ferté-Macé et environs	(0.01, 0.05]	(0.77, 0.88]	(0, 450]	(0.47, 0.57]	(0, 0.1]	0	1
Mundolsheim et environs	(0.01, 0.05]	(0.77, 0.88]	(0, 450]	(0.47, 0.57]	(0, 0.1]	1	1
Paris, 7e arrondissement	(0.01, 0.05]	(0.77, 0.88]	(1,800, 2,250]	(0.47, 0.57]	(0.1, 0.2]	0	1
Sainte-Geneviève et environs	(0.01, 0.05]	(0.77, 0.88]	(0, 450]	(0.47, 0.57]	(0, 0.1]	0	0
Cranves-Sales et environs	(0.01, 0.05]	(0.77, 0.88]	(0, 450]	(0.47, 0.57]	(0, 0.1]	0	0
Hem et environs	(0.01, 0.05]	(0.77, 0.88]	(0, 450]	(0.47, 0.57]	(0, 0.1]	0	1
Legé et environs	(0.01, 0.05]	(0.77, 0.88]	(0, 450]	(0.47, 0.57]	(0, 0.1]	0	1
Moûtiers et environs	(0.01, 0.05]	(0.77, 0.88]	(0, 450]	(0.47, 0.57]	(0, 0.1]	0	0
Paris, 7e arrondissement	(0.01, 0.05]	(0.77, 0.88]	(1,800, 2,250]	(0.47, 0.57]	(0.1, 0.2]	0	1
Craponne-sur-Arzon et environs	(0.01, 0.05]	(0.77, 0.88]	(0, 450]	(0.47, 0.57]	(0, 0.1]	0	0
Matched Group 2							
Nantes	(0.19, 0.22]	(0.66, 0.77]	(0, 450]	(0.47, 0.57]	(0.1, 0.2]	1	1
Alès	(0.19, 0.22]	(0.66, 0.77]	(0, 450]	(0.37, 0.47]	(0.2, 0.3]	1	1
Sin-le-Noble	(0.19, 0.22]	(0.66, 0.77]	(0, 450]	(0.47, 0.57]	(0.2, 0.3]	1	1
Grand-Couronne et environs	(0.19, 0.22]	(0.66, 0.77]	(0, 450]	(0.47, 0.57]	(0.1, 0.2]	1	1
Dreux	(0.19, 0.22]	(0.66, 0.77]	(0, 450]	(0.47, 0.57]	(0.2, 0.3]	1	1
Vosges	(0.19, 0.22]	(0.77, 0.88]	(0, 450]	(0.47, 0.57]	(0.1, 0.2]	0	0
Arras et environs	(0.19, 0.22]	(0.66, 0.77]	(0, 450]	(0.37, 0.47]	(0.1, 0.2]	1	1
Montargis et environs	(0.19, 0.22]	(0.66, 0.77]	(0, 450]	(0.37, 0.47]	(0.2, 0.3]	1	1
Marseille, 3e arrondissement	(0.19, 0.22]	(0.66, 0.77]	(450, 900]	(0.47, 0.57]	(0.1, 0.2]	1	1
Nantes	(0.19, 0.22]	(0.66, 0.77]	(0, 450]	(0.47, 0.57]	(0.1, 0.2]	1	1
Mâcon et environs	(0.19, 0.22]	(0.66, 0.77]	(0, 450]	(0.37, 0.47]	(0.1, 0.2]	1	1

Table 3: Two sample matched groups generated by FLAME on the application data described in Section 6. The columns are a subset of the covariates used for matching. Territory was not used for matching. Original covariates are continuous and were coarsened into 5 bins. Last election PS vote share was coarsened into 10 bins. Labels in the cells represent lower and upper bounds of the covariate bin each unit belongs to. The two groups have relatively good match quality overall.