SEMIPARAMETRIC ANALYSIS OF RANDOMISED EXPERIMENTS USING L-MOMENTS

LUIS ANTONIO F. ALVAREZ AND CIRO BIDERMAN

Abstract. L-moments are linear combinations of order statistics that provide robust alternatives to standard moments. The estimation of parametric density models by matching sample L-moments is known to outperform maximum likelihood estimation in small samples from several distributions. Recently, it has been further shown that, by varying the number of L-moments with sample size and weighting these acoordingly, one is able to construct an estimator that outperforms MLE in small samples, and yet does not underperform asymptotically. Methods to automatically select the number of L-moments have also been developed. Given their good statistical properties and computational simplicity, it is expected that extending L-moment-based approaches to estimation to semi- and nonparametric settings may be able to produce computationally convenient estimators with good statistical properties. In this paper, we undertake this task by extending the L-moment approach to the estimation of semiparametric models of treatment effects in randomised trials. Recently, Athey et al. (2021) introduced semiparametric models as a convenient tool for the analysis of experiments with heavy-tailed data. We show that, in their setting, a "plug-in" L-moment estimator produces an efficient estimator without requiring further corrections. For flexible parametrisations of treatment effects, our estimator is also computationally acttrative, as it can be obtained by solving a quadratic program. We also discuss how to perform specification testing and moment selection. As an application, we apply our methods to a randomised experiment conducted by a large car-hailing service in São Paulo, which randomised discounts to some of its active users. The goal of the experiment was to understand whether short-run changes in the prices of car rides starting or ending at train stations could lead to long-run changes in the demand for public transportation due to learning effects. For the largest discounts randomised, we observe large effects on bimodal rides during the weeks the discount was in place and no effects thereafter. As for unimodal (car-only) rides, we observe a negative effect after the discount is over, which persists over more than a month. We introduce a simple learning model that is able to rationalise these results.

1. Introduction

Introduced by Hosking (1990), L-moments are linear combinations of order statistics that provide robust alternatives to standard moments. The estimation of parametric density models by

Alvarez: Doctoral candidate in Statistics, University of São Paulo.

Biderman: Professor, São Paulo School of Business Administration, FGV.

E-mail addresses: alvarez@ime.usp.br, ciro.biderman@fgv.br.

Key words and phrases. L-moments, semiparametric models, randomised controlled trials, learning effects.

*This article is a reworked version of Chapter 4 of Alvarez's PhD thesis.

matching sample L-moments – a procedure known in the literature as "method of L-moments" –, has been shown to outpeform maximum likelihood estimation in small samples from several distributions.[1] The number of L-moments is typically set equal to the dimension of the parameter space, though, which leads to inefficient estimators in larger samples. Recently, Alvarez (2022) has shown that, by varying the number of L-moments with the sample size and weighting these properly, one is able to construct a "generalised method of L-moments estimator" that is able to outperform maximum likelihood estimation in small samples, and yet preserve asymptotic efficiency. The latter fact is due to a convenient property of L-moments, first discussed in Hosking (1990): L-moments characterise distributions with finite first moments, a property not generally enjoyed by standard moments (Billingsley, 2012, Example 30.2). Alvarez (2022) then provides two methods to automatically select the number of L-moments used in estimation: one method based on higher-order expansions of the target-estimand, which are used to compute a high-order mean-squared error that is then minimised by choosing the appropriate number of L-moments; and another method based on ℓ_1-regularised estimation of the optimal moment weighting matrix. The author verifies the finite sample gains of both approaches in Monte Carlo simulations, and favours the second approach due to its computational simplicity.

Given the attractiveness of L-moments both from computational and statistical viewpoints, it appears fruitful to understand whether L-moment-based approaches are able to produce statistically efficient estimators in semi- and nonparametric contexts where semi/nonparametric maximum likelihood may be computationally prohibitive. In this paper, we undertake this approach in one such setting. Recently, Athey et al. (2021) introduced semiparametric models as a convenient tool for the analysis of randomised experiments. As argued by the authors, in several settings – especially in randomised experiments conducted by tech companies –, outcomes are heavy-tailed, and treatment effects, even if small, may be of economic interest. In these settings, a standard comparison of means tends to perform poorly, and sensible parametric restrictions on the treatment effect distribution, whilst still leaving the control group distribution unspecified, may entail large efficiency gains by reducing the model's efficiency bound. In light of these expected gains, Athey et al. (2021) propose efficient estimators of semiparametric models of treatment effects. These estimators are obtained by adding a cross-fitted estimate of the model's efficient influence function to a first-step estimator. In contrast, we show that, in the same setting of Athey et al. (2021), a "plug-in" version of the parametric L-moment estimator of Alvarez (2022) is efficient without requiring further corrections. Moreover, for flexible parametrisations of treatment effects, our proposed estimator can be obtained by solving a quadratic program, which is quite convenient from the computational viewpoint. We also propose a specification test based on overidentifying restrictions and briefly discuss how to select the L-moments used in estimation.

[1]Examples include the generalised extreme value, generalised Pareto and generalised exponential distributions. See Alvarez (2022) and references therein.

As an application, we use our method in estimating treatment effects in a randomised experiment conducted by a car-hailing service in São Paulo. The experiment provided short-term discounts to rides starting on or ending in subway/train stations. The goal of the experiment was to understand whether short-term changes in prices could lead to long-run changes in the demand for public transportation due to *learning effects*. Our semiparametric specifications allows us to better estimate treatment effects and, as a consequence, to draw sharper inferences. For the largest randomised discounts, we find a strong substitution effect away from unimodal ("car-only") rides after the discount expires, and no effects on bimodal rides after the discount expires. Even though such effects could be interpreted as a large substitution effect – coupled with the assumption that there is low downward-substitutability in bimodal rides –, the magnitude of the substitution and their persistence over a month after the discount is over appears to be better interpreted through the lenses of a simple learning model, which we introduce in Appendix B.

The remainder of this paper is organised as follows. Section 2 briefly reviews the estimation of parametric models by means of L-moments as discussed in Alvarez (2022). In Section 3, we introduce the setting of Athey et al. (2021) and discuss the properties of our proposed estimator in this setup. Section 4 presents the results of our empirical application. Section 5 concludes.

2. Parametric estimation with L-moments

Consider a scalar random variable Y with distribution function F and finite first moment. For $r \in \mathbb{N}$, Hosking (1990) defines the r-th L-moment as:

$$\lambda_r \coloneqq \int_0^1 Q_Y(u) P^*_{r-1}(u) du \,, \tag{1}$$

where $Q_Y(u) \coloneqq \inf\{y \in \mathbb{R} : F(y) \geq u\}$ is the quantile function of Y, and $P^*_r(u) = \sum_{k=0}^{r} (-1)^{r-k} \binom{r}{k} \binom{r+k}{k} u^k$ are shifted Legendre polynomials.[2] Expanding the polynomials and using the quantile representation of a random variable (Billingsley, 2012, Theorem 14.1), we arrive at the equivalent expression:

$$\lambda_r = r^{-1} \sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k} \mathbb{E}[Y_{(r-k):r}] \,, \tag{2}$$

where, $Y_{(r-k):r}$ is the $(r-k)$-th order statistic of a random sample from F with r observations. Equation (2) motivates our description of L-moments as linear combinations of order statistics. Notice that the first L-moment corresponds to the expected value of Y.

[2]Legendre polynomials are defined by applying the Gramm-Schmidt orthogornalisation process to the polynomials $1, x, x^2, x^3 \ldots$ defined on $[-1, 1]$ (Kreyszig, 1989, p. 176-180). If P_r denotes the r-th Legendre polynomial, shifted Legendre polynomials are related to the standard ones through the affine transformation $P^*_r(u) = P_r(2u - 1)$ (Hosking, 1990).

To see how L-moments may offer "robust" alternatives to conventional moments, it is instructive to consider, as in Hosking (1990), the second L-moment. In this case, we have:

$$\lambda_2 = \frac{1}{2}\mathbb{E}[Y_{2:2} - Y_{1:2}] = \frac{1}{2}\int\int \left(\max\{y_1, y_2\} - \min\{y_1, y_2\}\right) F(dy_1)F(dy_2) = \frac{1}{2}\mathbb{E}|Y_1 - Y_2|\,,$$

where Y_1 and Y_2 are independent copies of Y. This is a measure of dispersion. Indeed, comparing it with the variance, we have:

$$\mathbb{V}[Y] = \mathbb{E}[(Y - \mathbb{E}[Y])^2] = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 = \frac{1}{2}\mathbb{E}[(Y_1 - Y_2)^2]\,,$$

from which we note that the variance puts more weight to larger differences.

Next, we discuss the estimation of parametric models based on matching L-moments. Suppose that F belongs to a parametric family of distribution functions $\{F_\theta : \theta \in \Theta\}$, where $\Theta \subseteq \mathbb{R}^d$ and $F = F_{\theta_0}$ for some $\theta_0 \in \Theta$. Let $l_r(\theta) \coloneqq \int_0^1 P^*_{r-1}(u)Q(u|\theta)du$ denote the theoretical r-th L-moment, where $Q(\cdot|\theta)$ is the quantile function associated with F_θ. Suppose we have access to an identically distributed sample, $Y_t \sim F$, $t = 1, \ldots, T$. Let $H^L(\theta) \coloneqq (l_1(\theta), l_2(\theta), \ldots, l_L(\theta))'$, and $\hat{H}^L$ be a vector stacking sample estimators for the first L L-moments.[3]. Researchers then usually estimate θ by solving:

$$H^d(\theta) - \hat{H}^d = 0\,.$$

As argued in Alvarez (2022), such estimator is usually inefficient asymptotically. Alternatively, the author considers the estimator:

$$\hat{\theta} \in \arg\inf_{\theta\in\Theta}(H^L(\theta) - \hat{H}^L)'W^L(H^L(\theta) - \hat{H}^L)\,, \tag{3}$$

for a (possibly estimated) weighting matrix W^L. In a framework where L varies with T, Alvarez (2022) shows that, under the optimal choice of weighting matrix, the estimator is able to outperform maximum likelihood estimation in small samples[4] and still retain asymptotic efficiency. The author then proposes methods to automatically select the number of L-moments. These are based on either higher-order expansions of the target-estimand, which are then minimised by the choice of L, similarly to existing approaches discussed in the GMM literature (Donald et al., 2009; Okui, 2009); or on ℓ_1-regularised estimation of the optimal combination matrix, which adapts the approach of Luo et al. (2015) to the L-moment setting. In Monte Carlo simulations, both approaches tend to perform well, and Alvarez (2022) favours the latter on computational grounds.

[3]Estimators can either be based on empirical analogs of (1) or (2). See Alvarez (2022) for a discussion.

[4]When the goal is to estimate a quantile of a distribution function, reductions in the root mean squared error can be as large as 30% for popular distributions such as the generalised extreme value or generalised Pareto.

3. Semiparametric analysis of randomised experiments

Consider a setting where there is an outcome of interest Y and a binary treatment $D \in \{0, 1\}$. For a population of interest, we define the random variables $(Y(0), Y(1))$ as the **potential outcomes** (Imbens and Rubin, 2015), where $Y(d)$ specifies what would occur if a subject randomly drawn from the population is assigned treatment status $D = d$.[5] The distribution of $(Y(0), Y(1))$ reflects the distribution of potential outcomes in the population. The distribution of treatment effects in the population is given by $\tau = Y(1) - Y(0)$; and the average treatment effect is given by $\beta \coloneqq \mathbb{E}[Y(1) - Y(0)]$. We consider the goal of experimentation to be to conduct inference on β.

In randomised experiments, a random sample of N individuals is drawn from the population; and treatment is assigned randomly to N_1 individuals, independently from their potential outcomes. Denoting by D_i the treatment status indicator of the i-th individual in the sample, we have that the observed outcome is $Y_i = D_i \cdot Y_i(1) + (1 - D_i) \cdot Y_i(0)$, where, by the sampling assumption, $(Y_i(0), Y_i(1)) \overset{\text{iid}}{\sim} (Y(0), Y(1))$. The researcher observes a resulting sample $\{(D_i, Y_i)\}_{i=1}^N$. The **fundamental problem of causal inference** is that the researcher does not observe both potential outcomes $(Y_i(0), Y_i(1))$ simultaneously, so individual effects $\tau_i = Y_i(1) - Y_i(0)$ are generally unidentifiable. It is possible however, to estimate the average treatment effect by:

$$\hat{\beta} = \frac{\sum_{i=1}^N D_i Y_i}{N_1} - \frac{\sum_{i=1}^N (1 - D_i) Y_i}{N_0} \tag{4}$$

i.e. a comparison of means between treatment and control units. Under random assignment of the treatment, this estimator is unbiased for β, and its variance is given by $\mathbb{V}[Y(0)]/N_0 + \mathbb{V}[Y(1)]/N_1$.

Is $\hat{\beta}$ the "best" we can do in randomised experiments? If the distribution of potential outcomes in the population is left unspecified (except for regularity conditions), then the answer is **yes**, in the sense that the estimator achieves the semiparametric efficiency bound (Newey, 1990) of the problem as $N_1, N_0 \to \infty$. However, as argued by Athey et al. (2021), in several contexts, this result is not enough. Especially in experimental settings deployed by big tech companies, treatment effects are small, and yet **economically significant**. These settings are characterised by heavy-tailed distributions of the outcome distribution, which leads to the variance of estimates based on (4) being quite large. The combination of small effects and large variances leads to low power in detecting effects based on (4), which renders the estimator quite unappealing.[6]

[5] Underlying this definition of potential outcome is a non-interference assumption, which postulates that the treatment status of an individual depends only on her assignment, and not on the treatment status of the remaining individuals in the population. The analysis of experimental (and observational) studies under different patterns of interference – where there can be spillovers from one unit's status to the other –, is an active topic of research in Statistics. See Sävje et al. (2021) and references therein.

[6] Indeed, as exemplified in Athey et al. (2021,page 1): "For example, Lewis and Rao (2015) analyze challenges with statistical power in experiments designed to measure digital advertising effectiveness. They discuss a hypothetical experiment where the average expenditure per potential customer is \$7, with a standard deviation of \$75, and where

In light of the preceding observations, Athey et al. (2021) propose estimating β using semiparametric methods. Their idea is to leave the distribution (quantile function) of potential outcomes in the absence of treatment, $F_{Y(0)}$ ($Q_{Y(0)}$), unspecified; and to parametrise the distribution in the treatment group as:

$$Q_{Y(1)}(u) = G(Q_{Y(0)}(u); \theta_0)\,, \tag{5}$$

where $\theta_0 \in \Theta \subseteq \mathbb{R}^d$. The authors focus on constructing efficient estimators, which achieve the semiparametric efficiency bound of the problem. We propose to analyse their environment through the lenses of L-moment-based estimation, which has been shown to work well in finite samples in the parametric seeting. In particular, we propose to estimate (5) using a plug-in version of the L-moment estimator (3), where we plug a nonparametric estimator of $Q_{Y(0)}$ (the quantile function in the control group) on (5) and estimate θ by minimising:

$$\hat{\theta}_{\text{plug}} \in \text{arginf}_{\theta\in\Theta} \left(\int_{\underline{p}}^{\overline{p}} [\hat{Q}_{Y(1)}(u) - G(\hat{Q}_{Y(0)}(u); \theta)] \boldsymbol{P}^L(u) du \right)' W^L \left(\int_{\underline{p}}^{\overline{p}} [\hat{Q}_{Y(1)}(u) - G(\hat{Q}_{Y(0)}(u); \theta)] \boldsymbol{P}^L(u) du \right), \tag{6}$$

where $\hat{Q}_{Y(d)}$, $d \in \{0, 1\}$ denotes the empirical quantile function in the group with $D = d$; $\boldsymbol{P}^L(u) = (P_1(u), P_2(u), \ldots, P_L(u))'$ is a vector with the first L elements of an orthonormal sequence in $L^2[0, 1]$, and $0 \leq \underline{p} < \overline{p} \leq 1$ are trimming constant. This estimator nests a "plug-in" L-moment estimator as a particular case, in which case we set $P_l = \sqrt{2l-1} P^*_{l-1}$ and take $0 = \underline{p} < \overline{p} = 1$. It also generalises the L-moment approach by allowing different types of orthonormal sequences to be used.

The plug-in approach (6) will lead to a larger variance than in the case where $F_{Y(0)}$ is known. Indeed, proceeding similarly as in Alvarez (2022), we are able to write, under differentiability conditions analogous to those in Proposition 2 Alvarez (2022), and as L,$N_1 \wedge N_0 \to \infty$, $(N_1 \wedge N_0)/N_j \to c_j$, $j \in \{0, 1\}$:

$$\begin{aligned} &\sqrt{N_0 \wedge N_1}(\hat{\theta}_{\text{plug}} - \theta_0) = \\ &-(\nabla_{\theta'} h_1^L(\theta_0)' \Omega^L \nabla_{\theta'} h_1^L(\theta_0))^{-1} \nabla_{\theta'} h_1^L(\theta_0)' \Omega^L \left[\sqrt{c_1}\sqrt{N_1} \left(\int_{\underline{p}}^{\overline{p}} (\hat{Q}_{Y(1)}(u) - Q_{Y(1)}(u)) \boldsymbol{P}^L(u) du \right) - \right. \\ &\left. \sqrt{c_0}\sqrt{N_0} \left(\int_{\underline{p}}^{\overline{p}} \partial_q G(Q_{Y(0)}(u); \theta_0) (\hat{Q}_{Y(0)}(u) - Q_{Y(0)}(u)) \boldsymbol{P}^L(u) du \right) \right] + o_P(1) \end{aligned} \tag{7}$$

an average treatment effect of \$0.35 (0.005 of a standard deviation) would be substantial in the sense of being highly profitable for the company. In that example, an experiment with power for a treatment effect of \$0.35 equal to 80%, and a significance level for the test of means of 0.05, would require a sample size of 1.4 million customers. As a result confidence intervals for the average effect of an advertisement are likely to include zero even if the average effect were substantively important, even with large sample sizes, e.g., over a hundred thousand units."

where $h_1^L(\theta) = \int_{\underline{p}}^{\overline{p}} (\hat{Q}_{Y(1)}(u) - G(Q_{Y(0)}(u);\theta))$, and Ω^L is a nonstochastic $L \times L^{'}$ matrix such that $\|W^L - \Omega^L\|_2 = o_p(1)$, where $\|\cdot\|_2$ is the spectral norm. The first term in the above equation refers to the asymptotic linear representation that would have been obtained were $Q_{Y(0)}$ known. The additional term is due to nonparametric estimation of $Q_{Y(0)}$. This term will inflate the variance of $\hat{\theta}^{\text{plug}}$, vis-à-vis the case where $Q_{Y(0)}(u)$ is known.[7]

In spite of the additional term in the asymptotic linear representation of the estimator, we are able to show that our L-moment estimator, under optimal weights, a choice of functions $\{P_l\}$ that constitute orthonormal bases (such as rescaled Legendre polynomial $\sqrt{2l-1}P_{l-1}^*$) and $\underline{p} = 0 < 1 = \overline{p}$, is efficient, in the sense that it attains the semiparametric efficiency bound derived in Lemma 1 of Athey et al. (2021). A proof of this fact is sketched in Appendix A and proceeds similarly as the proof of the parametric case presented in Section 2.5 of Alvarez (2022). Optimal weights can be computed using either the Gaussian or Bahadur-Kiefer approximations discussed in Alvarez (2022) and the asymptotic linear representation (7). Indeed, it follows that, under the conditions of Theorem 3 of Csorgo and Revesz (1978) (stated as Theorem 1 in Alvarez (2022)), there exist independent Brownian bridges B_{N_0} and B_{N_1} such that:

$$\begin{aligned}\sqrt{N_0 \wedge N_1}(\hat{\theta}_{\text{plug}} - \theta_0) = \\ -(\nabla_{\theta'} h_1^L(\theta_0)' \Omega^L \nabla_{\theta'} h_1^L(\theta_0))^{-1} \nabla_{\theta'} h_1^L(\theta_0)' \Omega^L \left[\sqrt{c_1} \left(\int_{\underline{p}}^{\overline{p}} \frac{B_{N_1}(u)}{f_{Y(1)}(Q_{Y(1)}(u))} \boldsymbol{P}^L(u) du \right) - \\ \sqrt{c_0} \left(\int_{\underline{p}}^{\overline{p}} \partial_q G(Q_{Y(0)}(u);\theta_0) \frac{B_{N_0}(u)}{f_{Y(0)}(Q_{Y(0)}(u))} \boldsymbol{P}^L(u) du \right) \right] + o_P(1),\end{aligned} \tag{8}$$

provided that $\int_{\underline{p}}^{\overline{p}} \frac{1}{f_{Y(1)}(Q_{Y(1)}(u))^2} < \infty$ and $\int_{\underline{p}}^{\overline{p}} \frac{\partial G(Q_{Y(0)}(u);\theta_0)^2}{f_{Y(0)}(Q_{Y(0)}(u))^2} < \infty$. Representation (8) can be used to compute the obtimal weigting scheme:[8]

$$\Omega_L^* = \mathbb{V}\left[\sqrt{c_1} \left(\int_{\underline{p}}^{\overline{p}} \frac{B_{N_1}(u)}{f_{Y(1)}(Q_{Y(1)}(u))} \boldsymbol{P}^L(u) du \right) + \sqrt{c_0} \left(\int_{\underline{p}}^{\overline{p}} \partial_q G(Q_{Y(0)}(u);\theta_0) \frac{B_{N_0}(u)}{f_{Y(0)}(Q_{Y(0)}(u))} \boldsymbol{P}^L(u) du \right) \right]^{-},$$

and to conduct inference based on normal critical values. Alternatively, one may resort to a Bahadur-Kiefer approximation, which leads to an identical formula for the optimal weighting matrix, but motivates inference based on draws from the uniform distribution. We refer the reader to Chapter 2 of Alvarez (2022) on the merits of each approximation as a basis for inference.

Interestingly, our proposed estimator is asymptotically efficient without further corrections, which contrasts with the estimators proposed by Athey et al. (2021), which require estimating the efficient influence function via cross-fitting to correct a first-step estimator. We also expect

[7] In the text, we consider a parametrisation where $Q_{Y(1)}(u)$ depends on $Q_{Y(0)}$ solely through $Q_{Y(0)}(u)$. This coincides, up to notation, with the parametrisation in Athey et al. (2021). We could consider more general forms of dependence of $Q_{Y(1)}(u)$ on $Q_{Y(0)}$ by working with Gâteaux derivatives (Newey, 1994).

[8] We use A^- to denote the generalised inverse of matrix A.

our estimator to work well in practice, given existing Monte Carlo evidence in the fully parametric setting. Finally, we note that our plug-in approach is computationally efficient for flexible parametrisations of treatment effects. Indeed, if we consider:

$$G(Q_{Y(1)}(u);\theta) = \sum_{j=0}^{K} \theta_j u^j + Q_{Y(0)}(u), \tag{9}$$

then the plugin estimator solves a quadratic program, which can be computed efficiently.[9][10] In this specification, *quantile treatment effects*, which measure how the distributions of treated and untreated potential outcomes differ in the population, are given by $q(u) = Q_{Y(1)}(u) - Q_{Y(0)}(u) = \sum_{j=0}^{K} \theta_j u^j$. The average treatment effect is given by $\beta = \int_0^1 (Q_{Y(1)}(u) - Q_{Y(0)}(u))du = \sum_{j=0}^{K} \frac{\theta_j}{j+1}$.

Comment 1 (A statistic for overidentifying restrictions)**.** When $L > d$, we may consider a test based on overidentifying restrictions. Specifically, we consider the J statistic:

$$\begin{aligned}\hat{J} = (N_0 \wedge N_1) \left(\int_{\underline{p}}^{\overline{p}} [\hat{Q}_{Y(1)}(u) - G(\hat{Q}_{Y(0)}(u); \hat{\theta}_{\text{plug}})] \boldsymbol{P}^L(u) du \right)' W^L \times \\ \left(\int_{\underline{p}}^{\overline{p}} [\hat{Q}_{Y(1)}(u) - G(\hat{Q}_{Y(0)}(u); \hat{\theta}_{\text{plug}})] \boldsymbol{P}^L(u) du \right).\end{aligned} \tag{10}$$

Under the null that the model is correctly specified (i.e. that there exists $\theta \in \Theta$ such that $Q_{Y(1)}(u) = G(\theta; Q_{Y(0)}(u))$), and when an estimator of the optimal weighting scheme is used, it is possible to use the Gaussian strong approximation discussed above to show that the distribution of the test statistic can be approximated by a chi-squared with $L - d$ degrees of freedom, *provided that L grows sufficiently slowly* so that the Kolmogorov distance between the true distribution and the approximating distribution converge to zero. We refer the reader to Comment 2 in Alvarez (2022) for further details on the latter point. ■

Comment 2 (Selecting the number of L-moments used)**.** It is possible to adapt the moment selection procedures proposed in Chapter 3 of Alvarez (2022) in the fully parametric setting to our

[9]Observe that, since $G(Q_{Y(1)}(u);\theta)$ is a quantile function, we expect it to be non-decreasing. In some settings (e.g. Gourieroux and Jasiak (2008)), it is desirable that the *estimated* quantile function preserve monotonic. In our context, we may impose monotonicity in the estimation of (9) by including linear restrictions in the quadratic program that ensure the estimated quantile function is monotonic. Altrnatively, we may estimate the quantile function without restrictions, and monotonise it by using the rearrangement procedure in Chernozhukov et al. (2009).

[10]Another advantage of our estimation method under this polynomial specification is that the optimal weighting matrix can be computed without resorting to a first-step estimator of θ_0. Indeed, since $\partial_q G(Q_Y(0)(u);\theta_0) = 1$, an estimator of the optimal weighting matrix may be computed by estimating densities in the treatment and control groups nonparametrically and using the empirical quantile functions in each group. In our empirical application, we use the estimator of Cattaneo et al. (2020) to estimate the densities in each group nonparametrically.

semiparametric context. It should be noted, however, that, as discussed by the author, higher-order expansions of quantile estimators can only be obtained in an heuristic manner. This is due to nondifferentiability of the estimating function associated with the empirical quantile process. In these cases, one can appeal to the heuristic proposed by Phillips (1991), whereby the "derivative" of an indicator function is replaced by the Dirac delta function. As recently shown by Franguridi et al. (2021), this approach is able to deliver precisely those higher-order terms that are estimable. In a fully parametric context under iid data, Alvarez (2022) bypasses resorting to Phillips' heuristic by using a parametric bootstrap to account for higher-order moments of the empirical quantile process. However, in a semiparametric context, this solution is not available, and moment selection procedures based on higher-order expansions of the target estimand will require appealing to such heuristic.

As for the regularised estimation of the weighting matrix, such approach is immediately extendable to the semiparametric setting. Specifically, Alvarez (2022) provides high-level conditions that ensure validity of such approach, by allowing application of an approximation result in Luo et al. (2015). Nonetheless, appropriate algorithms for setting the ℓ_1 penalties are more difficult to establish in the semiparametric setting, for they require good estimators of the variance of the nonparametric estimators used in constructing the estimated weighting matrix. As a consequence, we leave construction of proper selection methods in the semiparametric setting for future versions of this paper. ■

4. Empirical application

We illustrate the applicability of our approach by using data from a randomised experiment in the municipality São Paulo. In 2018, a large ride-hailing company in Brazil randomised discounts to a subset of its users. The goal of the experiment was to understand whether short-run incentives to modal integration could alter long-run behavior via **learning effects**. Specifically, the platform asked for information from some of its active users and, among the respondents, randomised discounts to trips starting or ending at a subway/train station during two weeks between the end of November and beginning of December 2018. Users were randomised into three regimes: (i) a control group (which was not informed about the experiment); (ii) eligibility to two 20% discounts per day, limited to a total of 10 BRL discount per car ride; and (iii) eligibility to two 50% discounts per day, limited to 10 BRL per ride. The discounts were announced the day they started. We have access to the number of rides taken by each individual in each group on a biweekly basis in the fortnights leading up to, during and after the treatment. We also know how many of these rides started or ended in a subway-train station.

Classical consumer theory, where consumers are assumed to have full knowledge of the goods available to them, predicts two types of effects related to giving discounts: first, one would expect a **substitution** effect, whereby users reduce consumption in other goods and increase consumption

in the good affected by the discount. In our setting, such effect is expected to increase bimodal rides in the two weeks during which the discount is available, and reduce the number of "unimodal" rides in the same period. One would also expect a reduction of both types of rides in the weeks after the discount, due to intertemporal substitution (consumers shift consumption between periods). Nonetheless, given that transportation may be taken to be a necessary good, we would expect such effects to be quite small. On the other hand, standard theory also predicts **an income effect**: given that a consumer's basket is less costly, there is more available income to spend. This effect is expected to increase consumption in both types of rides both during and after treatment.

Finally, a third possible effect, which is not predicted by standard consumer theory, would be **a learning effect**. This effect hinges on the assumption that consumers do not fully know *ex-ante* the goods available to them. In our setting, where discounts are randomised among active users of the platform, this effect would be expected to occur among users that do not use public transportation. Inasmuch as the discounts lead them to learn more about this service, one would expect a long-term change in bimodal rides, even after the discount is over. Note that such change could be either positive or negative. We would also expect an oppositely-signed, long-term effect, on "unimodal" rides, as users shift from (to) "unimodal" rides to (from) "bimodal" ones.

In practice, a mix of the three types of effects is expected to occur. Since they act in counteracting directions, one would expect the overall effect to be small. In light of that, and given the heavy tailed nature of the distribution – the average number of rides by fortnight in the dataset, even after excluding those pairs of user/fortnights where no rides were taken, is 3.64, whereas the maximum is 57 –, a simple comparison of means is expected to perform poorly. We thus propose to analyse the experiment using the methodology introduced in this chapter.

To perform our analysis, we restrict our sample to users who completed at least one ride between the start of 2018 and up until the day before treatment was in place. With that, we expect to estimate effects among users who already know the car-riding services (so learning effects should not act towards increasing "unimodal" rides). This leaves us with $N_0 = 1,291$ control units, $N_{1,20\%} = 1,381$ units in the 20% discount treatment arm, and $N_{1,50\%} = 1,319$ in the 50% treatment arm. We analyse the effect of each treatment on bimodal and "unimodal" (total minus bimodal) rides, in the fortnights leading up to, on, and after treatment was in place. Since users were only alerted of the treatment the day it started, we expect no effects in the fortnight leading up to treatment. We contrast a simple comparison of means between treatment and control arms with parametrisation (9). We consider specifications with $K = 0, 1, 2, 3$ for "unimodal" rides. The "bimodal" outcome is lightly-tailed, so the difference in means tends to perform well. Indeed, Table 1 reports treatment effects estimates from differences in means for bimodal rides, in the fortnight prior to ($t = -1$), during ($t = 0$) and after ($t > 0$) treatment was in place. Standard errors are reported in parentheses. We observe a strong and significant increase in bimodal rides

on the fortnight during which treatment is in place, and no effects prior or after that. As expected, the 50% discount leads to a larger increase in bimodal rides than the 20% discount.

TABLE 1. Bimodal rides

20% discount

$t=-1$		$t=0$		$t=1$		$t=2$		$t=3$		$t=4$		$t=5$		$t=6$	
Mean control	Diff mean	Mean control	Diff mean	Mean control	Diff mean	Mean control	Diff mean	Mean control	Diff mean	Mean control	Diff mean	Mean control	Diff mean	Mean control	Diff mean
0.1084	-0.0100	0.1325	0.0471	0.1154	0.0157	0.0434	0.0037	0.1030	-0.0053	0.1038	-0.0046	0.0813	0.0258	0.1131	0.0086
(0.0140)	(0.0189)	(0.0181)	(0.0269)	(0.0154)	(0.0231)	(0.0070)	(0.0098)	(0.0149)	(0.0212)	(0.0151)	(0.0197)	(0.0132)	(0.0222)	(0.0145)	(0.0218)

50% discount

$t=-1$		$t=0$		$t=1$		$t=2$		$t=3$		$t=4$		$t=5$		$t=6$	
Mean control	Diff mean	Mean control	Diff mean	Mean control	Diff mean	Mean control	Diff mean	Mean control	Diff mean	Mean control	Diff mean	Mean control	Diff mean	Mean control	Diff mean
0.1084	-0.0061	0.1325	0.0950	0.1154	-0.0100	0.0434	0.0105	0.1030	-0.0090	0.1038	-0.0090	0.0813	0.0134	0.1131	0.0021
(0.0140)	(0.0189)	(0.0181)	(0.0281)	(0.0154)	(0.0210)	(0.0070)	(0.0120)	(0.0149)	(0.0200)	(0.0151)	(0.0199)	(0.0132)	(0.0193)	(0.0145)	(0.0206)

Tables 2 and 3 report results for unimodal rides using the difference in means estimator and the parametric specifications (7) with K ranging from 0 to 3. In using the L-moment estimator, we set $L = 10$. We also report the p-value from the specification test in Comment 1. A few patterns stand out: first, the parametric specifications allow us to obtain substantial reductions in standard errors (vis-à-vis the difference in means estimator), especially with $K = 0$ and $K = 1$.[11] This gain in precision allows us to make tighter inferences on treatment effects. Specifically, and discarding specifications for which the J-test rejects the null, we note that:

(1) For the 20%, systematic effects on unimodal rides do not appear to exist – estimated effects in the low-K specifications that survive the J-test are small, (relatively) precisely estimated and statistically insignificant.
(2) For the 50% discount, there is strong evidence of a **decrease** in unimodal rides in $t = 1$ and $t = 3$. There is also some evidence of a decrease in rides in $t = -1$ when using values of $K > 0$, which would suggest some anticipation effect, even though the discount was announced at $t = 0$. However, it should be noted that this effect at $t = -1$ does not survive in the constant treatment effect specification ($K = 0$), which also estimates effects much more precisely than other specifications. Similarly, there is some evidence of negative effects at $t = 5$ when using $K > 0$, though this effect does not survive the constant treatment effect specification.

All in all, the results appear compatible with the 50% discount producing large intertemporal substitution effects on unimodal rides. The fact that bimodal rides in other periods are not affected could be explained by the assumption that bimodal rides may be considered a necessary good, with low degree of downward substitutability. Negative effects on unimodal rides persist

[11]For values of K larger than 1, in some periods the standard error of the parametric estimator exceeds that of the difference in means. This is due to the fact that L is kept fixed. By increasing L, we are able to decrease the estimated standard error, though potentially at the cost of finite-sample bias. In future versions of this project, we hope that, by extending the selection methods discussed in Comment 2, we may be able to produce automatic methods to select L with good statistical properties.

over a month after the discount is over.[12] However, it is interesting to note that there does not appear to be a contemporaneous substitution effect. This could be explained by the income effect counterbalancing the substitution effect at $t = 0$. Such large income effect in the contemporaneous period is expected to occur in settings where the return to savings is low. Finally, note that we do not find evidence of learning effects towards bimodal rides.

Even though our results could be driven by standard substitution effects – coupled with a downward substitutability constraint on bimodal rides, which may be interpreted as reflecting the "necessary" character of this good –, one could also interpret them as being driven by a setting where users, by substantially increasing bimodal rides, learn about the quality of public transportation, and, being positively surprised, shift consumption away from unimodal rides towards purely public transportation trips, whilst keeping bimodal rides unchanged. In Appendix B, we introduce a learning model that is able to produce these results. This model may be able to better explain the persistence of strong and negative effects on unimodal rides, which may be incompatible with *a priori* reasonable values for the intertemporal elasticity of substitution of rides, depending on the relative price of bimodal and unimodal rides. As future steps of this research, we hope to obtain access to the monetary values of each ride, so that we could better understand which explanation is more appropriate. We also hope to further explore the model in Appendix B, so as to derive further testable implications on our available data.

[12]The fact that we dot not encounter significant effects on rides at $t = 2$, but at $t = 1$ and $t = 3$, could be justified by the fact that this fortnight corresponds to the last week of 2018 and the first week of 2019, when unimodal rides may be assumed difficult to substitute due to the holidays/end-of-year celebration.

TABLE 2. Unimodal rides: 20% discount

			$t=-1$			
	Mean control	Diff Means	$K=0$	$K=1$	$K=2$	$K=3$
Estimate	1.9899	-0.0377	0.0202	-0.1683	-0.1192	-0.1312
Std. Error	(0.0777)	(0.1148)	(0.0420)	(0.0987)	(0.1083)	(0.1091)
pval J-test			0.1139	0.2800	0.2841	0.2561
			$t=0$			
	Mean control	Diff Means	$K=0$	$K=1$	$K=2$	$K=3$
Estimate	2.1472	0.0961	0.0475	0.1121	0.1018	0.1015
Std. Error	(0.0860)	(0.1257)	(0.0455)	(0.1000)	(0.1146)	(0.1146)
pval J-test			0.8334	0.8110	0.7267	0.6181
			$t=1$			
	Mean control	Diff Means	$K=0$	$K=1$	$K=2$	$K=3$
Estimate	2.1526	-0.0519	0.0229	0.0802	0.0667	0.0693
Std. Error	(0.0900)	(0.1258)	(0.0437)	(0.1013)	(0.1125)	(0.1129)
pval J-test			0.3235	0.2687	0.1959	0.1340
			$t=2$			
	Mean control	Diff Means	$K=0$	$K=1$	$K=2$	$K=3$
Estimate	1.2208	0.0979	0.0301	0.2108	0.1386	0.1527
Std. Error	(0.0601)	(0.0853)	(0.0404)	(0.0913)	(0.1005)	(0.1020)
pval J-test			0.3187	0.6993	0.9210	0.9251
			$t=3$			
	Mean control	Diff Means	$K=0$	$K=1$	$K=2$	$K=3$
Estimate	1.8737	0.0046	-0.0045	0.0316	0.0284	0.0286
Std. Error	(0.0895)	(0.1253)	(0.0521)	(0.1133)	(0.1278)	(0.1288)
pval J-test			0.9995	0.9990	0.9969	0.9906
			$t=4$			
	Mean control	Diff Means	$K=0$	$K=1$	$K=2$	$K=3$
Estimate	1.9535	-0.0136	0.0648	0.1932	0.1424	0.1546
Std. Error	(0.0883)	(0.1267)	(0.0491)	(0.1152)	(0.1250)	(0.1273)
pval J-test			0.0137	0.0135	0.0113	0.0065
			$t=5$			
	Mean control	Diff Means	$K=0$	$K=1$	$K=2$	$K=3$
Estimate	1.9620	-0.0707	-0.0141	-0.0917	-0.0660	-0.0744
Std. Error	(0.0916)	(0.1299)	(0.0654)	(0.1270)	(0.1336)	(0.1349)
pval J-test			0.9800	0.9802	0.9770	0.9631
			$t=6$			
	Mean control	Diff Means	$K=0$	$K=1$	$K=2$	$K=3$
Estimate	2.1309	-0.0201	-0.0120	-0.0980	-0.0609	-0.0851
Std. Error	(0.0977)	(0.1358)	(0.0506)	(0.1256)	(0.1313)	(0.1351)
pval J-test			0.9641	0.9642	0.9820	0.9883

TABLE 3. Unimodal rides: 50% discount

			$t=-1$			
	Mean control	Diff Means	$K=0$	$K=1$	$K=2$	$K=3$
Estimate	1.9899	-0.1507	-0.0323	-0.2035	-0.1804	-0.1735
Std. Error	(0.0777)	(0.1067)	(0.0382)	(0.0973)	(0.1021)	(0.1035)
pval J-test			0.7931	0.9868	0.9902	0.9829
			$t=0$			
	Mean control	Diff Means	$K=0$	$K=1$	$K=2$	$K=3$
Estimate	2.1472	-0.0554	0.0040	-0.1070	-0.0558	-0.0655
Std. Error	(0.0860)	(0.1190)	(0.0419)	(0.1066)	(0.1130)	(0.1135)
pval J-test			0.8439	0.8909	0.9723	0.9872
			$t=1$			
	Mean control	Diff Means	$K=0$	$K=1$	$K=2$	$K=3$
Estimate	2.1526	-0.2049	-0.0487	-0.3115	-0.2299	-0.2475
Std. Error	(0.0900)	(0.1259)	(0.0426)	(0.1048)	(0.1108)	(0.1124)
pval J-test			0.0658	0.3854	0.8441	0.8642
			$t=2$			
	Mean control	Diff Means	$K=0$	$K=1$	$K=2$	$K=3$
Estimate	1.2208	-0.0767	-0.0118	-0.0997	-0.0838	-0.0762
Std. Error	(0.0601)	(0.0839)	(0.0395)	(0.0865)	(0.0966)	(0.0982)
pval J-test			0.9864	0.9984	0.9971	0.9957
			$t=3$			
	Mean control	Diff Means	$K=0$	$K=1$	$K=2$	$K=3$
Estimate	1.8737	-0.2960	-0.0803	-0.3768	-0.2032	-0.2172
Std. Error	(0.0895)	(0.1177)	(0.0506)	(0.1039)	(0.1245)	(0.1247)
pval J-test			0.0004	0.0107	0.0611	0.1419
			$t=4$			
	Mean control	Diff Means	$K=0$	$K=1$	$K=2$	$K=3$
Estimate	1.9535	-0.1613	0.0074	-0.0931	-0.0772	-0.0700
Std. Error	(0.0883)	(0.1230)	(0.0468)	(0.1113)	(0.1201)	(0.1223)
pval J-test			0.6302	0.6387	0.5455	0.4397
			$t=5$			
	Mean control	Diff Means	$K=0$	$K=1$	$K=2$	$K=3$
Estimate	1.9620	-0.2357	-0.0290	-0.3059	-0.2555	-0.2520
Std. Error	(0.0916)	(0.1257)	(0.0641)	(0.1253)	(0.1307)	(0.1326)
pval J-test			0.4859	0.9848	1.0000	1.0000
			$t=6$			
	Mean control	Diff Means	$K=0$	$K=1$	$K=2$	$K=3$
Estimate	2.1309	-0.2295	-0.0423	-0.2098	-0.1480	-0.1716
Std. Error	(0.0977)	(0.1304)	(0.0530)	(0.1293)	(0.1366)	(0.1396)
pval J-test			0.6826	0.8051	0.9210	0.9270

5. CONCLUDING REMARKS

This paper extends the L-moment-based estimation approach to the semiparametric models of treatment effects introduced by Athey et al. (2021). We show that a simple plug-in version of the estimator proposed by Alvarez (2022) is able to produce a statistically efficient and coputationally attractive estimator, without having to resort to the influence-function corrections discussed in

Athey et al. (2021). We then apply our methodology to data on a randomised experiment conducted in São Paulo by a ride-hailing company, where the goal was to understand whether short term changes in the price of bimodal rides could lead to long-term changes in the demand for public transportation. For the 50% discount, our results appear compatible with either strong intertemporal substitution effects away from unimodal rides and towards bimodal rides at the time of the discount; or with a learning effect that permanently shifts demand from unimodal rides towards purely public transportation trips. As future steps of this research project, we aim to fully extend the methods of selection of L-moments proposed by Alvarez (2022) to the semiparametric setting; and to better understand which of the two competing explanations is more plausible in our experimental setting.

References

Alvarez, L. (2022): "Inference in parametric models with many L-moments," Ph.D. thesis, Institute of Mathematics and Statistics, University of São Paulo, draft available at: `https://www.dropbox.com/s/ysglbdhyeb66wyj/l_moments_redux.pdf?dl=0`.

Athey, S., P. J. Bickel, A. Chen, G. W. Imbens, and M. Pollmann (2021): "Semiparametric Estimation of Treatment Effects in Randomized Experiments," .

Billingsley, P. (2012): *Probability and Measure*, Wiley.

Cattaneo, M. D., M. Jansson, and X. Ma (2020): "Simple Local Polynomial Density Estimators," *Journal of the American Statistical Association*, 115, 1449–1455.

Chernozhukov, V., I. Fernández-Val, and A. Galichon (2009): "Improving point and interval estimators of monotone functions by rearrangement," *Biometrika*, 96, 559–575.

Csorgo, M. and P. Revesz (1978): "Strong Approximations of the Quantile Process," *Ann. Statist.*, 6, 882–894.

Donald, S. G., G. W. Imbens, and W. K. Newey (2009): "Choosing instrumental variables in conditional moment restriction models," *Journal of Econometrics*, 152, 28–36, recent Adavances in Nonparametric and Semiparametric Econometrics: A Volume Honouring Peter M. Robinson.

Firpo, S., A. F. Galvao, C. Pinto, A. Poirier, and G. Sanroman (2021): "GMM quantile regression," *Journal of Econometrics*.

Franguridi, G., B. Gafarov, and K. Wuthrich (2021): "Conditional quantile estimators: A small sample theory," .

Gourieroux, C. and J. Jasiak (2008): "Dynamic quantile models," *Journal of Econometrics*, 147, 198–205.

Hosking, J. R. M. (1990): "L-Moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics," *Journal of the Royal Statistical Society: Series B (Methodological)*, 52, 105–124.

Imbens, G. W. and D. B. Rubin (2015): *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge University Press.

Kreyszig, E. (1989): *Introductory Functional Analysis with Applications*, Wiley.

Luo, Y. et al. (2015): "High-dimensional econometrics and model selection," Ph.D. thesis, Massachusetts Institute of Technology.

Newey, W. K. (1990): "Semiparametric efficiency bounds," *Journal of Applied Econometrics*, 5, 99–135.

——— (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382.

Okui, R. (2009): "The optimal choice of moments in dynamic panel data models," *Journal of Econometrics*, 151, 1–16.

Phillips, P. C. B. (1991): "A Shortcut to LAD Estimator Asymptotics," *Econometric Theory*, 7, 450–463.

Sävje, F., P. M. Aronow, and M. G. Hudgens (2021): "Average treatment effects in the presence of unknown interference," *The Annals of Statistics*, 49, 673 – 701.

Appendix A. Efficiency of semiparametric L-moment estimator

Let $N = N_0 + N_1$ denote the sample size. Consider the alternative (unfeasible) estimator:

$$\check{\theta} \in \text{argmin}_{\theta\in\Theta} \sum_{i\in\mathcal{S}_N}\sum_{j\in\mathcal{S}_N} (\frac{1}{\sqrt{p}_1}(\hat{Q}_{Y(1)}(i) - Q_{Y(1)}(\theta)) + \frac{1}{\sqrt{p}_0}\partial_q G \cdot (\hat{Q}_{Y(0)}(i) - Q_Y(0))) \times \kappa_{i,j} \times \left(\frac{1}{\sqrt{p}_1}(\hat{Q}_{Y(1)}(j) - Q_{Y(1)}(\theta)) + \frac{1}{\sqrt{p}_0}\partial_q G \cdot (\hat{Q}_{Y(0)}(j) - Q_Y(0))\right)$$

for a grid of S_N points $\mathcal{S}_N = \{s_1, s_2, \ldots, s_{S_N}\} \subseteq (0,1)$ and weights $\kappa_{i,j}$, $i,j \in \mathcal{S}_N$. Here, $Q_{Y(1)}(\theta) = G(Q_{Y(0)}(u), \theta)$ and $\partial_q G = \partial_q G(Q_{Y(0)}(u), \theta_0)$. We also set $p_l = \lim \frac{N_l}{N}$, $l \in \{0,1\}$, and we assume $p_l \in (0,1)$. Under some conditions, and as $N, S_N \to \infty$, the estimator has asymptotic linear representation as follows:

$$\sqrt{N}(\check{\theta} - \theta_0) = -(\partial_{\theta'} G'_{S_N} \boldsymbol{\kappa}_{S_N} \partial_{\theta'} G_{S_N})^{-1} \partial_{\theta'} G'_{S_N} \boldsymbol{\kappa}_{G_N} \Big[\frac{1}{\sqrt{p}_1} \boldsymbol{f_{Y(1)}}^{-1} * \sqrt{T} F_{Y(1),S_N} + \frac{\boldsymbol{\partial_q G}}{\sqrt{p}_0} * \boldsymbol{f_{Y(0)}}^{-1} * \sqrt{T} F_{Y(0),S_N}\Big] + o_p(1)$$

where $*$ denotes entry-by-entry multiplication. But then, taking $\mathcal{S}_N = \left\{\frac{1}{S_N+1}, \frac{2}{S_N+1}, \ldots, \frac{S_N}{S_N+1}\right\}$ and applying Lemma C.1. in Firpo et al. (2021), we are able to show that the variance of the estimator under optimal weights is:

$$\mathbb{V}^* = ((\partial G_{S_N} * (\mathbf{1}_d' \otimes [\boldsymbol{f_{Y(1)}}\sqrt{p}_1 + \boldsymbol{\partial_q G}^{-1} * \boldsymbol{f_{Y(0)}}\sqrt{p}_0)])' \Sigma_{S_N}^{-1} (\partial G_{S_N} * (\mathbf{1}_d' \otimes [\boldsymbol{f_{Y(1)}}\sqrt{p}_1 + \boldsymbol{\partial_q G}^{-1} * \boldsymbol{f_{Y(0)}}\sqrt{p}_0])))^{-1}$$

where

$$(\Sigma_{S_N}^{-1})_{s_i,s_j} = \mathbb{1}_{\{s_i=s_j\}} 2(S_N+1) - (\mathbb{1}_{\{s_i=s_{j+1}\}} + \mathbb{1}_{\{s_i=s_{j-1}\}})(S_N+1)$$

Proceeding similarly as in Section 2.5 of Alvarez (2022), we obtain that:

$$\lim_{N\to\infty} \left(\mathbb{V}^{*-1}\right)_{d_1,d_2} = \int_0^1 \left.\frac{dH_{d_1}(v)}{dv}\right|_{v=u} \left.\frac{dH_{d_2}(v)}{dv}\right|_{v=u} du$$

with $H_d(u) = \left(\sqrt{p_1} f_{Y(1)}(Q_{Y(1)}(v)) + \sqrt{p_0}\frac{f_{Y(0)}(Q_{Y(0)}(v))}{\partial_q G(Q_{Y(0)}(v);\theta_0)}\right) \partial_{\theta_d} G(Q_{Y(0)}(v), \theta_0)$. Next, observing that:

$$\frac{1}{f_{Y(1)}(Q_{Y(1)}(v))} = Q'_{Y(1)}(v) = \partial_q G(Q_{Y(0)}(v);\theta_0) \cdot \frac{1}{f_{Y(0)}(Q_{Y(0)}(v))}$$

we obtain that:

$$H_d(u) = \sqrt{p_1}\sqrt{p_0} f_{Y(1)}(Q_{Y(1)}(v))$$

And then, proceeding similarly as in the proof of Section 2.5 of Alvarez (2022), we conclude that:

$$\left(\mathbb{V}^{*-1}\right)_{d_1,d_2} = p_0 p_1 (I(\theta_0))_{d_1,d_2}$$

where $I(\theta_0)$ is the Fisher information matrix of the parametric model $\theta \mapsto f_{Y(1)}(y|\theta)$ that assumes $Q_{Y(0)}$ known. It then follows by Lemma 1 of Athey et al. (2021) that the estimator is asymptotically efficient, as it achieves the efficiency bound derived by the authors.

To conclude, we note that, when the $\{P_l\}_l$ constitute orthonormal **bases**, the estimator $\check{\theta}$ corresponds to a method-of L-moments estimator that uses **infinitely** many moments. We are then able to show, by reasoning similarly as in Section 2.5. of Alvarez (2022), that the estimator (6), which uses a finite but increasing number of L-moments, is also efficient, since under an appropriate choice of weights this estimator is asymptotically equivalent to $\check{\theta}$.

Appendix B. A learning model on the demand for public transportation

In this section, we introduce a simple model which is able to rationalise the findings in the paper. The model consists of a standard intertemporal choice problem with one additional ingredient: a learning mechanism, whereby increased demand for bimodal rides increases knowledge of the quality of public transportation in the future. The model produces two effects: first, a discount in an initial period may increase demand for bimodal rides in the same period due to a *learning bequest*, whereby an agent increases demand so as to have better information on the quality of the service

in the future. Moreover, for a given configuration of parameters, bimodal rides need not change by much, even if public transportation is revealed to be of good quality: in contrast, unimodal rides decrease when this occurs. The two effects combined may be better able to rationalise the strong and long-running negative substitution effects reported in the main text for unimodal rides, and the insignificant changes encountered in bimodal rides.

We consider a two-period problem ($t \in \{0, 1\}$) where an agent has to choose her share of consumption on unimodal rides (u), bimodal rides (b) and public transportation (c). Her instantaneous preferences at each period are given by:

$$S(u_t, b_t, c_t) = u_t^\alpha + \gamma A^\phi b_t^\alpha + A c_t^\alpha$$

for positive constants α, γ, ϕ and $\alpha < 1$. Observe that the parameter A governs the relative gain of choosing public transportation over unimodal rides. It also affects the relative gain of bimodal rides through the term A^ϕ.[13] We assume the agent does not observe A, but has a Gaussian prior over $\log(A)$, i.e. the agent assumes $\log(A) \sim N(\mu_0, \sigma_0^2)$. Consumption of bimodal rides and public transportation in period 0 produces an unbiased signal of the quality of rides, where the informativeness of the signal is inversely related to the amount consumed in period 0. Specifically, we assume that, afeter consuming (b_0, c_0) units in period zero, the agent observes a signal:

$$Y|\log(A) \sim N\left(\log(A), \frac{1}{h_1(b_0, c_0)}\right)$$

for an increasing, everywhere positive and differentiable mapping h_1. We assume that the agent maximises expected utility, given the information available to her. Specifically, and taking c as the numéraire, at period 0, the agent solves:

$$\begin{aligned} &\max_{u_0, b_0, c_0 \geq 0} \mathbb{E}[S(u_0, b_0, c_o) + \beta S(u_1(y, d), b_1(y, d), c_1(y, s))] \\ s.t. \quad &d = (1 + r)[w_0 - p_{0,b} b_0 - p_{0,u} u_0 - c_0] \end{aligned} \tag{11}$$

where β is the discount factor, d is the savings from period zero to one, w_0 is the income in period 0, and r is the interest rate. The random variables $u_1(y, d), b_1(y, d), c_1(y, s)$ are the solutions to the second period problem, i.e.:

$$\begin{aligned} &\max_{u_1, b_1, c_1 \geq 0} \mathbb{E}[S(u_1, b_1, c_1)|Y] \\ s.t. \quad &d + w_1 = p_{1,b} b_1 + p_{1,u} u_1 + c_1 \end{aligned} \tag{12}$$

We note that the model above nests a standard intertemporal choice problem with full knowledge of the quality of the service if we set $\sigma_0^2 = 0$.

[13] The term A^ϕ could be interpreted as a reduced form for the agent's problem of deciding the composition of bimodal rides between public transportation and car trips according to a Cobb-Douglas production function, where A enters the production function as a public-transportation-augmenting factor.

We begin by solving the model by backward induction. First, by known results on conjugate priors, we have that:

$$\log(A)|Y \sim N\left(\frac{h_0\mu_0 + h_1(b_0, x_0)Y}{h_0 + h_1(b_0, c_0)}, (h_0 + h_1(b_0, c_0))^{-1}\right)$$

where $h_0 = 1/\sigma_0^2$ is the precision of the prior. It then folows, by the properties of the lognormal distribution that:

$$\mathbb{E}[A|Y] = \exp\left(\frac{h_0\mu_0 + h_1(b_0, c_0)Y}{h_0 + h_1(b_0, c_0)} + \frac{(h_0 + h_1(b_0, c_0))^{-1}}{2}\right) \tag{13}$$

and, similarly:

$$\mathbb{E}[A^\delta|Y] = (\mathbb{E}[A|Y])^\delta \exp\left(\delta(\delta-1)\frac{(h_0 + h_1(b_0, c_0))^{-1}}{2}\right) = (\mathbb{E}[A|Y])^\delta \psi_\delta(b_0, c_0) \tag{14}$$

Next, we note that the first-order conditions on the second period problem entail:

$$\begin{aligned} u_1 &= \left(\frac{1}{p_{1,u}\mathbb{E}[A|Y]}\right)^{\frac{1}{1-\alpha}} c_1 \\ b_1 &= \left(\frac{\gamma\psi_\delta(b_0, c_0)}{p_{1,b}\mathbb{E}[A|Y]^{1-\delta}}\right)^{\frac{1}{1-\alpha}} c_1 \\ c_1 &= \frac{d + w_1}{p_{1,u}^{-\alpha/(1-\alpha)}\mathbb{E}[A|Y]^{-1/(1-\alpha)} + p_{1,b}^{-\alpha/(1-\alpha)}\left(\frac{\gamma\psi_\delta(b_0,c_0)}{p_{1,b}\mathbb{E}[A|Y]^{1-\delta}}\right)^{\frac{1}{1-\alpha}} + 1} \end{aligned} \tag{15}$$

A few properties show up in this optimisation. First, if $\delta < 1$, then, all else equal, an increase in the signal Y unambiguously increases c_1 and necessarily decreases u_1. As for the effect on bimodal rides, its sign depends on relative prices and the region $\mathbb{E}[A|Y]$ lies. Specifically, it is immediate from the above expression that:

$$\frac{db_1}{dY} \propto p_{1,u}^{-\alpha/(1-\alpha)}\delta\mathbb{E}[A|Y]^{(\alpha-\delta-1)/(1-\alpha)} - (1-\delta)\mathbb{E}[A|Y]^{(\alpha-\delta)/(1-\alpha)} \tag{16}$$

Observe that, for large enough values of $p_{1,u}$, the effect is negative. Alternatively, if $\delta \approx 1$, the effect is positive. Clearly, there exist combination of parameter-signals where the effect is arbitrarily small.

As for understanding the learning bequest, it is useful to consider some candidate choice $(\tilde{b}_0, \tilde{c}_0)$ which leads to a signal with precision $\tilde{h}_0$; and an alternative consumption bundle which produces a signal with precision $\check{h}_0 < \tilde{h}_0$. In this case, it can be shown that, from the viewpoint of the posterior distribution, observing a signal with precision $\check{h}_0$ is equivalent to observing **two independent** signals: one with precision $\tilde{h}_0$, and an additional signal with another precision. It then follows by iterated expectations and the properties of maximisation that, from the viewpoint of the agent in

period 0, choosing a more precise signal necessarily leads to higher expected utility in period 1. In this way, a discount on bimodal rides acts in period 0 as a subsidy to the learning mechanism, and we would expect an increase in bimodal rides at period 0 due to this reason (in addition to income and substitution effects).