Corinne Emmenegger

Regularized Double Machine Learning in Partially Linear Models with Unobserved Confounding

Corinne Emmenegger[1,†,*]; Peter Bühlmann[2]

[1]Seminar for Statistics, ETH Zurich, Zurich, Switzerland, emmenegger@stat.math.ethz.ch
[2]Seminar for Statistics, ETH Zurich, Zurich, Switzerland, peter.buehlmann@stat.math.ethz.ch

Abstract:
Double machine learning (DML) can be used to estimate the linear coefficient in a partially linear model with confounding variables. However, the standard DML estimator has a two-stage least squares interpretation and may yield overly wide confidence intervals. To address this issue, we present the regularization-selection *regsDML* method that leads to narrower confidence intervals but preserves coverage guarantees. We rely on DML to estimate nuisance parameters with arbitrary machine learning algorithms and combine it with a regularization and selection scheme. Our regsDML method is fully data driven and optimizes the estimated asymptotic mean squared error of the coefficient estimate. The regsDML estimator can be expected to converge at the parametric rate and to follow an asymptotic Gaussian distribution. Empirical examples demonstrate our theoretical and methodological developments. Software code for the regsDML method is available in the R-package `dmlalg`.

Keywords:
Double machine learning, endogenous partially linear model, regularization, semiparametric estimation, two-stage least squares.

1. Introduction:

We consider a structural equation model (SEM) whose equation of the response is given by the endogenous partially linear model (PLM)

$$Y \leftarrow X^T \beta_0 + g_Y(W) + h_Y(H) + \varepsilon_Y. \tag{1}$$

This PLM combines flexibility of the nonparametric function of W with ease of interpretation of the linear term of X. The variable H in (1) is unobserved and introduces endogeneity if it correlates with X and W. A common approach to cope with endogeneity uses two-stage least squares (TSLS) (Angrist et al., 1996) with an instrumental variable that does not appear on the right hand side of (1), which we call A. The variable ε_Y in (1) denotes a random error.
Chernozhukov et al. (2018) introduced "standard" double machine learning (DML) to estimate β_0 in a model similar to (1). The central ingredients are Neyman orthogonality and sample splitting with cross-fitting. These ingredients allow potentially biased machine learning (ML) estimates of

[†]We thank Matthias Löffler for constructive comments. The research of C. Emmenegger and P. Bühlmann was supported by the European Research Council under the Grant Agreement No 786461 (CausalStats - ERC-2017-ADG).

nuisance terms to be plugged into the estimating equation of β_0. Emmenegger and Bühlmann (2021) refine this DML procedure in the PLM (1). They only require the identifiability condition

$$\mathbb{E}\left[R_A(R_Y - R_X^T\beta_0)\right] = \mathbf{0} \tag{2}$$

for the adjusted variables $R_A := A - \mathbb{E}[A|W]$, $R_X := X - \mathbb{E}[X|W]$, and $R_Y := Y - \mathbb{E}[Y|W]$ instead of conditional moment restrictions. Moreover, the dimension of A may exceed the dimension of X in Emmenegger and Bühlmann (2021). This overidentification can lead to more efficient and more robust estimators.
Both DML estimators of β_0 in Chernozhukov et al. (2018) and Emmenegger and Bühlmann (2021) are asymptotically Gaussian distributed and converge at the parametric rate although ML algorithms are used to learn the nuisance terms. However, both have a TSLS interpretation. In TSLS estimation, the strength of the instruments can lead to nonexisting variance and overly wide confidence intervals. K-class estimators can sometimes reduce this variance (Theil, 1961; Rothenhäusler et al., 2021; Jakobsen and Peters, 2020). There is a large literature on the presented concepts, and more references are given in Emmenegger and Bühlmann (2021).
We present the regularization-selection method regsDML (Emmenegger and Bühlmann, 2021) to reduce the potentially excessive standard deviation of DML. The reg**s**DML estimator **s**elects either the DML estimator or its regularized version regDML, depending on which one has a smaller standard deviation. The regularization parameter of regsDML is data driven. The coefficient estimator can be expected to converge at the parametric rate and to follow an asymptotic Gaussian distribution. The regsDML method focuses on statistical inference beyond point estimation with coverage guarantees in potentially complex partially linear models.

Overview: Section 2 presents the regularization schemes and supporting theory. Section 3 applies our method in a simulation study and a real data experiment. Section 4 concludes our work.

2. Methodology:

The regularization-only estimator regDML is obtained by regularizing DML and choosing a data-driven regularization parameter. Subsequently, we introduce it for a fixed regularization parameter $\gamma \geq 0$. Consider the operator $P_{R_A}(\cdot) := \mathbb{E}\left[\,\cdot\, R_A^T\right]\mathbb{E}\left[R_A R_A^T\right]^{-1} R_A$ that projects linearly onto the adjusted term $R_A = A - \mathbb{E}[A|W]$. Recall the adjustments $R_X = X - \mathbb{E}[X|W]$ and $R_Y = Y - \mathbb{E}[Y|W]$. The regularized population coefficient b^γ optimizes

$$b^\gamma := \underset{\beta\in\mathbb{R}^d}{\arg\min}\, \mathbb{E}\left[\left((\mathrm{Id} - P_{R_A})(R_Y - R_X^T\beta)\right)^2\right] + \gamma\,\mathbb{E}\left[\left(P_{R_A}(R_Y - R_X^T\beta)\right)^2\right], \tag{3}$$

where Id denotes the identity operator. This objective function is form-wise analogous to the one used in anchor regression (Rothenhäusler et al., 2021) or K-class regression (Theil, 1961). If $\gamma = 1$, ordinary least squares (OLS) of R_Y on R_X is performed. If $\gamma = 0$, then R_A is partialled out or adjusted for. If $\gamma = \infty$, TSLS of R_Y on R_X with the instrument R_A is performed, and b^γ coincides with β_0 from (1). If a general $\gamma > 1$ is considered, b^γ interpolates between OLS and TSLS.

We estimate b^γ with double machine learning. Let N iid observations $\{S_i = (A_i, X_i, W_i, Y_i)\}_{i\in[N]}$ of $S = (A, X, W, Y) \in \mathbb{R}^{q+d+v+1}$ from the SEM (1). They are concatenated row-wise into $\boldsymbol{A} \in \mathbb{R}^{N\times q}$, $\boldsymbol{X} \in \mathbb{R}^{N\times d}$, $\boldsymbol{W} \in \mathbb{R}^{N\times v}$, and $\boldsymbol{Y} \in \mathbb{R}^N$. The dimension v of W may grow with N, but the dimensions q, d, and 1 of A, X, and Y, respectively, are fixed. Also β_0 is of fixed dimension d.

First, the data is split into $K \geq 2$ disjoint sets $I_1, \ldots, I_K$. For simplicity, we assume these sets are of equal cardinality $n = \frac{N}{K}$, but the cardinalities may differ in practice due to rounding.
The conditional expectations $m_A^0(W) := \mathbb{E}[A|W]$, $m_X^0(W) := \mathbb{E}[X|W]$, and $m_Y^0(W) := \mathbb{E}[Y|W]$ in R_A, R_X, and R_Y, respectively, act as nuisance parameters and are estimated with ML algorithms. For each $k \in [K]$, they are estimated by $\hat{m}_A^{I_k^c}$, $\hat{m}_X^{I_k^c}$, and $\hat{m}_Y^{I_k^c}$, respectively, with data from the complement I_k^c of I_k. Then, the adjustments $\widehat{R}_{A,i}^{I_k} := A_i - \hat{m}_A^{I_k^c}(W_i)$, $\widehat{R}_{X,i}^{I_k} := X_i - \hat{m}_X^{I_k^c}(W_i)$, and $\widehat{R}_{Y,i}^{I_k} := Y_i - \hat{m}_Y^{I_k^c}(W_i)$ for $i \in I_k$ are evaluated on I_k. These adjustments are concatenated row-wise into $\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k} \in \mathbb{R}^{n \times q}$, $\widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k} \in \mathbb{R}^{n \times d}$, and $\widehat{\boldsymbol{R}}_{\boldsymbol{Y}}^{I_k} \in \mathbb{R}^n$, respectively.
The K iterates are assembled to form the estimator

$$\hat{b}^\gamma := \left(\frac{1}{K} \sum_{k=1}^{K} (\widehat{\boldsymbol{R}}_{\widetilde{\boldsymbol{X}}}^{I_k})^T \widehat{\boldsymbol{R}}_{\widetilde{\boldsymbol{X}}}^{I_k} \right)^{-1} \frac{1}{K} \sum_{k=1}^{K} (\widehat{\boldsymbol{R}}_{\widetilde{\boldsymbol{X}}}^{I_k})^T \widehat{\boldsymbol{R}}_{\widetilde{\boldsymbol{Y}}}^{I_k} \tag{4}$$

of b^γ, where $\widehat{\boldsymbol{R}}_{\widetilde{\boldsymbol{X}}}^{I_k} := (\mathbb{1} + (\sqrt{\gamma} - 1)\Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}}) \widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k}$ and $\widehat{\boldsymbol{R}}_{\widetilde{\boldsymbol{Y}}}^{I_k} := (\mathbb{1} + (\sqrt{\gamma} - 1)\Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}}) \widehat{\boldsymbol{R}}_{\boldsymbol{Y}}^{I_k}$ are projected terms, where $\Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}} := \widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k} \left((\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k})^T \widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k} \right)^{-1} (\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k})^T$ denotes the orthogonal projection matrix onto the column space of $\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}$, and where $\mathbb{1}$ denotes the identity matrix. Therefore, $\hat{b}^\gamma$ is obtained from a finite sample version of (3) by replacing P_{R_A} by $\Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}}$.
The closed-form expression (4) of $\hat{b}^\gamma$ resembles an OLS scheme. It is also possible to perform K individual OLS regressions of $\widehat{\boldsymbol{R}}_{\boldsymbol{Y}}^{I_k}$ on $\widehat{\boldsymbol{R}}_{\widetilde{\boldsymbol{X}}}^{I_k}$ for $k \in [K]$ and average the resulting coefficients. Both schemes are asymptotically equivalent, but the presented scheme enhances stability of $\hat{b}^\gamma$.

The K sample splits are random. To reduce the effect of this randomness, the overall procedure is repeated $\mathcal{S}$ times, and a correction term is added to the variance estimates to account for the random splitting. The results are assembled with the median as suggested in Chernozhukov et al. (2018). A summary of this procedure is given in Algorithm 1 in Emmenegger and Bühlmann (2021).

We have the following description of the asymptotic behavior of the estimator $\hat{b}^\gamma$.

Theorem 2.1. *(Emmenegger and Bühlmann, 2021, Theorem 4.1) Let $\gamma \geq 0$. Suppose Assumption G.5 in Emmenegger and Bühlmann (2021) holds, and denote the true nuisance parameter by $\eta^0 := (m_A^0, m_X^0, m_Y^0)$. Then, $\hat{b}^\gamma$ is approximately linear and centered Gaussian. More precisely,*

$$\sqrt{N}\sigma^{-1}(\gamma)(\hat{b}^\gamma - b^\gamma) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \overline{\psi}(S_i; b^\gamma, \eta^0) + o_P(1) \xrightarrow{d} \mathcal{N}(0, \mathbb{1}_{d \times d}) \quad (N \to \infty)$$

for some variance-covarince matrix $\sigma(\gamma)$ and some $\overline{\psi}$ uniformly over laws P of $S = (A, W, X, Y)$.

Assumption G.5 in Emmenegger and Bühlmann (2021) specifies regularity conditions. The L^2-norms of the ML estimation errors need to decay fast enough: for $k \in [K]$, the error terms $e_A^k := \|m_A^0(W) - \hat{m}_A^{I_k^c}(W)\|_2$, $e_X^k := \|m_X^0(W) - \hat{m}_X^{I_k^c}(W)\|_2$, and $e_Y^k := \|m_Y^0(W) - \hat{m}_Y^{I_k^c}(W)\|_2$ need to satisfy the product relations $(e_A^k)^2 \ll N^{-\frac{1}{2}}$, $e_X^k(e_Y^k + e_X^k) \ll N^{-\frac{1}{2}}$, and $e_A^k(e_Y^k + e_X^k) \ll N^{-\frac{1}{2}}$. This allows the individual error terms to be of larger order than $N^{-\frac{1}{2}}$. In particular, they may be of order $N^{-\frac{1}{4}}$, which allows us to use almost arbitrary ML algorithms; see Chernozhukov et al. (2018).
The asymptotic variance $\sigma^2(\gamma)$ in Theorem 2.1 can be consistently estimated (Emmenegger and Bühlmann, 2021, Theorem H.3). For $\gamma < \infty$, the asymptotic variance $\sigma^2(\gamma)$ is typically smaller than the one of the DML estimator.

Furthermore, the proof of Theorem 2.1 uses Neyman orthogonality of the underlying score functions, which makes them insensitive to inserting biased ML estimators of the nuisance parameters. Our score functions are Neyman orthogonal because their Gateaux derivative vanishes at η^0. This property neither depends on the distribution of S nor on the true unknown β_0 and η^0.

Subsequently, we introduce a data-driven scheme to choose the regularization parameter γ, and we present the regularization-selection scheme regsDML.
For simplicity, we assume $d = 1$. We choose the data-driven regularization parameter estimator

$$\hat{\gamma}' := a_N \cdot \operatorname*{arg\,min}_{\gamma \geq 0} \frac{1}{N} \hat{\sigma}^2(\gamma) + |\hat{b}^\gamma - \hat{\beta}|^2,$$

which optimizes the estimated asymptotic mean squared error (MSE) of $\hat{b}^\gamma$. The term $\hat{\sigma}^2(\gamma)$ is the consistent estimator of $\sigma^2(\gamma)$ from Emmenegger and Bühlmann (2021, Theorem H.3). The term $|\hat{b}^\gamma - \hat{\beta}|^2$ is a plug-in estimator of the squared population bias $|b^\gamma - \beta_0|^2$, where the DML estimator $\hat{\beta}$ from Emmenegger and Bühlmann (2021) replaces β_0. The deterministic multiplication factor a_N may be chosen arbitrarily, but it needs to diverge to $+\infty$ as $N \to \infty$. However, we observed that $a_N = \log(\sqrt{N})$ works well in practice. This multiplicative factor ensures that the population bias term $|b^{\hat{\gamma}'} - \beta_0|$ with $\hat{\gamma}'$ vanishes at the rate $o_P(N^{-\frac{1}{2}})$. Thus, we can argue that $\hat{b}^{\hat{\gamma}'}$ is approximately Gaussian distributed, which allows us to construct approximately valid confidence intervals. More precisely, we expect $\sqrt{N}(\hat{b}^{\hat{\gamma}'} - \beta_0) \approx \mathcal{N}\big(0, \sigma^2(\hat{\gamma}')\big)$ whenever N is sufficiently large; see Emmenegger and Bühlmann (2021). This argument is not entirely rigorous because $\hat{\gamma}'$ is estimated from all the data. Both $\hat{b}^{\hat{\gamma}'}$ and $\hat{\beta}$ have the same asymptotic MSE behavior, but $\hat{b}^{\hat{\gamma}'}$ may exhibit substantially better finite sample properties.

We call $\hat{b}^{\hat{\gamma}'}$ the regDML (regularized DML) estimator. The regularization-selection method reg**s**DML **s**elects between the DML estimator $\hat{\beta}$ and regDML based on whose variance is smaller. It can be expected that regsDML concentrates in a $N^{-\frac{1}{2}}$ neighborhood of β_0 and asymptotically follows a Gaussian distribution as does the DML estimator $\hat{\beta}$.

3. Result:

This section illustrates the performance of regsDML, regDML, and DML in a simulation study and on a real dataset. We use $K = 2$ sample splits and $\mathcal{S} = 100$ overall repetitions and estimate the nuisance parameters with random forests. Our implementation is available in the R-package `dmlalg` (Emmenegger, 2021).

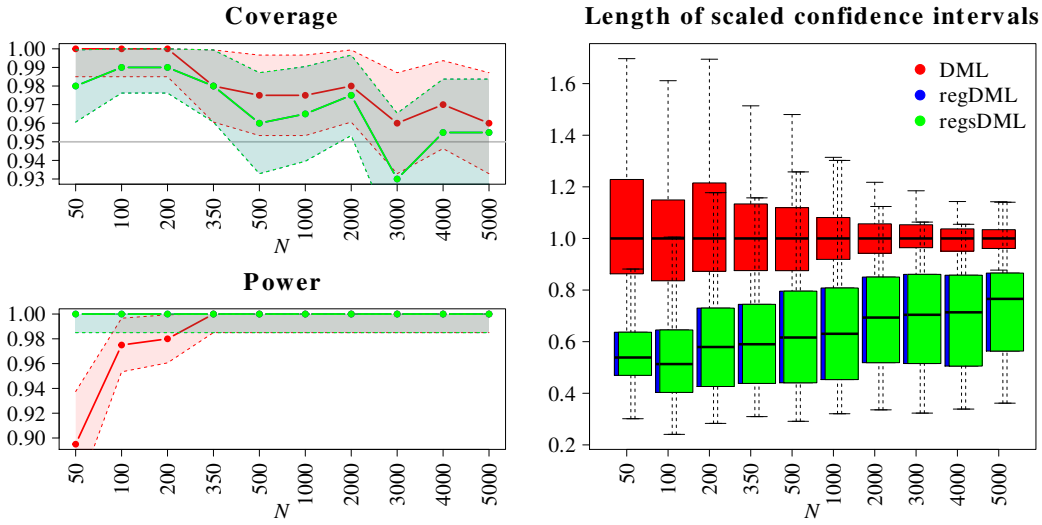
First, we consider a simulation study. We generate data from the overidentified SEM ($\beta_0 = 1$)

$$\begin{aligned}
&(\varepsilon_{A_1}, \varepsilon_{A_2}, \varepsilon_{W_1}, \varepsilon_{W_2}, \varepsilon_H, \varepsilon_X, \varepsilon_Y) \sim \mathcal{N}_7(\mathbf{0}, \mathbb{1}), && W_2 \leftarrow \varepsilon_{W_2}, \\
&A_1 \leftarrow \mathbb{1}_{\{\varepsilon_{A_1} \leq 0\}}, && H \leftarrow 2\mathbb{1}_{\{\sin(\pi W_1) \cdot \tanh(W_2) \geq 0\}} + \varepsilon_H, \\
&A_2 \leftarrow -4A_1 + \varepsilon_{A_2}, && X \leftarrow \tfrac{3}{2}A_1 - \tfrac{1}{2}A_2 + \tanh(H) - 2\mathbb{1}_{\{W_1 \geq 0, W_2 \leq 0\}} + \varepsilon_X, \\
&W_1 \leftarrow 2A_2 + \varepsilon_{W_1}, && Y \leftarrow X + \mathbb{1}_{\{W_2 \leq 0\}} + \sin(\pi H) + \varepsilon_Y
\end{aligned} \tag{5}$$

from Emmenegger and Bühlmann (2021). This SEM contains step functions and interaction terms. The identifiability condition (2) is satisfied, so that DML is asymptotically Gaussian.

Figure 1 illustrates our simulation results. The lengths of the regsDML confidence intervals are about 50% to 80% the lengths of DML's. Nevertheless, the coverage of regsDML remains around the nominal 95% level. No coverage region falls below the 95% level marked by the gray line. The power of DML is lower for small sample sizes N. As N increases, regsDML starts to resemble DML's behavior but continues to produce shorter confidence intervals. Thus, regsDML (and also its regularization-only version regDML) is a highly effective method to increase the power and sharpness of statistical inference and to keep the type I error and coverage under control.
Simulation results with $\beta_0 = 0$ in the SEM (5) are similar and presented in Figure 11 in Emmenegger and Bühlmann (2021).

Figure 1: (Emmenegger and Bühlmann, 2021, Figure 8) The results come from 200 simulation runs from the SEM (5) for a range of sample sizes N with $K = 2$ sample splits and $\mathcal{S} = 100$ overall repetitions. The nuisance parameters are estimated with random forests consisting of 500 trees whose minimal node size is 5. The figure displays the coverage of two-sided confidence intervals for β_0, power for two-sided testing of the hypothesis $H_0 : \beta_0 = 0$, and scaled lengths of two-sided confidence intervals of DML (red), regDML (blue), and regsDML (green), all at level 95%. At each N, the lengths of the confidence intervals are scaled with the median length from DML. The shaded regions in the coverage and the power plots represent 95% confidence bands with respect to the 200 simulation runs. The blue and green lines are indistinguishable in the left panel.



Second, we consider a real data example. We estimate the linear effect β_0 of institutions on the economic performance of countries following the work of Acemoglu et al. (2001), Chernozhukov et al. (2018), and Emmenegger and Bühlmann (2021). We adjust nonlinearly for the latitude and some binary geographic information. Simultaneity may be present because countries with better institutions achieve a greater level of income and vice versa. However, mortality rates of the first European settlers serve as a source of exogenous variation in institutions. The dataset contains $N = 64$ observations. Please see Emmenegger and Bühlmann (2021) for further details.
We use $K = 2$ sample splits and $\mathcal{S} = 100$ overall repetitions and estimate the nuisance parameters with random forests consisting of 1000 trees whose minimal node size is 5. The DML point estimator of β_0 is 0.739, its standard deviation is 0.459, and its two-sided 95% confidence interval is $[-0.161, 1.639]$. The regsDML point estimator of β_0 is 0.688, its standard deviation is 0.229, and its two-sided 95% confidence interval is $[0.239, 1.136]$. The DML estimate is not significant because

its confidence interval contains 0. The regsDML estimate is significant and has a smaller standard deviation than DML. Note that regsDML falls within the DML confidence interval.
Chernozhukov et al. (2018) use the same machine learners to analyze this data, but obtain a significant DML point estimator due to a smaller standard deviation. However, they implicitly assume an additional homoscedasticity condition on the errors $R_Y - R_X^T\beta_0$, which is questionable.

4. Discussion and Conclusion:

We regularized double machine learning (DML) in overidentified partially linear models (PLMs) with endogenous variables to perform inference for the linear parameter. Standard DML methods can lead to overly wide confidence intervals due to their two-stage least squares (TSLS) interpretation. This effect is particularly pronounced if the confounding is strong.
We presented a regularization scheme, regDML, and a regularization-selection scheme, regsDML. The latter selects between DML and regDML depending on whose standard deviation is smaller. For finite sample sizes, regsDML leads to drastically shorter confidence intervals than DML. Nevertheless, coverage guarantees for β_0 remain: both regDML and regsDML are expected to concentrate in an $N^{-\frac{1}{2}}$ neighborhood of β_0 and to follow a Gaussian distribution asymptotically.
Depending on the strength of the confounding, regsDML may inherit additional bias from the biased DML estimator. Emmenegger and Bühlmann (2021, Section E) present examples with strong and reduced confounding to demonstrate the coverage behavior of DML and regsDML.
Although a wide range of machine learners can be employed to estimate the nuisance parameters, additive splines can estimate more precise results than random forests if the underlying structure is additive in good approximation, especially if the sample size is small.
The regsDML methodology can be used with the implementation that is available in the R-package `dmlalg` (Emmenegger, 2021).

References

D. Acemoglu, S. Johnson, and J. A. Robinson. The colonial origins of comparative development: An empirical investigation. *The American Economic Review*, 91(5):1369–1401, 2001.

J. D. Angrist, G. W. Imbens, and D. B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

C. Emmenegger. *dmlalg: Double machine learning algorithms*, 2021. R-package to be published.

C. Emmenegger and P. Bühlmann. Regularizing double machine learning in partially linear endogenous models, 2021. Preprint arXiv:2101.12525.

M. E. Jakobsen and J. Peters. Distributional robustness of K-class estimators and the PULSE, 2020. Preprint arXiv:2005.03353.

D. Rothenhäusler, N. Meinshausen, P. Bühlmann, and J. Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(2):215–246, 2021.

H. Theil. *Economic forecasts and policy*, volume 15 of *Contributions to economic analysis*. North-Holland Publishing Company, Amsterdam, 2 edition, 1961.