Distributional robustness of K-class estimators and the PULSE

Jakobsen, Martin Emil; Peters, Jonas

Published in:
The Econometrics Journal

DOI:
10.1093/ectj/utab031

Publication date:
2022

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Distributional robustness of K-class estimators and the PULSE

MARTIN EMIL JAKOBSEN AND JONAS PETERS

Department of Mathematical Sciences, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen, Denmark.

Email: m.jakobsen@math.ku.dk, jonas.peters@math.ku.dk

First version received: 1 March 2021; final version accepted: 29 July 2021.

Summary: While causal models are robust in that they are prediction optimal under arbitrarily strong interventions, they may not be optimal when the interventions are bounded. We prove that the classical K-class estimator satisfies such optimality by establishing a connection between K-class estimators and anchor regression. This connection further motivates a novel estimator in instrumental variable settings that minimizes the mean squared prediction error subject to the constraint that the estimator lies in an asymptotically valid confidence region of the causal coefficient. We call this estimator PULSE (p-uncorrelated least squares estimator), relate it to work on invariance, show that it can be computed efficiently, as a data-driven K-class estimator, even though the underlying optimization problem is nonconvex, and prove consistency. We evaluate the estimators on real data and perform simulation experiments illustrating that PULSE suffers from less variability. There are several settings, including weak instrument settings, where it outperforms other estimators.

Keywords: Causality, distributional robustness, instrumental variables.

JEL codes: C01, C20, C26.

1. INTRODUCTION

Learning causal parameters from data has been a key challenge in many scientific fields and has been a long-studied problem in econometrics (e.g., Simon, 1953; Wold, 1954; Goldberger, 1972). Many years after the groundbreaking work by Peirce (1883) and Fisher (1935), causality again plays an increasingly important role in machine learning and statistics, two research areas that are most often considered part of mathematics or computer science (e.g., Spirtes et al., 2000; Pearl, 2009; Imbens and Rubin, 2015; Peters et al., 2017). Even though current research in mathematics and computer science on the one hand and econometrics on the other, does not develop independently, we believe that there is a lot of potential for more fruitful interaction between these two fields. Differences in the language have emerged, which can make communication difficult, but the target of inference, the underlying principles, and the methodology in both fields are closely related. This paper establishes a link between two developments in these fields: K-class estimation, which aims at estimation of causal parameters with good statistical properties, and invariance principles, which are used to build methods that are robust with respect to distributional shifts. This connection allows us to prove distributional robustness guarantees for K-class estimators and motivates a new estimator, PULSE. We summarize our main results in Section 1.2.
1.1. Related work

Given causal background knowledge, causal parameters can be estimated when taking into account confounding effects between treatment and outcome. Several related techniques have been suggested to tackle that problem, including variable adjustment (Pearl, 2009), propensity score matching (Rosenbaum and Rubin, 1983), inverse probability weighting (Horvitz and Thompson, 1952), and G-computation (Robins, 1986).

If some of the relevant variables have not been observed, one may instead use exogenous variation in the data to infer causal parameters; for instance, in the setting of instrumental variables (e.g., Wright, 1928; Imbens and Angrist, 1994; Newey, 2013; Wang and Tchetgen, 2018). Limited information estimators leverage instrumental variables to conduct single equation inference. An example of such methods is the two-stage least squares estimators (TSLS) developed by Theil (1953). Instead of minimizing the residual sum of squares as done by the ordinary least square (OLS) estimator, the TSLS minimizes the sample covariance between the instruments and regression residuals. TSLS estimators are consistent, but are known to have suboptimal finite sample properties, e.g., they only have moments up to the degree of over-identification (Mariano, 1972). Kadane (1971) shows that under suitable conditions, the mean squared error of TSLS might even be larger than the one of OLS if the sample size is small (more precisely and using the notation introduced below, if \(0 \leq n - q \leq 2(3 - (q_2 - d_1))\), where \(q_2 - d_1\) is the degree of over-identification). This result is another indication that under certain conditions, it might be beneficial to use the OLS for regularization. Another method of inferring causal parameters in structural equation models (SEMs) is the limited information maximum likelihood (LIML) estimator due to Anderson and Rubin (1949). Theil (1958) introduced K-class estimators, which contain OLS, TSLS, and the LIML estimator as special cases. This class of estimators is parametrized by a deterministic or stochastic parameter \(\kappa \in [0, \infty)\) that depends on the observational data. Under mild regularity conditions a member of this class is consistent and asymptotically normally distributed if \((\kappa - 1)\) and \(\sqrt{n(\kappa - 1)}\) converge, respectively, to zero in probability when \(n\) tends to infinity; see, e.g., Mariano (1975, 2001). While the LIML does not have moments of any order, it shares the same asymptotic normal distribution with TSLS. Based on simulation studies, Anderson (1983) argues that, in many practically relevant cases, the normal approximation to a finite-sample estimator is inadequate for TSLS but a useful approximation in the case of LIML. Using Monte Carlo simulations, Hahn et al. (2004) recommend that the no-moment estimator LIML should not be used in weak instrument situations, where Fuller estimators have a substantially smaller mean squared error (MSE). The Fuller estimators (Fuller, 1977) form a subclass of the K-class estimators based on a modification to the LIML, which fixes the no-moment problem while maintaining consistency and asymptotic normality. Kiviet (2020) proposes a modification to the OLS estimator that makes use of explicit knowledge of the partial correlation between the covariates and the unobserved noise in \(Y\). Andrews and Armstrong (2017) propose an unbiased estimator, based on knowledge of (1) the sign of the first stage regression and (2) the covariance matrix of the reduced form errors, that is less dispersed than TSLS. Judge and Mittelhammer (2012) consider an affine combination of the OLS and TSLS estimators, which, again, yields a modification in the space of estimators. We prove that our proposed estimator, PULSE, can also be written as a data-driven K-class estimator. As such, it minimizes a convex combination of the OLS and TSLS loss functions and can, in general, not be written as a convex combination of the estimators.

All of the above methods exploit background knowledge; for example, in the form of exogeneity of some of the variables. If no such background knowledge is available, it may still be possible,
under additional assumptions, to infer the causal structure, e.g., represented by a graph, from observational (or observational and interventional) data. This problem is sometimes referred to as causal discovery. Constraint-based methods assume that the underlying distribution is Markov and faithful with respect to the causal graph and perform conditional independence tests to infer (parts of) the graph; see, e.g., Spirtes et al. (2000). Score-based methods assume a certain statistical model and optimize (penalized) likelihood scores; see, e.g., Chickering (2002). Some methods exploit a simple form of causal assignments, such as additive noise (e.g., Shimizu et al., 2006; Peters et al., 2014) and others are based on exploiting invariance statements (e.g., Meinshausen et al., 2016; Peters et al., 2016). Many of such methods assume causal sufficiency, i.e., that all causally relevant variables have been observed, but some versions exist that allow for hidden variables; see, e.g., Claassen et al. (2013) and Spirtes et al. (1995).

Recent works in the fields of machine learning and computational statistics (e.g., Schölkopf et al., 2012; Heinze-Deml and Meinshausen, 2017; Pfister et al., 2019) investigate whether causal ideas can help to make machine learning methods more robust. The reasoning is that causal models are robust against any intervention in the following sense. Consider a target or response variable $Y$ and covariates $X_1, \ldots, X_p$. If we regress $Y$ on the set $X_S, S \subseteq \{1, \ldots, p\}$, of direct causes, then this regression function $x \mapsto E[Y|X_S = x]$ does not change when intervening on any of the covariates (which is sometimes referred to as ‘invariance’). This statement can be proved using the local Markov property (Lauritzen, 1996), for example, but the underlying fundamental principle has been discussed already several decades ago; most prominently using the terms ‘autonomy’ or ‘modularity’ (Haavelmo, 1944; Aldrich, 1989). As a result, causal models of the form $x \mapsto E[Y|X_S = x]$ may perform well in prediction tasks, where, in the test distribution, the covariates have been intervened on. If, however, training and test distributions coincide, a model focusing only on prediction and the estimand $x \mapsto E[Y|X = x]$ may outperform a causal approach.

The two models described above (OLS and the causal model) formally solve a minimax problem on distributional robustness. Consider therefore an acyclic linear SEM over $(Y, X)$ with observational distribution $F$. Details on SEMs and interventions can be found in Online Appendix S1. Assume that the assignment for $Y$ equals $Y := \gamma_0^T X + \varepsilon_Y$ for some $\gamma_0 \in \mathbb{R}^d$. The variables corresponding to nonzero entries in $\gamma_0^T X$ are called the parents of $Y$, and $\varepsilon_Y$ is assumed to be independent of these parents. Then, the mean squared prediction error when considering the observational distribution is not necessarily minimized by $\gamma_0$, that is, in general, we have $\gamma_0 \neq \gamma_{\text{OLS}} := \arg \min_{\gamma} E_F \left[ (Y - \gamma^T X)^2 \right]$. Intuitively, we may improve the prediction of $Y$ by including other variables than the parents of $Y$, such as its descendants. When considering distributional robustness, we are interested in finding a $\gamma$ that minimizes the worst case expected squared prediction error over a class of distributions, $\mathcal{F}$, that is,

$$\arg \min_{\gamma} \sup_{F \in \mathcal{F}} E_F \left[ (Y - \gamma^T X)^2 \right].$$

(1.1)

If we observe data from all different distributions in $\mathcal{F}$ (and know which data point comes from which distribution), we can tackle this optimization directly (Meinshausen and Bühlmann, 2015). But estimators of equation (1.1) may be available even if we do not observe data from each distribution in $\mathcal{F}$. The true causal coefficient $\gamma_0$, for example, minimizes equation (1.1) when $\mathcal{F}$ is the set of all possible (hard) interventions on $X$ (e.g., Rojas-Carulla et al., 2018). The OLS solution is optimal when $\mathcal{F}$ only contains the training distribution. In this sense, the OLS solution and the true causal coefficient constitutes the end points of a spectrum of estimators that are prediction optimal under a certain class of distributions.
Intuitively, models trading off causality and predictability may perform well in situations where the test distribution is only moderately different from the training distribution. Anchor regression by Rothenhäuser et al. (2021) is one approach formalizing this intuition in a linear setup; see Section 2.2 for details. Similarly to an instrumental variable setting, one assumes the existence of exogenous variables called $A$ (for anchor) that may or may not act directly on the target $Y$. The proposed estimator minimizes a convex combination of the residual sum of squares and the TSLS loss function and is shown to be prediction optimal in the sense of equation (1.1) for a class $\mathcal{F}$ containing interventions on the covariates up to a certain strength; this strength depends on a regularization parameter: the weight that is used in the convex combination of anchor regression. Other approaches (Magliacane et al., 2018; Rojas-Carulla et al., 2018; Pfister et al., 2021) search over different subsets $S$ and aim to choose sets that are both invariant and predictive.

1.2. Summary and contributions

This paper contains two main contributions: a distributional robustness property of K-class estimators with fixed $\kappa$-parameter and a novel estimator for causal coefficients called the $p$-uncorrelated least squares estimator (PULSE). The following two sections summarize our contributions.

1.2.1. Distributional robustness of K-class estimators. In Section 2 we show that anchor regression is closely related to K-class estimators. In particular, we prove that, for a restricted subclass of models, K-class estimators can be written as anchor regression estimators. For this subclass, this directly implies a distributional robustness property of K-class estimators. We then prove a similar robustness property for general K-class estimators with a fixed penalty parameter, and show that these properties hold even if the model is misspecified.

Consider a possibly cyclic linear SEM over the variables $(Y, X, H, A)$ of the form

$$[Y \ X^\top \ H^\top] := [Y \ X^\top \ H^\top]B + A^\top M + \epsilon^\top,$$

subject to regularity conditions that ensure that the distribution of $(Y, X, H, A)$ is well-defined. Here, $B$ and $M$ are constant matrices, the random vectors $A$ and $\epsilon$ are defined on a common probability space $(\Omega, \mathcal{F}, P)$, $Y$ is the endogenous target for the single equation inference, $X$ are the observed endogenous variables, $H$ are hidden endogenous variables, and $A$ are exogenous variables independent from the unobserved noise innovations $\epsilon$.

SEMs allow for the notion of interventions, i.e., modelling external manipulations of the system. In this work, we are only concerned with interventions on the exogenous variables $A$ of the form $do(A := v)$. Because $A$ is exogeneous, these interventions can be defined as follows: they change the distribution of $A$ to that of a random vector $v$. The interventional distribution of the variables $(Y, X, H, A)$ under the intervention $do(A := v)$ is given by the simultaneous distribution of $(X_v, Y_v, H_v, v)$ generated by the SEM

$$[Y_v \ X_v^\top \ H_v^\top] := [Y_v \ X_v^\top \ H_v^\top]B + v^\top M + \epsilon^\top.$$

Thus, the intervention does not change any of the original structural assignments of the endogenous variables. Instead, the change in the distribution of the exogeneous variable propagates through the system. We henceforth let $E^{do(A := v)}$ denote the expectation with respect to the interventional distribution of the system under the intervention $do(A := v)$. More details on interventions can be found in Online Appendix S1.

© The Author(s) 2022.
Let \((Y, X, H, A)\) consist of \(n\) row-wise independent and identically distributed copies of the random vector \((Y, X, H, A)\) and consider the single equation of interest
\[
Y = X\gamma_0 + A\beta_0 + H\eta_0 + \varepsilon_Y = X\gamma_0 + A\beta_0 + \tilde{U}_Y.
\]
The K-class estimator with parameter \(\kappa\) using nonsample information that only \(Z_n \subset [X A]\) have nonzero coefficients in the target equation of interest is given by
\[
\hat{\theta}_n^\kappa(\kappa) = (Z_n^\kappa(I - \kappa P_A^\perp)Z_n)^{-1}Z_n^\kappa(I - \kappa P_A^\perp)Y,
\]
where \(P_A^\perp\) is the projection onto the orthogonal complement of the column space of \(A\). For a fixed \(\kappa \in [0, 1)\) K-class estimators can be represented by a penalized regression problem
\[
\hat{\theta}_n^\kappa(\kappa) = \arg\min_{\alpha} l_{\text{OLS}}(\alpha) + \kappa/(1 - \kappa) l_{\text{IV}}(\alpha),
\]
where \(l_{\text{OLS}}\) and \(l_{\text{IV}}\) are the empirical OLS and TSLS loss functions, respectively. This representation and the ideas of Rothenhäusler et al. (2021) allow us to prove that K-class estimator converges to a coefficient that is minimax optimal when considering all distributions induced by a certain set of interventions of \(A\). More specifically, we show that for a fixed \(\kappa\) and regardless of identifiability,
\[
\hat{\theta}_n^\kappa(\kappa) \toP \arg\min_{\alpha} \sup_{v \in C(\kappa)} E_{d(A; v)}\left[(Y - \alpha^T Z_n)^2\right],
\]
where \(C(\kappa) := \{v : \Omega \to \mathbb{R}^n : \text{Cov}(v, \varepsilon) = 0, E[vv^\top] \leq \frac{1}{1 - \kappa} E[AA^\top]\}\). The argmin on the right-hand side minimizes the worst case prediction error when considering interventions up to a certain strength (measured by the set \(C(\kappa)\)). This objective becomes relevant when we consider a response variable with several covariates and aim to minimize the mean squared prediction error of future realizations of the system of interest that do not follow the training distribution.

The above result says that if the new realizations correspond to (unknown) interventions on the exogenous variables that are of bounded strength, K-class estimators with fixed \(\kappa \in (0, 1)\) minimize the worst case prediction performance and, in particular, outperform the true causal parameter and the least squares solution (see also Figure S8.1 in Online Appendix S8.1). For \(\kappa\) approaching one, we recover the guarantee of the causal solution and for \(\kappa\) approaching zero, the set of distributions contains the training distribution. The above minimax property therefore adds to the discussion whether nonconsistent K-class estimators with penalty parameter not converging to one can be useful; see, e.g., Dhrymes (1974).

1.2.2. The PULSE estimator. Section 3 contains the second main contribution in this work. We propose a novel data-driven K-class estimator for causal coefficients, which, as mentioned above, we call PULSE. As above, we consider a single endogenous target in an SEM (or simultaneous equation model) and aim to predict it from observed predictors that are with a priori (nonsample) information known to be either endogenous or exogenous. The PULSE estimator can be written in several equivalent forms. It can, first, be seen as a data-driven K-class estimator
\[
\hat{\lambda}_n^*/(1 + \lambda_n^*) = \arg\min_{\alpha} l_{\text{OLS}}^n(\alpha) + \lambda_n^* l_{\text{IV}}^n(\alpha),
\]
where
\[
\lambda_n^* := \inf \left\{ \lambda > 0 : \text{testing Corr}(A, Y - Z \hat{\alpha}_n^\kappa(\lambda/(1 + \lambda))) = 0 \right\},
\]
for some pre-specified level of the hypothesis test \(p_{\text{min}} \in (0, 1)\). In other words, the PULSE estimator outputs the K-class estimator closest to the OLS while maintaining a nonrejected test of uncorrelatedness. In principle, PULSE can be used with any testing procedure. The choice of

© The Author(s) 2022.
test, however, may influence the difficulty of the resulting optimization problem. In this paper, we investigate PULSE in connection with a specific class of hypothesis tests that, for example, contain the test of Anderson and Rubin (1949). For these hypothesis tests we develop an efficient and provably correct optimization method that is based on binary line search and quadratic programming.

We show that our estimator can, second, be written as the solution to a constrained optimization problem. To that end, define the primal problems

\[
\hat{\alpha}_{\text{Pr}}(t) := \arg \min_{\alpha} \ln \text{OLS}(\alpha) \text{ subject to } \ln \text{IV}(\alpha) \leq t.
\]

For the choice \( t^*_n := \sup \{ t : \text{testing Corr}(A, Y - Z\hat{\alpha}_{\text{Pr}}(t)) = 0 \text{ yields a } p\text{-value } \geq p_{\min} \} \), we provide a detailed analysis proving that \( \hat{\alpha}_{\text{K}}(\lambda^*_n/(1 + \lambda^*_n)) = \hat{\alpha}_{\text{Pr}}(t^*_n) \).

For the testing procedure proposed in this paper, we show that, third, PULSE can be written as

\[
\arg \min_{\alpha} \ln \text{OLS}(\alpha; Y, Z) \text{ subject to } \alpha \in A_n(1 - p_{\min}),
\]

where \( A_n(1 - p_{\min}) \) is the nonconvex acceptance region for our test of uncorrelatedness.

This third formulation allows for a simple interpretation of our estimator: among all coefficients (not restricted to K-class estimators) that do not yield a rejection of uncorrelatedness, we choose the one that yields the best prediction. If the acceptance region is empty it outputs a warning indicating a possible model misspecification or an assumption violation to the user (in that case, one can formally output another estimator such as TSLS or Fuller, yielding PULSE well-defined).

In the just-identified setup, the TSLS estimator solves a normal equation which is equivalent to setting a sample covariance between the instruments and the resulting prediction residuals to zero; it then corresponds to \( t = 0 \). For this (and the over-identified) setting, we prove that PULSE is a consistent estimator for the causal coefficient.

The TSLS does not have a finite variance if there is insufficient degree of over-identification, for example. In particular for weak instruments, this usually comes with poor finite sample performance. In such cases, however, the acceptance region of uncorrelatedness is usually large. This yields a weak constraint in the optimization problem and the PULSE will be closer to the OLS, which in certain settings suffers from less variability (see, e.g., Hahn et al., 2004; Hahn and Hausman, 2005). In simulations we indeed see that, similarly to other data-driven K-class estimators that are pulled towards the OLS, such as Fuller estimators, the PULSE comes with beneficial finite sample properties compared to TSLS and LIML.

Unlike other estimators, such as LIML or the classical TSLS, the PULSE is well-defined in under-identified settings, too. Here, its objective is still to find the best predictive solution among all parameters that do not reject uncorrelatedness. Uncorrelatedness to the exogenous variable is sometimes referred to as invariance. The idea of choosing the best predictive among all invariant models has been investigated in several works (e.g., Magliacane et al., 2018; Rojas-Carulla et al., 2018; Pfister et al., 2021) with the motivation to find models that generalize well (in particular, with respect to interventions on the exogenous variables). Existing methods, however, focus on selecting subsets of variables and then consider least squares regression of the response variable onto the full subset. PULSE can recover such type of solutions if they are indeed optimal, but it also allows us to search over coefficients that are different from least squares regression for sets of variables. Consequently, PULSE allows us to find solutions in situations where the above
methods would not find any invariant subsets, which may often be the case if there are hidden variables (see Online Appendix S8.3 for an example).

We show in a simulation study that there are several settings in which PULSE outperforms existing estimators both in terms of MSE ordering and several one-dimensional scalarizations of the MSE. More specifically, we show that PULSE can outperform the TSLS and Fuller estimators in weak instrument situations, for example, where Fuller estimators are known to have good MSE properties; see, e.g., Stock et al. (2002) and Hahn et al. (2004).

Implementation of PULSE and code for experiments (R) are available on GitHub.1

2. ROBUSTNESS PROPERTIES OF K-CLASS ESTIMATORS

In this section we consider K-class estimators (Theil, 1958; Nagar, 1959) and show a connection with anchor regression of Rothenhäusler et al. (2021). In Section 2.3.1 we establish the connection in models where we use a priori information that there are no included exogenous variables in the target equation of interest. In Section 2.3.2 we then show that general K-class estimators can be written as the solution to a penalized regression problem. In Section 2.3.3 we utilize this representation and the ideas of Rothenhäusler et al. (2021) to prove a distributional robustness guarantee of general K-class estimators with fixed \( \kappa \in [0,1) \), even under model misspecification and nonidentifiability. Proofs of results in this section can be found in Online Appendix S3.

2.1. Setup and assumptions

Denote the random vectors \( Y \in \mathbb{R}, X \in \mathbb{R}^d, A \in \mathbb{R}^q, H \in \mathbb{R}^r \) and \( \varepsilon \in \mathbb{R}^{d+1+r} \) by the target, endogenous regressor, anchors, hidden and noise variables, respectively. Let further \((Y, X, H, A)\) be generated by the possibly cyclic structural equation model (SEM)

\[
\begin{bmatrix} Y \\ X^T \\ H^T \end{bmatrix} := \begin{bmatrix} Y \\ X^T \\ H^T \end{bmatrix} B + A^T M + \varepsilon^T,}

(2.1)

for some random vectors \( \varepsilon \perp A \) and constant matrices \( B \) and \( M \). Let \((Y, X, H, A)\) consist of \( n \geq \min\{d, q\} \) row-wise independent and identically distributed copies of the random vector \((Y, X, H, A)\). Solving for the endogenous variables we get the structural and reduced form equations \(
[Y X H] \Gamma = AM + \varepsilon 
\)
and \([Y X H] = A \Pi + \varepsilon \Gamma^{-1}, \)
where \( \Gamma := I - B \) and \( \Pi := M \Gamma^{-1} \). Assume without loss of generality that \( \Gamma \) has a unity diagonal, such that the target equation of interest is given by

\[
Y = X \gamma_0 + A \beta_0 + H \eta_0 + \varepsilon_Y = Z \alpha_0 + \hat{U}_Y,}

(2.2)

where \((1, -\gamma_0, -\eta_0) \in \mathbb{R}^{(1+d+r)}, \beta_0 \in \mathbb{R}^q \) and \( \varepsilon_Y \) are the first columns of \( \Gamma \), \( M \) and \( \varepsilon \in \mathbb{R}^{d+q} \) respectively, \( Z := [X A], \alpha_0 = (\gamma_0, \beta_0) \in \mathbb{R}^{d+q} \) and \( \hat{U}_Y := H \eta_0 + \varepsilon_Y \).

The possible dependence between the noise \( \hat{U}_Y \) and the endogenous variables, i.e., the influence by hidden variables, generally renders the standard OLS approach for estimating \( \alpha_0 \) inconsistent. Instead, one can make use of the components in \( A \) that have vanishing coefficient in equation (2.2) for consistent estimation. In the remainder of this work, we disregard any a priori (nonsample) information not concerning the target equation. The question of identifiability of \( \alpha_0 \) has been studied extensively (Frisch, 1938; Haavelmo, 1944; Koopmans et al., 1950) and more recent overviews can be found in, e.g., Fisher (1966), Greene (2003), and Didelez et al. (2010).

1 https://github.com/MartinEmilJakobsen/PULSE.
We will use the following assumptions concerning the structure of the SEM:

Assumption 2.1 (Global Assumptions). (a) \((Y, X, H, A)\) is generated in accordance with the SEM in equation (2.1); (b) \(\rho(B) < 1\) where \(\rho(B)\) is the spectral radius of \(B\); (c) \(\epsilon\) has jointly independent marginals \(\epsilon_1, \ldots, \epsilon_d + 1, \ldots\); (d) \(A\) and \(\epsilon\) are independent; (e) No variable in \(Y, X\) and \(H\) is an ancestor of \(A\), that is, \(A\) is exogenous; (f) \(E[\|\epsilon\|_2^2], E[\|A\|_2^2] < \infty\); (g) \(E[\epsilon] = 0\); (h) \(\text{Var}(A) > 0\), i.e., the variance matrix of \(A\) is positive definite; (i) \(A^TA\) is almost surely of full rank;

Assumption 2.2 (Finite Sample Assumptions). (a) \(Z^T Z_s\) is almost surely of full rank; (b) \(A^TZ_s\) is almost surely of full column rank. (c) \(X^TX\) is almost surely of full rank;

Assumption 2.3 (Population Assumptions). (a) \(\text{Var}(Z_s) > 0\), i.e., the variance matrix of \(Z_s\) is positive definite; (b) \(E[AZ_s^T]\) is of full column rank.

We will henceforth assume that Assumption 2.1 always holds. This assumption ensures that the SEM and that the TSLS objectives are well-defined. In the above assumptions, \(Z_s\) and \(Z_s\) are generic placeholders for a subset of endogenous and exogenous variables from \([X^T A]^T\) and \([X A]\), respectively, which should be clear from the context in which they are used. Both Assumption 2.1 (i) and Assumption 2.2 (c) hold if \(X\) and \(A\) have density with respect to Lebesgue measure, which in turn is guaranteed by Assumption 2.1 (d) if \(A\) and \(\epsilon\) have density with respect to Lebesgue measure. Assumption 2.1 (h) and 2.1 (i) implies that the instrumental variable objective functions introduced below are almost surely well-defined, and Assumption 2.2 (c) yields that the OLS solution is almost surely well-defined. Assumption 2.1 (f) implies that \(Y, X\) and \(H\) all have finite second moments. For Assumption 2.3 (b) and 2.2 (b), it is necessary that \(q \geq \text{dim}(Z_s)\), i.e., that the setup must be just- or over-identified; see Section 3.1.

2.2. Distributional robustness of anchor regression

Rothenh"ausler et al. (2021) proposes a method, called anchor regression, for predicting the endogenous target variable \(Y\) from the endogenous variables \(X\). The collection of exogenous variables \(A\), called anchors, are not included in that prediction model. Anchor regression trades off predictability and invariance by considering a convex combination of the OLS loss function and the TSLS (IV) loss function using the anchors as instruments. More formally, we define

\[
I_{\text{OLS}}(\gamma; Y, X) := E(Y - \gamma^TX)^2, \tag{2.3}
\]

\[
I_{\text{IV}}(\gamma; Y, X, A) := E(A(Y - \gamma^TX)^TE(AA^T)^{-1}E(A(Y - \gamma^TX)),
\]

\[
I_{\text{OLS}}^n(\gamma; Y, X) := n^{-1}(Y - X\gamma)^T(Y - X\gamma), \tag{2.4}
\]

\[
I_{\text{IV}}^n(\gamma; Y, X, A) := n^{-1}(Y - X\gamma)^TP_A(Y - X\gamma), \tag{2.5}
\]

as the population and finite sample versions of the loss functions. \(P_A = A(A^TA)^{-1}A^T\) is the orthogonal projection onto the column space of \(A\). To simplify notation, we omit the dependence on \(Y, X, A, X\) or \(Y\) when they are clear from a given context. For a penalty parameter \(\lambda > -1\),

© The Author(s) 2022.
the anchor regression coefficients are defined as
\[ \gamma_{AR}(\lambda) : = \arg\min_{\gamma \in \mathbb{R}^d} \{ \text{OLS}(\gamma) + \lambda \text{IV}(\gamma) \}, \]
\[ \hat{\gamma}_{AR}^{\beta}(\lambda) : = \arg\min_{\gamma \in \mathbb{R}^d} \{ \text{OLS}(\gamma) + \lambda \text{IV}(\gamma) \}. \] (2.6)

The estimator \( \hat{\gamma}_{AR}^{\beta}(\lambda) \) consistently estimates the population estimand \( \gamma_{AR}(\lambda) \) and minimizes prediction error while simultaneously penalizing a transformed sample covariance between the anchors and the resulting prediction residuals. Unlike the TSLS estimator, for example, the anchor regression estimator is almost surely well-defined under the rank condition of Assumption 2.2 (c), even if the model is under-identified, that is, there are less exogenous than endogenous variables. The solution to the empirical minimization problem of anchor regression is given by
\[ \hat{\gamma}_{AR}^{\beta}(\lambda) = [X^T(I + \lambda P_A)X]^{-1}X^T(I + \lambda P_A)Y, \] (2.7)
which follows from solving the normal equation of equation (2.6).

The motivation of anchor regression is not to infer a causal parameter. Instead, for a fixed penalty parameter \( \lambda \), the estimator is shown to possess a distributional or interventional robustness property: the estimator is optimal when predicting under interventions on the exogenous variables that are below a certain intervention strength. By theorem 1 of Rothenh¨ausler et al. (2021) it holds that
\[ \gamma_{AR}(\lambda) = \arg\min_{\gamma \in \mathbb{R}^d} \sup_{v \in C(\lambda)} \mathbb{E} \left[ (Y - \gamma^T X)^2 \right], \]
where \( C(\lambda) := \{ v : \Omega \rightarrow \mathbb{R}^q : \text{Cov}(v, \varepsilon) = 0, \mathbb{E}(vv^\top) \leq (\lambda + 1)\mathbb{E}(AA^\top) \} \).

### 2.3. Distributional robustness of K-class estimators

We now introduce the limited information estimators known as K-class estimators (Theil, 1958; Nagar, 1959) used for single equation inference. Suppose that we are given nonsample information about which components of \( \gamma_0 \) and \( \beta_0 \) of equation (2.2) are zero. We can then partition \( X = [X_s, X_{-s}] \in \mathbb{R}^{n \times (d_1 + d_2)} \), \( A = [A_s, A_{-s}] \in \mathbb{R}^{n \times (q_1 + q_2)} \) and \( Z = [Z_s, Z_{-s}] = [X_s, A_s, X_{-s}, A_{-s}] \) with \( Z_s \in \mathbb{R}^{n \times ((d_1 + q_1) + (d_2 + q_2))} \), where \( X_{-s} \) and \( A_{-s} \) correspond to the variables for which our nonsample information states that the components of \( \gamma_0 \) and \( \beta_0 \) are zero, respectively. We call the variables corresponding to \( A_s \) included exogenous variables. Similarly, we write \( \gamma_0 = (\gamma_{0,s}, \gamma_{0,-s}), \beta_0 = (\beta_{0,s}, \beta_{0,-s}) \) and \( \alpha_0 = (\alpha_{0,s}, \alpha_{0,-s}) = (\gamma_{0,s}, \beta_{0,s}, \gamma_{0,-s}, \beta_{0,-s}) \). The structural equation of interest then reduces to \( Y = X_s \gamma_{0,s} + X_{-s} \gamma_{0,-s} + A_s \beta_{0,s} + A_{-s} \beta_{0,-s} + \hat{U}_Y = Z_s \alpha_{0,s} + \hat{U}_Y \), where \( \hat{U}_Y = X_{-s} \gamma_{0,-s} + A_{-s} \beta_{0,-s} + H \eta_0 + \varepsilon_Y \). In the case that the nonsample information is indeed correct, we have that \( \hat{U}_Y = \hat{U}_Y = H \eta_0 + \varepsilon_Y \). When well-defined, the K-class estimator with parameter \( \kappa \in \mathbb{R} \) for a simultaneous estimation of \( \alpha_{0,s} \) is given by
\[ \hat{\alpha}_{K}^{\beta}(\kappa; Y, Z_s, A) = (Z_s^T(I - \kappa P_{A,s}^\perp)Z_s)^{-1}Z_s^T(I - \kappa P_{A,s}^\perp)Y, \] (2.8)
where \( I - \kappa P_{A,s}^\perp = I - \kappa (I - P_A) = (1 - \kappa)I + \kappa P_A \).

Comparing equations (2.7) and (2.8) suggests a close connection between anchor regression and K-class estimators for inference of structural equations with no included exogenous variables. In the following subsections, we establish this connection and subsequently extend the distributional robustness property to general K-class estimators.

© The Author(s) 2022.
2.3.1. K-class estimators in models with no included exogenous variables. Assume that, in addition to Assumption 2.1, we have the nonsample information that \( \beta_0 = 0 \), that is, no exogenous variable in \( A \) directly affects the target variable \( Y \). By direct comparison we see that the K-class estimator for \( \kappa < 1 \) coincides with the anchor regression estimator with penalty parameter \( \lambda = \kappa/(1 - \kappa) \), i.e., \( \hat{\gamma}_K^\alpha(\kappa) = \hat{\gamma}_K^\alpha(\kappa/(1 - \kappa)) \). Equivalently, we have \( \hat{\gamma}_K^\alpha(\kappa/(1 + \lambda)) \) for any \( \lambda > -1 \). As such, the K-class estimator, for a fixed \( \kappa < 1 \), inherits the following distributional robustness property:

\[
\hat{\gamma}_K^n(\kappa) \xrightarrow{n \to \infty} \underset{\gamma \in \mathbb{R}^q}{\arg\min} \sup_{\gamma \in C(\kappa/(1 - \kappa))} E^{d(A = v)}[\alpha - \gamma^\top X]^2, \tag{2.9}
\]

where \( C(\kappa/(1 - \kappa)) = \{ v : \Omega \to \mathbb{R}^q : \text{Cov}(v, \varepsilon) = 0, E[v\varepsilon^\top] \leq \frac{1}{\lambda} E[AA^\top] \} \). This statement holds by theorem 1 of Rothenh"au
er et al. (2021).

In an identifiable model with \( P \lim_{n \to \infty} \kappa = 1 \) we have that \( \hat{\gamma}_K^n(\kappa) \) consistently estimates the causal parameter; see e.g., Mariano (2001). For such a choice of \( \kappa \), the robustness above is just a weaker version of what the causal coefficient can guarantee. However, the above result in equation (2.9) establishes a robustness property for fixed \( \kappa < 1 \), even in cases where the model is not identifiable. Furthermore, since we did not use the nonsample information that \( \beta_0 = 0 \) was true, the robustness property is resilient to model misspecification in terms of excluding included exogenous variables from the target equation, which generally also breaks identifiability.

2.3.2. The K-class estimators as penalized regression estimators. We now show that general K-class estimators can be written as solutions to penalized regression problems. The first appearance of such a representation is, to the best of our knowledge, due to McDonald (1977) building upon previous work of Basmann (1960a, b). Their representation, however, concerns only the endogenous part \( \gamma \). We require a slightly different statement and will show that the entire K-class estimator of \( \alpha_{0,s} \), i.e., the simultaneous estimation of \( \gamma_{0,s} \) and \( \beta_{0,s} \), can be written as a penalized regression problem. Let therefore \( l_{IV}(\alpha; Y, Z_s, A) \), \( l_{IV}^n(\alpha; Y, Z_s, A) \) and \( l_{OLS}(\alpha; Y, Z_s) \), \( l_{OLS}^n(\alpha; Y, Z_s) \) denote the population and empirical TLS and OLS loss functions as defined in equations (2.3) to (2.4). That is, the TLS loss function for regressing \( Y \) on the included endogenous and exogenous variables \( Z_s \) using the exogeneity of \( A \) and \( A_{-s} \) as instruments and the OLS loss function for regressing \( Y \) on \( Z_s \). We define the K-class population and finite sample loss functions as an affine combination of the two loss functions above. That is,

\[
l_K(\alpha; \kappa, Y, Z_s, A) = (1 - \kappa)l_{OLS}(\alpha; Y, Z_s) + \kappa l_{IV}(\alpha; Y, Z_s, A), \tag{2.10}
\]

\[
l_K^n(\alpha; \kappa, Y, Z_s, A) = (1 - \kappa)l_{OLS}^n(\alpha; Y, Z_s) + \kappa l_{IV}^n(\alpha; Y, Z_s, A). \tag{2.11}
\]

**Proposition 2.1.** Consider one of the following scenarios: 1) \( \kappa < 1 \) and 2.2 (a) holds, or 2) \( \kappa = 1 \) and 2.2 (b) holds. The estimator minimizing the empirical loss function of equation (2.11) is almost surely well-defined and coincides with the K-class estimator of equation (2.8). That is, it almost surely holds that

\[
\hat{\alpha}_K^n(\kappa; Y, Z_s, A) = \arg\min_{\alpha \in \mathbb{R}^{d_l+1}} l_K^n(\alpha; \kappa, Y, Z_s, A). \tag{2.12}
\]

Assuming \( \kappa \neq 1 \), we can rewrite equation (2.12) to

\[
\hat{\alpha}_K^n(\kappa; Y, Z_s, A) = \arg\min_{\alpha \in \mathbb{R}^{d_l+1}} \{l_{OLS}^n(\alpha; Y, Z_s) + \frac{\kappa}{1 - \kappa} l_{IV}^n(\alpha; Y, Z_s, A)\}. \tag{2.13}
\]
Thus, K-class estimators seek to minimize the OLS loss for regressing $Y$ on $Z_*$, while simultaneously penalizing the strength of a transform on the sample covariance between the prediction residuals and collection of exogenous variables $A$.

In the following section, we consider a population version of the above quantity. If we replace the finite sample Assumption 2.2 with the corresponding population Assumption 2.3, we get that the minimization estimator of the empirical loss function of equation (2.11) is asymptotically well-defined. Furthermore, we now prove that whenever the population assumptions are satisfied, then, for any fixed $\kappa \in [0, 1]$, $\hat{\alpha}^{\text{p}}(\kappa; Y, Z_*, A)$ converges in probability towards the population K-class estimand.

**Proposition 2.2.** Consider one of the following scenarios: 1) $\kappa \in [0, 1)$ and 2.3 (a) holds, or 2) $\kappa = 1$ and 2.3 (b) holds. It holds that $(\hat{\alpha}^{\text{p}}(\kappa; Y, Z_*, A))_{n \geq 1}$ is an asymptotically well-defined sequence of estimators. Furthermore, the sequence consistently estimates the well-defined population K-class estimand. That is,

$$\hat{\alpha}^{\text{p}}(\kappa; Y, Z_*, A) \overset{P}{\rightarrow} \alpha_{K}(\kappa; Y, Z_*, A) := \arg \min_{\alpha \in \mathbb{R}^{d+q}} \alpha_{K}(\alpha; Y, Z_*, A).$$

**2.3.3. Distributional robustness of general K-class estimators.** We are now able to prove that the general K-class estimator possesses a robustness property similar to the statements above. It is prediction optimal under a set of interventions on all exogenous $A$ up to a certain strength.

**Theorem 2.1.** Let Assumption 2.1 hold. For any fixed $\kappa \in [0, 1)$ and $Z_* = (X_*, A_*)$ with $X_* \subseteq X$ and $A_* \subseteq A$, we have, whenever the population K-class estimand is well-defined, that

$$\alpha_{K}(\kappa; Y, Z_*, A) = \arg \min_{\alpha \in \mathbb{R}^{d+q}} \sup_{v \in C(\kappa)} E^{\text{do}(A := v)} \left[ (Y - \alpha^\top Z_*)^2 \right],$$

where $C(\kappa) := \{v : \Omega \to \mathbb{R}^q : \text{Cov}(v, \varepsilon) = 0, E[vv^\top] \leq \frac{1}{1-\kappa} E[AA^\top] \}$.

Here, $E^{\text{do}(A := v)}$ denotes the expectation with respect to the distribution entailed under the intervention $\text{do}(A := v)$ (see Section 1.2.1 and Online Appendix S1) and $(\Omega, \mathcal{F}, P)$ is the common background probability space on which $A$ and $\varepsilon$ are defined.

In other words, among all linear prediction methods of $Y$ using $Z_*$ as predictors, the K-class estimator with parameter $\kappa$ has the lowest possible worst case mean squared prediction error when considering all interventions on the exogenous variables $A$ contained in $C(\kappa)$. As $\kappa$ approaches one, the estimator is prediction optimal under a class of arbitrarily strong interventions in the direction of the variance of $A$. (Here, $\kappa$ is arbitrary but fixed; the statement does not cover data-driven choices of $\kappa$, such as LIML or Fuller.) The above result is a consequence of the relation between anchor regression and K-class estimators. The special case $A_* = \emptyset$ is a consequence of Theorem 1 by Rothenhäusler et al. (2021). Our proof follows similar arguments, but additionally allows for $A_* \neq \emptyset$.

The property in Theorem 2.1 has a decision-theoretic interpretation (see Chamberlain, 2007 for an application of decision theory in IV models based on another loss function). Consider a response $Y$, covariates $Z_*$ and a distribution (specified by $\theta$) over $(Y, Z_*)$, and the squared loss $\ell(Y, Z, \alpha) := (Y - \alpha^\top Z_*)^2$. Then, assuming finite variances, for each distribution the risk $E_{\theta}(Y - \alpha^\top Z_*^2)$ is minimized by the (population) OLS solution $\alpha = \alpha_{\theta} := \text{cov}_{\theta}(Z_*)^{-1}\text{cov}_{\theta}(Z_*, Y)$. In the setting of Theorem 2.1, we are given a distribution over $(Y, Z_*)$, specified by $\theta$, but we are interested in minimizing the risk $E_{\theta, v}(Y - \alpha_{\theta, v}^\top Z_*^2)$ for another distribution that is induced by an intervention

\[\text{the minimization estimator of the empirical loss function of equation (2.11) is asymptotically well-defined.}

Furthermore, we now prove that whenever the population assumptions are satisfied, then, for any fixed $\kappa \in [0, 1]$, $\hat{\alpha}^{\text{p}}(\kappa; Y, Z_*, A)$ converges in probability towards the population K-class estimand. That is,

$$\hat{\alpha}^{\text{p}}(\kappa; Y, Z_*, A) \overset{P}{\rightarrow} \alpha_{K}(\kappa; Y, Z_*, A) := \arg \min_{\alpha \in \mathbb{R}^{d+q}} \alpha_{K}(\alpha; Y, Z_*, A).$$

2.3.3. Distributional robustness of general K-class estimators. We are now able to prove that the general K-class estimator possesses a robustness property similar to the statements above. It is prediction optimal under a set of interventions on all exogenous $A$ up to a certain strength.

**Theorem 2.1.** Let Assumption 2.1 hold. For any fixed $\kappa \in [0, 1)$ and $Z_* = (X_*, A_*)$ with $X_* \subseteq X$ and $A_* \subseteq A$, we have, whenever the population K-class estimand is well-defined, that

$$\alpha_{K}(\kappa; Y, Z_*, A) = \arg \min_{\alpha \in \mathbb{R}^{d+q}} \sup_{v \in C(\kappa)} E^{\text{do}(A := v)} \left[ (Y - \alpha^\top Z_*)^2 \right],$$

where $C(\kappa) := \{v : \Omega \to \mathbb{R}^q : \text{Cov}(v, \varepsilon) = 0, E[vv^\top] \leq \frac{1}{1-\kappa} E[AA^\top] \}$.

Here, $E^{\text{do}(A := v)}$ denotes the expectation with respect to the distribution entailed under the intervention $\text{do}(A := v)$ (see Section 1.2.1 and Online Appendix S1) and $(\Omega, \mathcal{F}, P)$ is the common background probability space on which $A$ and $\varepsilon$ are defined.

In other words, among all linear prediction methods of $Y$ using $Z_*$ as predictors, the K-class estimator with parameter $\kappa$ has the lowest possible worst case mean squared prediction error when considering all interventions on the exogenous variables $A$ contained in $C(\kappa)$. As $\kappa$ approaches one, the estimator is prediction optimal under a class of arbitrarily strong interventions in the direction of the variance of $A$. (Here, $\kappa$ is arbitrary but fixed; the statement does not cover data-driven choices of $\kappa$, such as LIML or Fuller.) The above result is a consequence of the relation between anchor regression and K-class estimators. The special case $A_* = \emptyset$ is a consequence of Theorem 1 by Rothenhäusler et al. (2021). Our proof follows similar arguments, but additionally allows for $A_* \neq \emptyset$.

The property in Theorem 2.1 has a decision-theoretic interpretation (see Chamberlain, 2007 for an application of decision theory in IV models based on another loss function). Consider a response $Y$, covariates $Z_*$ and a distribution (specified by $\theta$) over $(Y, Z_*)$, and the squared loss $\ell(Y, Z, \alpha) := (Y - \alpha^\top Z_*)^2$. Then, assuming finite variances, for each distribution the risk $E_{\theta}(Y - \alpha^\top Z_*^2)$ is minimized by the (population) OLS solution $\alpha = \alpha_{\theta} := \text{cov}_{\theta}(Z_*)^{-1}\text{cov}_{\theta}(Z_*, Y)$. In the setting of Theorem 2.1, we are given a distribution over $(Y, Z_*)$, specified by $\theta$, but we are interested in minimizing the risk $E_{\theta, v}(Y - \alpha_{\theta, v}^\top Z_*^2)$ for another distribution that is induced by an intervention...
and specified by \((\theta, \nu)\). The above result states that the K-class estimator minimizes a worst case risk when considering all \(\nu \in C(\kappa)\).

Theorem 2.1 makes use of the language of SEMs in that it yields the notion of interventions.\(^2\) As such, the result can be rephrased using other causal frameworks. The crucial assumptions are the exogeneity of \(A\) and the linearity of the system. Furthermore, the result is robust with respect to several types of model misspecifications that break the identifiability of \(\alpha_0\), such as excluding included endogenous or exogenous predictors or the existence of latent mediators between the exogenous variables and the target; see Remark S7.1 in Online Appendix S7.

3. THE P-UNCORRELATED LEAST SQUARE ESTIMATOR

We now introduce the PULSE. As discussed in Section 1.2, PULSE allows for different representations. In this section we start with the third representation and show the equivalence of the other representations afterwards.

Consider predicting the target \(Y\) from endogenous and possibly exogenous regressors \(Z\). Let therefore \(\mathcal{H}_0(\alpha)\) denote the hypothesis that the prediction residuals using \(\alpha\) as a regression coefficient is simultaneously uncorrelated with every exogenous variable, that is, \(\mathcal{H}_0(\alpha) : \text{Corr}(A, Y - \alpha^T Z) = 0\). This hypothesis is in some models under certain conditions equivalent to the hypothesis that \(\alpha\) is the true causal coefficient. One of these conditions is the rank condition Assumption 3.5 introduced below, also known as the rank condition for identification; see Wooldridge (2010).\(^3\)

The two-stage least square (TSLS) estimator exploits the equivalence between the causal coefficient and the zero correlation between the instruments and the regression residuals. Here, one minimizes a sample covariance between the instruments and the regression residuals: we can write
\[
\ell_n^{\text{IV}}(\alpha; Y, Z, A) = \|\text{Cov}_n(A, Y - \alpha^T Z)\|^2_{(n^{-1}A^TA)^{-1}}
\]
when \(A\) is mean zero.\(^3\) In the just-identified setup the TSLS estimator yields a sample covariance that is exactly zero and is known to be unstable, in that it has no moments of any order. Intuitively, the constraint of vanishing sample covariance may be too strong.

Let \(T(\alpha; Y, Z, A)\) be a finite sample test statistic for testing the hypothesis \(\mathcal{H}_0(\alpha)\) and let p-value\((T(\alpha; Y, Z, A))\) denote the p-value associated with the test of \(\mathcal{H}_0(\alpha)\). We then define the PULSE as
\[
\hat{\alpha}_n^{\text{PULSE}}(p_{\text{min}}) = \alpha_n \in \left\{ \alpha \mid \|\text{Cov}_n(A, Y - \alpha^T Z)\|^2_{(n^{-1}A^TA)^{-1}} \leq p_{\text{min}}, \text{subject to } \text{p-value}(T(\alpha; Y, Z, A)) \geq p_{\text{min}} \right\},
\]
where \(p_{\text{min}}\) is a pre-specified level of the hypothesis test. In other words, we aim to minimize the mean squared prediction error among all coefficients which yield a p-value for testing \(\mathcal{H}_0(\alpha)\) that does not fall below some pre-specified level-threshold \(p_{\text{min}} \in (0, 1)\), such as \(p_{\text{min}} = 0.05\). That is, the minimization is constrained to the acceptance region of the test, i.e., a confidence region for the causal coefficient in the identified setup. Among these coefficient, we choose the solution that is closest to the OLS solution.\(^4\)

\(^2\) In particular, we have not considered the SEM as a model for counterfactual statements.

\(^3\) \(\| \cdot \|_{(n^{-1}A^TA)^{-1}}\) is the norm induced by the inner product \(\langle x, y \rangle = x^T(n^{-1}A^TA)^{-1}y\).

\(^4\) Here, closeness is measured in the OLS distance: we define the OLS norm via \(\|\alpha\|^2_{\text{OLS}} := \ell_n^{\text{OLS}}(\alpha + \hat{\alpha}_n^{\text{OLS}})\) and let \(\hat{\alpha}_n^{\text{OLS}}\) be the OLS estimator. This defines a norm (rather than a semi-norm) if \(Z^TZ\) is nondegenerate. Minimizing \(\ell_n^{\text{OLS}}(\alpha) = \|Y - Z\alpha\|^2_{\text{OLS}} = (\alpha - \hat{\alpha}_n^{\text{OLS}})^T Z^T Z (\alpha - \hat{\alpha}_n^{\text{OLS}}) + \|Y - Z\hat{\alpha}_n^{\text{OLS}}\|^2_{\text{OLS}}\) is equivalent to minimizing \(\|\alpha - \hat{\alpha}_n^{\text{OLS}}\|_{\text{OLS}}^2\).

© The Author(s) 2022.
Thus, PULSE allows for an intuitive interpretation. We will see in the experimental section that it has good finite sample performance, in particular for weak instruments. Unlike other estimators, such as LIML, the above estimator is well-defined in the under-identified setup, too.\(^5\) In such cases, PULSE extends existing literature that aims to trade off predictability and invariance, but that so far has been restricted to search over subsets of variables (see Sections 1.2.2 and Online Appendix S8.3). To maintain consistency of the estimator, the chosen test must have asymptotic power of one.

In this paper, we propose a class of significance tests that contains, e.g., the Anderson–Rubin test (Anderson and Rubin, 1949). While the objective function in equation (3.14) is quadratic in \(\alpha\), the resulting constraint is, in general, nonconvex. In Section 3.5, we develop a computationally efficient procedure that provably solves the optimization problem at low computational cost. Other choices of tests are possible, too, but may result in even harder optimization problems.

In Section 3.1 we briefly introduce the setup and assumptions. In Section 3.2 we specify a class of asymptotically consistent tests for \(H_0(\alpha)\). In Section 3.3 we formally define the PULSE estimator. In Section 3.4 we show that the PULSE estimator is well-defined by proving that it is equivalent to a solvable convex quadratically constrained quadratic program, which we denote by the primal PULSE. In Section 3.5 we utilize duality theory and derive an alternative representation, which we denote by the dual PULSE. This representation yields a computationally feasible algorithm and shows that the PULSE estimator is a K-class estimator with a data-driven \(\kappa\). Proofs of results in this section can be found in Online Appendix S5 unless stated otherwise.

3.1. Setup and assumptions

In the following sections we again let \((Y, X, H, A)\) consist of \(n \geq \min\{d, q\}\) row-wise independent and identically distributed copies of \((Y, X, H, A)\) generated in accordance with the SEM in equation (2.1). The structural equation of interest is

\[
Y = \gamma^T_X X + \eta^T_Y H + \beta^T_A A + \varepsilon_Y.
\]

Assume that we have some nonsample information about which \(d_2 = d - d_1\) and \(q_2 = q - q_1\) coefficients of \(\gamma_0\) and \(\beta_0\), respectively, are zero. As in Section 2, we let the subscript \(\ast\) denote the variables and coefficients that are nonzero according to the nonsample information but, to simplify notation, we drop the \(\ast\) subscript from \(Z, Z\) and \(\alpha_0;\) so, we write \(Z = [X^T_\ast A^T_\ast]^T \in \mathbb{R}^{d_1 + q_1}; Z = [X, A]\) \(\in \mathbb{R}^{n \times (d_1 + q_1)}\) and \(\alpha_0 := (\gamma^T_{0, \ast}, \beta^T_{0, \ast})^T \in \mathbb{R}^{d_1 + q_1}.\) That is, \(Y = \alpha_0^T Z + U_Y, \) where \(U_Y = \alpha_{0, \ast}^T Z_{\ast} + \eta_0^T H + \varepsilon_Y.\)

We define a setup as being under-, just-, and over-identified by the degree of over-identification \(q_2 - d_1\) being negative, equal to zero, and positive, respectively. That is, the number of excluded exogenous variables \(A_{\ast}\) being less, equal, or larger than the number of included endogenous variables \(X_{\ast}\) in the target equation.

We assume that the global assumptions of Assumption 2.1 from Section 2.1 still hold. Furthermore, we will make use of the following situational assumptions.

**Assumption 3.1.** (a) \(A \perp \!\!\!\!\perp U_Y;\) (b) \(E[A] = 0.\)

**Assumption 3.2.** \(\varepsilon\) has nondegenerate marginals.

\(^5\) The PULSE estimator is defined for finite samples, but the following deliberation may help to build intuition: in an under-identified IV setting, minimizing \(I_{\text{OLS}}(Y)\) under the constraint that \(l_Y(\gamma) = 0,\) can be seen as choosing, under all causal models compatible with the distribution, the model with the least amount confounding – when using \(E(Y - \gamma^T X)^2 - E(Y - \gamma^T_{\text{OLS}} X)^2\) as a measure for confounding.
ASSUMPTION 3.3. (a) $Z^\top Z$ is of full rank; (b) $A^\top Z$ is of full rank.

ASSUMPTION 3.4. $[Z \ Y]$ is of full column rank.

ASSUMPTION 3.5. $E[AZ^\top]$ is of full rank.

Assumption 3.1 (a) holds if our nonsample information is true and the instrument set $A$ is independent of all unobserved endogenous variables $H_i$ which directly affect the target $Y$. This holds, for example, if the latent variables are source nodes; that is, they have no parents in the causal graph of the corresponding SEM. Assumption 3.1 (b) can be achieved by centring the data. Strictly speaking, this introduces a weak dependence structure in the observations, which is commonly ignored. Alternatively, one can perform sample splitting. For more details on this assumption and the possibility of relaxing it, see Remark 3.1. Assumption 3.3 (a) ensures that $K$-class estimators for $\kappa < 1$ are well-defined, regardless of the over-identification degree. In the under-identified setup, Assumption 3.3 (b) yields that there exists a subspace of solutions minimizing $\ln IV(\alpha)$. In the just- and over-identified setup, this assumption ensures that $\ln IV(\alpha)$ has a unique minimizer given by the TSLS estimator $\hat{\alpha}_n^{TSLS} = (Z^\top P A Z)^{-1}Z^\top P A Y$. Assumption 3.4 is used to ensure that the OLS objective function $\alpha \mapsto \ln OLS(\alpha; Y, Z)$ is strictly positive, such that division by this function is always well-defined. Assumptions 3.2 and 3.5 ensure that various limiting arguments are valid. In the just- and over-identified setup, Assumption 3.5 is known as the rank condition for identification.

3.2. Testing for vanishing correlation

We now introduce a class of tests for the null hypothesis $H_0(\alpha): \text{Corr}(A, Y - Z\alpha) = 0$ that have point-wise asymptotic level and point-wise asymptotic power. These tests will allow us to define the corresponding PULSE estimator. When Assumption 3.4 holds, we can define $T_n^c: \mathbb{R}^{d_1 + q_1} \to \mathbb{R}$ by

$$T_n^c(\alpha) := c(n) \frac{n_{IV}(\alpha)}{n_{OLS}(\alpha)} = c(n) \frac{\|P_A(Y - Z\alpha)\|_2^2}{\|Y - Z\alpha\|_2^2},$$

where $c(n)$ is a function that will typically scale linearly in $n$. Let us denote the $1 - p$ quantile of the central Chi-Squared distribution with $q$ degrees of freedom by $Q_{\chi^2_q}(1 - p)$. By standard limiting theory we can test $H_0(\alpha)$ in the following manner.

LEMMA 3.1 (LEVEL AND POWER OF THE TEST). Let Assumptions 3.1, 3.2, and 3.4 hold and assume that $c(n) \sim n$ as $n \to \infty$. For any $p \in (0, 1)$ and any fixed $\alpha$, the statistical test rejecting the null hypothesis $H_0(\alpha)$, if $T_n^c(\alpha) > Q_{\chi^2_q}(1 - p)$, has point-wise asymptotic level $p$ and point-wise asymptotic power of 1 against all alternatives as $n \to \infty$.

REMARK 3.1. Assumption 3.1, $E[A] = 0$, is important for the test statistic to be asymptotic pivotal under the null hypothesis; that is, to ensure that the asymptotic distribution of $T_n^c(\alpha)$ does not depend on the model parameters except for $q$. We can drop this assumption if we change the null hypothesis to $H_0(\alpha): E[A(Y - Z^\top \alpha)] = 0$ and add the assumption that $E[U^2] = 0$. Furthermore, if we are in the just- or over-identified setup and Assumption 3.5 holds, both of these hypotheses are, under their respective assumptions, equivalent to $\tilde{H}_0(\alpha): \alpha = \alpha_0$. That is, the test in Lemma 3.1 becomes an asymptotically consistent test for the causal coefficient.
Depending on the choice of \( c(n) \), this class contains several tests, some of which are well known. With \( c(n) = n - q + Q_{\chi^2}(1 - p_{\text{min}}) \), for example, one recovers a test that is equivalent to the asymptotic version of the Anderson–Rubin test (Anderson and Rubin, 1950). We make this connection precise in Remark S7.2 in Online Appendix S7. The Anderson–Rubin test is robust to weak instruments in the sense that the limiting distribution of the test statistic under the null-hypothesis is not affected by weak instrument asymptotics; see, e.g., Staiger and Stock (1997) and Stock et al. (2002). For weak instruments, the confidence region may be unbounded with large probability; see Dufour (1997). Moreira (2009) shows that the test suffers from loss of power in the over-identified setting.

To simplify notation, we will from now on work with the choice \( c(n) = n \) and define the acceptance region with level \( p_{\text{min}} \in (0, 1) \) as

\[
A_n(1 - p_{\text{min}}) = \{ \alpha \in \mathbb{R}^{d_1 + q_1} : T_n(\alpha) \leq Q_{\chi^2}(1 - p_{\text{min}}) \},
\]

where \( T_n(\alpha) \) corresponds to the choice \( c(n) = n \).

### 3.3. The PULSE estimator

For any level \( p_{\text{min}} \in (0, 1) \), we formally define the PULSE estimator of equation (3.14) by letting the feasible set be given by the acceptance region \( A_n(1 - p_{\text{min}}) \) of \( H_0(\alpha) \) using the test of Lemma 3.1. That is, we consider

\[
\hat{\alpha}_{\text{PULSE}}(p_{\text{min}}) := \arg \min_{\alpha} \ln \text{OLS}(\alpha) \text{ subject to } T_n(\alpha) \leq Q_{\chi^2}(1 - p_{\text{min}}). \tag{3.15}
\]

In general, this is a nonconvex optimization problem (Boyd and Vandenberghe, 2004) as the constraint function is nonconvex; see the blue contours in Figure 1(left). From Figure 1(right) we see that in the given example the problem nevertheless has a unique and well-defined solution: the smallest level set of \( \ln \text{OLS} \) with a nonempty intersection of the acceptance region \( \{ \alpha : T_n(\alpha) \leq Q_{\chi^2}(1 - p_{\text{min}}) \} \) intersects with the latter region in a unique point. In Section 3.4, we prove that this is not a coincidence: equation (3.15) has a unique solution that coincides with the solution of a strictly convex, quadratically constrained quadratic program (QCQP) with a data-dependent constraint bound. In Section 3.5, we further derive an equivalent Lagrangian dual problem. This has two important implications: (1) it allows us to construct a computationally efficient procedure to compute a solution of the nonconvex problem above; and (2) it shows that the PULSE estimator can be written as K-class estimators.

Estimators with similar constraints, albeit different optimization objectives, have been studied by Gautier et al. (2018). In Remark S7.3 in Online Appendix S7 we briefly discuss the connection to pre-test estimators. Furthermore, any method for inverting the test (see, e.g., Davidson and MacKinnon, 2014), yields a valid confidence set including the proposed point estimator (given that the method outputs the point estimator when the acceptance region is empty).

### 3.4. Primal representation of PULSE

We now derive a QCQP representation of the PULSE problem, which we call the primal PULSE. For all \( t \geq 0 \) define the empirical primal minimization problem (Primal.\( t.n \)) by

\[
\min_{\alpha} \quad l^n_{\text{OLS}}(\alpha; Y, Z) \quad \text{subject to} \quad l^n_{\text{IV}}(\alpha; Y, Z, A) \leq t. \tag{3.16}
\]

\[6\] Weak instrument asymptotics is a model scheme where the instrument strength tends to zero at a rate of \( n^{-1/2} \), i.e., the reduced form structural equation for the endogenous variables is given by \( X = \Lambda n^{-1/2} \Pi_X + \varepsilon \Gamma_X^{-1} \).
Figure 1. Illustrations of the level sets of \( \ln \text{OLS} \) (red contours), the proposed test statistic \( T_n \) (blue contours), and \( \ln \text{IV} \) (green contours) in a just-identified setup. The example is generated with a two dimensional anchor \( A = (A_1, A_2) \), one of which is included, and one included endogenous variable \( X \), i.e., \( Y = \alpha_1 X + \alpha_2 A_1 + H + \varepsilon \) with \((\alpha_1, \alpha_2) = (1, 1)\). Both illustrations show level sets from the same setup, but they use different scales. The black text denotes the level of the test-statistic contours. In this setup, the PULSE constraint bound, the rejection threshold of the test with \( p_{\text{min}} = 0.05 \), is \( Q_{\chi^2}(0.95) \approx 5.99 \). The blue level sets of \( T_n \) are nonconvex. The sublevel set of the test, corresponding to the acceptance region, is illustrated by the blue area. In the right plot, we see that the smallest level set of \( \ln \text{OLS} \) that has a nonempty intersection with the \( Q_{\chi^2}(1 - p_{\text{min}}) \)-sublevel set of \( T_n \) is a singleton (black dot, \( t^* \)). This shows that in this example the PULSE problem is solvable and has a unique solution. The \( \ln \text{IV} \) level set that intersects this singleton is exactly the \( t^*_{\text{Pulse}} \)-level set of \( \ln \text{IV} \), illustrating the statement of Theorem 3.1 in that the primal PULSE with that choice of \( t \) solves the PULSE problem. The black line visualizes the solutions \{\( \hat{\alpha}_n \) : \( t \in D_{\text{Pr}} \)\}. The black points and corresponding text labels indicate which constraint bound \( t \) yields the specific point. In general, the class of primal solutions does not coincide with the class of convex combinations of the OLS and the TSLS estimators.

We drop the dependence of \( Y, Z \), and \( A \) and refer to the objective and constraint functions as \( I_{\text{OLS}}(\alpha) \) and \( I_{\text{IV}}(\alpha) \). The following lemma shows that under suitable assumptions these problems are solvable, strictly convex QCQP problems satisfying Slater’s condition.

**Lemma 3.2 (Unique Solvability of the Primal).** Let Assumption 3.3 hold. It holds that \( \alpha \mapsto I_{\text{OLS}}(\alpha) \) and \( \alpha \mapsto I_{\text{IV}}(\alpha) \) are strictly convex and convex, respectively. Furthermore, for any \( t > \inf \alpha I_{\text{IV}}(\alpha) \) it holds that the constrained minimization problem (Primal.\( t \).n) has a unique solution and satisfies Slater’s condition. In the under- and just-identified setup the constraint bound requirement is equivalent to \( t > 0 \) and in the over-identified setup to \( t > I_{\text{IV}}(\hat{\alpha}_{\text{TSLS}}) \), where \( \hat{\alpha}_{\text{TSLS}} = (Z^TPAZ)^{-1}Z^TPAY \).

We restrict the constraint bounds to \( D_{\text{Pr}} := (\inf \alpha I_{\text{IV}}(\alpha), I_{\text{IV}}(\hat{\alpha}_{\text{OLS}})) \). Considering \( t \) that are larger than \( \inf \alpha I_{\text{IV}}(\alpha) \) ensures that the problem (Primal.\( t \).n) is uniquely solvable and furthermore...
that Slater’s condition is satisfied (see Lemma 3.2 above). Slater’s condition will play a role in Section 3.5 when establishing a sufficiently strong connection with its corresponding dual problem for which we can derive a (semi-)closed form solution. Constraint bounds greater than or equal to $l_{IV}^n(\hat{\alpha}_{OLS}^n)$ yield identical solutions. Whenever well-defined, let $\hat{\alpha}_{Pr}^n : D_{Pr} \rightarrow \mathbb{R}^{d_l+q_1}$ denote the constrained minimization estimator given by the solution to the (Primal or equal to problem for which we can derive a (semi-)closed form solution. Constraint bounds greater than in Section 3.5 when establishing a sufficiently strong connection with its corresponding dual that Slater’s condition is satisfied (see Lemma 3.2 above). Slater’s condition will play a role

\[ \hat{\alpha}_{Pr}^n(t) := \arg \min_{\alpha} l_{IV}^n(\alpha) \] subject to $l_{IV}^n(\alpha) \leq t. \] (3.17)

We now prove that for a specific choice of $t$, the PULSE and the primal PULSE yield the same solutions. Define $t_n^*(p_{\min})$ as the data-dependent constraint bound given by

\[ t_n^*(p_{\min}) := \sup\{t \in (\inf l_n^p(\alpha), l_n^p(\hat{\alpha}_{OLS}^n)) : T_n(\hat{\alpha}_{Pr}^n(t)) \leq Q_{\chi_2^2}(1 - p_{\min})\}. \] (3.18)

If $t_n^*(p_{\min}) > -\infty$ or equivalently $t_n^*(p_{\min}) \in D_{Pr}$ we define the primal PULSE problem and its solution by (Primal,$t_n^*(p_{\min})$) and $\hat{\alpha}_{Pr}^n(t_n^*(p_{\min}))$. The following theorem yields conditions for when the solutions to the primal PULSE and PULSE problems coincide.

**Theorem 3.1 (Primal Representation of PULSE).** Let $p_{\min} \in (0, 1)$ and Assumptions 3.3 and 3.4 hold and assume that $t_n^*(p_{\min}) > -\infty$. If $T_n(\hat{\alpha}_{Pr}^n(t_n^*(p_{\min}))) \leq Q_{\chi_2^2}(1 - p_{\min})$, then the PULSE problem has a unique solution given by the primal PULSE solution. That is,

\[ \hat{\alpha}_{PULSE}(p_{\min}) = \hat{\alpha}_{Pr}^n(t_n^*(p_{\min})). \]

We show that $t_n^*(p_{\min}) > -\infty$ is a sufficient condition for $T_n(\hat{\alpha}_{Pr}^n(t_n^*(p_{\min}))) \leq Q_{\chi_2^2}(1 - p_{\min})$ in the proof of Theorem 3.2. The sufficiency of $t_n^*(p_{\min}) > -\infty$ is postponed to the latter proof as it easily follows from the dual representation. Hence, we have shown that finding the PULSE estimator, i.e., finding a solution to the nonconvex PULSE problem, is equivalent to solving the convex QCQP primal PULSE for a data-dependent choice of $t_n^*(p_{\min})$. However, $t_n^*(p_{\min})$ is still unknown. Figure 1 shows an example of the equivalence in Theorem 3.1. Figure 1(right) shows that the level set of $l_{IV}(\alpha) = r^*(p_{\min})$ intersects the optimal level curve of $l_{OLS}^n(\alpha)$ in the same point given by minimizing over the constraint $T_n(\alpha) \leq Q_{\chi_2^2}(1 - p_{\min})$.

The set of solutions to the primal problem $[\hat{\alpha}_{Pr}^n(t) : t \in D_{Pr}]$ can in the just- and over-identified setup be visualized as an (in general) nonlinear path in $\mathbb{R}^{d_l+q_1}$ between the TLSLS estimator ($t = l_{IV}^n(\hat{\alpha}_{TLSLS}^n)$) and the OLS estimator ($t = l_{IV}^n(\hat{\alpha}_{OLS}^n)$) (see also Rothenhäusler et al., 2021). Theorem 3.1 yields that the PULSE estimator ($t = t_n^*(p_{\min})$) then seeks the estimator ‘closest’ to the OLS estimator along this path that does not yield a rejected test of simultaneous vanishing correlation between the resulting prediction residuals and the exogenous variables $A$; see Figure 1. The path of possible solutions is not necessarily a straight line (see black line); thus, in general, the PULSE estimator is different from the affine combination of OLS and TLSLS estimators studied by, e.g., Judge and Mittelhammer (2012).

In the under-identified setup, the TLSLS end point corresponding to $t = \min_n l_{IV}^n(\alpha)$ is instead given by the point in the IV solution space $\{\alpha \in \mathbb{R}^{d_l+q_1} : l_{IV}^n(\alpha) = 0\}$ with the smallest mean squared prediction residuals.

### 3.5. Dual representation of PULSE

In this section, we derive a dual representation of the primal PULSE problem, which we will denote the dual PULSE problem. This specific dual representation allows for the construction of
a binary search algorithm for the PULSE estimator and yields that PULSE is a member of the K-class estimators with stochastic $\kappa$-parameter.

For any penalty parameter $\lambda \geq 0$ we define the dual problem (Dual.$\lambda$.n) by

$$\text{minimize } l^n_{\text{OLS}}(\alpha) + \lambda l^n_{\text{IV}}(\alpha).$$

Whenever Assumption 3.3 (a) holds, i.e., $Z^T Z$ is of full rank, then for any $\lambda \geq 0$ the solution to (Dual.$\lambda$.n) coincides with the K-class estimator with $\kappa = \lambda/(1 + \lambda) \in [0, 1)$; see Proposition 2.1. That is,

$$\hat{\alpha}_K^n(\kappa) = (Z^T (I + \lambda P_A)Z)^{-1} Z^T (I + \lambda P_A) Y,$$

solves (Dual.$\lambda$.n). Henceforth, let $\hat{\alpha}_K^n(\lambda)$ denote the solution to (Dual.$\lambda$.n), i.e., in a slight abuse of notation we will denote the solution to (Dual.$\lambda$.n) by $\hat{\alpha}_K^n(\lambda)$, such that $\hat{\alpha}_K^n(\lambda) = \hat{\alpha}_K^n(\kappa)$ for $\kappa = \lambda/(1 + \lambda)$. We refer to these two representations as the K-class estimator with penalty parameter $\lambda$ and parameter $\kappa$, respectively. The usage of $\kappa$ or $\lambda$ as argument should clarify which notation we refer to.

Under Assumption 3.3 (b) we have that the infimum of $l^n_{\text{IV}}(\alpha)$ is attainable (see the proof of Lemma 3.2). Hence, let the solution space for the minimization problem min$_{\alpha} l^n_{\text{IV}}(\alpha)$ be given by

$$M_{\text{IV}} := \arg \min_{\alpha} l^n_{\text{IV}}(\alpha) = \{\alpha \in \mathbb{R}^{d_1 + q_1} : l^n_{\text{IV}}(\alpha) = \inf_{\alpha'} l^n_{\text{IV}}(\alpha')\}.$$ (3.20)

In the under-identified setup ($q_2 < d_1$), $M_{\text{IV}}$ is a $(d_1 - q_2)$-dimensional subspace of $\mathbb{R}^{d_1 + q_1}$ and in the just- and over-identified setup it holds that $M_{\text{IV}} = \{\hat{\alpha}^n_{\text{TSLS}}\}$.

We now prove that, in the generic case, K-class estimators for $\lambda \in [0, \infty)$ are different from the TSLS estimator. This result may not come as a surprise, but we include it as we need the result later and have not found it elsewhere.

**Lemma 3.3 (K-class estimators and TSLS differ).** Assume that we are in the just- or over-identified setup and $n > q$. Furthermore, assume that $\epsilon$ has density with respect to Lebesgue measure and that the coefficient matrix $B$ of the SEM in equation (2.1) is lower triangular. If the rank conditions of Assumption 3.3 hold almost surely, then it almost surely holds that all K-class estimators with penalty parameter $\lambda \in [0, \infty)$ differ from the TSLS estimator, i.e.,

$$\hat{\alpha}^n_{\text{TSLS}} \notin \{\hat{\alpha}_K^n(\lambda) : \lambda \geq 0\}.$$ (3.21)

We conjecture that the corresponding statement holds in the under-identified setup and without the lower triangular assumption on $B$, too. That is, $M_{\text{IV}} \cap \{\hat{\alpha}_K^n(\lambda) : \lambda \geq 0\} = \emptyset$ holds almost surely. We therefore introduce this as an assumption.

**Assumption 3.6.** No K-class estimator $\hat{\alpha}_K^n(\kappa)$ with $\kappa \in [0, 1)$ is a member of $M_{\text{IV}}$.

Furthermore, when imposing that Assumption 3.6 holds, we also have that the K-class estimators differ from each other.

**Corollary 3.1 (K-class estimators differ).** Let Assumptions 3.3 and 3.6 hold. If $\lambda_1, \lambda_2 \geq 0$ with $\lambda_1 \neq \lambda_2$, then $\hat{\alpha}_K^n(\lambda_1) \neq \hat{\alpha}_K^n(\lambda_2)$.

The above corollary is proven as Corollary S4.1 in Online Appendix S4. We now show that the class of K-class estimators with penalty parameter $\lambda \geq 0$, i.e., $\kappa \in [0, 1)$, coincides with the class of constrained minimization estimators that minimize the primal problems with constraint bounds $t > \min_a l^n_{\text{IV}}(\alpha)$.
LEMMA 3.4 (CONNECTING THE PRIMAL AND DUAL). If Assumptions 3.3, 3.4, and 3.6 hold, then both of the following statements hold: (a) for any $t \in D_{\Pr}$, there exists a unique $\lambda(t) \geq 0$ such that $(\text{Primal}_t.n)$ and $(\text{Dual}_\lambda.t.n)$ have the same unique solution; (b) for any $\lambda \geq 0$, there exists a unique $t(\lambda) \in D_{\Pr}$ such that $(\text{Primal}_t(\lambda).n)$ and $(\text{Dual}_\lambda.n)$ have the same unique solution.

Lemma 3.4 tells us that, under appropriate assumptions, $[\hat{\alpha}_K^n(\kappa) : \kappa \in [0, 1)] = [\hat{\alpha}_K^n(\lambda) : \lambda \geq 0] = [\hat{\alpha}_K^n(t) : t \in D_{\Pr}]$. In other words, we have recast the K-class estimators with $\kappa \in [0, 1]$ as the class of solutions to the primal problems previously introduced. That the minimizers of $l^p_{\Gamma}(\alpha)$ are different from all the K-class estimators with penalty $\lambda \geq 0$ (or $\kappa \in [0, 1]$) guarantees that when representing a K-class problem in terms of a constrained optimization problem it satisfies Slater’s condition.

We are now able to show the main result of this section. The PULSE estimator $\hat{\alpha}_{\text{PULSE}}^n(p_{\min})$ solves a K-class problem $(\text{Dual}_\lambda.n)$ and can therefore be seen as a K-class estimator with a data-dependent parameter. To see this, let us define the dual PULSE penalty parameter, i.e., the dual analogue of the primal PULSE constraint $t^*_n(p_{\min})$ as

$$\lambda^*_n(p_{\min}) := \inf\{\lambda \geq 0 : T_n(\hat{\alpha}_K^n(\lambda)) \leq Q_{\chi^2}(1 - p_{\min})\}. \quad (3.21)$$

If $\lambda^*_n(p_{\min}) < \infty$, we define the dual PULSE problem by $(\text{Dual}_\lambda^*_n(p_{\min}).n)$ with solution $\hat{\alpha}_K^n(\lambda^*_n(p_{\min})) = \arg\min_{\alpha \in F'(\epsilon_1)} l^n_{\text{OLS}}(\alpha) + \lambda^*_n(p_{\min}) l^n_{\text{IV}}(\alpha)$.

THEOREM 3.2 (DUAL REPRESENTATION OF PULSE). Let $p_{\min} \in (0, 1)$ and Assumptions 3.3, 3.4, and 3.6 hold. If $\lambda^*_n(p_{\min}) < \infty$, then it holds that $t^*_n(p_{\min}) > -\infty$ and $\hat{\alpha}_K^n(\lambda^*_n(p_{\min})) = \hat{\alpha}_{\text{PULSE}}^n(t^*_n(p_{\min})) = \hat{\alpha}_{\text{PULSE}}^n(p_{\min})$.

Thus, the PULSE estimator seeks to minimize the K-class penalty $\lambda$, i.e., to pull the estimator along the K-class path $[\hat{\alpha}_K^n(\lambda) : \lambda \geq 0]$ as close to the OLS estimator as possible. Furthermore, the statement implies that the PULSE estimator is a K-class estimator with data-driven penalty $\lambda^*_n(p_{\min})$ or, equivalently, parameter $\kappa = \lambda^*_n(p_{\min})/(1 + \lambda^*_n(p_{\min}))$. Given a finite dual PULSE penalty parameter $\lambda^*_n(p_{\min})$ we can, by utilizing the closed form solution of the K-class problem, represent the PULSE estimator in the following form:

$$\hat{\alpha}_{\text{PULSE}}^n(p_{\min}) = \hat{\alpha}_K^n(\lambda^*_n(p_{\min})) = (Z^T(1 + \lambda^*_n(p_{\min})P_A)Z)^{-1}Z^T(I + \lambda^*_n(p_{\min})P_A)Y.$$ 

However, to the best of our knowledge, $\lambda^*_n(p_{\min})$ has no known closed form, so the above expression cannot be computed in closed form either. In Section 3.5.1, we prove that the PULSE penalty parameter $\lambda^*_n(p_{\min})$ can be approximated with arbitrary precision by a simple binary search procedure.

The following lemma provides a necessary and sufficient (in practice, checkable) condition for when the PULSE penalty parameter $\lambda^*_n(p_{\min})$ is finite.

LEMMA 3.5 (INFEASIBILITY OF THE DUAL REPRESENTATION). Let $p_{\min} \in (0, 1)$ and Assumptions 3.3, 3.4, and 3.6 hold. In the under- and just-identified setup we have that $\lambda^*_n(p_{\min}) < \infty$. In the over-identified setup it holds that $\lambda^*_n(p_{\min}) < \infty \iff T_n(\hat{\alpha}_{\text{TSLS}}^n) < Q_{\chi^2}(1 - p_{\min})$. This is not guaranteed to hold as the event that $\mathcal{A}_0(1 - p_{\min}) = \emptyset$ can have positive probability.

Thus, under suitable regularity assumptions Lemma 3.5 yields that our dual representation of the PULSE estimator always holds in the under- and just-identified setup. It furthermore yields a sufficient and necessary condition for the dual representation to be valid in the over-identified setup, namely that the TSLS is in the interior of the acceptance region. Furthermore, this condition is possibly violated in the over-identified setup with nonnegligible probability.
3.5.1. Binary search for the dual parameter. The key insight allowing for a binary search procedure for \(\lambda^*_n(p_{\text{min}})\) is that the mapping \(\lambda \mapsto T_n(\hat{\alpha}_n(\lambda))\) is monotonically decreasing.

**Lemma 3.6 (Monotonicity of the losses and the test statistic).** When Assumption 3.3 holds, the maps \([0, \infty) \ni \lambda \mapsto \ln \text{OLS}(\hat{\alpha}_n(\lambda))\) and \([0, \infty) \ni \lambda \mapsto \ln \text{IV}(\hat{\alpha}_n(\lambda))\) are monotonically increasing and monotonically decreasing, respectively. Consequently, if Assumption 3.4 holds, we have that the map \([0, \infty) \ni \lambda \mapsto T_n(\hat{\alpha}_n(\lambda))\) is monotonically decreasing. Furthermore, if Assumption 3.6 also holds, these monotonicity statements can be strengthened to strictly decreasing and strictly increasing.

The above lemma is proven as Lemma S4.1 in Online Appendix S4. If the OLS solution is not strictly feasible in the PULSE problem, then \(\lambda^*_n(p_{\text{min}})\) is the smallest penalty parameter for which the test statistic reaches a p-value of exactly \(p_{\text{min}}\); see Lemma S4.2 in Online Appendix S4.

We propose the binary search algorithm presented in Algorithm 1 in Online Appendix S2, that can approximate a finite \(\lambda^*_n(p_{\text{min}})\) with arbitrary precision.\(^8\) We terminate the binary search (see line 2) if \(\lambda^*_n(p_{\text{min}})\) is not finite, in which case we have no computable representation of the PULSE estimator. We now prove that Algorithm 1 achieves the required precision and is asymptotically correct.

**Lemma 3.7.** Let \(p_{\text{min}} \in (0, 1)\) and Assumptions 3.3 and 3.4 hold. If \(\lambda^*_n(p_{\text{min}}) < \infty\), then \(\lambda^*_n(p_{\text{min}})\) can be approximated with arbitrary precision by the binary search Algorithm 1, that is, \(\text{BinarySearch}(N, p_{\text{min}}) - \lambda^*_n(p_{\text{min}}) \to 0\), as \(N \to \infty\).

### 3.6. Algorithm and consistency

The dual representation of the PULSE estimator is not guaranteed to be well-defined in the over-identified setup. In particular, it is not well-defined if the TSLS is outside the interior of the acceptance region (which corresponds to a p-value of less than or equal to \(p_{\text{min}}\)). In this case, we propose to output a warning. This can be helpful information for the user, since it may indicate a model misspecification. For example, if the true relationship is in fact nonlinear, and one considers an over-identified case (e.g., by constructing different transformations of the instrument), even the TSLS may be rejected when erroneously considering a linear model; see Keane (2010) and Mogstad and Wiswall (2010). For any \(p_{\text{min}} \in (0, 1)\) we can still define an always well-defined modified PULSE estimator \(\hat{\alpha}_{\text{PULSE}+}(p_{\text{min}})\) as \(\hat{\alpha}_{\text{PULSE}+}(p_{\text{min}})\) if the dual representation is feasible, and some other consistent estimator \(\hat{\alpha}_{\text{ALT}}\) (such as the TSLS, LIML, or Fuller estimator) otherwise. That is, we define

\[
\hat{\alpha}^n_{\text{PULSE}+}(p_{\text{min}}) := \begin{cases} 
\hat{\alpha}^n_{\text{PULSE}}(p_{\text{min}}), & \text{if } T_n(\hat{\alpha}^n_{\text{TSLS}}) < Q_{\chi^2_q}(1 - p_{\text{min}}) \\
\hat{\alpha}^n_{\text{ALT}}, & \text{otherwise}.
\end{cases}
\]

Similarly to the case of an empty rejection region, we also output a warning for the case when the OLS estimator is accepted. This may, but does not have to, indicate weak instruments. Thus, we have the algorithm presented as Algorithm 2 in Online Appendix S2 for computing the PULSE+ estimator.

We now prove that the PULSE+ estimator consistently estimates the causal parameter in the just- and over-identified setting. Assume that we choose a consistent estimator \(\hat{\alpha}^n_{\text{ALT}}\) (under

\(^8\) It is possible to improve this algorithm in the under- and just-identified setup by initializing \(\ell_{\text{max}}\) as the quantity given by equation (S5.17) in the proof of Lemma 3.5. This initialization removes the need for the first while loop in lines 4–6.
standard regularity assumptions, this is satisfied for the TSLS).\textsuperscript{9} We can then show that, under mild conditions, the PULSE$^+$ estimator, too, is a consistent estimator of $\alpha_0$.

**Theorem 3.3 (Consistency of PULSE$^+$).** Consider the just- or over-identified setup and let $p_{\text{min}} \in (0, 1)$. If Assumptions 3.1, 3.3, 3.4, 3.5, and 3.6 hold almost surely for all $n \in \mathbb{N}$ and $\hat{\alpha}_{\text{ALT}}^n$ consistently estimates $\alpha_0$, then $\hat{\alpha}_{\text{PULSE}^+}(p_{\text{min}}) \xrightarrow{p} \alpha_0$, when $n \to \infty$.

We believe that a similar statement also holds in the under-identified setting; see Online Appendix S8.3.

4. SIMULATION EXPERIMENTS

In Online Appendix S8 we conduct an extensive simulation study investigating the finite sample behaviour of the PULSE estimator. The concept of weak instruments is central to our analysis. An introduction to weak instruments can be found in Online Appendix S10. Here, we give a brief overview of the study and the observations.

4.1. Distributional robustness

The theoretical results on distributional robustness proved in Section 2 translate to finite data. The experiments in Online Appendix S8.1 shows that, even for small sample sizes, K-class estimators outperform both OLS and TSLS for a certain range of interventions, matching the theoretical predictions with increasing sample size. In Online Appendix S8.3 we also consider an under-identified setting.

4.2. Estimating causal effects

When focusing on the estimation of a causal effect in an identified setting, our simulations show that there are several settings where PULSE outperforms the Fuller and TSLS estimators in terms of MSE. In univariate simulation experiments, such settings are characterized by weakness of instruments and weak confounding (endogeneity). The characterization becomes more involved in multivariate settings, but is similar in that PULSE outperforms all other methods for small confounding strengths, an effect amplified by the weakness of instruments. Below we detail the univariate simulation setup and refer the reader to Online Appendix S8 for further details and the multivariate simulation experiments mentioned above.

4.2.1. Univariate model. We first compared performance measures of the estimators in a univariate instrumental variable model. As seen in Hahn and Hausman (2002) and Hahn et al. (2004), we consider structural equation models of the form

\[ A := A \in \mathbb{R}^q, \quad X := A^\top \xi + U_X \in \mathbb{R}, \quad Y := X \gamma + U_Y \in \mathbb{R}, \]

where $A \sim \mathcal{N}(0, I)$ and $A_{\parallel}(U_X, U_Y)$ with $(U_X^\top, U_Y^\top) \sim \mathcal{N}\left((0,0), \left(\begin{array}{c} 1 \\ \rho \end{array}\right)\right)$. Furthermore, we let $\gamma = 1$ and $\xi^\top = (\xi, \ldots, \xi) \in \mathbb{R}^q$, where $\xi > 0$ is chosen according to the theoretical $R^2$-coefficient. We consider the following simulation scheme: for each $q \in \{1, 2, 3, 4, 5, 10, 20, 30\}$, $\rho \in [0, 1]$, $\xi > 0$ is chosen according to the theoretical $R^2$-coefficient.

\textsuperscript{9} This holds as $\hat{\alpha}_{\text{TSLS}}^n = \alpha_0 + \left(n^{-1}Z^\top A(n^{-1}A^\top A)^{-1}n^{-1}A^\top Z)^{-1}n^{-1}Z^\top A(n^{-1}A^\top A)^{-1}n^{-1}A^\top U_Y \right.$.
\[
\{0.1, 0.2, ..., 0.9\}, \quad R^2 \in \{0.0001, 0.001, 0.01, 0.1, 0.3\} \quad \text{and} \quad n \in \{50, 100, 150\},
\]
we simulate \(n\)-samples from the above system and calculate the OLS, TSLS, Fuller(1), Fuller(4), and PULSE (\(p_{\text{min}} = 0.05\)) estimates; see Online Appendix S8.2.1.

Figure 2 contains illustrations of the relative change in square-root mean squared error (RMSE) estimated from 15000 repetitions. On the horizontal axis we have plotted the average first stage F-test as a measure of weakness of instruments; see Online Appendix S10 for further details. A test for \(H_0 : \bar{\xi} = 0\), i.e., for the relevancy of instruments, at a significance level of 5\% has different rejection thresholds in the range \([1.55, 4.04]\) depending on \(n\) and \(q\). The vertical dashed line corresponds to the smallest rejection threshold of 1.55 and the dotted line corresponds to the ‘rule of thumb’ threshold of 10. Note that the lowest possible negative relative change is \(-1\) and a positive relative change means that PULSE is better.

In Online Appendix S11, further illustrations of, e.g., the relative change in mean bias and variance of the estimators are presented. We also conducted the simulations for setups with combinations of \(\gamma \in \{-1, 0\}\), components of \(\bar{\xi}\) chosen negatively, with random flipped sign in each coordinate, and for negative \(\rho\) (not shown but available in the folder ‘Plots’ in the code repository). The results with respect to MSE are similar to those shown in Figure 2, while the bias comparison changes depending on the setup.

We observe that there are settings in which the PULSE is superior to TSLS, Fuller(1), and Fuller(4) in terms of MSE. This is particularly often the case in weak instrument settings (\(\hat{E}_N(G_n) < 10\)) for low confounding strength (\(\rho \leq 0.2\)). Furthermore, as we tend towards the weakest instrument setting considered, we also see a gradual shift in favour of PULSE for higher confounding strengths. In these settings with weak instruments and low confounding we also see that OLS is superior to the PULSE in terms of MSE. However, for large confounding setups PULSE is superior to OLS in terms of both bias and MSE, and this superiority increases as the instrument strength increases. The PULSE is generally more biased than the Fuller and TSLS
estimators, but less biased than OLS. However, in the settings with weak instruments and low confounding, the bias of PULSE and OLS is comparable. In summary, the PULSE is, in these settings, more biased but its variance is so small that it is MSE superior to the Fuller and TSLS estimators.

5. EMPIRICAL APPLICATIONS

We now consider three classical instrumental variable applications; see Albouy (2012) and Buckles and Hungerman (2013) for discussions on the underlying assumptions.

(i) ‘Does compulsory school attendance affect schooling and earnings?’ by Angrist and Krueger (1991). This paper investigates the effects of education on wages. The endogenous effect of education on wages are remedied by instrumenting education on quarter of birth indicators.

(ii) ‘Using geographic variation in college proximity to estimate the return to schooling’ by Card (1993). This paper also investigates the effects of education on wages. In this paper, education is instrumented by proximity to college indicator.

(iii) ‘The colonial origins of comparative development: An empirical investigation’ by Acemoglu et al. (2001). This paper investigates the effects of extractive institutions (proxied by protection against expropriation) on the gross domestic product (GDP) per capita. The endogeneity of the explanatory variables are remedied by instrumenting protection against expropriation on early European settler mortality rates.

We have applied the different estimators OLS, TSLS, PULSE, and Fuller to the classical data sets Acemoglu et al. (2001), Angrist and Krueger (1991), and Card (1993). All models considered in Angrist and Krueger (1991) and Card (1993), where we estimate the effect of years of education on wages using quarter of birth and proximity to colleges as instruments, respectively, the OLS estimates are not rejected by our test statistic, and PULSE outputs the OLS estimates; see Online Appendix S9 for further details. This may be either due to weak endogeneity (weak confounding) or that the test has insufficient power to reject the OLS estimates due to weak instruments or severe over-identification.

5.1. Acemoglu et al. (2001)

The dataset of Acemoglu et al. (2001) consists of 64 observations, each corresponding to a different country for which mortality rate estimates encountered by the first European settlers are available. The endogenous target of interest is log GDP per capita (in 1995). The main endogenous regressor in the dataset is an index of expropriation protection (averaged over 1985–1995), i.e., protection against expropriation of private investment by the respective governments. The average expropriation protection is instrumented by the settler mortality rates. We consider eight models, M1–M8, which correspond to the models presented in column (1)–(8) in table 4 of Acemoglu et al. (2001). Model M1 is given by the reduced form structural equations

\[
\begin{align*}
\log \text{GDP} &= \text{avexpr} \cdot \gamma + \mu_1 + U_1, \\
\text{avexpr} &= \log \text{em4} \cdot \delta + \mu_2 + U_2,
\end{align*}
\]

where avexpr is the average expropriation protection, em4 is the settler mortality rates, \(\mu_1\) and \(\mu_2\) are intercepts, and \(U_1\) and \(U_2\) are possibly correlated, unobserved noise variables. In model M2 we additionally introduce an included exogenous regressor describing the country latitude.
Table 1 shows the OLS and TSLS estimates (which replicate the values from the study), as well as the Fuller(4) and PULSE estimates for the linear effect of the average expropriation protection on log GDP. In model M1, for example, we see that the PULSE estimate suggests that the average expropriation risk linear effect on log GDP is 0.6583, which is 26% larger than the OLS estimate but 34% smaller than the TSLS estimate. In models M5–M8, the OLS estimates are not rejected by the Anderson–Rubin test, so the PULSE estimates coincide with the OLS estimates.

We can also use this example to illustrate the robustness property of K-class estimators; see Theorem 2.1. Even though interventional data are not available, we can consider the mean squared prediction error when holding out the observations with the most extreme values of the instrument. Depending on the degree of generalization, we indeed see that the PULSE and Fuller tend to outperform OLS or TSLS in terms of mean squared prediction error on the held out data; see Online Appendix S9.3 for further details.

6. SUMMARY AND FUTURE WORK

We have proven that a distributional robustness property similar to the one shown for anchor regression (Rothenhäusler et al., 2021) fully extends to general K-class estimators of possibly nonidentifiable structural parameters in a general linear structural equation model that allows for latent endogenous variables. We have further proposed a novel estimator for structural parameters in linear structural equation models. This estimator, called PULSE, is derived as the solution to a minimization problem, where we seek to minimize mean squared prediction error constrained to a confidence region for the causal parameter. Even though this region is nonconvex, we have shown that the corresponding optimization problem allows for a computationally efficient algorithm that approximates the above parameter with arbitrary precision using a simple binary search.
procedure. In the under-identified setting, this estimator extends existing work in the machine learning literature that considers invariant subsets or the best predictive sets among them: PULSE is applicable even in situations when no invariant subsets exist. We have proven that this estimator can also be written as a K-class estimator with data-driven $\kappa$-parameter, which lies between zero and one. Simulation experiments show that in various settings with weak instruments and weak confounding, PULSE outperforms other estimators, such as the Fuller(4) estimator. We thus regard PULSE as an interesting alternative for estimating causal effects in instrumental variable settings. It is easy to interpret and automatically provides the user feedback in case that the OLS is accepted (which may be an indication that the instruments are too weak) or that the TSLS is outside the acceptance region (which may indicate a model misspecification). We have applied the different estimators to classical data sets and have seen that, indeed, K-class estimators tend to be more distributionally robust than OLS or TSLS.

There are several further directions that we consider worthwhile investigating. This includes better understanding of finite sample properties and, for the identified setups, the study of loss functions other than MSE. It would be helpful, in particular with respect to real world applications, to understand to what extent similar principles can be applied to models allowing for a time structure of the error terms. We believe that the simple primal form of PULSE could make it applicable for model classes that are more complex than linear models (see also Christiansen et al., 2020). Our procedure can be combined with other tests, and it could be interesting to find efficient optimization procedures for tests that are robust with respect to weak instruments, such as Kleibergen’s K-statistic (Kleibergen, 2002) for example. In an under-identified setting, the causal parameters are not identified, but the solutions obtained by optimizing predictability under invariance might be promising candidates for models that generalize well to distributional shifts.

ACKNOWLEDGEMENTS

We are grateful to Trine Boomsma, Peter Bühlmann, Rune Christiansen, Steffen Lauritzen, Nicolai Meinshausen, Whitney Newey, Cosma Shalizi, and Nikolaj Thams for helpful discussions. We thank the editor and two anonymous referees for helpful and constructive comments. MEJ and JP were supported by a research grant (CF17-0883) from The Carlsberg Foundation; JP was, in addition, supported by a research grant (18968) from VILLUM FONDEN.

REFERENCES


© The Author(s) 2022.


© The Author(s) 2022.


**SUPPORTING INFORMATION**

Additional Supporting Information may be found in the online version of this article at the publisher’s website:

Online Appendix
Replication Package

*Co-editor Victor Chernozhukov handled this manuscript.*