King's Research Portal

[Link to publication record in King's Research Portal](#)

Stein-like Estimators for Causal Mediation Analysis in Randomized Trials

Cedric E. Ginestet[1], Richard Emsley[1], and Sabine Landau[1]

Abstract

Causal mediation analysis aims to estimate natural direct and natural indirect effects under clearly specified assumptions. Traditional mediation analysis based on Ordinary Least Squares (OLS) assumes an absence of unmeasured causes to the putative mediator and outcome. When these assumptions cannot be justified, Instrumental Variables (IV) estimators can be used in order to produce an asymptotically unbiased estimator of the mediator-outcome link, commonly referred to as a Two-Stage Least Squares (TSLS) estimator. Such bias removal, however, comes at the cost of variance inflation. A Semi-Parametric Stein-Like (SPSL) estimator has been proposed in the literature that strikes a natural trade-off between the unbiasedness of the TSLS procedure and the relatively small variance of the OLS estimator. The SPSL has the advantage of allowing for a direct estimation of its shrinkage parameter. In this paper, we demonstrate how this Stein-like estimator can be implemented in the context of the estimation of natural direct and natural indirect effects of treatments in randomized controlled trials. The performance of the competing methods is studied in a simulation study, in which both the strength of hidden confounding and the strength of the instruments are independently varied. These considerations are motivated by a trial in mental health evaluating the impact of a primary care-based intervention to reduce depression in the elderly.

Keywords

Causal mediation analysis, Instrumental variables, Stein estimator, Randomized trials, Two-stage least squares

[1]Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London.

Corresponding author:
Cedric E. Ginestet, Department of Biostatistics and Health Informatics Institute of Psychiatry, Psychology and Neuroscience King's College London, PO20, 16 De Crespigny Park, London SE5 8AF, UK.
Email: cedric.ginestet@kcl.ac.uk

Introduction

Mediation analysis has become a popular approach to data analysis in a variety of disciplines. This approach permits to study alternative causal paths linking an experimental factor of interest with a particular outcome[1]. It has been especially successful in the context of mental health, where psychologists and psychiatrists are particularly interested in the mechanisms of action of a given treatment. These mechanisms are usually studied with respect to certain intermediate variables that are likely to be related to the personality, cognition and social environment of the individuals that are taking part in the study.

In mental health, we are often concerned with evaluating the effect of psychological therapy on clinical outcome, with respect to certain intermediate variables. When the indirect effect of the treatment through the intermediate variable is of interest, such a variable is referred to as a *target mediator*. By contrast, when we are controlling for the intermediate variable, and the primary interest of the study lies in estimating the direct effect of treatment on the outcome; we refer to such a variable as a *nuisance mediator*. Often, the distinction between a target and a nuisance mediator depends on whether or not the mediator constitutes an alternative form of treatment. This is the case in the PROSPECT data set that motivates this study, in which the effect of psychotherapy is mediated by adherence to a course of anti-depressant medication.

Several theoretical frameworks have been proposed for studying mediation from a causal perspective. Such approaches tend to build upon the foundational work of Baron and Kenny[2] (1986), who have established the basis of mediation analysis. This framework has then been formalized in order to allow for causal inference. The first formalization of causal mediation analysis was given by Robins and Greenland[3] (1992); and several variants have been proposed in the literature, including the works of Pearl[4] (2001), Rubin[5] (2004), and VanderWeele[6] (2008). In the paper at hand, we will describe causal mediation in terms of potential outcomes, using the notation and the set of assumptions adopted by Imai, Keele and Yamamoto[7] (2010). Throughout this article, we will assume that the outcome of interest is continuous. In this setting, the main estimands of interest are the natural direct and natural indirect effects, denoted NDE and NIE respectively. In trials, such quantities can be estimated without bias, under the assumption that the intermediate variable is exogenous in the model for the outcome. (A predictor of the outcome variable is said to be *exogenous*, whenever it is not correlated with the error term in the model, and *endogenous*, otherwise.)

In practice however, this exogeneity assumption can be difficult to justify, due to the likely presence of baseline variables that are common causes of the intermediate and the clinical outcome variable. One of the proposed solutions to this problem has been the use of instrumental variables (IVs), which can be combined with mediation analysis, in order to draw causal inference. (See Lynch et al.[8] (2008) and Ten Have et al.[9] (2012) for a review of causal mediation analysis.) In this article, we will specifically focus on the use of interaction terms as instruments, constructed by interacting the experimental factor with the baseline covariates. Note, however, that our methods can readily be generalized to other IVs.

The most common estimator using IVs is the Two-Stage Least Squares (TSLS)[10], which relies on further assumptions about the behavior of the candidate instruments. Under these additional assumptions, the asymptotic properties of the TSLS estimator are well-understood. Provided that the instruments solely affect the outcome through the endogenous variable of interest, the TSLS estimator is guaranteed to be asymptotically unbiased[10]. For finite sample sizes, however, the decrease in bias associated with the use of this estimator, will lead to an increase in variance. In particular, there may be situations in which the

variance increase of the TSLS estimator does not warrant preferring that estimator over the potentially biased Ordinary Least Squares (OLS) estimator.

In this paper, we follow the lead of Judge and Mittelhammer[11] (2004), who have constructed a combined estimator, which strikes a trade-off between the OLS and the TSLS estimators, by minimizing the Mean Squared Error (MSE) of the resulting combined estimator. This method closely resembles the so-called Stein estimator, originally introduced by James and Stein[12] (1961), and made popular by Efron[13] (1973). Stein estimators have anticipated some of the central ideas of Bayesian statistics, by shifting the main focus of statistical analysis from minimizing an estimator's unbiasedness to minimizing an estimator's MSE. These ideas are best articulated within the language of decision theory. (See (author?)[14], for an introduction to decision theory.) From this perspective, the MSE can be formalized as a loss function, and the optimal estimator is the quantity that minimizes that function. These estimators bear some similarities with other families of estimators that rely on loss functions, whether or not these are articulated within a Bayesian framework. See, for instance, the minimum expected loss (MELO) estimator[15,16]. From a decision-theoretic perspective, the MSE can be formalized as a loss function, and the optimal estimator is the estimator that minimizes that function. The Semi-Parametric Stein-Like (SPSL) estimator is defined as an affine combination of the OLS and TSLS estimators; where the shrinkage parameter controlling the respective contributions of the OLS and TSLS estimators can be estimated from the data, under the assumption that the TSLS estimator is asymptotically unbiased.

The main contributions of this paper are twofold. Firstly, we provide the first use of the SPSL estimator in the context of causal mediation analysis. The SPSL will here be compared with standard estimators, including the OLS and TSLS estimators for estimating the effect of endogenous intermediate variables in causal mediation; where OLS estimation here corresponds to the standard Baron–Kenny approach. The Baron–Kenny framework generally assumes that the mediators are continuous, whereas our approach enables us to accommodate binary mediators. Note, however, that binary mediators can only be incorporated at the cost of a loss in efficiency of the resulting estimators. The asymptotic behaviors of the family of SPSL estimators have recently been well-studied[17–20]. The SPSL estimator has been used to investigate Local Average Treatment Effects (LATEs) in dose-response models[21]. However, to the best of the authors' knowledge, this family of estimator has not been used in the context of causal mediation analysis, when the estimation of the causal path from the intermediate variable to the outcome is potentially biased, due to unmeasured confounding.

Secondly, we generalize the SPSL estimator by allowing for the selection of a subset of parameters that affects the optimization of the shrinkage parameter. Indeed, in many circumstances, one is solely interested in the estimation of a particular set of estimands, and it is therefore convenient to be able to restrict the dependence of the shrinkage parameter on the MSE of a subset of target estimands. In this paper, we implement such a restriction by introducing a projection matrix, which permits to restrain the estimation of the shrinkage parameter to a subset of the parameters of interest, such as the direct effect of treatment, for instance. This provides a generalization of the closed-form formula for the SPSL, originally derived by Judge and Mittelhammer[11].

Our use of the SPSL estimator for causal mediation analysis is motivated by a clinical trial in mental health. The Prevention of Suicide in Primary Care Elderly: Collaborative Trial, more concisely referred to as PROSPECT[22], is a randomized controlled trial, which tested the effect of a primary care intervention on major risk factors for suicide in an elderly population, and in which the intermediate variable is whether or not patients are taking antidepressant medication. This particular study has served as a

motivating example for several causal mediation analyses previously published in the literature, including studies by Ten Have et al. (2007)[23], Emsley et al. (2010)[24], and Small (2012)[25]. In the paper at hand, we replicate some of these previous results, and compare them with the performance of the SPSL estimator for this data set.

The PROSPECT data set is an unusual example of a mediation analysis, since the main estimand of interest is the NDE. That is, we wish to evaluate whether or not the psychotherapeutic intervention affects the outcome, after having controlled for the effect of taking antidepressant medication. This should be contrasted with most other mediation studies, in which one is typically interested in estimating the NIE –that is, the effect of the target mediator on the outcome. In contradistinction, the intermediate variable in the PROSPECT data may be regarded as a nuisance mediator, which is solely of secondary interest to the trialists.

The paper is organized as follows. In the first section, we introduce the causal estimands of interest, and describe how such parameters can be estimated using the OLS, the TSLS and the SPSL estimator. The performance of our three competing estimators for causal mediation analysis is then evaluated by means of a Monte Carlo simulation study, in the second section. The methods are then applied to a re-analysis of the PROSPECT data set, in the third section; and we close with a discussion of the limitations and further generalizations of such estimators in our final section. The proofs of the main results in the paper are deferred to an appendix.

## Causal Mediation Analysis

Causal Estimands

The sample data are assumed to have been collected as part of a clinical trial, in which R_i denotes randomized treatment offer to the ith subject. The clinical outcome of interest, denoted by Y_i, is a continuous post-randomization variable, and M_i is the putative mediator under investigation, which is also a post-randomization variable. The mediator is here assumed to have been measured before the outcome. The mediator may be either binary or continuous, albeit note that the use of binary mediators in our framework, entails a considerable loss in efficiency. In addition, there are also k pre-randomization (or baseline) variables, denoted by a random vector, X_i, such that $X_i := (X_{i1}, \ldots, X_{ik})'$. Without any loss of generality, these baseline variables may also be either binary or continuous.

For every $r \in \{0, 1\}$, and for every $m \in \mathbb{R}$, the potential outcome $Y_i(r, m)$ is defined as the outcome that would be observed for the ith subject, if R_i and M_i were to take values r and m, respectively. Several possible mechanisms have been proposed in the literature that allow the potential outcomes, $Y_i(r, m)$, to take different values according to different choices of r and m[23,25]. Similarly, the potential mediator, $M_i(r)$, is defined as the value taken by the mediator in the ith subject, when the value of R is r. The aforementioned observed outcomes and observed mediators are then defined as a function of the potential outcomes and potential mediators, such that we have $Y_i := Y_i(R_i, M_i)$, and $M_i := M_i(R_i)$, respectively.

For every subject, every r, and every m, the potential outcomes are given the following structural model,

$$Y_i(r, m) := Y_i(0, 0) + \beta_{R,i} r + \beta_{M,i} m, \tag{1}$$

with $Y_i(0, 0) := \boldsymbol{\beta}_X' X_i + \omega_i$, and $\mathbb{E}[\omega_i] = 0$. and where the parameters, $\beta_{M,i}$ and $\beta_{R,i}$, can vary between subjects reflecting treatment effect and mediator effect heterogeneity, respectively.
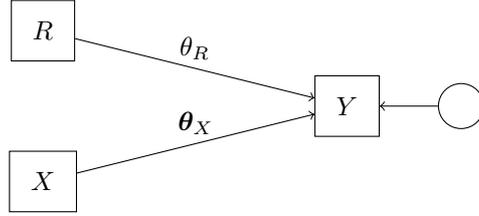
Figure 1. Graphical representation of the model of the total effect (TE) of R on Y, as described in equation (4); where the empty circle denotes an error term. Since subjects have been randomly assigned to the different levels of treatment allocation, R; it follows that R is an exogenous predictor of Y.

The parameters in model (1) can thus be interpreted in the following manner. Given a subject i, the parameter, $\beta_{M,i}$, denotes the effect caused by a unit increase in the mediator on the outcome, holding treatment level at r. Similarly, $\beta_{R,i}$ should be interpreted as the effect of treatment on the outcome, while holding the mediator constant at level m. Finally, we will respectively denote by $\beta_M := \mathbb{E}[\beta_{M,i}]$ and $\beta_R := \mathbb{E}[\beta_{R,i}]$, the average causal effect of the mediator and the average effect of the treatment on the outcome.

Using our definitions of the observed outcome, Y_i, and of the counterfactual, $Y_i(0,0)$; we obtain the following linear model for the observed outcome,

$$Y_i = \boldsymbol{\beta}'_X X_i + \beta_R R_i + \beta_M M_i + \varepsilon_i, \tag{2}$$

for every $i = 1, \ldots, n$; in which $\boldsymbol{\beta}_X$ is a k-dimensional column vector of unknown parameters containing an intercept, and where the error terms comprise the individual deviations from the average causal effect,

$$\varepsilon_i := \omega_i + \big(\beta_{M,i} - \beta_M\big)M_i + \big(\beta_{R,i} - \beta_R\big)R_i. \tag{3}$$

where recall that $\omega_i = Y_i(0,0) - \mathbb{E}[Y_i(0,0)|X_i]$.

Treating the mediator as unobserved, the potential outcomes can be described by the following structural model,

$$Y_i(r) = \boldsymbol{\theta}'_X X_i + \theta_{R,i} r + \xi_i,$$

where as before, $\theta_R := \mathbb{E}[\theta_{R,i}]$ denotes the average effect of treatment offer on the outcome, and with $\mathbb{E}[\xi_i] = 0$. This then leads to the following model for the observed outcomes,

$$Y_i = \boldsymbol{\theta}'_X X_i + \theta_R R_i + \nu_i, \qquad \text{with } \nu_i := \xi_i + (\theta_{R,i} - \theta_R)R_i. \tag{4}$$

This model is represented in Figure 1. Observe that in this model, the parameter, θ_R, corresponds to the Intention-To-Treat (ITT) effect, assuming no missing data.

For continuous mediators, we can translate our choice of notation, into the conventional Baron-Kenny notation[2]. If we were to represent the average causal effect of R on M by γ_R; we could then adopt the following notation, $a := \gamma_R$, $b := \beta_M$, $c' := \beta_R$, and $c := \theta_R$.

The main estimands of interest will be the natural direct effect (NDE) and the natural indirect effect (NIE). For continuous outcomes, the total effect (TE) can be decomposed such that

$$\text{TE} := \mathbb{E}\big[Y_i(1) - Y_i(0)\big] = \mathbb{E}\big[Y_i(1, M_i(1)) - Y_i(0, M_i(0))\big].$$
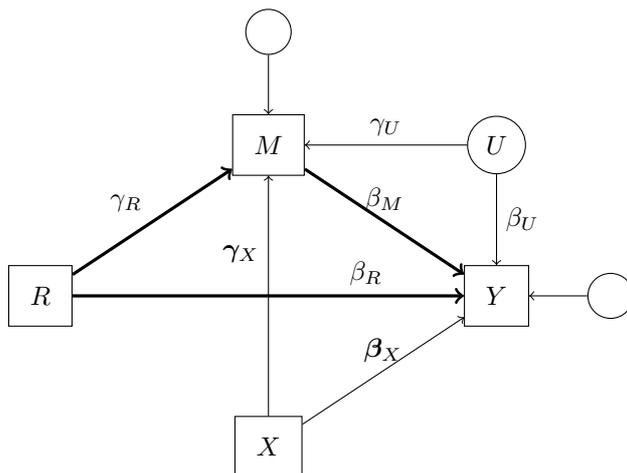
Figure 2. Graphical representation of the mediation model described in equation (2) in the presence of a confounder, U; where empty circles denote error terms. For continuous mediators, we have the following correspondence between the above notation and the standard Baron-Kenny notation: $a = \gamma_R$, $b = \beta_M$, and $c' = \beta_R$; in which R and M denote treatment offer and the mediator, respectively. The three paths of interest in mediation investigations, have been emphasized in bold.

In our notation, TE corresponds to effect of treatment offer on the outcome, according to the structural model in equation (4), such that $\text{TE} = \theta_R$. The NDE, on the other hand, is defined as follows,

$$\text{NDE} := \mathbb{E}[Y_i(1, M_i(0)) - Y_i(0, M_i(0))] = \beta_R.$$

Finally, for continuous Y_i's, the NIE can be expressed as a difference between these two estimands. Formally, this gives

$$\text{NIE} := \mathbb{E}[Y_i(1, M_i(1)) - Y_i(1, M_i(0))] = \theta_R - \beta_R.$$

This expression for the NIE with continuous outcomes is convenient, because it covers both continuous and binary mediators, although binary mediators should be used with caution, as aforementioned.

OLS Estimator

Observe that since both the baseline covariates, X_i's, and the randomization variable, R_i, are exogenous, it follows that parameters, γ_R and θ_R, can be unbiasedly estimated using OLS. However, there is no guarantee that the effect of the mediator on the outcome is not confounded by an unmeasured variable. Therefore, a naive OLS estimator of the parameter, β_M, may be biased. Similarly, the OLS estimator of β_R may also be biased due to the endogeneity of the mediator in Equation (2). A simplified version of such a causal mediation model, in the presence of a confounder, U_i's, has been represented in Figure 2. In Figure 2, β_R and β_M are biased due to unmeasured confounding, since the intermediate variable, M_i, is endogenous in the model of the Y_i's, in this figure.

Various sets of assumptions can be used in order to conduct causal mediation analysis. For the OLS estimator, we will use a set of assumptions referred to as sequential ignorability[7]. For every $r \in \{0, 1\}$, and every $m \in \mathbb{R}$, sequential ignorability assumes that

(OLS–1) $Y_i(r, m) \perp R_i \,|\, X_i$.

(OLS–2) $M_i(r) \perp R_i \,|\, X_i$.

(OLS–3) $Y_i(r, m) \perp M_i \,|\, X_i$.

These assumptions respectively state the following: ignorable treatment assignment in terms of the outcome, given the covariates, (OLS–1); ignorable treatment assignment in terms of the mediator, given covariates, (OLS–2); and ignorable mediator assignment, given covariates (OLS–3).

Observe that, whenever the R_i's correspond to random allocation to treatment offer, as in our motivating trial, it then follows that conditions (OLS–1) and (OLS–2) are automatically satisfied. The fact that (OLS–1) holds in our setting, allows us to unbiasedly estimate the causal effect of treatment offer on the outcome, denoted by θ_R; using an OLS estimator, denoted by $\widetilde{\theta}_R$. Similarly, the fact that (OLS–2) holds permits us to unbiasedly estimate the causal effect of treatment offer on the mediator, denoted by γ_R, using an OLS estimator, denoted by $\widetilde{\gamma}_R$.

Moreover, it is additionally assumed, for regulatory reasons, that the following strict inequalities hold, $\mathbb{P}(R = r | X = x) > 0$, and $\mathbb{P}(M = m | R = r, X = x) > 0$. Under the model for the potential outcomes described in equation (1), the third assumption of sequential ignorability given in (OLS–3) can be reformulated as follows,

$$Y_i(0, 0), \beta_{M,i}, \beta_{R,i} \perp M_i \,|\, X_i.$$

Therefore, under sequential ignorability, the three random variables on the RHS of equation (1) are assumed to be conditionally independent of the mediator, given the values of the baseline covariates.

In the absence of unmeasured confounders between the M_i's and the Y_i's, sequential ignorability holds, and one can estimate the NIE and NDE by computing the OLS estimator of the direct effect of treatment offer on the outcome, denoted β_R. For convenience, the parameters of interest in the model described in equation (2) will be collectively denoted as a vector,

$$\boldsymbol{\beta} := (\boldsymbol{\beta}'_X, \beta_M, \beta_R)'.$$

Similarly, all the variables in this model will be expressed as the random vector,

$$V_i := (X'_i, M_i, R_i)',$$

where recall that X_i represents a k-dimensional column vector of baseline covariates including an intercept, whereas M_i and R_i are real-valued random variables, although M_i is also allowed to be binary. Thus, each V_i is a $(k + 2)$-dimensional random vector. In addition, a set of n observations from the Y_i's will be denoted by the vector $\mathbf{y}$, while a set of n realizations from the V_i's will take the form of a matrix of order $n \times (k + 2)$, denoted $\mathbf{V}$.

Under the further assumption that the matrix $\mathbb{E}[V_i V'_i]$ is full-rank, we can compute the OLS estimator.

(OLS–4) $\text{rank}(\mathbb{E}[V_i V'_i]) = k + 2$.

The OLS estimator for $\boldsymbol{\beta}$, which is uniquely given by the vector that minimizes the empirical MSE, and takes the form, $\widetilde{\boldsymbol{\beta}} := (\mathbf{V}'\mathbf{V})^{-1}(\mathbf{V}'\mathbf{y})$. Moreover, the empirical variance of this estimator is given by $\widehat{\sigma}^2(\mathbf{V}'\mathbf{V})^{-1}$, where $\widehat{\sigma}^2$ is defined as $(\mathbf{y} - \mathbf{V}\widetilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{V}\widetilde{\boldsymbol{\beta}})/(n - k - 2)$. Moreover, the vector of parameters for the total effects, $\boldsymbol{\theta} := (\boldsymbol{\theta}'_X, \theta_R)'$, from equation (4) can also be estimated using OLS, thereby producing the following estimators of the natural effects: $\widetilde{\mathrm{NDE}} := \widetilde{\beta}_R$, and $\widetilde{\mathrm{NIE}} := \widetilde{\theta}_R - \widetilde{\beta}_R$. (Hence, observe that albeit we are estimating the full vector of parameters, $\boldsymbol{\beta}$; the sole element of interest in this vector for estimating NDE and NIE is β_R.)

TSLS Estimator

In the presence (or suspected presence) of unmeasured confounders, different assumptions are required in order to estimate the parameters of interest without bias. In the data at hand, although allocation to treatment has been randomized, both M_i and Y_i are post-randomization variables, which may be affected by common causes. Therefore, one cannot guarantee that the path from the mediator to the outcome has not been confounded by an unobserved variable. When unmeasured confounders affect the relationship between the outcome and the mediator, as illustrated in Figure 2, the third portion of sequential ignorability, (OLS–3) does not hold, and further assumptions are hence required to ensure that such a model is identifiable.

Several groups of researchers have used instruments that are defined as interactions between certain baseline variables and random assignment to treatment [23,26,27]. Such choices of IVs require a particular set of assumptions, which ensure that the resulting variables constitute valid instruments. In this paper, we will consider a variant of the conditions described by Small [25]. These assumptions apply to general mediation models that make use of such interaction terms as instruments. For consistency with the previous literature on this topic, we will also adopt some of the notation used by Small [25], throughout the rest of this section. However, we should emphasize that SPSL estimation in causal mediation, is not restricted to the use of interaction terms as instruments.

Here, we supplemented the model for the observed outcome, Y_i's, with a predictive model for the observed continuous (or binary) mediator, M_i's, such that we obtain the system of equations that has also been illustrated graphically in Figure 3,

$$\begin{aligned} Y_i &= \boldsymbol{\beta}'_X X_i + \beta_R R_i + \beta_M M_i + \varepsilon_i, \\ M_i &= \boldsymbol{\gamma}'_X X_i + \gamma_R R_i + \boldsymbol{\gamma}'_{RX} R_i X_i + \delta_i; \end{aligned} \tag{5}$$

with $\delta_i := \mathbb{E}[M_i|X_i, R_i,] - M_i$, and $\mathbb{E}[\delta_i] = 0$; and where the ε_i's and the δ_i's are assumed to be independent. Furthermore, note that the M_i's are here modelled linearly, despite the fact that this variable may be binary. This does not pose a problem per se, as long as the instruments, $R_i X_i$'s, are predictive of the M_i's. That is, mis-specification of the functional form of this model (e.g. as a linear regression, when, in reality, this is a logistic regression), while leading to difficulties interpreting the gamma's; does not affect the estimation of the parameters of interest, which are the beta's in the model for the Y_i's, since TSLS estimation solely requires a correct specification of the model for the outcomes. Thus, the error terms, δ_i's, of the linear model for the mediator need not be normally distributed.

For convenience, we will define the set of instruments as the following vectors,

$$Z_i := (X'_i, R_i, R_i X'_i)',$$
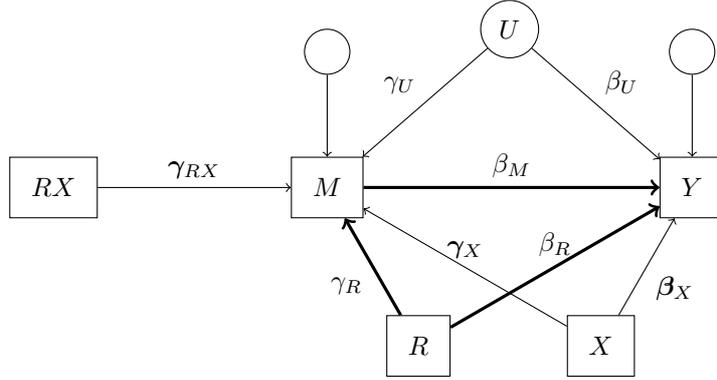
Figure 3. Graphical representation of the instrumented mediation model described in equation (5), in which the relationship between the mediator, M, and the outcome Y, is confounded by the presence of an unknown variable U; where, as before, empty circles denote error terms. The interaction instrument, RX, is here used to handle the endogeneity of M; while the randomization variable, R, and the baseline covariates, X, are all assumed to be exogenous. As in Figure 2, the three links defining the main causal mediation model, have been emphasized in bold.

where each such Z_i is a $(2k + 1)$-dimensional column vector. Equipped with this notation, we can then state the assumptions required to guarantee the validity of the $R_i X_i$'s as instruments. We will assume that the model in Equation (5) is the true model, and that the following conditions hold for every subject,

(TSLS–1) $M_i \perp \beta_{M,i} | R_i, X_i$.

(TSLS–2) $\beta_R = \mathbb{E}[\beta_{R,i} | X_i]$, and $\beta_M = \mathbb{E}[\beta_{M,i} | X_i]$, for every X_i.

(TSLS–3) $\mathbb{E}[Z_i Z_i']$, and $\mathbb{E}[Z_i V_i']$ are full-rank.

(TSLS–4) $\mathbb{C}\text{ov}(V_i, Z_i) \neq \mathbf{0}$.

Here, (TSLS–1) should be interpreted as the independence of the individual mediator effects with the mediator. Condition (TSLS–4) is commonly referred to in the literature on causal inference, as the *relevance* of the IVs. In addition, observe that assumption (TSLS–2) is weaker than the ones made by previous authors, who have used interaction terms as instruments, and who have assumed homogeneous treatment effects[23,26,27], such that the $\beta_{M,i}$'s, and $\beta_{R,i}$'s are assumed to be identical for all subjects. Here, by contrast, we have only required these parameters to have identical conditional expectations given the X_i's, as stated in condition (TSLS–2).

Also, note that this set of assumptions slightly differs from the one described by Small[25], since we have replaced the assumption that this author refers to as (IV–A1), by an assumption on the ranks of the matrices $\mathbb{E}[Z_i Z_i']$, and $\mathbb{E}[Z_i V_i']$, which we refer to as (TSLS–3). The latter assumption is here expressed in terms of the ranks of the expectations of the cross-products of the vector of instruments, and the vector of covariates. This condition is a relatively weak requirement that guarantees the identifiability of the resulting TSLS estimator[10]. However, further work will be needed to clarify whether an assumption of the form (TSLS–3) is weaker, stronger or simply equivalent to the corresponding assumption made by Small (2012).

We can now show that the corresponding TSLS estimator weakly converges to the target vector of the parameters of interest. Firstly, following Small[25], we demonstrate that the above assumptions are sufficient to guarantee the exogeneity of the $R_i X_i$'s in model (5). A proof of this proposition has been relegated to the appendix.

Proposition 1. *Under assumptions (TSLS–1) and (TSLS–2), and under the assumption that the R_i's are exogenous with respect to the Y_i's in model* (5)*, we have* $\mathbb{C}\mathrm{ov}(R_i X_i, \varepsilon_i) = \mathbf{0}$.

In the context of trials, observe that the exogeneity of the R_i's is automatically satisfied. It then follows that the TSLS estimator, $\widehat{\boldsymbol{\beta}}$, can be computed with respect to the matrix $\widehat{\mathbf{V}}$, such that $\widehat{\boldsymbol{\beta}} := (\widehat{\mathbf{V}}'\widehat{\mathbf{V}})^{-1}(\widehat{\mathbf{V}}'\mathbf{y})$, where $\widehat{\mathbf{V}}$ denotes the projected matrix of the variables in the second-stage equation with respect to the matrix of instruments, $\mathbf{Z}$. Analogously to the OLS, the variance of the estimator is then given by $\widetilde{\sigma}^2(\mathbf{V}'\mathbf{V})^{-1}$, where $\widetilde{\sigma}^2$ is defined as $(\mathbf{y} - \mathbf{V}\widehat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{V}\widehat{\boldsymbol{\beta}})/(n - k - 2)$. The consistency of the TSLS estimator, can then immediately be derived.

Proposition 2. *Under conditions (TSLS–1) to (TSLS–4), and under the assumptions that both the R_i's and the X_i's are exogenous with respect to the Y_i's in model* (5)*; we have* $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$.

As before, the proof of this proposition is provided in the appendix. It then suffices to plug in this estimator of $\boldsymbol{\beta}$ in our definitions of the natural effects, in order to construct the TSLS estimators for these causal estimands, such that we obtain $\widehat{\mathrm{NDE}} := \widehat{\beta}_R$, and $\widehat{\mathrm{NIE}} := \widetilde{\theta}_R - \widehat{\beta}_R$; where note that $\widetilde{\theta}_R$ is still estimated using OLS, since the randomization variable, R, is assumed to be exogenous with respect to the mediator, M. Moreover, observe that these TSLS estimators of the NDE and NIE solely rely on the TSLS estimator of β_R.

SPSL Estimator

As we have seen, the OLS and the TSLS estimators satisfy competing, yet complementary demands. Under assumptions (OLS–1), (OLS–2), and (OLS–4), the OLS will be asymptotically efficient but possibly biased, whereas under assumptions (TSLS–1) to (TSLS–4), the TSLS will be asymptotically unbiased but relatively inefficient. Thus, it is natural to try to strike a trade-off between these two estimators, by considering affine combinations of the form

$$\bar{\boldsymbol{\beta}}_\alpha := \alpha\widehat{\boldsymbol{\beta}} + (1 - \alpha)\widetilde{\boldsymbol{\beta}},$$

where recall that $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$ denote the TSLS and OLS estimators, respectively. and where α needs not be comprised between 0 and 1, but may take any real values. This family of estimators are sometimes referred to as semi-parametric Stein-like (SPSL) estimators, for reasons which will become clear in the sequel[11].

In this framework, the shrinkage parameter, α, is commonly selected as the value that minimizes an empirical estimate of the MSE of $\bar{\boldsymbol{\beta}}_\alpha$. However, in many circumstances, it may be desirable to optimize such a trade-off with respect to a subset of the parameters of interest. This may be achieved by pre-multiplying the vectors of estimators and estimands with the matrix of an orthogonal projection, which will select the particular subset of parameters that one wishes to emphasize. That is, given a projection,

$\mathbf{P}$, we may consider the MSE of the vector

$$\mathbf{P}(\bar{\boldsymbol{\beta}}_\alpha - \boldsymbol{\beta}) = (\mathbf{P}\bar{\boldsymbol{\beta}}_\alpha - \mathbf{P}\boldsymbol{\beta}).$$

The shrinkage parameter, α, is defined as the value that minimizes the trace of the MSE of that projected vector, which is given by

$$\operatorname{tr} \operatorname{MSE}(\mathbf{P}\bar{\boldsymbol{\beta}}_\alpha) := \operatorname{tr} \mathbb{E}\big[\mathbf{P}(\bar{\boldsymbol{\beta}}_\alpha - \boldsymbol{\beta})(\bar{\boldsymbol{\beta}}_\alpha - \boldsymbol{\beta})'\mathbf{P}'\big].$$

The use of a projection in this setting can be regarded as a generalization of the original SPSL framework introduced by Judge and Mittelhammer[11]. Before turning to the minimization of that quantity, we describe a particular decomposition of the MSE of the SPSL estimator.

Using $\mathbf{P}\bar{\boldsymbol{\beta}}_\alpha = \alpha\mathbf{P}\widehat{\boldsymbol{\beta}} + (1-\alpha)\mathbf{P}\widetilde{\boldsymbol{\beta}}$, one can show that the MSE of $\mathbf{P}\bar{\boldsymbol{\beta}}_\alpha$ can be decomposed into a weighted combination of the MSEs for the projected OLS and TSLS estimators. That is, for every α, and every projection, $\mathbf{P}$, we obtain,

$$\operatorname{MSE}(\mathbf{P}\bar{\boldsymbol{\beta}}_\alpha) = \alpha^2 \operatorname{MSE}(\mathbf{P}\widehat{\boldsymbol{\beta}}) + \alpha(1-\alpha)\operatorname{CSE}(\mathbf{P}\widehat{\boldsymbol{\beta}}, \mathbf{P}\widetilde{\boldsymbol{\beta}}) + (1-\alpha)^2 \operatorname{MSE}(\mathbf{P}\widetilde{\boldsymbol{\beta}}), \tag{6}$$

where the *cross sum of squares*, $\operatorname{CSE}(\mathbf{P}\widehat{\boldsymbol{\beta}}, \mathbf{P}\widetilde{\boldsymbol{\beta}})$ is defined as $\mathbb{E}[\mathbf{P}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{P}']$. The theoretical parameter, α, controlling the respective contribution of the OLS and TSLS estimators is then defined as the following minimizer,

$$\alpha := \underset{\alpha \in \mathbb{R}}{\operatorname{argmin}} \operatorname{tr} \operatorname{MSE}(\mathbf{P}\bar{\boldsymbol{\beta}}_\alpha). \tag{7}$$

This parameter can be shown to be available in closed-form. This follows from the fact that the MSE of $\mathbf{P}\bar{\boldsymbol{\beta}}_\alpha$ is a convex function of α. In the following proposition, for every estimator $\boldsymbol{\beta}^\dagger$, the quantity $(\operatorname{tr} \operatorname{MSE}(\boldsymbol{\beta}^\dagger))^{1/2}$ is referred to as the trace RMSE of $\boldsymbol{\beta}^\dagger$. A proof of this proposition is provided in the appendix.

Proposition 3. *For every n, and every $\mathbf{P}$; the parameter α from equation (7) is*

$$\alpha = \frac{\operatorname{tr}(\operatorname{MSE}(\mathbf{P}\widehat{\boldsymbol{\beta}}) - \operatorname{CSE}(\mathbf{P}\widehat{\boldsymbol{\beta}}, \mathbf{P}\widetilde{\boldsymbol{\beta}}))}{\operatorname{tr}(\operatorname{MSE}(\mathbf{P}\widehat{\boldsymbol{\beta}}) - 2\operatorname{CSE}(\mathbf{P}\widehat{\boldsymbol{\beta}}, \mathbf{P}\widetilde{\boldsymbol{\beta}}) + \operatorname{MSE}(\mathbf{P}\widetilde{\boldsymbol{\beta}}))}.$$

If, in addition, the random vectors, $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$ are elementwise squared-integrable, then α is unique whenever the trace RMSEs of $\mathbf{P}\widehat{\boldsymbol{\beta}}$ and $\mathbf{P}\widetilde{\boldsymbol{\beta}}$ are not equal.

In order to estimate the shrinkage parameter from the data, we need to construct a consistent estimator of the bias of $\mathbf{P}\bar{\boldsymbol{\beta}}_\alpha$. Indeed, the MSE of that estimator can be decomposed as follows,

$$\operatorname{MSE}(\mathbf{P}\bar{\boldsymbol{\beta}}_\alpha) = \mathbb{V}\mathrm{ar}(\mathbf{P}\bar{\boldsymbol{\beta}}_\alpha) + \mathbb{B}\mathrm{ias}^2(\mathbf{P}\bar{\boldsymbol{\beta}}_\alpha),$$

where $\mathbb{B}\mathrm{ias}^2(\mathbf{P}\bar{\boldsymbol{\beta}}_\alpha) := (\mathbb{E}[\mathbf{P}\bar{\boldsymbol{\beta}}_\alpha] - \mathbf{P}\boldsymbol{\beta})(\mathbb{E}[\mathbf{P}\bar{\boldsymbol{\beta}}_\alpha] - \mathbf{P}\boldsymbol{\beta})'$. In general, the second term in the latter equation will not be directly available. Nonetheless, one can show that the assumptions that were made to guarantee the validity of the instruments, will also be sufficient to provide us with a consistent estimator of the bias of $\mathbf{P}\bar{\boldsymbol{\beta}}_\alpha$. Indeed, since by proposition 2, we have seen that the TSLS estimator converges in probability to the true parameter, $\boldsymbol{\beta}$; it follows that this particular estimator can be used in the place of

the true parameter in order to produce a consistent estimator of the bias of $\mathbf{P}\bar{\boldsymbol{\beta}}_\alpha$. That is, we can define the empirical bias of the projected SPSL estimator as follows,

$$\widehat{\mathbb{B}\mathrm{ias}}(\mathbf{P}\bar{\boldsymbol{\beta}}_\alpha) := \mathbf{P}\bar{\boldsymbol{\beta}}_\alpha - \mathbf{P}\widehat{\boldsymbol{\beta}}.$$

The CSE from proposition 3 can be estimated in an analogous fashion. Therefore, the consistency of the TSLS estimator guarantees the consistency of the SPSL estimator.

The choice of terminology for this family of estimator can be justified by observing that the expression for α in proposition 3 bears some similarities with the theory of Stein estimators[13]. Indeed, the empirical version of the formula for the shrinkage parameter can be expressed as follows,

$$\widehat{\alpha} = \frac{\mathrm{tr}(\widehat{\mathbb{V}\mathrm{ar}}(\mathbf{P}\widehat{\boldsymbol{\beta}}) - \widehat{\mathrm{CSE}}(\mathbf{P}\widehat{\boldsymbol{\beta}}, \mathbf{P}\widetilde{\boldsymbol{\beta}}))}{||\mathbf{P}(\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}})||^2},$$

where $|| \cdot ||$ denotes the L_2-norm on $\mathbb{R}^{k+2}$, with respect to the empirical joint distribution of the data. Using this expression, we can then formulate the SPSL estimator as a weighted deviation from the unbiased TSLS estimator, shrunk toward the OLS estimator,

$$\bar{\boldsymbol{\beta}}_{\widehat{\alpha}} = \widehat{\boldsymbol{\beta}} - \frac{\widehat{\tau}}{||\mathbf{P}(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}})||^2}(\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}),$$

in which $\widehat{\tau} := \mathrm{tr}(\widehat{\mathbb{V}\mathrm{ar}}(\mathbf{P}\widehat{\boldsymbol{\beta}}) - \widehat{\mathrm{CSE}}(\mathbf{P}\widehat{\boldsymbol{\beta}}, \mathbf{P}\widetilde{\boldsymbol{\beta}}))$, and where observe that we have made implicit the dependence of the LHS in the latter equation on $\mathbf{P}$. Indeed, $\bar{\boldsymbol{\beta}}_{\widehat{\alpha}}$ is solely dependent on the projection, $\mathbf{P}$, through the value of $\widehat{\alpha}$, since we have $\widehat{\alpha} = \widehat{\tau}/||\mathbf{P}(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}})||^2$.

The relationship between the SPSL estimator and the traditional Stein estimators has been studied by previous authors. See Judge and Mittelhammer[20], for instance. One can also observe that under the additional assumption that the random vectors, $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$, are elementwise squared integrable; it follows that we can obtain a central limit theorem for the SPSL estimator dependent on $\mathbf{P}$. This would generalize a previous result by Judge and Mittelhammer[20] for the standard SPSL estimator.

As for the OLS and TSLS estimators, the natural causal effects of the experimental manipulation onto the outcome, can be estimated using the components of the SPSL estimator, $\bar{\boldsymbol{\beta}}_{\widehat{\alpha}}$, such that we obtain $\overline{\mathrm{NDE}} := \bar{\beta}_R$, and $\overline{\mathrm{NIE}} := \widetilde{\theta}_R - \bar{\beta}_R$; where note that, as for the TSLS natural effects, the quantity $\widetilde{\theta}_R$ is still estimated using the OLS estimator.

For the analysis of the PROSPECT data set, since the estimations of both the NDE and the NIE rely on this quantity, it follows that the main parameter of interest is β_R. We have here arranged the variables in this model according to $V_i = (X_i', M_i, R_i)'$. Thus, the projection matrix, $\mathbf{P}$, will be defined as a null matrix with a single non-null value in the last element of its diagonal (that is, $P_{ij} = 0$ holds every element in $\mathbf{P}$, apart from $P_{k+2,k+2} = 1$); thereby estimating the shrinkage parameter solely on the basis of the respective values taken by $\widetilde{\beta}_R$ and $\widehat{\beta}_R$.

Simulations

We now present a simulation study, which compares the OLS and TSLS with the combined estimator, SPSL. We generate data from a confounded mediation model augmented with an instrumental variable.

The design of this simulation experiment is partly motivated by the model fitted to the PROSPECT data set analyzed in the sequel. Note, however, that in our simulations, the mediator is assumed to be continuous, whereas that same variable is dichotomous in the PROSPECT data. The effect of treatment on the endogenous mediator is allowed to vary according to the values taken by the baseline variables. Apart from this source of variation, the effects are assumed to be homogeneous in these simulations.

Mediation Models

Our objective in constructing our simulation model is twofold. Firstly, we wish to be able to control the degree of endogeneity of the mediator, as well as the strength of the instrument; such that both factors can be varied independently of each other. Secondly, we will also require the variances of the response, Y_i's, and of the intermediate variable, M_i's, to be equal to 1, to be able to interpret the size of the effect on a standardized scale.

As represented in Figure 3, we formulate the following structural model for the clinical outcome,

$$Y_i = \beta_X X_i + \beta_R R_i + \beta_M M_i + \beta_U U_i + \varepsilon_i;$$

for every $i = 1, \ldots, n$. (Note that, contrary to the model in Equation (2), the ε_i's in this simulation model are uncorrelated with the U_i's.) As previously mentioned, in order facilitate interpretability, we will fix the variance of the response variable to be equal to 1 for all scenarios. The variance of the intermediate variable, M_i's, will also be constrained to be unity. Both of these objectives will be achieved by controlling the variances of the error terms, ε_i's in the above model; and δ_i's in the following model for the intermediate variable,

$$M_i = \gamma_X X_i + \gamma_R R_i + \gamma_{RX} R_i X_i + \gamma_U U + \delta_i.$$

(Note again that the δ_i's in the above simulation model for the mediator are uncorrelated with the U_i's.) The variance of the δ_i's is defined as a function of the parameters in the equation for the M_i's, such that $\sigma_\delta^2(\gamma_X, \gamma_R, \gamma_{RX}, \gamma_U) := \mathbb{V}\mathrm{ar}(\delta)$. This function will be defined in the sequel. For convenience, we will simulate a single baseline covariate, denoted by X_i. This baseline covariate is given the following distribution, $X_i \overset{\text{iid}}{\sim} N(0, 2)$; where the variance was arbitrarily fixed to two, in order to simplify some of our computations. In addition, the experimental factor is drawn from a Bernoulli distribution, taking the form, $R_i \overset{\text{iid}}{\sim} \mathrm{Bern}(1/2)$. Finally, the unmeasured confounder is also generated from a unit normal distribution, such that $U_i \overset{\text{iid}}{\sim} N(0, 1)$.

In this model, the X_i's are assumed to be independent of other observed baseline variables, such that $X_i \perp R_i$; and the confounders, denoted by U_i's, are assumed to solely affect the relationship between the outcome and the mediator, such that we also have $U_i \perp X_i, R_i, R_i X_i$. These assumptions, combined with our constraints on the variances of the Y_i's and the M_i's, can be used to compute a range of possible values for the parameters of interest. A description of the specific computations involved in this derivation has been relegated to an appendix. (See Appendix B, for the details of the computation of the variance of the error terms, σ_ε^2 and σ_δ^2.) Throughout these simulations, the parameters controlling the effect of the X_i's and R_i's have been set to $\gamma_X := 1/4$, and $\gamma_R := 1/\sqrt{2}$, respectively. These choices of parameters correspond to small to moderate effect sizes. For convenience, we have further set the coefficients of the structural model for the Y_i's to take the same value, $\beta_X = \beta_R = \beta_M = \beta_U = 1/4$. It then follows that in order to guarantee $\sigma_\varepsilon^2 > 0$, we need to choose γ_U, as satisfying $\gamma_U \leq 1/2$, as well as, $\gamma_X + \gamma_{RX} \leq 1/2$.

The two main factors that are manipulated in this simulation study are the degree of confounding of the mediator, and the strength of the instrument. These simulation factors are respectively quantified using the correlation of the intermediate variable, M_i's, with the confounders, U_i's; and with the instruments, $R_i X_i$'s. Owing to our choice of normalization, these two correlations can be expressed as follows,

$$\mathrm{Cor}(M_i, U_i) = \eta, \qquad \text{and} \qquad \mathrm{Cor}(M_i, R_i X_i) = \kappa;$$

where it can be verified that $\eta = \gamma_U$, and $\kappa = \gamma_X + \gamma_{RX}$. Under the additional constraint that both σ_ε^2 and σ_δ^2 are positive, it follows that we can select η to take values in the set $\{0.0, 0.25, 0.50\}$, which represent different choices for the degree of confounding (none, moderate, and strong on a correlation scale); and κ to take values in the set $\{0.01, 0.25, 0.50\}$, which represent different choices for the strength of the instrument (weak, moderate, and strong also on a correlation scale). Observe that the correlation between the mediator and its instrument, κ, must be non-zero; in order to ensure that the TSLS estimator is well-identified in all scenarios. As a result of these choices of simulation parameters, the true NDE and the true NIE were $1/4$ and $1/2$, respectively, throughout all our simulations.
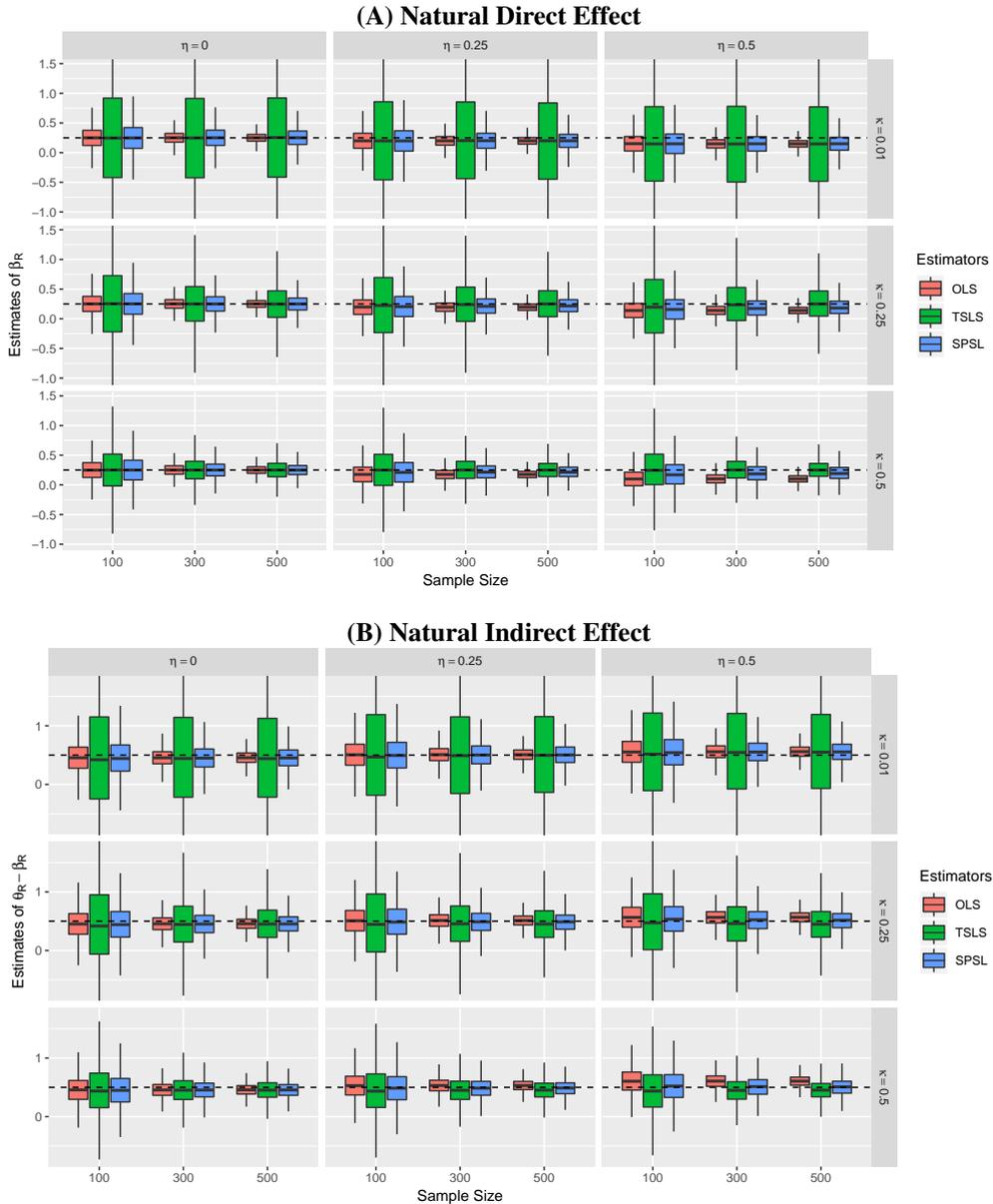
Evaluation of the Estimators

We generated 10^5 Monte Carlo samples from the aforementioned model, under combinations of the three values taken by η, the three values taken by κ; and the three different sample sizes typical of mental health trials, $n \in \{100, 300, 500\}$. Altogether, this produced a total of $270,000$ distinct synthetic data sets.

The OLS, TSLS and SPSL estimators of the NIE and NDE were computed as follows. Firstly, for each data set, we computed the OLS estimator, $\widetilde{\theta}_R$ of the total effect of R on the outcome Y. This corresponds to estimating the non-mediated model presented in Equation (4), and illustrated in Figure 1. Observe that the OLS estimator, $\widetilde{\theta}_R$, is identical for all methods of estimation. Indeed, the estimation of the total effect in this model is assumed to be unbiased, since subjects have been randomly allocated to the levels of the experimental factor, R.

Secondly, we fitted the instrumented mediation model, corresponding to the diagram in Figure 3, for the three different estimation procedures. This produced the OLS, TSLS and SPSL estimators for β_R, which corresponds to the estimator of the NDE. The NIE estimator could then be obtained by subtracting that estimate from $\widetilde{\theta}_R$. The performances of these estimators were compared by computing the empirical root MSE over the 10^5 Monte Carlo samples generated in each scenario. These RMSEs are reported in Figure 5.

Simulation Results

Consider the distribution of the values taken by the three estimators of interest in Figure 4. These are reported for the two causal estimands under scrutiny: NIE, $\widetilde{\theta}_R - \beta_R^\dagger$, and NDE, $\beta_R^\dagger$, in which $\beta_R^\dagger$ may represent either the OLS, TSLS or SPSL estimators. As expected, for both the NIE and NDE, the OLS was more likely to be biased when η was large, and the TSLS was more likely to exhibit a high variance when κ was small. Increases in sample size tended to result in better precision for all estimators. This trend was particularly noticeable for the TSLS estimator, using both continuous and binary mediators. The overall performances of these estimators were also compared using their respective RMSEs. These have been reported in Figure 5. There were substantial differences between the behavior of the TSLS and SPSL in these two simulation models.

(A) Natural Direct Effect

(B) Natural Indirect Effect

Figure 4. Monte Carlo distributions of estimators' values of the three estimators of interest under the simulation scenarios described in Equation (5), for the NDE, β_R, and NIE, $\theta_R - \beta_R$, in panels (A) and (B), respectively. The simulations are reported for different degrees of confounding, and varying levels of instrument's strength, measured by η and κ, respectively. These results are based on 10^5 iterations in each condition. The dashed lines indicate the values of the true NDE and NIE, in panels (A) and (B), respectively.
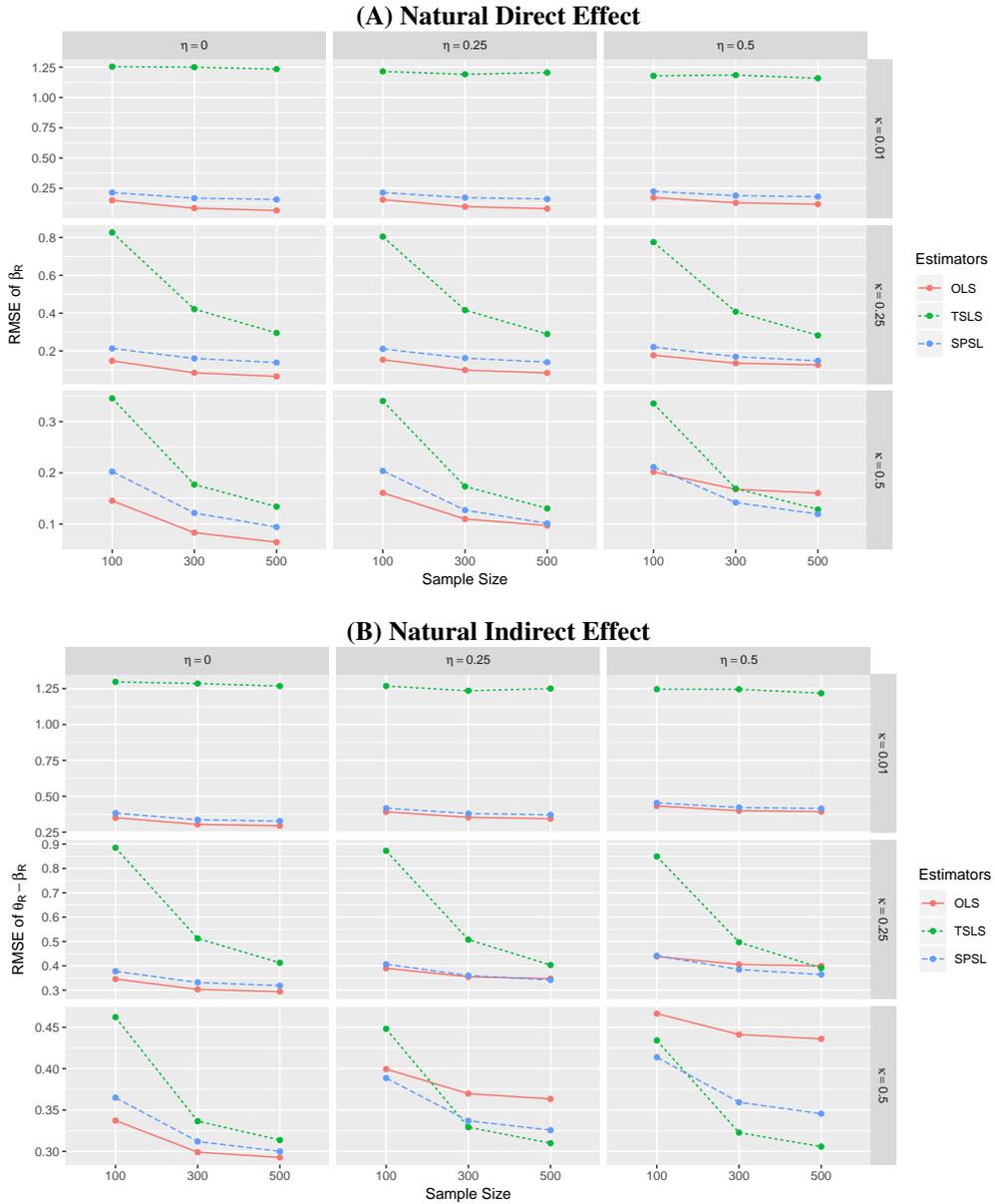
Figure 5. Continuous Mediator Model: Monte Carlo estimates of the root mean squared errors (RMSEs) of the three estimators of interest under the simulation scenarios described in equation (5), for the NDE, β_R, and NIE, $\theta_R - \beta_R$, in panels (A) and (B), respectively. The simulations are reported for different degrees of confounding, and varying levels of instrument's strength, measured by η and κ, respectively. These results are based on 10^5 iterations in each condition.

Under the continuous mediator model, the patterns exhibited by the NDE and NIE were very similar. For weak instrumental variables, i.e. $\kappa = 0.01$, the RMSE of the TSLS estimator was high in comparison to the ones of the OLS and SPSL estimators, due to the large variance of the TSLS. In these scenarios, the Stein-like estimator's RMSE was almost identical to the one of the OLS. By contrast, when the instrumental variables were strongly predictive of the endogenous variable, i.e. $\kappa = 0.5$, the RMSEs of the different estimators varied with the amount of bias. This trend is particularly noticeable in the last row of Figure 5(A). Fixing the correlation between the mediator, M, and the confounder, U, to be $\eta = 0.5$; one can observe that for small values of κ, the RMSE of the OLS is optimal, whereas for large values of κ, the RMSE of the TSLS is optimal; while the SPSL strikes a trade-off between these two counterparts irrespective of the values taken by κ.

In practice, we can usually evaluate the strength of a set of instruments, by computing the F-test of the equation for the M_i's. In our simulation, this corresponds to having some knowledge of κ. However, it is generally not possible to obtain any information about the degree of confounding, η. These simulations have therefore demonstrated that the SPSL outperforms its counterparts in a *global* sense –that is, when we 'average' the performances of these estimators over different values of η. Intuitively, this approach bears some similarities with the Bayesian framework for model averaging, in which the degree of unmeasured confounding, η, is treated as a source of uncertainty.

PROSPECT Study

We here re-analyze a randomized controlled trial known as PROSPECT[22]. This study tested the impact of a primary care intervention on reducing major risk factors for suicide in late life. Patients were recruited from 20 different primary care practices on the East coast of the United-States, over a 16-month period. The intervention consisted in two major components[22]. Firstly, the physicians followed a clinical algorithm specifically designed for treating geriatric depression. Secondly, the treatment was managed and adjusted by depression care managers. This primary care intervention was compared to a treatment as usual (TAU) condition.

Mediation Model

The main question of interest here is to investigate whether the intent-to-treat effect of the intervention on the 4-month Hamilton Depression Rating Scale (HDRS) was due to a direct effect of treatment allocation, after excluding the indirect effect mediated through taking antidepressant medication. Thus, the intermediate variable in this study should be regarded as a nuisance mediator, since we are primarily interested in the direct effect. The PROSPECT mediation study is therefore unusual, in the sense that the main effect of interest in the present analysis, is not the indirect effect or NIE as in most mediation studies.

The subjects' scores on the HDRS after a four-month follow-up is the main outcome under scrutiny. The instrumental variables for the mediator were defined as the set of interaction terms between the randomized intervention and the baseline covariates. This particular choice of instruments has been proposed previously[23], and we are here following this choice for comparability; see also Small[25] for a discussion of the use of interaction terms as instruments in the context of causal mediation. The instruments were found to be good predictors of the endogenous mediator; and explained about 50% of the variance in that variable using the R^2. Fitting a linear regression with taking prescribed antidepressant
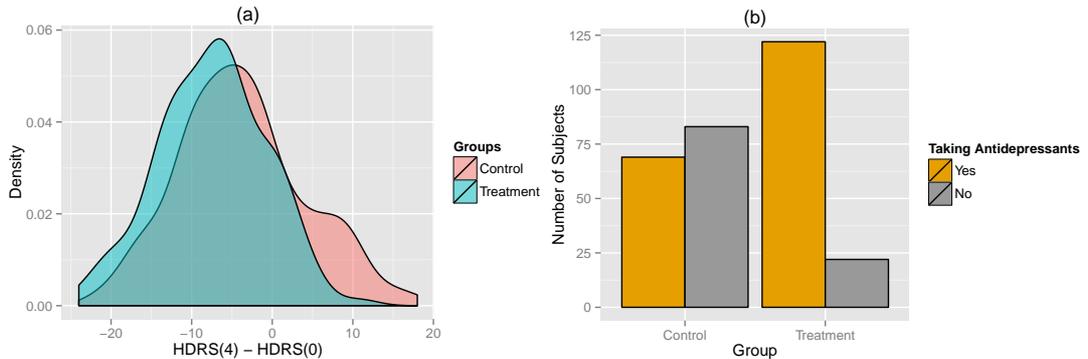
Figure 6. Descriptive statistics for the PROSPECT data set. In panel (a), we have provided histograms of the difference, HDRS(4) − HDRS(0), for both the control and treatment groups, where HDRS(0) and HDRS(4) denote the Hamilton Depression Rating Scale (HDRS) at baseline and after a four-month follow-up. In panel (b), the barplots represent the distribution of patients according to whether or not they have been taking antidepressants, which here corresponds to the intermediate variable, M, reported by treatment groups, R.

medication as the response, and the instruments as predictors, resulted in a highly significant F-statistic ($F = 9.10, \mathrm{df}_1 = 6, \mathrm{df}_2 = 282, p < 0.001$), thereby justifying our choice of instruments for this study. See also a similar analysis of the weak instrument bias in this study in Emsley et al. (2010)[24].

In order to compare the strength of the instruments in the PROSPECT data set, with the results of the Monte Carlo simulations, we have additionally computed the value of κ for this data set. Contrary to our simulation studies, however, note that we are here considering several instruments –that is, an entire vector of baseline variables, interacted with the treatment under scrutiny. Therefore, the definition used for the simulation study, $\mathrm{Cor}(M_i, R_i X_i) = \kappa$, such that $\kappa = \gamma_X + \gamma_{RX}$, cannot be directly used. Instead, we have calculated the multiple correlation of the mediator, M_i, with all the instruments used in the study. This produced a large correlation coefficient, $R = 0.436$, which approximately corresponds to the large κ values that we have investigated in the preceding simulations, and which we have described as referring to a strong set of instruments.

The baseline variables included HDRS scores at baseline, denoted HDRS(0), a binary variable denoting suicide ideation, past medication use (i.e. whether or not patients had been using past medication for dementia and other conditions but excluding psychotropic treatment for depression), and antidepressant use (i.e. specifically whether or not patients had been using antidepressant medication in the past). Moreover, the model also included two dummy variables, which controlled for the three different collection sites that were used in the study. Descriptive statistics for the main variables of interest in this study, have been reported in Figure 6.

Results of Re-analysis

The results of this re-analysis are reported in Table 1. The natural direct and natural indirect effects have been computed for the three estimators of interest. The OLS estimates and their standard errors were found to be approximately identical to the ones reported by Ten Have and colleagues in a previous analysis of the same data set[23].

Table 1. Re-analysis of the PROSPECT data set[a], in which the outcome variable is the Hamilton Depression Rating Scale at four-month, HDRS(4), the main intervention is the primary care intervention of interest, whereas the mediator is taking antidepressant medication. Three estimators of interest are here compared. These include the Ordinary Least Squares (OLS), Two-Stage Least Squares (TSLS), and the Semi-Parametric Stein-Like (SPSL) estimators. Bootstrapped standard errors[b] for all estimators are denoted in parentheses.

Variables in Model for Y	OLS	TSLS	SPSL
Randomization & Mediator:			
R : Primary Care Intervention	$-2.66\,(0.96)$	$-2.38\,(1.41)$	$-2.63\,(1.17)$
M : Antidepressant Medication	$-1.24\,(1.11)$	$-1.95\,(2.60)$	$-1.30\,(1.84)$
Baseline Covariates:			
X_1 : HDRS(0)	$0.62\,(0.07)$	$0.62\,(0.07)$	$0.62\,(0.07)$
X_2 : Suicide Ideation	$1.25\,(0.96)$	$1.25\,(0.96)$	$1.25\,(1.00)$
X_3 : Past Medication Use	$1.48\,(1.07)$	$1.59\,(1.10)$	$1.49\,(1.03)$
X_4 : Antidepressant Use	$-0.14\,(0.40)$	$-0.07\,(0.44)$	$-0.13\,(0.43)$
X_5 : Second Collection Site	$-0.46\,(0.98)$	$-0.50\,(0.97)$	$-0.47\,(0.99)$
X_6 : Third Collection Site	$-2.13\,(1.05)$	$-2.05\,(1.16)$	$-2.12\,(1.13)$
Causal Effects:			
NDE: Natural Direct Effect	$-2.66\,(0.96)$	$-2.38\,(1.41)$	$-2.63\,(1.17)$
NIE: Natural Indirect Effect	$-0.48\,(1.26)$	$-0.76\,(1.60)$	$-0.51\,(1.42)$
Shrinkage Parameter:			
SPSL's $\widehat{\alpha}$[c]	$--$	$--$	0.075

[a] Complete cases, for whom all measures were available, $n = 296$.
[b] The SEs for all estimators are based on $1{,}000$ bootstrap iterations.
[c] The estimated shrinkage used in the computation of the SPSL estimator, where the optimal shrinkage is estimated using a projection matrix, **P**, which is specified to be a unit matrix with a single one in its diagonal, corresponding to the offer of treatment variable, β_R.

In this paper, we have introduced a generalization of the SPSL estimator, which includes the use of a projection matrix, **P**. Such a matrix permits to restrict the computation of the shrinkage parameter, α, to a specific subset of variables. In the case of PROSPECT, the intermediate variable in this study is treated as a nuisance mediator, in the sense that the main focus of the analysis lies in estimating the NDE. Therefore, one can select a projection matrix, **P**, that emphasizes the estimation of the NDE, as opposed to optimizing the estimator with respect to all the parameters in the model. This can be done by specifying **P** to be a unit matrix with a single one in its diagonal, corresponding to the offer of treatment, β_R. The results for the projected SPSL estimator are reported in the third column of Table 1. The values of the SPSL and associated shrinkage estimator did not markedly differ, when using an identity matrix (results not shown); thereby indicating that the amount of shrinkage exerted by the SPSL estimator was mostly determined by the OLS and TSLS estimates of β_R, the main estimand of interest.

As expected, the values taken by the NDE and NIE under the SPSL framework were located between the ones of the OLS and TSLS estimators. Similarly, the standard errors (SEs) of the SPSL estimator was also found to strike a trade-off between the SE of the OLS estimator and the SE of the TSLS estimator, for both the NDE and NIE. The shrinkage parameter of the SPSL estimator was found to be close to zero, $\widehat{\alpha} = 0.075$. Thus, the statistical properties of the OLS estimator were favored over the corresponding properties of the TSLS estimator. This suggested that, despite the strength of the instruments used in this study, these instruments generated an increase in the variance of the TSLS estimator, which penalized the use of the TSLS. Consequently, the use of the TSLS was not deemed warranted for this model.

Conclusions

We have here demonstrated the usefulness of the SPSL estimator in the context of causal mediation analysis. This implementation has also generalized some of the previous uses of this family of estimators, by restricting the estimation of the shrinkage parameter to a subset of the parameters of interest. Although weak instrument bias can usually be estimated from the data; the degree of unmeasured confounding is, by definition, unknowable. In situations in which there is little or weak confidence in the instruments used, the SPSL estimator can be employed to mitigate any detrimental increase in variance that would result from a naive application of the TSLS. This appears to be especially true for binary mediators, since these are likely to yield inefficient TSLS estimators.

Furthermore, the SPSL estimator possesses desirable asymptotic properties. Under standard assumptions on the properties of the instruments, the SPSL estimator is indeed asymptotically unbiased. It also has the advantage of being directly estimable from the data. Moreover, the shrinkage parameter used in combining the OLS and TSLS estimators may be of special interest. This parameter can be interpreted as a gauge that measures the usefulness of the instruments, in terms of gains in MSE. That is, a very low value for $\widehat{\alpha}$ indicates that the OLS is preferable over the TSLS, and therefore that the corresponding instruments mostly contribute to increasing the variance of the combined estimator, without substantial gains in terms of unbiasedness.

Throughout this paper, we have assumed that the IVs of interest were valid instruments. In particular, we have required that each IV only affects the clinical outcome through the intermediate variable. Moreover, these assumptions have been tailored to the case in which the instruments are constructed by interacting some of the baseline variables with treatment offer, following the work of Small[25]. Such assumptions are particularly important in our context, since the asymptotic unbiasedness of the SPSL estimator solely holds, when the TSLS estimator is also guaranteed to be asymptotically unbiased. Note, however, that the SPSL framework is more widely applicable, and could be used with instruments that are not necessarily composed of interaction terms.

It is of special interest to consider the behavior of the SPSL estimator, when some of our assumptions fail to be satisfied. Let us first focus on some of the aforementioned OLS and TSLS assumptions. We can evaluate how the violation of these assumptions would impact on the behavior of the SPSL estimator. In the first instance, consider condition (OLS–4), which requires that $\mathbb{E}[X_i' X_i]$ should be full-rank; or equivalently that the OLS estimator is identified. If such an assumption were to fail, then the condition number of the matrix, $\mathbb{E}[X_i' X_i]$, would be very high, and consequently the determinant of its inverse would be very large. As a result, this would produce a large OLS variance, possibly tending towards

infinity. Therefore, everything else being equal, the failure of (OLS–4) would be likely to put the OLS at a disadvantage, in comparison to the TSLS, in the construction of SPSL.

Similarly, when considering the TSLS estimator, one can also predict the consequences of the violation of certain assumptions. Consider (TSLS–3), for instance. This assumption requires that the matrices, $\mathbb{E}[Z_i'Z_i]$ and $\mathbb{E}[Z_i'X_i]$, are both full-rank; which guarantees that the TSLS estimator is identified. If this condition were to fail, we would obtain a very large variance for the TSLS estimator. As a result, a failure of assumption (TSLS–3) would then lead to the SPSL estimator being shrank toward the value of the OLS estimator. Moreover, the TSLS would also suffer if condition (TSLS–4) were to fail. This assumption requires that the instruments, Z_i's, are relevant, in the sense that they should be correlated with the intermediate variables, M_i's. If this assumption were to be violated, the instruments would solely contribute to TSLS by increasing its variance; thereby making it more likely for the combined estimator to favor the OLS estimator over its TSLS counterpart.

Observe that all of the assumptions that we have made in this paper have also been posited by Small (2012)[25], in his investigations of the properties of IVs which are defined as interaction terms between baseline variables and the experimental variable. This choice of IVs corresponds to the instruments that we have used in the PROSPECT data set. In this setting, the IV assumptions could be subjected to a sensitivity analysis, as demonstrated by Small[25]; and we refer the reader to this paper for further details on the type of sensitivity analysis that can be conducted, when using interaction terms as instruments. However, further research will be needed to generalize these sensitivity analyses to the case of the SPSL estimator.

The SPSL could straightforwardly be applied to other causal estimands. It has been used to optimize the estimation of Local Average Treatment Effects (LATEs)[21], and could be implemented in other settings. Moreover, such methods could also accommodate other families of estimators, such as the jackknife IV estimator (JIVE)[28], for instance. Further research may also concentrate on improving the applicability of the present methods to data sets with binary outcomes, by using a more sophisticated approach than the one presented in the paper at hand. This may involve the use of generalized structural equation models (gSEM), which would be expected to handle binary mediators as well as binary outcomes, with greater efficiency. Several authors have proposed methodological frameworks for allowing the use of instruments in this context[29,30]; and the Stein-like estimators could be adapted to generalized linear models using the approaches advocated by these authors.

Appendix A: Proofs of Propositions

Proof of Proposition 1. The error term, ε_i, has been defined in equation (3), and can be seen to be the sum of three distinct random variables centered at zero. Thus, we solely need to consider whether or not the interaction terms, R_iX_i's, are uncorrelated with each of the summands composing the ε_i's in equation (3). Indeed, whenever a random variable is pairwise uncorrelated with a set of random variables, it is also uncorrelated with the sum of these variables. We will therefore consider the three summands of ε_i in turn. These are $A_{i1} := \big(Y_i(0,0) - \mathbb{E}[Y_i(0,0)|X_i]\big)$, $A_{i2} := \big(\beta_{M,i} - \beta_M\big)M_i$ and $A_{i3} := \big(\beta_{R,i} - \beta_R\big)R_i$.

Furthermore, observe that the covariance of ε_i with the interaction term is a vector of order $(k \times 1)$. Hence, the proposition is proved, if we are able to show that for each of the jth component of R_iX_i, we

have

$$\mathbb{C}\mathrm{ov}(R_i X_{ij}, \varepsilon_i) = \mathbb{C}\mathrm{ov}\left(R_i X_{ij}, \sum_{l=1}^{3} A_{il}\right) = 0,$$

where $j = 1, \ldots, k$. Thus, we fix an arbitrary component, say $R_i X_{ij}$, in the sequel; and consider its covariance with each of the three summands of ε_i.

Firstly, for A_{i1}, observe that the covariance $\mathbb{C}\mathrm{ov}(R_i X_{ij}, A_{i1})$ can be expressed as the difference, $\mathbb{E}[R_i X_{ij} A_{i1}] - \mathbb{E}[R_i X_{ij}] \cdot \mathbb{E}[A_{i1}]$. Using the tower rule, we have

$$\mathbb{E}[A_{i1}] = \mathbb{E}\big[Y_i^{0,0} - \mathbb{E}[Y_i^{0,0}|X_i]\big] = \mathbb{E}\big[Y_i^{0,0}\big] - \mathbb{E}\big[\mathbb{E}[Y_i^{0,0}|X_i]\big] = 0,$$

where for convenience, we have defined the shorthand, $Y_i^{0,0} := Y_i(0,0)$. It then suffices to consider the quadratic term, $\mathbb{E}[R_i X_{ij} A_{i1}]$, which simplifies as follows,

$$\mathbb{E}\big[R_i X_{ij}(Y_i^{0,0} - \mathbb{E}[Y_i^{0,0}|X_i])\big] = \mathbb{E}\big[R_i\big] \cdot \mathbb{E}\big[X_{ij}(Y_i^{0,0} - \mathbb{E}[Y_i^{0,0}|X_i])\big],$$

using the fact that the R_i's have been randomized. Through another application of the tower rule, the second term on the RHS of the latter equation gives

$$\mathbb{E}\big[\mathbb{E}[X_{ij}(Y_i^{0,0} - \mathbb{E}[Y_i^{0,0}|X_i])|X_i]\big] = \mathbb{E}\big[X_{ij}(\mathbb{E}[Y_i^{0,0}|X_i] - \mathbb{E}[Y_i^{0,0}|X_i])\big] = 0.$$

Secondly, considering the covariance of $R_i X_{ij}$ with the second summand of the error term, A_{i2}; we can again apply the tower rule in order to obtain

$$\begin{aligned}
\mathbb{E}\big[(\beta_{M,i} - \beta_M)M_i\big] &= \mathbb{E}\big[\mathbb{E}[(\beta_{M,i} - \beta_M)M_i|R_i, X_i]\big] \\
&= \mathbb{E}\big[\mathbb{E}[(\beta_{M,i} - \beta_M)|R_i, X_i] \cdot \mathbb{E}[M_i|R_i, X_i]\big],
\end{aligned}$$

where the second equality is a consequence of assumption (A3). Moreover, the first term inside the expectation on the RHS of the latter equation becomes,

$$\mathbb{E}[(\beta_{M,i} - \beta_M)|R_i, X_i] = \mathbb{E}[(\beta_{M,i} - \beta_M)|X_i] = \big(\mathbb{E}[\beta_{M,i}|X_i] - \beta_M\big) = 0,$$

using in turn, the fact that the R_i's are randomized, and assumption (A2). Thus, the covariance, $\mathbb{C}\mathrm{ov}(R_i X_{ij}, A_{i2})$, reduces to the quadratic term $\mathbb{E}[R_i X_{ij} A_{i2}]$. However, using the tower rule, this quantity can be expressed as

$$\begin{aligned}
\mathbb{E}[R_i X_{ij}(\beta_{M,i} - \beta_M)M_i] &= \mathbb{E}\big[\mathbb{E}[R_i X_{ij}(\beta_{M,i} - \beta_M)M_i|R_i, X_i]\big] \\
&= \mathbb{E}\big[R_i X_{ij} \cdot \mathbb{E}[(\beta_{M,i} - \beta_M)M_i|R_i, X_i]\big] \\
&= \mathbb{E}\big[R_i X_{ij} \cdot \mathbb{E}[(\beta_{M,i} - \beta_M)|R_i, X_i] \cdot \mathbb{E}[M_i|R_i, X_i]\big] \\
&= 0,
\end{aligned}$$

where the third equality follows from assumption (A3).

Thirdly, the covariance, $\mathbb{C}\mathrm{ov}(R_i X_{ij}, A_{i3})$ with $A_{i3} = (\beta_{R,i} - \beta_R)R_i$, can similarly be simplified by applying the fact that the R_i's are randomized, such that

$$\mathbb{E}[(\beta_{R,i} - \beta_R)R_i] = \mathbb{E}[(\beta_{R,i} - \beta_R)] \cdot \mathbb{E}[R_i] = \big(\mathbb{E}[\beta_{R,i}] - \beta_R\big) \cdot \mathbb{E}[R_i] = 0,$$

which follows from our definition of β_R. Thus, the covariance, $\mathbb{Cov}(R_i X_{ij}, A_{i3})$, reduces to the quadratic term, $\mathbb{E}[R_i X_{ij} A_{i3}]$, which can be expressed as

$$
\begin{aligned}
\mathbb{E}[R_i X_{ij}(\beta_{R,i} - \beta_R)R_i] &= \mathbb{E}[R_i^2] \cdot \mathbb{E}[X_{ij}(\beta_{R,i} - \beta_R)] \\
&= \mathbb{E}[R_i^2] \cdot \mathbb{E}\big[\mathbb{E}[X_{ij}(\beta_{R,i} - \beta_R)|X_i]\big] \\
&= \mathbb{E}[R_i^2] \cdot \mathbb{E}\big[X_{ij}\mathbb{E}[(\beta_{R,i} - \beta_R)|X_i]\big] \\
&= 0,
\end{aligned}
$$

where the last equality follows from the first part of assumption (A2).

Proof of Proposition 2. We can first invoke proposition 1, which guarantees that $\mathbb{Cov}(R_i X_i, \varepsilon_i) = \mathbf{0}$, for every subject. Moreover, since the baseline variables, X_i's, are assumed to be exogenous, it also follows that $\mathbb{Cov}(Z_i, \varepsilon_i) = \mathbf{0}$, since we have defined the Z_i's as $(X_i', R_i, R_i X_i')'$. Then, the proof of the consistency of $\widehat{\boldsymbol{\beta}}$ proceeds in a standard fashion. See chapter 5 of Woodridge (2002)[10] for details.

Proof of Proposition 3. The optimal value of α is obtained after minimizing $f_\alpha := \mathrm{MSE}(\mathbf{P}\bar{\boldsymbol{\beta}}_\alpha)$. We will expand the trace of this criterion as was done in equation (6), such that

$$
\mathrm{tr}\, f_\alpha = \mathrm{tr}(\alpha^2 M_1 + 2\alpha(1 - \alpha)C + (1 - \alpha)^2 M_2),
$$

with $M_1 := \mathrm{MSE}(\mathbf{P}\widetilde{\boldsymbol{\beta}})$, $C := \mathrm{CSE}(\mathbf{P}\widetilde{\boldsymbol{\beta}}, \mathbf{P}\widehat{\boldsymbol{\beta}})$, and $M_2 := \mathrm{MSE}(\mathbf{P}\widehat{\boldsymbol{\beta}})$, respectively. Commuting the derivative operator with the trace, we obtain

$$
\mathrm{tr}(\partial f/\partial \alpha) = 2\alpha\, \mathrm{tr}(M_2 - 2C + M_1) - 2\, \mathrm{tr}(M_2 - C).
$$

Setting this expression to zero and solving for α, yields $\alpha := \mathrm{tr}(M_2 - C)/\mathrm{tr}(M_2 - 2C + M_1)$, as required.

In addition, a second derivative test can be performed in order to show that such minimizer is, in fact, a unique global minimizer.

$$
\mathrm{tr}(\partial^2 f/\partial \alpha^2) = 2\, \mathrm{tr}(M_1 - 2C + M_2). \tag{8}
$$

By assumption, the random vectors, $\widetilde{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}$, are elementwise squared-integrable. Thus, the components, $\mathbb{E}[(\widetilde{\beta}_j - \beta_j)^2]$, of M_1 are finite. Hence, using the linearity of the trace, the MSE of $\mathbf{P}\widetilde{\boldsymbol{\beta}}$ can be treated as a sum of real numbers, thereby yielding the L^2-norm on $\mathbb{R}^{k+2}$, which we may denote by $||\mathbf{P}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})||$. The latter quantity will be referred to as the (trace) RMSE of $\widetilde{\boldsymbol{\beta}}$. By the same reasoning, it can be shown that C and M_2 corresponds to the inner product, $\langle \mathbf{P}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}), \mathbf{P}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\rangle$, and the squared norm, $||\mathbf{P}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})||^2$ on $\mathbb{R}^k$, respectively. Thus, equation (8) can now be expressed as follows,

$$
\mathrm{tr}(\partial^2 f/\partial \alpha^2) = 2\Big(||\mathbf{P}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})||^2 - 2\langle \mathbf{P}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}), \mathbf{P}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\rangle + ||\mathbf{P}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})||^2\Big).
$$

The Cauchy-Schwarz inequality can then be used to produce an upper bound,

$$
\langle \mathbf{P}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}), \mathbf{P}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\rangle \leq ||\mathbf{P}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})|| \cdot ||\mathbf{P}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})||.
$$

Finally, by completing the square, we obtain the following lower bound,

$$\text{tr}(\partial^2 f/\partial \alpha^2) \geq 2\Big(||\mathbf{P}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})|| - ||\mathbf{P}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})||\Big)^2 \geq 0,$$

for every $\mathbf{P}$, and where equality solely holds when the RMSEs of the two estimators, $\widetilde{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}$, are identical.

Appendix B: Simulation Model

This appendix demonstrates how the variances of the error terms, ε_i's and δ_i's, denoted by σ_ε^2 and σ_δ^2 respectively; can be obtained in closed form, under the constraints imposed upon our simulation model.

We have here assumed the X_i's to be exogenous, such that $X_i \perp R_i$. Moreover, the confounders, denoted by U_i's, have been assumed to solely affect the relationship between the outcome and the mediator, such that we also have $U_i \perp X_i, R_i, R_i X_i$. Consequently, the variance of the M_i's can be decomposed as follows,

$$\begin{aligned}
\mathbb{V}\text{ar}(M_i) &= \gamma_X^2 \, \mathbb{V}\text{ar}(X_i) + \gamma_R^2 \, \mathbb{V}\text{ar}(R_i) + \gamma_{RX}^2 \, \mathbb{V}\text{ar}(R_i X_i) + \gamma_U^2 \, \mathbb{V}\text{ar}(U_i) \\
&\quad + 2\gamma_X \gamma_{RX} \, \mathbb{C}\text{ov}(X_i, R_i X_i) + 2\gamma_R \gamma_{RX} \, \mathbb{C}\text{ov}(R_i, R_i X_i) + \sigma_\delta^2;
\end{aligned} \tag{9}$$

using $X_i \perp R_i$, and the fact that the U_i's are independent of all the other variables on the RHS of equation (9). It can also easily be seen that the mean and variance of the interaction variable $R_i X_i$ are respectively given by $\mathbb{E}[R_i X_i] = \mathbb{E}[R_i]\mathbb{E}[X_i] = 0$ and $\mathbb{V}\text{ar}(R_i X_i) = \mathbb{V}\text{ar}(R_i) \, \mathbb{V}\text{ar}(X_i) = 1$, by using the exogeneity of the X_i's, and the fact that the X_i's are centered at zero. By a similar argument, the two covariances in equation (9) can be simplified as follows,

$$\mathbb{C}\text{ov}(X_i, R_i X_i) = \mathbb{E}[R_i]\mathbb{E}[X_i^2] = 1, \quad \text{and} \quad \mathbb{C}\text{ov}(R_i, R_i X_i) = \mathbb{E}[R_i^2]\mathbb{E}[X_i] = 0;$$

since $R_i \in \{0, 1\}$, and therefore $\mathbb{E}[R_i^k] = \mathbb{E}[R_i]$, for every k. Hence, after fixing the variance of the M_i's at 1, we can express the variance of the δ_i's in terms of the remaining parameters in that structural equation, such that

$$\sigma_\delta^2(\boldsymbol{\gamma}) = 1 - \Big(2\gamma_X^2 + \frac{1}{4}\gamma_R^2 + \gamma_{RX}^2 + \gamma_U^2 + 2\gamma_{RX}\gamma_X\Big),$$

with $\boldsymbol{\gamma} := (\gamma_X, \gamma_R, \gamma_{RX}, \gamma_U)'$; and after using the Bernoulli distribution of the R_i's, which gives $\mathbb{V}\text{ar}(R_i) = 1/4$. Throughout the simulations, the parameters controlling the effect of the X_i's and R_i's have been set to $\gamma_X := 1/4$, and $\gamma_R := 1/\sqrt{2}$, respectively. This choice of parameters has been selected in order to simplify the expression for σ_δ^2, such that we obtain, $\sigma_\delta^2 = 0.75 - \gamma_{RX}^2 - \frac{1}{2}\gamma_{RX} - \gamma_U$.

Similarly, we can standardize the variance of the outcome variables, Y_i's. Given that the X_i's, R_i's and M_i's are cross-correlated, this produces a convoluted formula given by the following,

$$\begin{aligned}
\mathbb{V}\text{ar}(Y_i) &= \beta_X^2 \, \mathbb{V}\text{ar}(X_i) + \beta_R^2 \, \mathbb{V}\text{ar}(R_i) + \beta_M^2 \, \mathbb{V}\text{ar}(M_i) + \beta_U^2 \, \mathbb{V}\text{ar}(U_i) \\
&\quad + 2\beta_X \beta_M \, \mathbb{C}\text{ov}(X_i, M_i) + 2\beta_R \beta_M \, \mathbb{C}\text{ov}(R_i, M_i) \\
&\quad + 2\beta_M \beta_U \, \mathbb{C}\text{ov}(M_i, U_i) + \sigma_\varepsilon^2;
\end{aligned} \tag{10}$$

after applying $R_i \perp X_i$, and using the fact that the U_i's are independent of both the R_i's, and the X_i's. Equation (10) can be further simplified by using our choice of parametrization, which gives

$\mathbb{Cov}(X_i, M_i) = 2\gamma_X + \gamma_{RX}$, $\mathbb{Cov}(R_i, M_i) = \gamma_R/4$, and $\mathbb{Cov}(U_i, M_i) = \gamma_U$. Altogether, we therefore obtain a closed-form formula for the variance of the error terms of the Y_i's, expressed in terms of the model parameters, $\boldsymbol{\beta} := (\beta_X, \beta_R, \beta_M, \beta_U)$ and $\boldsymbol{\gamma}$. That is,

$$\sigma_\varepsilon^2(\boldsymbol{\beta}, \boldsymbol{\gamma}) = 1 - \left(2\beta_X^2 + \frac{1}{4}\beta_R^2 + \beta_M^2 + \beta_U^2 + C\right),$$

where $C := 2\beta_X\beta_M(2\gamma_X + \gamma_{RX}) + \beta_R\beta_M\gamma_R/2 + 2\beta_M\beta_U\gamma_U$. Therefore, we have obtained closed form formulas for both σ_δ^2 and σ_ε^2. These formulas have then be used to constrain the range of the parameters of interest, in the different simulation scenarios.

Acknowledgements

References

1. MacKinnon D. *Introduction to Statistical Mediation Analysis*. New York: Taylor and Francis Psychology Press, 2008.
2. Baron R and Kenny D. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statstical considerations. *Journal of Personality and Social Psychology* 1986; 51(6): 1173–1182.
3. Robins JM and Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology* 1992; 3(2): 143–155.
4. Pearl J. Direct and indirect effects. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., pp. 411–420.
5. Rubin DB. Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics* 2004; 31(2): 161–170.
6. VanderWeele TJ. Simple relations between principal stratification and direct and indirect effects. *Statistics and Probability Letters* 2008; 78(17): 2957–2962.
7. Imai K, Keele L and Yamamoto T. Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science* 2010; : 51–71.
8. Lynch KG, Cary M, Gallop R et al. Causal mediation analyses for randomized trials. *Health Services and Outcomes Research Methodology* 2008; 8(2): 57–76.
9. Ten Have TR and Joffe MM. A review of causal estimation of effects in mediation analyses. *Statistical Methods in Medical Research* 2012; 21(1): 77–107.
10. Wooldridge J. *Econometric analysis of cross-section and panel data*. London: MIT press, 2002.
11. Judge GG and Mittelhammer RC. A semiparametric basis for combining estimation problems under quadratic loss. *Journal of the American Statistical Association* 2004; 99(466): 479–487.

12. James W and Stein C. Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1. pp. 361–379.

13. Efron B and Morris C. Stein's estimation rule and its competitors: An empirical Bayes approach. *Journal of the American Statistical Association* 1973; 68(341): 117–130.

14. Berger J. *Statistical Decision Theory and Bayesian Analysis*. 2 ed. Springer Series in Statistics, London: Springer-Verlag, 1985.

15. Zellner A. Estimation of functions of population means and regression coefficients including structural coefficients: A minimum expected loss (MELO) approach. *Journal of Econometrics* 1978; 8(2): 127–158.

16. Zellner A and Park SB. Minimum expected loss (MELO) estimators for functions of parameters and structural coefficients of econometric models. *Journal of the American Statistical Association* 1979; 74(365): 185–193.

17. Mittelhammer RC and Judge GG. Combining estimators to improve structural model estimation and inference under quadratic loss. *Journal of econometrics* 2005; 128(1): 1–29.

18. Judge GG and Mittelhammer RC. *An information theoretic approach to econometrics*. Cambridge University Press, 2012.

19. Judge G and Mittelhammer R. A risk superior semiparametric estimator for over-identified linear models. *Advances in Econometrics* 2012; : 237–255.

20. Judge G and Mittelhammer R. A minimum mean squared error semiparametric combining estimator. *Advances in Econometrics* 2013; : 55–85.

21. Ginestet C, Emsley R and Landau S. Dose-response modeling in mental health using Stein-like estimators with instrumental variables. *Statistics in Medicine* 2017; 36(11): 1696–1714.

22. Bruce M, Ten Have T, Reynolds C et al. A randomized trial to reduce suicidal ideation and depressive symptoms in depressed older primary care patients: the PROSPECT study. *JAMA* 2004; 291(9): 1081–1091.

23. Ten Have TR, Joffe MM, Lynch KG et al. Causal mediation analyses with rank preserving models. *Biometrics* 2007; 63(3): 926–934.

24. Emsley R, Dunn G and White IR. Mediation and moderation of treatment effects in randomised controlled trials of complex interventions. *Statistical Methods in Medical Research* 2010; 19(3): 237–270.

25. Small D. Mediation analysis without sequential ignorability: Using baseline covariates interacted with random assignment as instrumental variables. *Journal of Statistical Research* 2012; 46(2): 91–103.

26. Dunn G and Bentall R. Modelling treatment-effect heterogeneity in randomized controlled trials of complex interventions. *Statistics in medicine* 2007; 26(26): 4719–4745.

27. Albert JM. Mediation analysis via potential outcomes models. *Statistics in medicine* 2008; 27(8): 1282–1304.

28. Angrist J, Imbens G and Krueger AB. Jackknife instrumental variables estimation. Technical Working Paper 172, National Bureau of Economic Research, 1995.

29. Clarke PS and Windmeijer F. Identification of causal effects on binary outcomes using structural mean models. *Biostatistics* 2010; 11(4): 756–770.

30. Clarke PS and Windmeijer F. Instrumental variable estimators for binary outcomes. *Journal of the American Statistical Association* 2012; 107(500): 1638–1652.