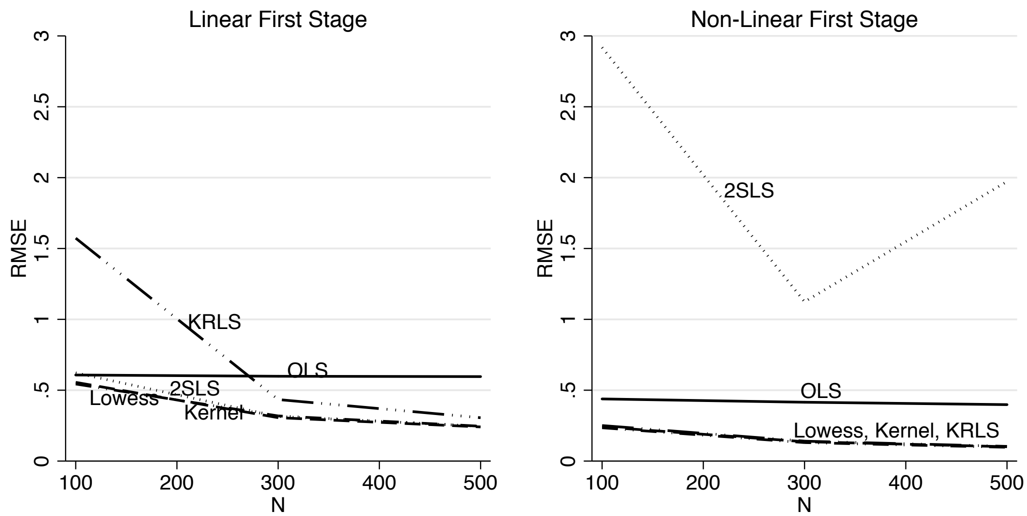
Online Appendix for:
Bridging the Gap Between Strength and Validity: How to increase the efficiency of weak continuous instruments

Abstract

Most instrumental variable analyses (IVs) employ a Two-Stage-Least-Squares (2SLS) estimator, whereby the predicted values of the instrumented variable (X_i) are generated by a linear regression of X_i on the instrument(s) Z_i. Theory often dictates monotone but not necessarily linear relationships, hence the effect of Z_i on X_i may not be properly depicted by a linear specification. We propose a way of increasing instrument strength and thus boosting efficiency in the first stage of IV estimation. In particular we show that the predicted values obtained from a series of local non-parametric smoothing techniques are better suited to capture this effect. Monte Carlo evidence suggests that non-parametric first-stage improves efficiency without violating the orthogonality of the errors guaranteed by the OLS. Bootstrapping can account for the uncertainty of the second-level estimates. We demonstrate the usefulness of the method with three empirical applications from development economics, international political economy and political psychology.

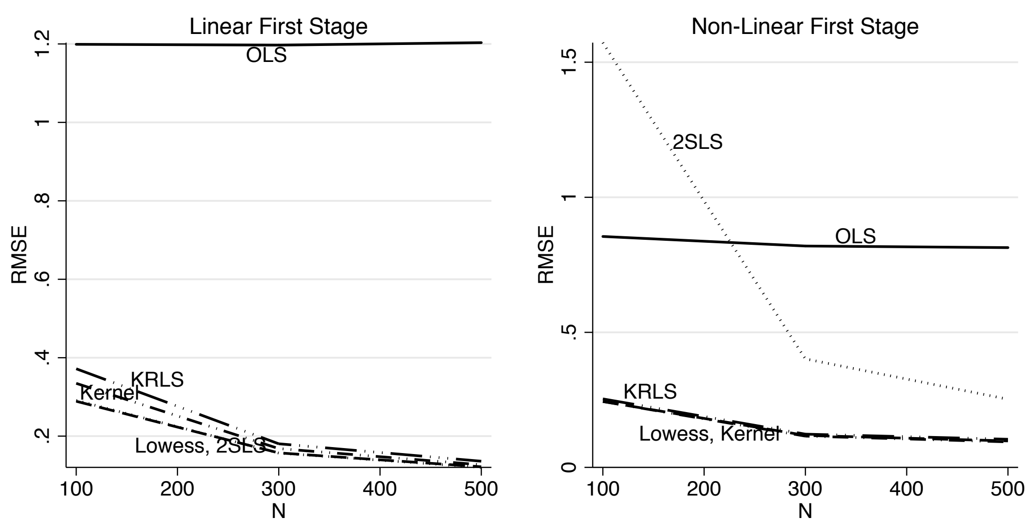
A.1: Replicating the full results from Figures 2 to 5 of the main text, adding the MC estimates for the Kernel Regularized Least Squares estimator.

Figure 1: Weak endogeneity and weak instrument, with KRLS



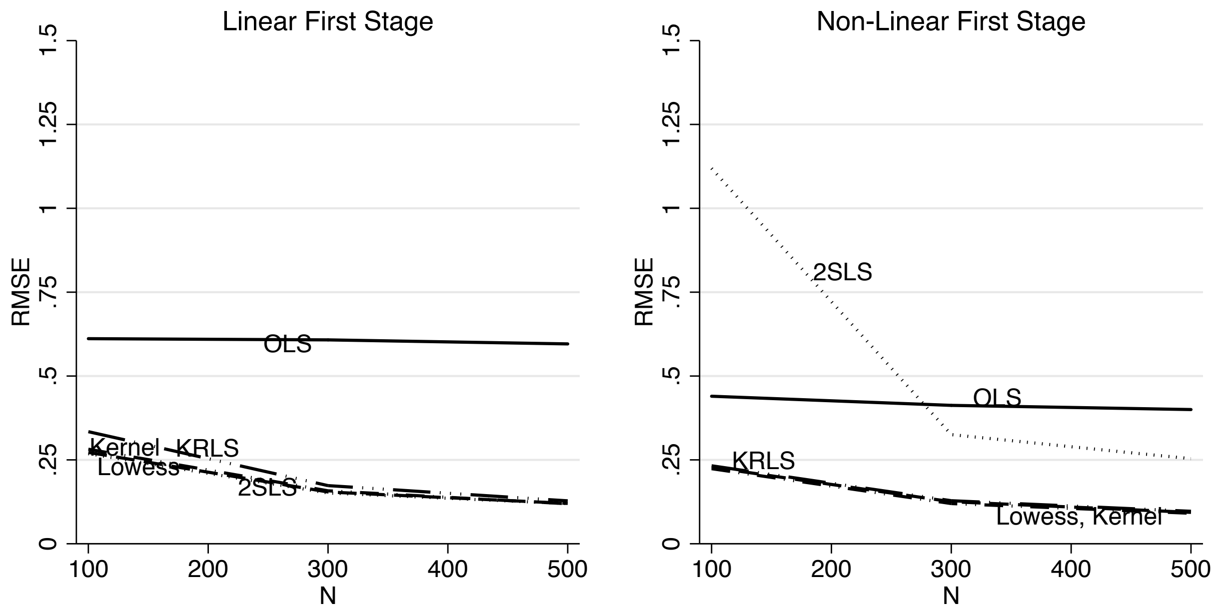
Note: Endogeneity: $corr(x, e) = 0.3$; Strength of instrument: $corr(x, z) = 0.3$; valid instrument: $corr(z, e) = 0$; $bw = 0.5$.

Figure 2: Strong endogeneity and strong instrument, with KRLS



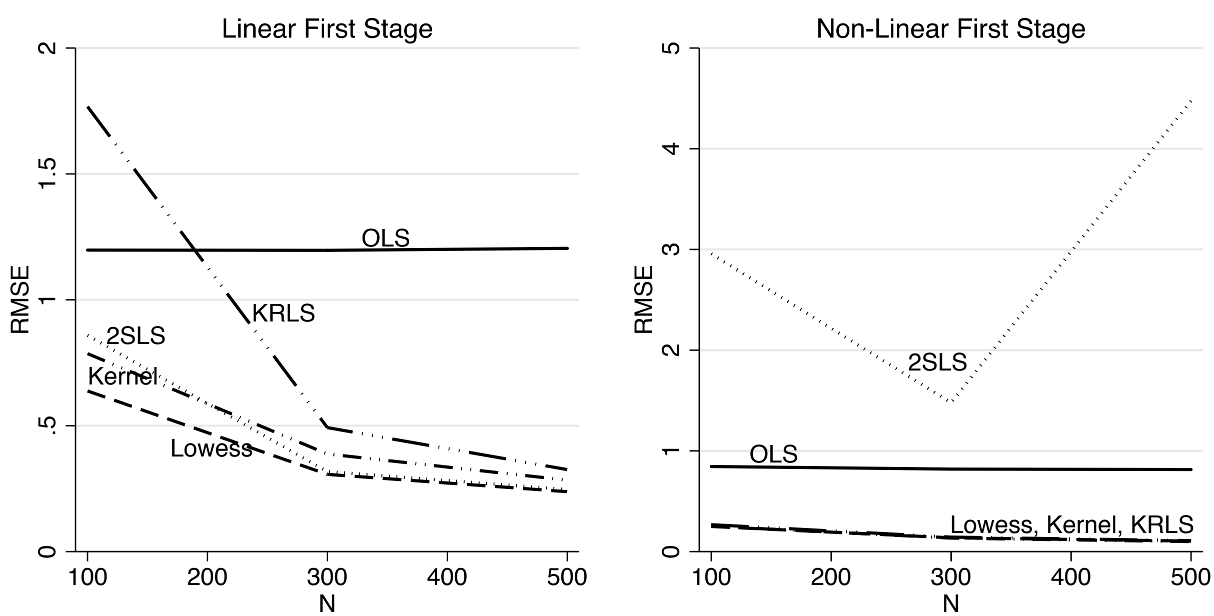
Note: Endogeneity: corr(x,e)=0.6; Strength of instrument: $corr(x, z) = 0.6$; valid instrument: $corr(z, e) = 0$; $bw = 0.5$.

Figure 3: Weak endogeneity and strong instrument, with KRLS



Endogeneity: $corr(x, e) = 0.3$; Strength of instrument: $corr(x, z) = 0.6$; valid instrument: $corr(z, e) = 0$; $bw = 0.5$.

Figure 4: Strong endogeneity and weak instrument, with KRLS



Endogeneity: $Corr(x, e) = 0.6$; Strength of instrument: $Corr(x, z) = 0.3$; valid instrument: $Corr(z, e) = 0$; $bw = 0.5$.

A.2: Performance of different estimators when instrument is invalid; both linear and non-linear relationship between Z_i and X_i.

Table 1: Invalid Instrument, Linear First Stage

n	Strength of Z_i: Corr(X, Z)	Endogeneity of X_i: Corr(X, e)	Validity of Z_i Corr(Z, e)	RMSE OLS	2SLS	IV-Lowess
100	0.3	0.3	0.3	0.602	2.244	1.784
500	0.3	0.3	0.3	0.6	2.03	1.972
1000	0.3	0.3	0.3	0.601	2.012	1.984
100	0.6	0.3	0.3	0.597	0.995	0.983
500	0.6	0.3	0.3	0.602	0.998	0.998
1000	0.6	0.3	0.3	0.6	1.001	1
100	0.3	0.6	0.3	1.194	2.74	1.945
500	0.3	0.6	0.3	1.201	2.026	2.019
1000	0.3	0.6	0.3	1.201	1.998	1.992
100	0.6	0.6	0.3	1.203	1.004	1.051
500	0.6	0.6	0.3	1.198	1	1.01
1000	0.6	0.6	0.3	1.201	1.001	1.005
100	0.3	0.3	0.6	0.609	5.559	3.324
500	0.3	0.3	0.6	0.603	4.086	3.916
1000	0.3	0.3	0.6	0.6	4.031	3.947
100	0.6	0.3	0.6	0.598	2.027	1.927
500	0.6	0.3	0.6	0.602	2.009	1.99
1000	0.6	0.3	0.6	0.599	2.004	1.996
100	0.3	0.6	0.6	1.202	4.473	3.502
500	0.3	0.6	0.6	1.2	4.038	3.896
1000	0.3	0.6	0.6	1.2	4.033	3.97
100	0.6	0.6	0.6	1.192	2.025	1.985
500	0.6	0.6	0.6	1.203	2.004	2
1000	0.6	0.6	0.6	1.2	2.005	2.002

Note: The span for Lowess Smoothing is set to 0.5.

A.3: Overfitting

Whenever non-parametric fitting techniques are considered, overfitting can pose a serious problem to inference. This problem equally applies to the proposed non-parametric first stage of a two stage instrumental variable model. In order to explore this potential caveat we analyze the performance of lowess and kernel when overfitting should occur with a high probability; that is when a) the correlation between instrument and instrumented variable is small and b) the chosen bandwidth is very small.

We use exactly the same set up and DGP of Monte Carlo experiments as above in the single excluded instrument case, but examine the point at which overfitting in the lowess or kernel first stage will lead to estimates that are outperformed by 2SLS. In this set of MCs we vary the following features of the DGP:

1. Number of observations: $n = [100, 500, 1000]$
2. Strength of the instrument Z_i: $Corr(X, Z) = [0.2, 0.4, 0.6] \rightarrow$ the problem of overfitting should be larger for weaker instruments.

Table 2: Invalid Instrument, Non-Linear First Stage

n	Strength of Z_i: Corr(X,Z)	Endogeneity of X_i: Corr(X,e)	Validity of Z_i Corr(Z,e)	RMSE OLS	RMSE 2SLS	RMSE IV-Lowess
100	0.3	0.3	0.3	0.431	21.402	0.329
500	0.3	0.3	0.3	0.407	12.034	0.223
1000	0.3	0.3	0.3	0.402	4.748	0.211
100	0.6	0.3	0.3	0.424	5.422	0.394
500	0.6	0.3	0.3	0.402	2.207	0.344
1000	0.6	0.3	0.3	0.401	2.09	0.342
100	0.3	0.6	0.3	0.843	50.791	0.341
500	0.3	0.6	0.3	0.807	8.448	0.229
1000	0.3	0.6	0.3	0.808	5.452	0.213
100	0.6	0.6	0.3	0.838	5.538	0.404
500	0.6	0.6	0.3	0.807	2.16	0.346
1000	0.6	0.6	0.3	0.801	2.059	0.342
100	0.3	0.3	0.6	0.432	44.246	0.496
500	0.3	0.3	0.6	0.407	25.856	0.399
1000	0.3	0.3	0.6	0.397	10.111	0.382
100	0.6	0.3	0.6	0.415	12.579	0.663
500	0.6	0.3	0.6	0.395	4.358	0.655
1000	0.6	0.3	0.6	0.4	4.144	0.659
100	0.3	0.6	0.6	0.808	66.258	0.478
500	0.3	0.6	0.6	0.795	21.274	0.382
1000	0.3	0.6	0.6	0.801	12.079	0.39
100	0.6	0.6	0.6	0.819	13.718	0.687
500	0.6	0.6	0.6	0.795	4.286	0.652
1000	0.6	0.6	0.6	0.806	4.096	0.667

Note: The span for Lowess Smoothing is set to 0.5.

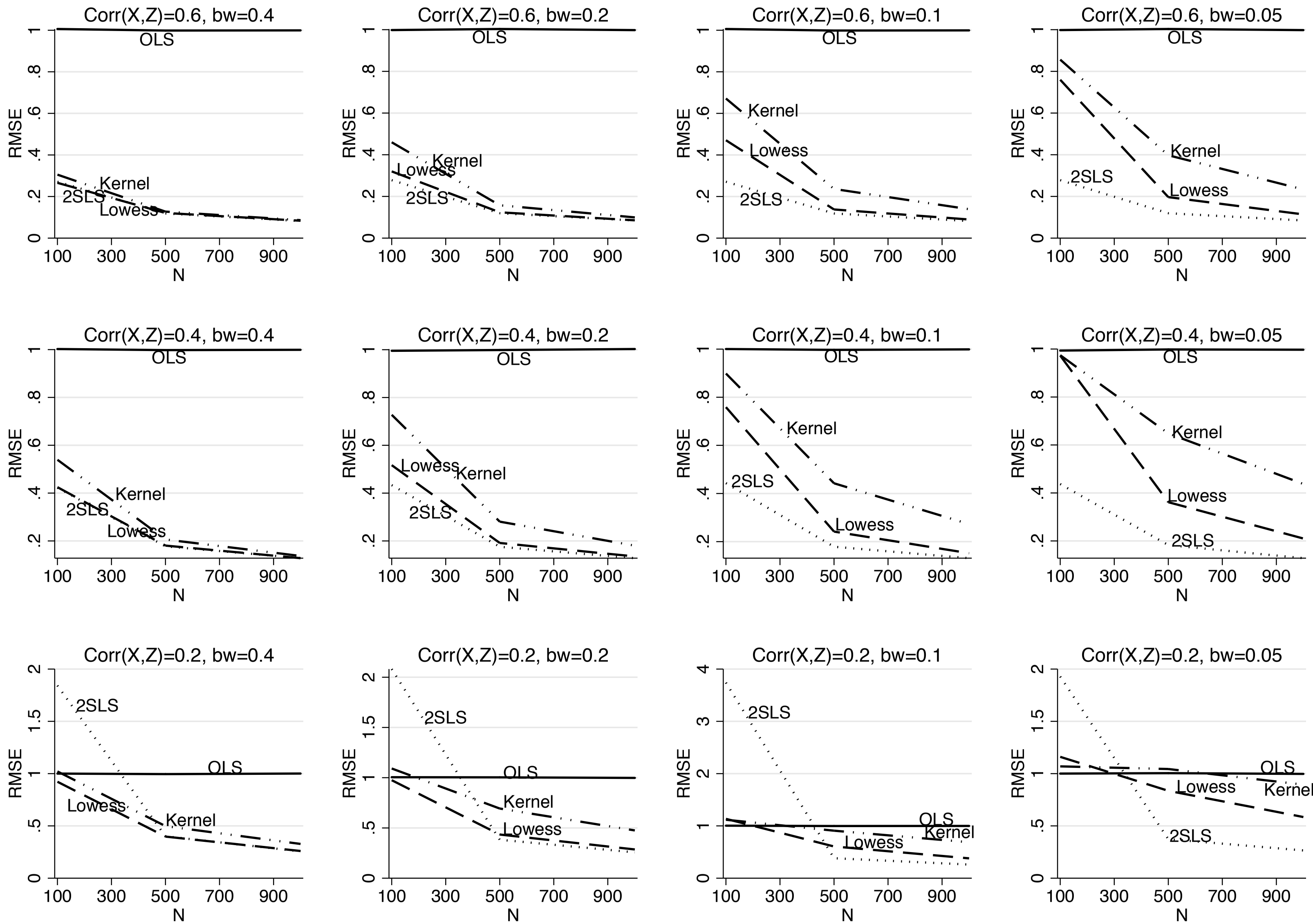
3. Linearity of the relationship between X_i and Z_i: $\gamma_2 = 0, \gamma_2 > 0 \rightarrow$ overfitting should be a bigger problem when Z_i exerts a linear effect on X_i.
4. Bandwidth of the lowess smoothing and kernel regression in the first stage (the α parameter): $bw = [0.05, 0.1, 0.2, 0.4] \rightarrow$ smaller bandwidths potentially lead to overfitting

In addition we assume the instrument Z_i to be valid (i.e. Corr(Z,e)=0) and we hold the endogeneity of X_i constant at a medium level (i.e. Corr(X,e)=0.5).

As expected, when the true relationship between the endogenous RHS variable X_i and the explanatory variable Z_i is linear, overfitting can be a problem. With very small bandwidth for non-parametric smoothing and weak instruments, a linear 2SLS estimator can outperform lowess or kernel smoothing especially when n is small (see Figure 5). Even in these cases the lowess smoother strictly outperforms the kernel. It seems safe to suggest that even if the true relationship between the outcome Y_i and the endogenous RHS variable X_i is linear, a lowess smoother with a bandwidth no smaller than 0.4 produces results that are at least as good as 2SLS estimates if n is large or better if n is small (see Figure 5).

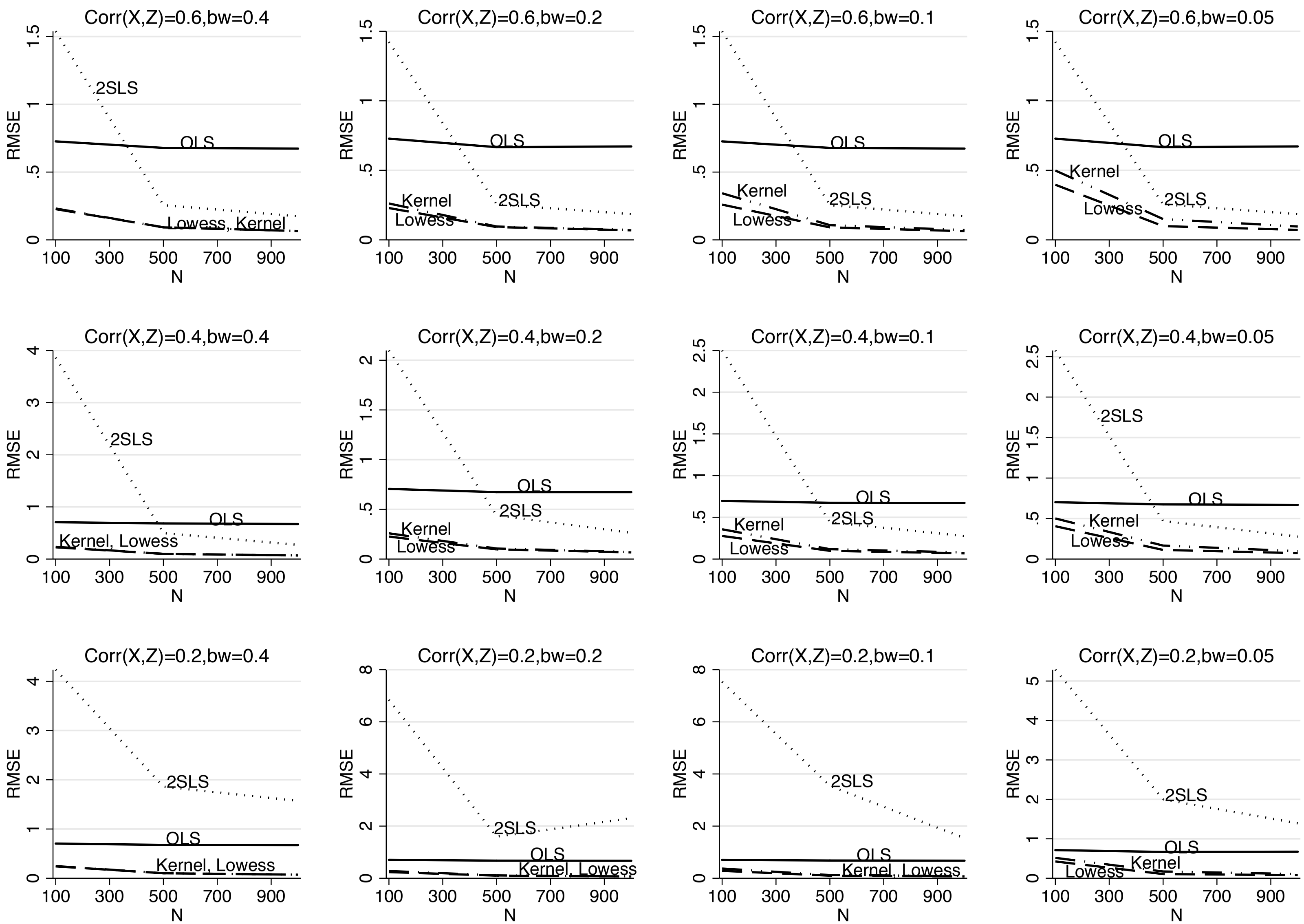
However, when the true relationship between X_i and Y_i is non-linear or even non-monotone, non-parametric versions of the two-stage IV model (with first stage lowess or kernel) always outperform the 2SLS estimator even in case the instrument is very weak

Figure 5: Overfitting – Effect of bandwidth and instrument strength on estimation of the endogenous RHS variable, linear case



Linear relationship between X_i and Y_i; Corr(X,e)=0.5; Corr(Z,e)=0.

Figure 6: Overfitting – Effect of bandwidth and instrument strength on estimation of the endogenous RHS variable, non-linear case



Quadratic relationship between X_i and Y_i; Corr(X,e)=0.5; Corr(Z,e)=0.

and the bandwidth is very small. In these cases lowess smoothing slightly outperforms kernel regression (6).

To sum up, overfitting presents a potential caveat of using lowess smoothing in the first stage especially when the true relationship between X_i and Z_i is linear. However, as the MC results show, IV-Lowess still outperforms 2SLS if we employ a reasonable bandwidth, e.g. no smaller than 0.4.

To be sure, simulated data is extremely well behaved as compared to observational data. We therefore recommend to use a range of values for the employed bandwidth (wherever possible) and/or regularization parameter λ and use cross-validation techniques to select the optimal bandwidth. In this way the trade-off between parsimony and complexity of the functional approximation can be resolved and bias inducing overfitting avoided.