Supplementary Material for “Learning heterogeneity in causal inference using sufficient dimension reduction”

Wei Luo

Center for Data Science, Zhejiang University.

Wenbo Wu

Department of Management Science and Statistics, The University of Texas at San Antonio.

Yeying Zhu

Department of Statistics and Actuarial Science, University of Waterloo.

In the Supplementary Material, we present the results for additional simulation studies complementary to the simulation studies in §7 of the main text, which include: first, the comparison between different sufficient dimension reduction methods, including ordinary least square, principal Hessian directions, the ensemble moment-based estimator, and the sparse ensemble moment-based estimator, in estimating the central causal effect subspace; second, the illustration of the effectiveness of the proposed approaches when applied to the same models but with $p = 20$. To avoid confusions, all the tables and figures in this supplementary file are labeled with “S”.

S1 Simulation study 1

In this section, we evaluate the ensemble moment-based estimator and its sparse modification in estimating the central causal effect subspace. To study the effectiveness of the ensemble strategy, we also include the individual estimators, i.e. ordinary least square and principal Hessian directions, in the comparison. Following the main text, we use Models 1 – 4 with $n = 600$ and $p = 10$, and simulate 200 samples independently for each model. The deviation between two linear spaces is again measured by $D(\cdot, \cdot)$. The results are summarized in Table 1.

From Table 1, for Models 1 and 2, ordinary least square is effective in recovering the central causal effect subspace and principal Hessian directions is noisy, whereas their performances switch for Model 3. Both estimators fail in Model 4. This matches with the monotonicity of the regression causal effect in these models. Clearly, the ensemble moment-based estimator captures the strength of both ordinary least square and principal Hessian directions, so that it performs similarly to the better of the latter two in Models 1 – 3, and recovers the entire central causal effect subspace in Model 4. Referring to the covariates' distributions in these models, the consistency of the ensemble moment-based estimator also illustrates its robustness to the violation of the linearity condition and the constant variance condition.

Model	OLS	PHD	ENS	ENSB
1	12.6	93.6	18.7	0
	(7.9)	(7.6)	(18.6)	(0)
2	20.5	97.6	23.3	11
	(6.1)	(0.8)	(6.9)	(10.5)
3	90.2	39.4	39.4	9.1
	(11)	(12.3)	(12.5)	(12.5)
4	100	93.7	26.2	0
	(0)	(8.8)	(6)	(0)

Table 1: Accuracy of estimating $\mathcal{S}_{E(\Delta Y|X)}$ when $p = 10$. The number in the top (bottom) of each cell is the average (standard deviation) of the deviation between the central causal effect subspace and its estimate over 200 replicates, multiplied by 100. "OLS" stands for ordinary least square, "pHd" for principal Hessian directions, "ENS" for the ensemble moment-based estimator, and "S-ENS" for the sparse ensemble moment-based estimator.

Since the cardinality of the active set is relatively small in all the models, the sparse ensemble moment-based estimator differs from its ordinary counterpart and consistently outperforms the latter. In particular, in a simple case like Model 1, it exactly recovers the central causal effect subspace in all the simulation samples. As mentioned in §5 of the main text, this phenomenon differs from the commonly seen cases in variable selection but is reasonable in the sufficient dimension reduction scenario, as the estimation accuracy is measured on the level of the central mean subspace.

S2 Simulation study 2

In this section, we study the performance of the proposed approaches when $p = 20$, using the same models and the same existing methods for reference as in §7 of the main text. We let $n = 600$ as in the main text when assessing the estimation accuracy for the central causal effect subspace, the regression causal effect, and variable selection. We raise n to 1000 when testing the heterogeneity of the regression causal effect, as we found that no tests in §7.3 of the main text performed well when $n = 600$ and $p = 20$, which is reasonable since the consistency of inference results usually require larger sample sizes.

Table 2 reports the estimation accuracy of the four aforementioned sufficient dimension reduction methods. As before, the ensemble moment-based estimator outperforms both ordinary least square and principal Hessian directions, and is further improved if we employ sparse sufficient dimension reduction.

Using the sample median intergraded absolute error defined in (17) of the main text, Table 3 records the performance of the proposed estimator of the regression causal effect. Although the accuracy slightly dropped compared with the cases when $p = 10$ (see Table 1 in the main text), the similar overall pattern

Model	OLS	PHD	ENS	ENSB
1	17.3	95.9	47	3.5
	(3.3)	(4.6)	(21.7)	(18.4)
2	31	98.7	53.9	29.9
	(6)	(1.8)	(26.4)	(30.9)
3	93.3	56.5	56.7	16
	(8)	(12.3)	(12.6)	(25.6)
4	100	95.8	36.4	7
	(0)	(4)	(9)	(7.1)

Table 2: Accuracy of estimating $\mathcal{S}_{E(\Delta Y|X)}$ when $p = 20$. The numbers in each cell are calculated in the same way as those in Table 1. All the abbreviations also follow those in Table 1.

can be observed. That is, except for the oracle estimator, the proposed estimator outperforms the others in all the models, especially when the individual outcome regressions are complex.

Table 4 records the performance of the proposed sparse ensemble moment-based estimator in variable selection. Again, it is consistent for all the models, and slightly outperforms the virtual twins method (Foster et al., 2011) in Model 3.

We next evaluate the effectiveness of the proposed chi-squared test for the homogeneity of the regression causal effect. As mentioned in the beginning of

Model	Oracle	S-ENS	LZG	GCL	GCQ	RF	GEL	GEQ	WML	WMQ
1	5.6	8.8	14.5	13.5	17.9	26.6	13.8	19	13.8	18.9
	(4)	(7.6)	(2.3)	(2.3)	(2.3)	(3.3)	(2.3)	(2.3)	(2.2)	(2.3)
2	8.8	19.6	19.5	53	19.3	57.6	22.7	23.1	22.9	23
	(5)	(8.1)	(2.3)	(7.1)	(2.1)	(4.5)	(3.3)	(3)	(3.3)	(3)
3	13.1	21.2	32.7	74.9	27.2	32.2	73.5	29.2	73	28.8
	(5.2)	(5.4)	(9.9)	(7.3)	(3.4)	(2.5)	(7.3)	(3.8)	(7.3)	(3.8)
4	19.3	20.3	60.9	122.5	85.6	109.2	123	86.4	123.1	86.3
	(3.4)	(7)	(3.9)	(4.8)	(6.6)	(13.7)	(5)	(6.6)	(5)	(6.5)

Table 3: Accuracy of estimating the regression causal effect when $p = 20$. The numbers in each cell are calculated in the same way as those in Table 1 of the main text. All the abbreviations also follow those in Table 1 of the main text.

the section, we now let $n = 1000$. The results are recorded in Table 5.

From Table 5, we again observe that the proposed test outperforms those proposed by Crump et al. (2008) in both the sensitivity and specificity. The same conclusion is suggested by Figure 1, which illustrates the distributions of the p-values when the robust and the ordinary chi-squared tests in Crump et al. (2008), and the proposed chi-squared test, are applied to each model.

Model	S-ENS		VT		dLasso		PHWD	
	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
1	0.965	0.004	1.000	0.003	1.000	0.312	1.000	0.299
2	0.988	0.062	1.000	0.000	1.000	0.317	1.000	0.333
3	1.000	0.059	1.000	0.134	0.445	0.142	1.000	0.321
4	0.998	0.001	1.000	0.002	0.642	0.202	1.000	0.271

Table 4: Accuracy of variable selection methods when $p = 20$. The number in each cell is calculated in the same way as those in Table 2 of the main text. All the abbreviations also follow those in Table 2 of the main text.

	1	3	4	5	6	7
SDR χ^2	93	89	100	99	100	100
r-Normal	100	69.5	100	84	10.5	1
r-χ^2	100	66	100	87	13	1
Normal	100	91.5	100	85	2.5	0
χ^2	100	90	100	89	2.5	0

Table 5: Percentage of correct decision by each test when $p = 20$. All the abbreviations also follow those in Table 3 of the main text.

References

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2008): “Nonparametric tests for treatment effect heterogeneity,” *The Review of Economics and Statistics*, 90, 389–405.

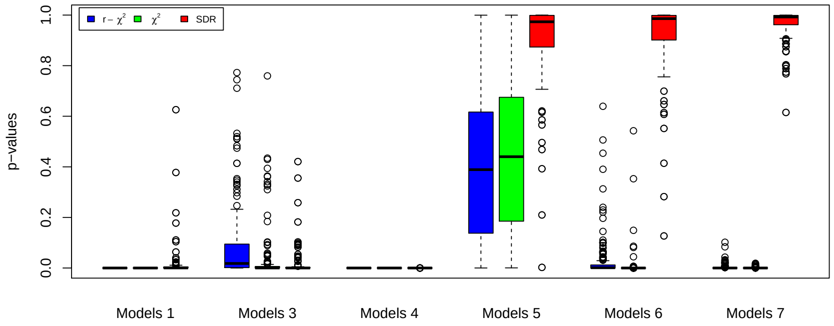

Figure 1: Distribution of p-values when $p = 20$. The box-plots of the p-values are drawn in the same way as those in Figure 1 of the main text.

Foster, J. C., J. M. G. Taylor, and S. J. Ruberg (2011): “Subgroup identification from randomized clinical trial data,” *Statistics in Medicine*, 30, 2867–2880.