Supporting Information for: Sparse least trimmed squares regression with compositional covariates for high dimensional data

Gianna Serafina Monti and Peter Filzmoser

1 Parameter selection

Section 2.3 of the manuscript describes the algorithms to reach the best subset for RobZS regression, for each fixed combination of the tuning parameters $\alpha \in [0, 1]$ and $\lambda \in [\varepsilon \cdot \lambda_{Max}, \lambda_{Max}]$, where $\varepsilon \geq 0$ controls the range. The parameter λ_{Max} is chosen such that all coefficients for a ZeroSum regression on the median-centered response and explanatory variables become zero. To select the optimal combination $(\alpha_{opt}, \lambda_{opt})$, leading to the optimal subset H_{opt}, a repeated K-fold CV procedure on a two-dimensional surface is adopted, with $K = 5$.

In K-fold CV the data are split into folds $V_1, \ldots, V_K$ of approximately equal size. The so-called out-of-fold observations $i \notin V_k$ which are not part of the k-th fold V_k, for $k \in \{1, \ldots, K\}$, are used for training the model, and in-fold observations $i \in V_k$ for evaluating it.

Note that we only consider samples of size h at this stage which are supposed to be outlier-free, and thus the derived prediction error criterion is robust. As criterion we use the root mean squared prediction error (RMSPE),

$$\text{RMSPE}(\alpha, \lambda) = \sqrt{\frac{1}{h} \sum_{k=1}^{K} \sum_{i \in V_k} \left(r_i(\hat{\boldsymbol{\beta}}(\alpha, \lambda)) \right)^2} \tag{1}$$

The 5-fold CV procedure is repeated five times in order to get higher stability for the resulting parameter selection. Thus, for each combination of tuning paramters, five RMSPE values from (1) are obtained. The chosen couple $(\alpha_{opt}, \lambda_{opt})$, over a grid of values α and λ, is the one giving the smallest CV error of the average of the five replications. When using only the Lasso penalty (thus $\alpha = 1$), we use the same procedure with repeated K-fold CV, but the computational cost reduces considerably as the search is only required for one parameter.

Here, λ_{Max} is chosen to get a model with full sparsity. In the simulations we considered 41 equally spaced values for α, and a grid of 40 values for λ.

2 Simulation results

In this section we report additional simulation results to empirically study the performance of the RobZS estimator with respect to different aspects:

(a) Section 2.1 presents more details of using different contamination schemes (see Section 3.1 of the main paper), with 10% and 20% of contamination;

(b) Section 2.2 investigates the effect of the proposed debiasing strategies (see Section 2.3 of the main paper), i.e. the rescaled, the relaxed and the hybrid RobZS estimator;

(c) Section 2.3 studies the behavior of RobZS when the sparsity changes;

(d) Section 2.4 uses the elastic-net penalty to compare ZS and RobZS.

2.1 Simulation results for different contamination schemes

The following simulation results refer to the sampling schemes described in Section 3.1. For the case of 10% contamination we also provide a comparison with the algorithm of Bates and Tibshirani (2019), here abbreviated by "ZS (B&T)". In their log-ratio lasso estimator they propose a fast approximate algorithm which is used here for comparison. Note that this algorithm does not return an optimized value of the tuning parameter λ, and thus we cannot report loss values. In general, this algorithm performs very similar to the ZS algorithm. A big difference is the excellent performance for the false positives (FP), but a much poorer performance for the false negatives (FN); the latter might be more important in applications.

2.1.1 Level of contamination: 10%

Tables SI-1, SI-2 , and SI-3 present the results for the low, moderate, and high-dimensional data configuration respectively, with $\rho = 0.5$. Table SI-4 reports the comparison results of selective performance, FP and FN, among different methods, scenarios, and parameter configurations.

		PE		PE (10%)		loss 1		loss 2		loss ∞	
		mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
(A)	Lasso	0.408	0.112	0.293	0.080	1.540	0.540	0.244	0.138	0.265	0.084
	ZS	**0.394**	0.103	**0.282**	0.078	**1.457**	0.494	**0.212**	0.121	**0.239**	0.073
	RobL	0.616	0.350	0.436	0.241	2.329	1.013	0.626	0.564	0.399	0.165
	RobZS	0.953	0.528	0.684	0.406	3.321	1.213	1.050	0.698	0.525	0.179
	SLTS	1.239	0.534	0.884	0.404	3.283	0.959	1.652	0.823	0.733	0.182
	ZS (B&T)	0.574	0.428	0.422	0.324						
(B)	Lasso	4.252	2.176	4.616	1.988	5.911	2.665	4.583	2.889	1.145	0.266
	ZS	4.058	1.898	4.473	2.031	5.552	2.080	4.150	2.412	1.110	0.263
	RobL	0.777	0.609	0.882	0.696	**2.509**	1.398	0.862	0.989	0.463	0.249
	RobZS	**0.713**	0.322	**0.811**	0.370	2.718	1.018	**0.709**	0.451	**0.426**	0.138
	SLTS	0.807	0.366	0.913	0.415	2.531	0.850	0.973	0.533	0.572	0.163
	ZS (B&T)	5.199	2.156	5.693	2.337						
(C)	Lasso	17.318	7.031	14.890	4.443	13.166	3.899	19.323	10.123	2.535	0.681
	ZS	16.885	6.449	15.353	4.277	12.013	3.019	15.064	5.915	2.211	0.493
	RobL	**0.743**	0.403	**0.838**	0.466	**2.535**	0.937	0.775	0.541	0.468	0.170
	RobZS	0.755	0.437	0.843	0.498	2.718	1.210	**0.725**	0.515	**0.434**	0.157
	SLTS	0.896	0.412	0.999	0.469	2.729	0.904	1.091	0.606	0.589	0.169
	ZS (B&T)	21.443	6.958	18.255	4.827						

Table SI-1: Means and standard deviations of various performance measures among different methods, based on 100 simulations. Parameter configuration: $(n,\, p)$=$(50, 30)$, $\rho = 0.5$.

		PE		PE (10%)		loss 1		loss 2		loss ∞	
		mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
(A)	Lasso	0.437	0.085	0.312	0.067	2.172	0.515	0.305	0.121	0.298	0.086
	ZS	**0.429**	0.088	**0.306**	0.071	**2.100**	0.582	**0.280**	0.111	**0.283**	0.076
	RobL	0.819	0.680	0.581	0.479	3.252	1.583	0.963	1.137	0.480	0.253
	RobZS	1.066	0.448	0.752	0.325	3.939	1.062	1.363	0.696	0.610	0.162
	SLTS	1.247	0.479	0.867	0.341	3.838	1.042	1.729	0.768	0.722	0.173
	ZS (B&T)	0.479	0.434	0.335	0.302						
(B)	Lasso	4.372	1.459	4.860	1.578	6.975	2.929	4.476	1.382	1.111	0.184
	ZS	4.216	1.291	4.682	1.402	6.572	2.172	4.344	1.208	1.122	0.203
	RobL	1.052	0.629	1.180	0.687	3.683	1.310	1.248	0.925	0.568	0.226
	RobZS	**0.845**	0.387	**0.949**	0.440	3.375	0.852	**0.930**	0.498	**0.499**	0.138
	SLTS	0.925	0.426	1.042	0.482	**3.192**	0.966	1.119	0.569	0.567	0.142
	ZS (B&T)	4.714	1.764	5.209	1.849						
(C)	Lasso	5.666	1.041	4.128	0.667	11.735	2.213	13.622	1.759	2.107	0.229
	ZS	9.814	1.794	7.537	1.269	10.835	1.900	9.691	1.374	1.782	0.245
	RobL	0.980	0.675	1.107	0.785	3.687	1.484	1.191	1.006	0.557	0.219
	RobZS	**0.773**	0.330	**0.866**	0.377	3.271	0.981	**0.897**	0.579	**0.482**	0.155
	SLTS	0.901	0.349	1.013	0.399	**3.329**	0.958	1.186	0.590	0.584	0.149
	ZS (B&T)	13.829	2.519	10.110	1.881						

Table SI-2: Means and standard deviations of various performance measures among different methods, based on 100 simulations. Parameter configuration: $(n,\, p)$=$(100, 200)$, $\rho = 0.5$.

		PE		PE (10%)		loss 1		loss 2		loss ∞	
		mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
(A)	Lasso	0.757	0.256	0.535	0.186	3.827	0.852	0.927	0.455	0.515	0.128
	ZS	**0.725**	0.233	**0.515**	0.167	**3.730**	0.884	**0.840**	0.417	**0.490**	0.142
	RobL	2.335	1.308	1.674	0.966	5.642	1.545	3.234	1.708	0.932	0.301
	RobZS	1.705	0.868	1.211	0.619	4.891	1.308	2.364	1.239	0.798	0.244
	SLTS	2.322	0.739	1.652	0.540	5.984	0.947	3.290	0.873	0.976	0.164
	ZS (B&T)	0.831	0.800	0.602	0.597						
(B)	Lasso	4.811	1.523	5.179	1.443	7.078	2.974	5.175	0.976	1.248	0.171
	ZS	4.847	1.392	5.239	1.387	7.324	2.795	5.175	0.956	1.242	0.183
	RobL	2.534	1.009	2.828	1.117	5.971	1.252	3.528	1.145	0.984	0.201
	RobZS	**1.379**	0.658	**1.548**	0.749	**4.365**	1.205	**1.927**	1.005	**0.710**	0.215
	SLTS	1.705	0.532	1.919	0.623	5.375	0.917	2.468	0.735	0.826	0.142
	ZS (B&T)	5.823	2.659	6.273	2.732						
(C)	Lasso	3.522	0.847	2.577	0.653	9.468	1.408	8.122	0.863	1.588	0.200
	ZS	7.538	1.432	5.537	0.989	9.912	1.398	8.169	0.939	1.619	0.182
	RobL	2.640	1.045	2.969	1.204	5.957	1.344	3.642	1.137	1.006	0.203
	RobZS	**1.435**	0.621	**1.614**	0.705	**4.531**	0.977	**1.965**	0.955	**0.731**	0.205
	SLTS	1.873	0.534	2.104	0.607	5.486	0.897	2.615	0.734	0.853	0.146
	ZS (B&T)	11.061	2.439	7.779	1.568						

Table SI-3: Means and standard deviations of various performance measures among different methods, based on 100 simulations. Parameter configuration: $(n,\, p)$=$(100, 1000)$, $\rho = 0.5$.

			Scenario					
			(A)		**(B)**		**(C)**	
(n, p)		**Method**	mean	sd	mean	sd	mean	sd
(50, 30)	FP	Lasso	11.64	3.558	5.16	5.013	10.35	2.911
		ZS	11.71	3.588	5.44	4.243	10.20	2.449
		RobL	11.93	3.927	10.84	4.867	10.46	3.922
		RobZS	15.81	4.869	14.19	4.849	14.05	5.435
		SLTS	6.95	2.393	6.77	2.304	7.83	2.336
		ZS (B&T)	**1.78**	1.244	**1.90**	1.925	**4.02**	1.859
	FN	Lasso	**0.00**	0.000	3.31	1.704	2.58	1.273
		ZS	**0.00**	0.000	2.91	1.429	2.72	0.944
		RobL	0.04	0.197	0.41	0.933	0.24	0.515
		RobZS	0.14	0.349	**0.07**	0.256	**0.14**	0.349
		SLTS	0.91	0.854	0.35	0.539	0.48	0.627
		ZS (B&T)	0.22	0.579	3.76	1.357	3.92	1.051
(100, 200)	FP	Lasso	35.23	11.247	14.81	13.763	22.98	13.276
		ZS	34.11	11.611	13.35	11.111	24.61	7.406
		RobL	36.20	14.613	30.64	12.826	32.92	15.255
		RobZS	35.26	12.435	33.43	11.229	34.28	11.568
		SLTS	21.35	5.364	21.86	5.472	23.08	6.250
		ZS (B&T)	**1.27**	1.246	**1.56**	1.725	**4.81**	1.745
	FN	Lasso	**0.00**	0.000	3.77	1.118	1.07	0.935
		ZS	**0.00**	0.000	3.60	1.181	2.99	0.893
		RobL	0.32	0.963	0.69	1.051	0.52	1.010
		RobZS	0.40	0.620	**0.22**	0.543	**0.21**	0.456
		SLTS	1.05	0.925	0.40	0.603	0.56	0.770
		ZS (B&T)	0.28	0.653	4.20	0.791	4.34	0.768
(100, 1000)	FP	Lasso	57.25	18.511	15.48	21.731	31.38	15.104
		ZS	57.64	17.505	17.64	20.104	32.11	10.087
		RobL	40.83	22.215	34.61	22.121	32.17	18.933
		RobZS	38.80	17.594	35.23	16.321	37.37	18.422
		SLTS	40.14	6.012	41.62	5.836	40.34	6.406
		ZS (B&T)	**1.70**	1.534	**2.07**	2.248	**4.27**	1.752
	FN	Lasso	0.28	0.533	4.76	0.889	1.37	0.812
		ZS	**0.16**	0.420	4.47	1.010	3.41	0.842
		RobL	2.64	1.755	3.05	1.123	3.23	1.162
		RobZS	1.53	1.381	**1.29**	1.131	**1.22**	1.079
		SLTS	2.71	0.856	1.94	0.789	1.97	0.834
		ZS (B&T)	0.72	1.064	4.59	0.866	4.71	0.656

Table SI-4: Comparison of selective performance among different methods, scenarios, and parameter configurations ($\rho = 0.5$).

2.1.2 Level of contamination: 20%

Tables SI-5, SI-6 , and SI-7 present the results for the low, moderate, and high-dimensional data configuration respectively with $\rho = 0.2$. Due to the high level of contamination we used for the comparison the prediction error, evaluated using a cleaned test set, and the trimmed prediction error with trim. levels equal to 0.2. Table SI-8 reports the comparison results of selective performance, FP and FN, among different methods, scenarios, and parameter configurations.

Tables SI-9, SI-10, SI-11, and SI-12 are related to a correlation structure with $\rho = 0.5$.

		PE		**PE (20%)**		**loss 1**		**loss 2**		**loss ∞**	
		mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
(A)	Lasso	0.407	0.125	0.236	0.075	1.260	0.565	0.168	0.110	0.218	0.064
	ZS	**0.399**	0.125	**0.231**	0.078	**1.190**	0.521	**0.147**	0.094	**0.196**	0.059
	RobL	0.680	0.395	0.407	0.255	1.984	0.990	0.460	0.413	0.346	0.141
	RobZS	1.002	0.613	0.602	0.417	3.073	1.205	0.831	0.657	0.440	0.161
	SLTS	1.495	0.862	0.904	0.507	2.966	1.099	1.361	0.870	0.654	0.208
(B)	Lasso	9.709	3.530	10.900	2.937	6.532	2.495	5.860	2.760	1.341	0.251
	ZS	9.352	3.340	10.709	2.958	6.384	2.321	5.613	2.627	1.312	0.265
	RobL	3.902	2.860	4.835	3.272	4.657	2.117	3.443	2.321	1.004	0.417
	RobZS	0.606	0.474	0.799	0.642	1.709	0.732	**0.349**	0.439	**0.296**	0.153
	SLTS	**0.589**	0.243	**0.781**	0.334	**1.637**	0.530	0.388	0.219	0.360	0.111
(C)	Lasso	25.381	6.659	14.271	4.063	15.394	3.901	21.654	10.409	2.424	0.771
	ZS	23.144	5.842	13.861	3.369	14.177	2.229	18.123	4.361	2.270	0.553
	RobL	1.592	0.659	2.067	0.866	3.586	0.947	1.572	0.646	0.716	0.179
	RobZS	**0.557**	0.241	**0.728**	0.340	1.665	0.824	**0.322**	0.317	**0.293**	0.126
	SLTS	0.609	0.262	0.797	0.357	**1.550**	0.595	0.403	0.284	0.369	0.128

Table SI-5: Means and standard deviations of various performance measures among different methods, based on 100 simulations. Parameter configuration: $(n,\ p)$=(50, 30), $\rho = 0.2$.

		PE		**PE (20%)**		**loss 1**		**loss 2**		**loss ∞**	
		mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
(A)	Lasso	0.379	0.074	0.222	0.046	1.520	0.525	0.159	0.072	0.211	0.067
	ZS	**0.368**	0.067	**0.217**	0.043	**1.420**	0.425	**0.138**	0.056	**0.191**	0.052
	RobL	0.626	0.583	0.368	0.368	2.133	1.300	0.425	0.588	0.312	0.168
	RobZS	1.124	0.476	0.660	0.293	3.515	0.946	0.988	0.506	0.521	0.148
	SLTS	1.120	0.538	0.649	0.318	3.003	1.020	1.035	0.600	0.547	0.165
(B)	Lasso	8.540	1.870	9.884	1.866	6.824	2.344	4.790	1.253	1.172	0.233
	ZS	8.486	1.925	9.802	1.950	6.825	2.324	4.704	1.248	1.151	0.251
	RobL	3.226	1.647	4.124	2.004	5.061	1.567	3.051	1.500	0.916	0.294
	RobZS	0.527	0.160	**0.703**	0.208	1.989	0.601	**0.290**	0.137	**0.280**	0.079
	SLTS	**0.524**	0.157	**0.703**	0.208	**1.860**	0.502	0.309	0.155	0.300	0.082
(C)	Lasso	5.776	1.160	3.482	0.691	10.825	1.543	11.265	1.343	1.737	0.162
	ZS	9.899	1.723	5.576	1.039	11.503	1.702	8.406	1.210	1.445	0.174
	RobL	1.511	0.547	1.978	0.743	3.918	1.210	1.387	0.545	0.685	0.153
	RobZS	**0.474**	0.140	**0.630**	0.205	1.701	0.583	**0.236**	0.152	**0.251**	0.084
	SLTS	0.528	0.172	0.701	0.239	**1.534**	0.504	0.341	0.202	0.327	0.093

Table SI-6: Means and standard deviations of various performance measures among different methods, based on 100 simulations. Parameter configuration: $(n,\ p)$=(100, 200), $\rho = 0.2$.

		PE		PE (20%)		loss 1		loss 2		loss ∞	
		mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
(A)	Lasso	0.505	0.138	0.296	0.081	2.253	0.755	0.301	0.162	0.287	0.080
	ZS	**0.485**	0.120	**0.286**	0.072	**2.119**	0.662	**0.268**	0.136	**0.266**	0.076
	RobL	3.039	1.755	1.757	1.012	5.604	1.949	3.140	1.871	0.897	0.370
	RobZS	1.668	1.130	0.955	0.656	4.134	1.649	1.619	1.289	0.615	0.286
	SLTS	2.647	1.072	1.506	0.597	5.549	1.312	2.685	1.062	0.868	0.211
(B)	Lasso	9.634	2.141	10.775	2.041	7.466	3.245	5.598	1.304	1.295	0.197
	ZS	9.751	2.038	10.898	1.980	7.679	3.412	5.673	1.202	1.293	0.208
	RobL	3.293	1.784	4.214	2.169	5.262	1.241	3.038	1.426	0.917	0.261
	RobZS	**0.900**	0.475	**1.177**	0.604	**3.023**	0.934	**0.699**	0.510	**0.416**	0.136
	SLTS	0.974	0.514	1.274	0.662	3.277	1.143	0.814	0.555	0.456	0.144
(C)	Lasso	3.466	0.834	2.250	0.542	8.253	0.889	6.613	0.666	1.288	0.134
	ZS	7.139	1.313	4.213	0.833	10.107	1.189	6.180	0.929	1.225	0.197
	RobL	2.816	0.968	3.709	1.286	5.379	1.358	2.700	0.980	0.895	0.165
	RobZS	**0.824**	0.410	**1.086**	0.553	3.013	0.969	**0.647**	0.448	**0.404**	0.141
	SLTS	1.000	0.474	1.322	0.649	**2.830**	1.062	0.893	0.536	0.498	0.156

Table SI-7: Means and standard deviations of various performance measures among different methods, based on 100 simulations. Parameter configuration: $(n,\ p)$=(100, 1000), $\rho = 0.2$.

			Scenario					
			(A)		(B)		(C)	
$(n,\ p)$		**Method**	mean	sd	mean	sd	mean	sd
(50, 30)	FP	Lasso	10.46	4.890	**4.23**	4.870	12.46	2.630
		ZS	10.44	4.509	4.34	4.479	11.62	2.473
		RobL	11.05	4.787	5.66	4.862	9.40	4.192
		RobZS	16.88	5.002	11.60	4.718	10.99	4.773
		SLTS	**7.01**	2.588	7.33	2.400	**5.55**	2.405
	FN	Lasso	**0.00**	0.000	4.03	1.514	2.42	1.224
		ZS	**0.00**	0.000	3.65	1.604	2.36	0.990
		RobL	0.02	0.200	2.45	2.022	0.88	0.769
		RobZS	0.07	0.256	0.04	0.243	**0.03**	0.223
		SLTS	0.48	0.611	**0.02**	0.141	**0.03**	0.171
(100, 200)	FP	Lasso	27.42	12.763	**11.27**	11.478	22.26	10.031
		ZS	26.19	10.670	11.71	10.994	29.08	7.804
		RobL	28.49	11.923	14.75	12.485	23.64	10.506
		RobZS	34.95	11.435	27.17	12.096	24.99	10.453
		SLTS	**20.22**	5.323	22.23	5.245	**9.74**	4.718
	FN	Lasso	**0.00**	0.000	3.91	1.349	0.73	0.566
		ZS	**0.00**	0.000	3.68	1.476	1.93	0.956
		RobL	0.07	0.293	2.18	1.642	0.80	0.752
		RobZS	0.07	0.256	**0.00**	0.000	**0.00**	0.000
		SLTS	0.22	0.484	**0.00**	0.000	0.02	0.141
(100, 1000)	FP	Lasso	45.53	18.607	**13.20**	16.909	23.09	9.484
		ZS	42.27	15.614	14.44	18.589	37.72	10.549
		RobL	40.04	20.747	24.02	17.580	27.62	13.510
		RobZS	41.25	17.076	37.59	14.063	41.34	14.195
		SLTS	**39.78**	7.624	38.06	7.799	**20.77**	8.546
	FN	Lasso	**0.00**	0.000	4.87	1.031	1.17	0.792
		ZS	**0.00**	0.000	4.72	1.155	2.35	0.833
		RobL	2.19	1.779	2.41	1.706	2.04	1.072
		RobZS	0.74	1.143	**0.07**	0.293	**0.07**	0.256
		SLTS	1.64	0.948	0.15	0.386	0.26	0.525

Table SI-8: Comparison of selective performance among different methods, scenarios, and parameter configurations ($\rho = 0.2$).

		PE		PE (20%)		loss 1		loss 2		loss ∞	
		mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
(A)	Lasso	0.421	0.118	0.250	0.083	1.704	0.675	0.286	0.166	0.280	0.090
	ZS	**0.391**	0.104	**0.231**	0.065	**1.523**	0.566	**0.232**	0.140	**0.252**	0.081
	RobL	0.668	0.440	0.398	0.257	2.441	1.164	0.664	0.591	0.427	0.185
	RobZS	0.927	0.373	0.550	0.227	3.493	0.945	1.074	0.540	0.529	0.148
	SLTS	1.250	0.475	0.747	0.299	3.358	0.904	1.677	0.713	0.730	0.161
(B)	Lasso	8.493	3.104	9.800	2.980	6.734	2.790	6.207	3.249	1.363	0.302
	ZS	8.083	2.526	9.575	2.739	6.354	2.341	5.704	2.768	1.345	0.276
	RobL	2.068	1.628	2.702	2.164	3.856	1.569	2.486	1.907	0.837	0.388
	RobZS	0.588	0.261	0.785	0.368	2.112	0.809	**0.456**	0.309	**0.347**	0.114
	SLTS	**0.583**	0.243	**0.777**	0.346	**1.882**	0.614	0.530	0.319	0.420	0.132
(C)	Lasso	26.858	8.908	15.546	4.543	16.323	5.348	30.367	17.613	3.181	0.938
	ZS	24.695	6.512	16.195	3.785	13.808	2.569	20.514	6.218	2.657	0.648
	RobL	1.537	0.722	2.040	0.977	4.036	1.081	1.959	0.937	0.759	0.204
	RobZS	**0.552**	0.225	**0.739**	0.314	**1.966**	0.810	**0.424**	0.282	**0.339**	0.120
	SLTS	0.634	0.285	0.854	0.404	2.022	0.747	0.674	0.439	0.480	0.168

Table SI-9: Means and standard deviations of various performance measures among different methods, based on 100 simulations. Parameter configuration: $(n,\ p)$=(50, 30), $\rho = 0.5$.

		PE		PE (20%)		loss 1		loss 2		loss ∞	
		mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
(A)	Lasso	0.444	0.108	0.257	0.073	2.229	0.714	0.317	0.155	0.294	0.080
	ZS	**0.434**	0.103	**0.251**	0.069	**2.116**	0.663	**0.296**	0.150	**0.283**	0.083
	RobL	0.765	0.680	0.444	0.400	3.104	1.413	0.818	0.967	0.440	0.218
	RobZS	1.132	0.500	0.654	0.311	4.038	1.111	1.461	0.774	0.635	0.191
	SLTS	1.264	0.442	0.733	0.276	3.910	1.018	1.763	0.744	0.724	0.165
(B)	Lasso	8.075	2.122	9.611	2.061	6.980	3.532	5.502	2.042	1.272	0.205
	ZS	7.851	1.734	9.403	1.865	6.898	2.676	5.346	1.455	1.264	0.204
	RobL	1.973	1.036	2.613	1.370	4.854	1.341	2.581	1.349	0.839	0.264
	RobZS	**0.581**	0.152	**0.779**	0.215	2.692	0.728	**0.529**	0.233	**0.383**	0.094
	SLTS	0.595	0.192	0.797	0.271	**2.554**	0.700	0.615	0.309	0.418	0.104
(C)	Lasso	6.487	1.365	3.662	0.835	11.582	2.080	14.509	1.863	2.210	0.294
	ZS	11.358	2.040	6.201	1.155	11.712	2.136	10.970	1.783	1.901	0.311
	RobL	1.556	0.581	2.049	0.810	4.680	1.313	1.974	0.844	0.739	0.173
	RobZS	**0.521**	0.211	**0.684**	0.276	2.318	0.764	**0.433**	0.329	**0.345**	0.114
	SLTS	0.605	0.245	0.796	0.325	**2.225**	0.793	0.680	0.419	0.456	0.129

Table SI-10: Means and standard deviations of various performance measures among different methods, based on 100 simulations. Parameter configuration: $(n,\ p)$=(100, 200), $\rho = 0.5$.

		PE		PE (20%)		loss 1		loss 2		loss ∞	
		mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
(A)	Lasso	0.686	0.205	0.407	0.137	3.575	0.736	0.811	0.419	0.476	0.135
	ZS	**0.663**	0.208	**0.389**	0.138	**3.434**	0.735	**0.755**	0.435	**0.468**	0.162
	RobL	2.461	1.214	1.400	0.698	5.676	1.393	3.517	1.621	0.988	0.287
	RobZS	1.658	0.836	0.977	0.504	4.792	1.362	2.336	1.307	0.778	0.264
	SLTS	2.238	0.736	1.310	0.458	5.903	1.095	3.197	0.979	0.961	0.161
(B)	Lasso	8.542	1.921	10.109	2.040	7.273	3.277	5.915	1.188	1.359	0.150
	ZS	8.619	2.026	10.158	2.062	7.403	3.543	5.953	1.244	1.363	0.150
	RobL	2.676	0.962	3.521	1.272	5.719	1.122	3.637	0.955	1.024	0.168
	RobZS	**1.135**	0.545	**1.498**	0.712	**3.977**	1.137	**1.477**	0.782	**0.649**	0.187
	SLTS	1.172	0.495	1.535	0.627	4.443	1.103	1.570	0.709	0.659	0.158
(C)	Lasso	3.749	0.913	2.304	0.610	9.211	0.971	8.730	0.988	1.656	0.238
	ZS	8.120	1.606	4.547	0.900	10.257	1.560	8.798	1.267	1.673	0.228
	RobL	2.501	0.808	3.306	1.106	5.662	1.225	3.363	0.889	0.981	0.143
	RobZS	**1.017**	0.497	**1.344**	0.680	**3.735**	1.008	**1.314**	0.764	**0.591**	0.193
	SLTS	1.145	0.441	1.524	0.617	3.913	1.074	1.660	0.687	0.692	0.156

Table SI-11: Means and standard deviations of various performance measures among different methods, based on 100 simulations. Parameter configuration: $(n,\ p)$=(100, 1000), $\rho = 0.5$.

			Scenario					
			(A)		(B)		(C)	
(n, p)		Method	mean	sd	mean	sd	mean	sd
(50, 30)	FP	Lasso	12.15	4.480	4.03	4.670	9.95	3.176
		ZS	11.66	3.906	**3.92**	4.007	9.52	2.359
		RobL	12.95	4.425	6.27	4.537	9.35	3.901
		RobZS	16.14	5.365	12.26	4.215	11.22	4.605
		SLTS	**7.08**	2.232	7.32	2.369	**5.91**	2.099
	FN	Lasso	0.01	0.100	4.52	1.403	2.18	1.258
		ZS	**0.00**	0.000	4.15	1.438	2.33	1.016
		RobL	0.08	0.307	1.88	1.996	1.20	1.101
		RobZS	0.15	0.411	**0.05**	0.261	**0.04**	0.197
		SLTS	0.81	0.825	0.12	0.327	0.28	0.533
(100, 200)	FP	Lasso	35.56	12.871	**9.11**	13.339	17.27	11.585
		ZS	33.66	12.849	9.47	11.585	24.57	7.178
		RobL	35.90	12.114	20.80	16.122	24.95	12.591
		RobZS	34.80	12.428	32.43	11.559	29.44	11.417
		SLTS	**21.62**	5.594	24.17	6.239	**13.34**	5.756
	FN	Lasso	**0.00**	0.000	4.70	1.106	1.50	0.927
		ZS	**0.00**	0.000	4.39	1.278	3.06	0.789
		RobL	0.23	0.827	2.22	1.618	1.31	1.143
		RobZS	0.53	0.745	**0.04**	0.197	**0.04**	0.243
		SLTS	1.02	0.953	0.08	0.273	0.13	0.442
(100, 1000)	FP	Lasso	56.67	18.029	**10.90**	17.712	**24.24**	13.446
		ZS	54.76	16.118	11.96	18.473	29.61	8.935
		RobL	**34.47**	20.586	22.79	18.578	24.45	14.422
		RobZS	37.91	17.545	35.23	14.602	36.76	15.428
		SLTS	40.38	7.318	42.45	6.452	26.15	7.488
	FN	Lasso	0.24	0.588	5.32	0.709	1.83	1.006
		ZS	**0.23**	0.633	5.13	0.895	3.50	0.948
		RobL	2.86	1.706	3.43	1.174	3.24	0.878
		RobZS	1.52	1.322	**0.65**	0.770	**0.70**	0.916
		SLTS	2.63	0.917	0.87	0.800	1.11	0.898

Table SI-12: Comparison of selective performance among different methods, scenarios, and parameter configurations ($\rho = 0.5$).

2.2 Effect of debiasing strategies

The RobZS estimator, as defined in Equation (10) of the main paper, is now called näive RobZS to distinguish this direct solution from the debiased ones. For an empirical comparison of the prediction accuracy of the näive, rescaled, relaxed and hybrid RobZS estimators, we consider a Monte Carlo simulation with 50 repetitions, generating the training and test data in a non-contaminated scenario, i.e. scenario (A), according to the simulation setting outlined in the main paper, but considering an elastic-net penalty.

Table SI-13 and Figure SI-1 summarize the prediction errors (means and standard deviations) among the different debiasing strategies and parameter configurations .

	$n = 50$, $p = 30$		$n = 100$, $p = 200$		$n = 100$, $p = 1000$	
RobZS	mean	sd	mean	sd	mean	sd
näive	0.986	0.598	1.884	1.342	3.728	1.726
rescaled	1.024	0.580	1.791	1.271	3.673	1.814
relaxed	0.996	0.597	2.467	1.272	4.063	1.541
hybrid	**0.831**	0.414	**1.137**	0.581	**2.329**	1.485

Table SI-13: Means and standard deviations of the prediction error among different debiasing strategies and parameter configurations ($\rho = 0.2$) based on 50 simulations.

It can be seen that hybrid RobZS outperforms the näive RobZS by providing higher predictive accuracy in all simulation settings. The reduction in prediction error in the low, medium, and high dimensional setting are 16%, 40% and 38%, respectively.

Figure SI-1 shows also box plots related to the empirical distribution of the optimal α parameters associated to the näive RobZS solution. It can be seen that in low dimension ($n > p$) the RobZS solution produces, as expected, sparse solutions. In high dimensional setting, the näive RobZS tends to select a large number of false positives, the optimal α parameters are in fact close to zero, thus the two-step hybrid rescaling helps a lot to remedy this distortion.

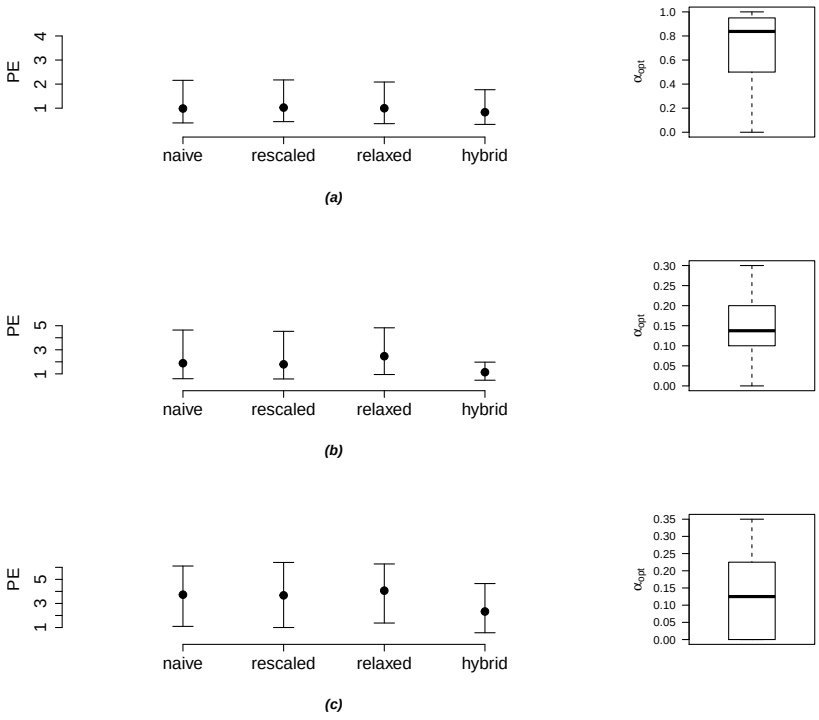

Figure SI-1: Average prediction error of different estimators, näive, rescaled, relaxed and hybrid RobZS, in scenario (A), over all simulation runs. The error bars extend from the 5% to the 95% quantile. The box plots refer to the optimal α parameter associated to the näive RobZS solution. Parameter configuration: *(a)* $n = 50$, $p = 30$, $\rho = 0.2$, *(b)* $n = 100$, $p = 200$, $\rho = 0.2$, *(c)* $n = 100$, $p = 1000$, $\rho = 0.2$

2.3 Simulations with varying sparsity

The same simulation setting as in Section 3.1 of the main paper is used, but here the sparsity of the regression coefficient vector $\boldsymbol{\beta}$ is changed. Originally, this vector consisted of the entries $\beta_1 = 1$, $\beta_2 = -0.8$, $\beta_3 = 0.6$, $\beta_6 = -1.5$, $\beta_7 = -0.5$, $\beta_8 = 1.2$ and $\beta_j = 0$ for $j \in \{1, \ldots, p\} \setminus \{1, 2, 3, 6, 7, 8\}$. Here, we modify in turn two zero entries by values $+1$ and -1, and thus in each step, two more active variables in the models are generated. We use $n = 50$, $p = 50$, $\rho = 0.2$, and the number of active variables is changed from 6 to 26. Figure SI-2 presents the resulting means over 10 replications in each step (solid lines). The dashed lines are the means plus/minus two times the standard errors from the replications. The left plot panels are for non-contaminated data, i.e. scenario (A), the right panels for 10% contamination using scenario (C). The red lines are for ZS, the blue lines for RobZS. In the uncontaminated setting we can see that RobZS loses precision if the sparsity gets lower (more active variables). This loss is based on an increasing level of FN (true active variables are not identified). Also, FP should decrease with an increasing number of active variables, since FP refers to the number of noise (inactive) variables that are not identified. For ZS, the PE increases only slightly, and this is based on much less sparser models: FN is always close to zero, and FP even increases at the beginning.

In the contaminated scenario (right column of Figure SI-2), RobZS shows essentially the same behavior as in the uncontaminated case. However, ZS is clearly affected by the outliers, which is visible in terms of a much higher prediction error, and a clear increase for the FN. Only with a higher number of active variables (less sparsity in teh model), FN and FP are comparable for the non-robust and the robust method, which reflects the difficulty of these methods to identify the correct model. It can be concluded that higher sparsity is connected to more reliable model identification.

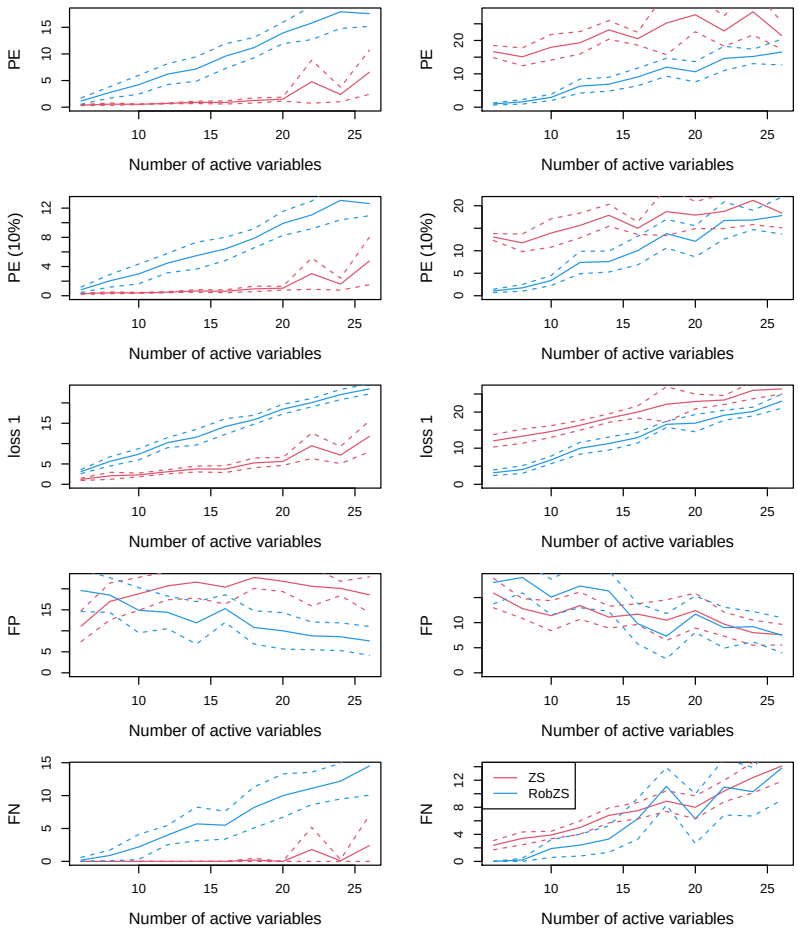

Figure SI-2: Performance of the ZS (red) and RobZS (blue) estimators by increasing the number of active variables. Left column for scenario (A), right column for scenario (C). Here, $n = 50$, $p = 50$, $\rho = 0.2$; shown are means (solid lines) plus/minus two standard errors derived from 10 simulation replications at each step.

2.4 Simulations using the elastic-net penalty

So far, all simulations were done just with a lasso penalty, except the investigation of the debiasing procedures in Section 2.2. The purpose of the following simulations is a direct comparison of ZS and RobZS by making use of their elastic-net penalties. We follow the strategy for the tuning parameter selection as mentioned in Section 1.

The simulation setting is different here. We consider two sets of highly correlated variables which are relevant for explaining the response, and a set of weakly correlated variables which does not contribute to the explanation of the response. In each of the 100 simulation replications, we generate $n = 50$ observations, and $p = 80$. The two blocks of highly correlated variables consist of 8 variables each, and they were (independently) generated as explained in Section 3.1 of the main paper, but with $\rho = 0.9$. The remaining block of 64 variables is also generated in the same way, but with $\rho = 0.2$. The entries of the variables of this last block in the vector $\boldsymbol{\beta}$ are all zeros, thus these are the noise variables, whereas the entries for the first blocks are alternating with $+1$ and -1. We used the simulation scenarios (A) (no contamination) and (B) (vertical outliers), and the results are presented in Table SI-14. Here we are also interested in the resulting optimized values of the tuning parameter α of the elastic-net penalty (last rows). It can be seen that in both scenarios, RobZS leads to a much lower value compared to ZS. If α gets close to zero, the penalization corresponds essentially to a ridge penalty, and the variable selection property is lost. This can also be seen in the lower values of FN, but simultaneously in higher values of FP for RobZS. Still RobZS is very competitive concerning the prediction error, and in the contaminated setting it clearly outperforms ZS.

		Scenario			
		(A)		**(B)**	
	Method	mean	sd	mean	sd
PE	ZS	1.655	0.397	3.279	1.520
	RobZS	1.665	0.326	1.640	0.392
PE (10%)	ZS	1.187	0.301	3.621	1.679
	RobZS	1.187	0.255	1.828	0.466
loss 1	ZS	16.89	1.130	17.75	2.603
	RobZS	17.02	0.823	17.03	1.009
loss 2	ZS	14.96	0.857	16.44	1.266
	RobZS	15.41	0.528	15.41	0.478
loss ∞	ZS	1.023	0.053	1.035	0.084
	RobZS	1.044	0.040	1.046	0.045
FP	ZS	25.6	20.2	15.7	21.8
	RobZS	42.3	24.1	42.3	26.0
FN	ZS	9.6	4.9	12.7	5.5
	RobZS	5.2	5.7	4.7	5.8
α	ZS	0.48	0.34	0.41	0.32
	RobZS	0.10	0.13	0.06	0.07

Table SI-14: Comparison ZS and RobZS by making use of their elastic-net penalties, for uncontaminated (A) and contaminated (B) data. The last two rows correspond to the optimized parameter α of the penalty.

3 Additional Application Results

In this section we report further results related to the application to human gut microbiome data.

Figure SI-3 shows the mean regression coefficients over all CV models by Lasso, ZeroSum, RobL, RobZS, and SLTS. The vertical dashed lines indicate bacterial taxa where the difference between ZeroSum and RobZS is greater than 0.01. Such differences may appear because of outliers particularly in these variables.

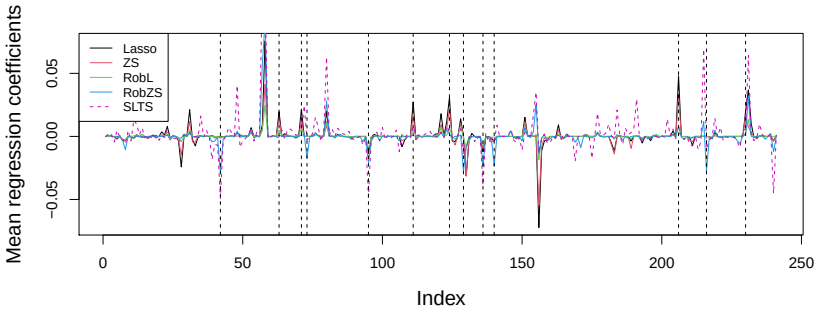

Figure SI-3: Analysis of gut microbiome data. Mean regression coefficients over all CV replications by Lasso, ZeroSum, RobL, RobZS, and SLTS.

A non-robust method may wrongly understand those values as important signals and include the variables with higher importance in the model. On the other hand, outliers may also lead to a masking effect (Maronna *et al.*, 2006), which can cause that crucial variables appear as less important in the classical fit.

References

Bates, S. and Tibshirani, R. (2019). Log-ratio lasso: Scalable, sparse estimation for log-ratio models. *Biometrics*, **75**(2), 613–624.

Maronna, R., Martin, R., and Yohai, V. (2006). *Robust Statistics*. John Wiley & Sons, Ltd.