**Supplementary**

**Post-estimation Shrinkage in Full and Selected Linear Regression Models in**

**Low-Dimensional Data Revisited**

Edwin Kipruto1, Willi Sauerbrei1

1Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center - University of Freiburg, Stefan-Meier-Street 26, 79104 Freiburg, Germany

**Corresponding Author**

Edwin Kipruto

Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center - University of Freiburg,

Stefan-Meier-Street 26, 79104 Freiburg, Germany

Email: <edwin.kipruto@uniklinik-freiburg.de>

1. **Simulation protocol from [1]**

**Table A1**. Summary of the simulation design following the ADEMP structure

|  |  |
| --- | --- |
| Aims (section 2.1) | * To compare variable selection methods using different tuning parameters (CV, AIC and BIC) or initial estimates in terms of model selection and prediction. * To assess the usefulness of post-estimation shrinkage in the prediction of classical variable selection methods and compare the results with penalized methods. * To compare the amount of shrinkage of regression coefficients of penalized and post-estimation shrinkage methods. * To assess the performance of different methods in the presence of relatively many noise variables, in larger sample size, in relatively high correlation and when *R2* approaches one. |
| Data generating mechanism (section 2.2) | **Training/development dataset**   * where *p* = 15 and is equal to the correlation coefficient between covariate and * where and   True regression coefficients (β) for 15 covariates  βA: 1.5, 0, 1, 0, 1, 0, 0.5, 0, 0.5, 0, 0.5, 0, -0.5, 0, 0  βB: 1.5, 0, 0.5, 0, 0.5, 0, 0.25, 0, 0.25, 0, 0.25, 0, -0.25, 0, 0 –(modified βA)  βC: 1,0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0 –From [21]  βD: 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0 –From [21]  Correlation structure ()  C1: Taken from [20]–low collinearity  C2: Autoregressive structure with = 0.3|i-j|-low collinearity  C3 Autoregressive structure with = 0.8|i-j| –moderate collinearity  C4: Adapted from real study, body fat data–high collinearity  ***R****2* **and sample size (*n*)**  *R2*={0.20, 0.30, 0.50, 0.71}; *n* ={100, 400}  **Number of scenarios (full factorial design) and simulation runs**  scenarios  N = 2,000 simulation repetitions per scenario  **Test dataset**   * New simulations with the same design as training dataset ().   Additional analysis  Additional analysis will be conducted with βA, C1, n = (400, 800) and a subset of *R2*={0.30, 0.50, 0.71, 0.8, 0.9} |
| Estimand/target of analysis (section 2.3) | * Selection status of each covariate and identification of the true model * Shrinkage factors for each regression estimate * Model prediction errors |
| Methods (section 2.4) | 1. Variable selection methods  |  |  |  | | --- | --- | --- | | Method | Tuning parameters | Initial estimates | | Lasso | 10-fold CV, AIC & BIC | N/A | | Garrote | 10-fold CV, AIC & BIC | OLS, ridge and lasso | | Alasso\* | 10-fold CV, AIC & BIC | OLS, ridge and lasso | | Rlasso\* | 10-fold CV, AIC & BIC | N/A | | Best subset | 10-fold CV, AIC & BIC | N/A | | BE\* | 10-fold CV, AIC & BIC | N/A |  1. Post-estimation shrinkage methods:   (i) Global [10], (ii) parameterwise [9] and (iii) Breiman’s method [5]  Estimation method: (i) leave-one-out CV and (ii) 10-fold CV |
| Performance measures (section 2.5) | * Inclusion and exclusion of variables: FNR & FPR–subsection 2.5.1 * classification of models: Probabilities–subsection 2.5.1 * Prediction accuracy: Model error (ME) –subsection 2.5.2 * Variability of ME within and between scenarios - section 5 in S1 File |

1. **Summary of simulation results of full models**

**Table A2**. Full models. Summary of prediction performance of post-estimation shrinkage and ridge regression in full models

|  |  |
| --- | --- |
| **Method** | **Summary** |
| Global shrinkage | * Tends to perform well in very low SNR levels and small sample sizes (Figure 4). * Performed poorly than NPWS, QPWS and ridge in moderate to high SNR (Figures 2, 4, and A3). * Outperformed NPWS in small sample sizes with low SNR (Figures 3, and 4). |
| PWS | * Performed better than the OLS only in very small sample sizes (Figure 4) and high correlation with low SNR levels where OLS estimates are known to be highly variable (Figures 2 and A3, bottom panels). * Overall, the results of PWS were unsatisfactory. |
| NPWS and QPWS | * Both performed similarly in most scenarios * Generally, NPWS and QPWS outperformed PWS and global shrinkage in most scenarios, except for very low SNR levels where global shrinkage was better (see Figures 2, 4, A3, and A4). * Outperformed ridge in low correlated settings, especially in moderate to high SNR (greater than 2) (Figures 2, 4, and A3). |
| Ridge | * Outperformed PWS and global shrinkage in most scenarios even in low correlated settings (Figures 2, 4, and A3). * Outperformed all post-estimation shrinkage methods in almost all scenarios with high correlation and low SNR levels (Figures 2, 3 and 4) |
| All post-estimation shrinkage (except PWS) | * Yielded estimated RR and RTE values smaller than those of OLS models in most scenarios (Figures 2, 3, 4, A3, and A4) * Exhibited greatest superiority over OLS in low SNR, high correlation and small sample sizes (Figures 2, 3, 4, A3 and A4). * Exhibited inferior predictive accuracy compared to ridge in scenarios with high correlation, very low SNR, or small sample sizes with SNR levels ranging from low to moderate (SNR<2). |

1. **Plots of results for full Models**
   1. **Shrinkage factors**

****

**Figure A1**. Full models. Distribution of shrinkage factors (truncated at -1) for global, PWS, NPWS, and QPWS in full models under low correlation () with p = 15 covariates, n = 100 and a beta-type A for SNR = 0.25 (upper panel), 1 (middle panel) and 6 (lower panel). Note that global shrinkage can estimate negative shrinkage factors in low SNR (upper panel).



**Figure A2**. Full models. Average shrinkage factors with one standard error band for global, PWS, NPWS, and QPWS in full models under high correlation () with p = 15, n = 100 and a beta-type A distribution for SNR levels of 0.25 (upper panel), 1 (middle panel) and 6 (lower panel). A comparison with low correlation shows that more shrinkage is applied to weak effects in high correlation.

* 1. **Prediction performance**

**3.2.1 Beta-type D distribution**



**Figure A3.** Full models. Prediction performance of methods in full models under low correlation (upper panel) and high correlation (low panel) settings, with n = 100, p = 15 covariates, and a beta-type D distribution.



**Figure A4**. Full models. Relative test error as a function of sample size (50, 100, and 400) for different methods in both low (left panel) and high (right panel) correlation settings. These scenarios include SNR values of 0.12 (upper panel) and 1 (lower panel) with p = 15 covariates, following a beta-type D (βD) distribution.

**4. Variable Selection**

**4.1 Shrinkage behaviour of a noise variable**



**Figure A5**. Selected model. Comparison of shrinkage factors against their regression estimates for a noise variable (x2) selected by best subset selection approach in small (n=50, left panel), moderate (n=100, middle panel), and large (n = 400, right panel) sample sizes. The analysis is conducted with an SNR of 1, low correlation (), and p = 15 covariates following beta-type A (βA) distribution.

**4.2 Prediction performance**

**4.2.1 Beta-type A distribution**



**Figure A6**. Selected models. Relative test error and average number of nonzero coefficients as functions of SNR, in both low (upper panel) and high (lower panel) correlation settings, with n = 50, p = 15 covariates, and beta-type A (βA) distribution.



**Figure A7**. Selected models. Relative test error and average number of nonzero coefficients as functions of SNR, in both low (upper panel) and high (lower panel) correlation settings, with n = 400, p = 15 covariates, and beta-type A (βA) distribution.

**4.2.2 Beta-type D distribution**



**Figure A8**. Selected models. Relative test error and average number of nonzero coefficients as functions of SNR, in both low (upper panel) and high (lower panel) correlation settings, with n = 100, p = 15 covariates, and beta-type D distribution.



**Figure A9**. Selected models. Relative test error and number of nonzero coefficients as functions of SNR, in both low (upper panel) and high (lower panel) correlation settings, with n = 50, p = 15 covariates, and beta-type D distribution.

**Reference**

* + - 1. Kipruto, E., and Sauerbrei, W. (2022). Comparison of variable selection procedures and investigation of the role of shrinkage in linear regression-protocol of a simulation study in low-dimensional data. *Plos one*. 17(10): e0271240.