Shrinkage Estimators for Uplift Regression

Krzysztof Rudaś[1,2] ✉ and Szymon Jaroszewicz[2]

[1] Warsaw University of Technology,
[2] Institute of Computer Science, Polish Academy of Sciences

Abstract. Uplift modeling is an approach to machine learning which allows for predicting the net effect of an action (with respect to not taking the action). To achieve this, the training population is divided into two parts: the treatment group, which is subjected to the action, and the control group, on which the action is not taken. Our task is to construct a model which will predict the difference between outcomes in the treatment and control groups conditional on individual objects' features. When the group assignment is random, the model admits a causal interpretation. When we assume linear responses in both groups, the simplest way of estimating the net effect of the action on an individual is to build two separate linear ordinary least squares (OLS) regressions on the treatment and control groups and compute the difference between their predictions. In classical linear models improvements in accuracy can be achieved through the use of so called shrinkage estimators such as the well known James-Stein estimator, which has a provably lower mean squared error than the OLS estimator. In this paper we investigate the use of shrinkage estimators in the uplift modeling problem. Unfortunately direct generalization of the James-Stein estimator does not lead to improved predictions, nor does shrinking treatment and control models separately. Therefore, we propose a new uplift shrinkage method where estimators in the treatment and control groups are shrunk jointly so as to minimize the error in the predicted net effect of the action. We prove that the proposed estimator does indeed improve on the double regression estimator.

1 Introduction

Selecting observations which should become targets for an action, such as a marketing campaign or a medical treatment, is a problem of growing importance in machine learning. Typically, the first step is to predict the effect of the action (response) using a model built on a sample of individuals subjected to the action. A new observation is classified as suitable for the action if the predicted response is above a certain threshold. Unfortunately this approach is not correct because the response that would have been observed had the action not been taken is ignored.

To clarify the problem, let us give a simple example. Suppose that we are owners of a shop which sells chocolate bars. In order to increase the sales of this product we give discounts to customers. Consider two cases. The first customer would have spent \$100 on chocolate after receiving the discount and \$95 in a situation when it was not given to him. The second one will spend \$50 and \$10, respectively. When we base our predictions only on a sample of customers subjected to the action (i.e. given the discount), we will prefer the first customer, but when we compare the amounts of money spent in cases of receiving and not receiving the discount, we will be disposed to send it to the second customer.

Clearly, the proper way to select targets for an action is to consider the difference between y^T, the response in case the individual is subjected to the action (treated) and y^C, the response when the individual is not subjected to the action (control). Unfortunately these two pieces of information are never available to us simultaneously. Once we send the discount we cannot 'unsend' it. This is known as the Fundamental Problem of Causal Inference [28].

Uplift modeling offers a solution to this problem based on dividing the population into two parts: treatment: subjected to the action and control on which the action is not taken. This second group is used as background thanks to which it is possible to partition the treatment response into a sum of two terms. The first is the response, which would have been observed if the treated objects were, instead, in the control group. The second is the additional effect observed only in the treatment group: the effect of the action. Based on this partition it is possible to construct a model predicting the desired difference between responses in the treatment and control groups [22].

Let us now introduce the notation used throughout the paper. We begin by describing the classical ordinary least squares regression. Only facts needed in the remaining part of the paper are given, full exposition can be found e.g. in [2]. We will assume that the predictor variables are arranged in an $n \times p$ matrix X and the responses are given in an n-dimensional vector y. We assume that y is related to X through a linear equation

$$y = X\beta + \varepsilon,$$

where β is an unknown coefficient vector and ε is a random noise vector with the usual assumptions that $\mathrm{E}\,\varepsilon_i = 0$, $\mathrm{Var}\,\varepsilon_i = \sigma^2$ and the components of ε are independent of each other. Moreover, we will make the assumption that the matrix X is fixed, which is frequently made in regression literature [2]. Our goal is to find an estimator of β which, on new test data X_{test}, y_{test}, achieves the lowest possible *mean squared error*

$$MSE(\hat{\beta}) = \mathrm{E}\,\|y_{test} - X_{test}\hat{\beta}\|^2, \tag{1}$$

where $\hat{\beta}$ is some estimator of β, and the expectation is taken over ε_{test} and $\hat{\beta}$. The most popular estimator is the Ordinary Least Squares (OLS) estimator obtained by minimizing the training set MSE $\|y - X\hat{\beta}\|^2$, given by

$$\hat{\beta} = (X'X)^{-1}X'y, \tag{2}$$

where $'$ denotes matrix transpose. In the rest of the paper $\hat{\beta}$ without additional subscripts will always denote the OLS estimator. It is well known that $\hat{\beta}$ is unbiased, $\mathrm{E}\,\hat{\beta} = \beta$, and its covariance matrix is $\sigma^2(X'X)^{-1}$ [2].

Let us now move to the case of regression in uplift modeling which is based on two training sets: treatment and control. We will adopt the convention that quantities related to the treatment group are denoted with superscript T, quantities related to the control group with a superscript C, and quantities related to the uplift itself with superscript U. Thus, in our context we will have two training sets X^T, y^T and X^C, y^C. Additionally denote $X = [X^{T'}|X^{C'}]'$, $y = [y^{T'}|y^{C'}]'$, i.e. the dataset obtained by concatenating treatment and control data records.

In this paper we will make an assumption (frequently made in statistical literature when linear models are considered), that responses in both groups are linear:

$$\begin{aligned} y^C &= X^C\beta^C + \varepsilon^C, \\ y^T &= X^T\beta^T + \varepsilon^T = X^T\beta^C + X^T\beta^U + \varepsilon^T. \end{aligned}$$

The additional effect observed in the treatment group, $X^T\beta^U$, is the quantity of interest and our goal, therefore, is to find an estimator of β^U. The easiest way to obtain such an estimator is to construct separate Ordinary Least Squares (OLS) estimators of β^T and β^C on treatment and control groups respectively, and to calculate the difference between them:

$$\hat{\beta}^U_d = \hat{\beta}^T - \hat{\beta}^C. \tag{3}$$

This estimator is called the *double regression estimator* [9]. It is easy to show that the estimator is an unbiased estimator of β^U [9].

In classical regression analysis there are several ways of lowering the predictive error of the ordinary least squares model by reducing its variance at the expense of introducing bias [2]. One class of such estimators are *shrinkage estimators* which scale the ordinary least squares estimate $\hat{\beta}$ by a factor $\alpha < 1$. The best known of such estimators is the James-Stein estimator [27]. Another choice is a class of shrinkage estimators based on minimizing predictive MSE [19].

The goal of this paper is to find shrinkage estimators for uplift regression, whose accuracy is better than that of the double regression estimator. We may shrink the treatment and control coefficients separately obtaining the following general form of uplift shrinkage estimator

$$\hat{\beta}^U_{\alpha^T,\alpha^C} = \alpha^T\hat{\beta}^T - \alpha^C\hat{\beta}^C.$$

with an appropriate choice of α^T and α^C.

We introduce two types of such estimators, the first following the James-Stein approach, the second the MSE minimization approach. For each type we again introduce to sub-types: one in which treatment and control shrinkage factors α^T and α^C are found independently of each other (these are essentially double shrinkage models) and another which in shrinkage factors are estimated jointly in order to produce the best possible estimates of β^U. We demonstrate experimentally that MSE minimization based shrinkage with joint optimization of α^T and α^C gives the best uplift shrinkage estimator. We also formally prove that under certain assumptions it dominates the double regression estimator $\hat{\beta}^U_d$.

1.1 Literature overview

We will now review the literature on uplift modeling. Literature related to shrinkage estimators will be discussed in Section 2.

Uplift modeling is a part of broader field of causal discovery and we will begin by positioning it within this field. The goal of causal discovery is not predicting future outcomes, but, instead, modeling the effects of interventions which directly change the values of some variables [20]. One can distinguish two general approaches to causal discovery: one based on purely observational data [20,26] and another one, in which the action being analyzed has actively been applied to a subgroup of the individuals.

Only the second approach is relevant to this paper. Large amount of related research has been conducted in the social sciences. However, their main research focus is on the cases where treatment assignment is nonrandom or biased [7,5]. Examples of methods used are propensity score matching or weighting by inverse probability of treatment [7,5]. Unfortunately, the success of those method depends on untestable assumptions such as 'no unmeasured confounders'. Only random treatment assignment guarantees that the causal effect is correctly identified. Most of those methods use double regression and do not try to improve the estimator itself. Uplift modeling differs from those methods since it is focused on obtaining the best possible estimate of an action's effect based on a randomized trial.

Most uplift modeling publications concern the problem of classification. The first published methods were based on decision trees [22,24]. They used modified splitting criteria to maximize difference in responses between the two groups. Similar methods have been devised under the name of estimating heterogenous treatment effects [1,8]. Later works extend these methods to ensembles of trees [4,25]. Work on linear uplift models includes approaches based on class variable transformation [15,10,11] used with logistic regression and approaches based on Support Vector Machines [13,29,14]. These methods can be used only with classification problems. Uplift regression methods were proposed in [9]. The paper also contained a theoretical analysis comparing several regression models.

The paper is organized as follows. In Section 2 we discuss shrinkage estimators used in classical linear regression. In Section 3 we derive four uplift shrinkage estimators and prove that, under certain assumptions, one of them dominates the double regression model. In Section 4 we evaluate the proposed estimators on two real-life datasets and conclude in Section 5.

2 Shrinkage estimators for linear regression

We now present a short review of shrinkage estimators for classical ordinary least squares models which is sufficient for understanding the results in Section 3.

2.1 James-Stein estimator

The famous James-Stein estimator has been presented in early 60s [27]. The authors proved that it allows for obtaining estimates with lower mean squared error than maximum likelihood based methods, which came as a shock to the statistical community. More specifically, let $Z \sim N(\mu, I)$ be a p-dimensional random vector whose mean μ is to be estimated based on a single sample z. The best unbiased estimator is $\hat{\mu} = z$. However, it can be proven [27] that the estimator

$$\hat{\mu}_{JS1} = \left(1 - \frac{(p-2)}{||\hat{\mu}||^2}\right)\hat{\mu} \tag{4}$$

has a lower mean squared error $\mathrm{E}\,\|\hat{\mu}_{JS1} - \mu\|^2 \leq \mathrm{E}\,\|\hat{\mu} - \mu\|^2$. The biggest gain is achieved for $\mu = 0$ and decreases when the norm of μ becomes large. To mitigate this effect, a modified shrinkage estimator was proposed by Efron [18]:

$$\hat{\mu}_{JS2} = \left(1 - \frac{(p-3)}{||\hat{\mu} - \overline{\hat{\mu}}||^2}\right)(\hat{\mu} - \overline{\hat{\mu}}) + \overline{\hat{\mu}}, \tag{5}$$

where $\overline{\hat{\mu}} = (\frac{1}{n}\sum_{i=1}^{p}\hat{\mu}_i)(1, \dots, 1)'$ is a column vector with each coordinate equal to the mean of $\hat{\mu}$'s coordinates.

The James-Stein estimator can be directly applied to the OLS estimator of regression coefficients $\hat{\beta}$, after taking into account their covariance matrix $\sigma^2(X'X)^{-1}$ [3, Chapter 7]:

$$\hat{\beta}_{JS1} = \left(1 - \frac{(p-2)}{\hat{\beta}'(\sigma^2(X'X)^{-1})^{-1}\hat{\beta}}\right)\hat{\beta}. \tag{6}$$

If σ^2 is unknown, we can substitute the usual estimate $\hat{\sigma^2} = \frac{r'r}{n-p}$, where r is the vector of residuals. It can be shown that $\hat{\beta}_{JS1}$ has smaller predictive error than the standard OLS estimator [3, Chapter 7]. Adapting the trick given in Equation 5 we get yet another estimator

$$\hat{\beta}_{JS2} = \left(1 - \frac{(p-3)}{(\hat{\beta} - \overline{\hat{\beta}})'(\sigma^2(X'X)^{-1})^{-1}(\hat{\beta} - \overline{\hat{\beta}})}\right)(\hat{\beta} - \overline{\hat{\beta}}) + \overline{\hat{\beta}}, \tag{7}$$

where $\overline{\hat{\beta}}$ is defined analogously to $\overline{\hat{\mu}}$ above. This form will be used to obtain a shrinked uplift regression estimator.

2.2 Shrinkage estimators based on optimizing predictive MSE

In [19] Ohtani gives an overview of another family of shrinkage estimators which improve on the OLS estimator. Their form is similar to the James-Stein estimator, but the shrinkage parameter is now obtained by minimizing the predictive mean squared error. Such estimators were first described in [6]. In our paper we use the following shrinkage factor proposed in [23]

$$\alpha = \frac{\beta'(X'X)\beta}{\beta'(X'X)\beta + \sigma^2}. \tag{8}$$

Notice that it depends on the unknown true coefficient vector β and error variance σ^2. Using the standard practice [19] of substituting OLS estimates $\hat{\beta}$ and $\hat{\sigma^2} = \frac{r'r}{n-p}$ (where r is the residual vector) we obtain an operational estimator

$$\hat{\beta}_{SMSE} = \frac{\hat{\beta}'(X'X)\hat{\beta}}{\hat{\beta}'(X'X)\hat{\beta} + \frac{r'r}{n-p}}\hat{\beta}, \tag{9}$$

where $SMSE$ stands for **S**hrinkage based on minimizing **MSE**. In [12] the MSE of this estimator was computed and sufficient conditions for it to dominate the OLS estimator were provided.

3 Shrinkage estimators for uplift regression

In this section we present the main contribution of this paper: shrinkage estimators for uplift regression. We begin by deriving James-Stein style estimators and later derive versions based on minimizing predictive MSE.

3.1 James-Stein uplift estimators

The most obvious approach to obtaining a shrinked uplift estimator is to use two separate James-Stein estimators in Equation 3, in place of OLS estimators $\hat{\beta}^T$ and $\hat{\beta^C}$. We obtain the following uplift shrinkage estimator

$$\hat{\beta}^U_{JSd} = \hat{\beta}^T_{JS2} - \hat{\beta}^C_{JS2}. \tag{10}$$

The d in the subscript indicates a 'double' model. This approach is fairly trivial and one is bound to ask whether it is possible to obtain a better estimator by directly shrinking the estimator $\hat{\beta}^U_d$ given by 3. To this end we need to estimate the variance of $\hat{\beta}^U_d$ and apply it to Equation 7.

We note that $\hat{\beta}^U_d$ is the difference of two independent random vectors, so the variance of $\hat{\beta}^U_d$ is the sum of variances of $\hat{\beta}^T$ and $\hat{\beta}^C$

$$\operatorname{Var}\hat{\beta}^U_d = {\sigma^T}^2 (X^{T\prime}X^T)^{-1} + {\sigma^C}^2 (X^{C\prime}X^C)^{-1}.$$

Substituting the usual estimators of ${\sigma^T}^2$ and ${\sigma^C}^2$ in the expression above and using it in Equation 7 we obtain following estimator:

$$\hat{\beta}^U_{JS} = \frac{(p-3)}{(\hat{\beta}^U_d - \overline{\hat{\beta}^U_d})' V^{-1} (\hat{\beta}^U_d - \overline{\hat{\beta}^U_d})}(\hat{\beta}^U_d - \overline{\hat{\beta}^U_d}) + \overline{\hat{\beta}^U_d}, \tag{11}$$

where $V = \frac{r^{T\prime}r^T}{n^T-p}(X^{T\prime}X^T)^{-1} + \frac{r^{C\prime}r^C}{n^C-p}(X^{C\prime}X^C)^{-1}$, r^T, r^C are OLS residuals in, respectively, treatment and control groups, and $\overline{\hat{\beta}^U_d} = (\frac{1}{n}\sum_{i=1}^{p}(\hat{\beta}^U_d)_i)(1, \ldots, 1)'$.

3.2 MSE minimizing uplift estimators

We may also adapt the MSE-minimizing variant of shrinkage estimators [19] to the uplift modeling problem. The first approach is to use the shrinkage estimator given in Equation 9 separately for β^T and β^C and construct a double uplift estimator:

$$\hat{\beta}^U_{SMSEd} = \hat{\beta}^T_{SMSE} - \hat{\beta}^C_{SMSE}. \tag{12}$$

The d in the subscript indicates a ‘double’ model.

Another possibility is to estimate α^T, α^C jointly such that the mean squared prediction error is minimized. This is an entirely new method and the main contribution of this paper. Recall from Section 1 that the general shrinked double uplift estimator is

$$\hat{\beta}^U_{\alpha^T,\alpha^C} = \alpha^T(X^{T'}X^T)^{-1}X^{T'}y^T - \alpha^C(X^{C'}X^C)^{-1}X^{C'}y^C,$$

where α^T and α^C are the shrinkage factors. Since there is no explicit value of ‘uplift response’ which can be observed we will define the analogue of the MSE as $\mathrm{E}\,\|X_{test}\beta^U - X_{test}\hat{\beta}^U_{\alpha^T,\alpha^C}\|^2$ where β^U is the true parameter vector, X_{test} is some test data, and the expectation is taken over $\hat{\beta}^U_{\alpha^T,\alpha^C}$. We have

$$\begin{aligned}
&\mathrm{E}\,\|X_{test}\beta^U - X_{test}\hat{\beta}^U_{\alpha^T,\alpha^C}\|^2\\
&= \mathrm{E}\,\mathrm{Tr}\left\{(\beta^U - \hat{\beta}^U_{\alpha^T,\alpha^C})'X'_{test}X_{test}(\beta^U - \hat{\beta}^U_{\alpha^T,\alpha^C})\right\}\\
&= \mathrm{E}\,\mathrm{Tr}\left\{(X'_{test}X_{test})(\beta^U - \hat{\beta}^U_{\alpha^T,\alpha^C})(\beta^U - \hat{\beta}^U_{\alpha^T,\alpha^C})'\right\}\\
&= \mathrm{Tr}\left\{(X'_{test}X_{test})\left(\mathrm{Var}\,\hat{\beta}^U_{\alpha^T,\alpha^C} + (\beta^U - \mathrm{E}\,\hat{\beta}^U_{\alpha^T,\alpha^C})(\beta^U - \mathrm{E}\,\hat{\beta}^U_{\alpha^T,\alpha^C})'\right)\right\}\\
&= (\mathrm{E}\,\hat{\beta}^U_{\alpha^T,\alpha^C} - \beta^U)'(X'_{test}X_{test})(\mathrm{E}\,\hat{\beta}^U_{\alpha^T,\alpha^C} - \beta^U)\\
&\quad + \mathrm{Tr}\left\{(X'_{test}X_{test})\left((\alpha^T)^2\,\mathrm{Var}\,\hat{\beta}^T + (\alpha^C)^2\,\mathrm{Var}\,\hat{\beta}^C\right)\right\},
\end{aligned} \tag{13}$$

where $\mathrm{E}\,\hat{\beta}^U_{\alpha^T,\alpha^C} = \alpha^T\beta^T - \alpha^C\beta^C$, the second equality is obtained by changing the multiplication order within the trace, and the third follows from the bias-variance decomposition. Variance of $\hat{\beta}^U_{\alpha^T,\alpha^C}$ can be decomposed since it is the sum of two independent components. Differentiating with respect to α^T and equating to zero we get

$$\begin{aligned}
0 = {} & 2\alpha^T\beta^{T'}(X'_{test}X_{test})\beta^T - 2\alpha^C\beta^{T'}(X'_{test}X_{test})\beta^C\\
& - 2\beta^{T'}(X'_{test}X_{test})\beta^U + 2\alpha^T\,\mathrm{Tr}((X'_{test}X_{test})\,\mathrm{Var}(\hat{\beta^T})).
\end{aligned}$$

Analogously for α^C we obtain

$$\begin{aligned}
0 = {} & 2\alpha^C\beta^{C'}(X'_{test}X_{test})\beta^C - 2\alpha^T\beta^{T'}(X'_{test}X_{test})\beta^C\\
& + 2\beta^{C'}(X'_{test}X_{test})\beta^U + 2\alpha^C\,\mathrm{Tr}((X'_{test}X_{test})\,\mathrm{Var}(\hat{\beta^C})).
\end{aligned}$$

Denote $W = X'_{test}X_{test}$. We can write the above system of equations for α^T and α^C in matrix form

$$\begin{bmatrix} {\beta^T}'W\beta^T + \operatorname{Tr}(W \operatorname{Var}(\widehat{\beta^T})) & -{\beta^T}'W\beta^C \\ -{\beta^C}'W\beta^T & {\beta^C}'W\beta^C + \operatorname{Tr}(W \operatorname{Var}(\widehat{\beta^C})) \end{bmatrix} \begin{bmatrix} \alpha^T \\ \alpha^C \end{bmatrix} = \begin{bmatrix} {\beta^T}'W\beta^U \\ -{\beta^C}'W\beta^U \end{bmatrix}.$$

Unfortunately we don't know true values of β^T and β^C so we have to replace them with their OLS estimators. Moreover, we cannot also use the test dataset while constructing the estimator. Therefore, in accordance with the fixed X assumption (see Section 1) we take $X_{test} = X$. Finally, we denote $V^T = (\hat{\sigma}^T)^2(X'X)({X^T}'X^T)^{-1}$ and $V^C = (\hat{\sigma}^C)^2(X'X)({X^C}'X^C)^{-1}$ to obtain an operational system of equations

$$\begin{bmatrix} \widehat{\beta^T}'X'X\widehat{\beta^T} + \operatorname{Tr} V^T & -\widehat{\beta^T}'X'X\widehat{\beta^C} \\ -\widehat{\beta^C}'X'X\widehat{\beta^T} & \widehat{\beta^C}'X'X\widehat{\beta^C} + \operatorname{Tr} V^C \end{bmatrix} \begin{bmatrix} \hat{\alpha}^T \\ \hat{\alpha}^C \end{bmatrix} = \begin{bmatrix} \widehat{\beta^T}'X'X\hat{\beta}^U_d \\ -\widehat{\beta^C}'X'X\hat{\beta}^U_d \end{bmatrix}. \quad (14)$$

Finally we are ready to define our shrinkage uplift regression estimator:

Definition 1 *Assume that $\widehat{\beta^T}$ and $\widehat{\beta^C}$ are OLS regression estimators built respectively on the treatment and control groups. Denote by $\widehat{\alpha^T}$ and $\widehat{\alpha^C}$ the solutions to the system of Equations 14. Then the estimator*

$$\hat{\beta}^U_{SMSE} = \widehat{\alpha^T}\widehat{\beta^T} - \widehat{\alpha^C}\widehat{\beta^C}$$

is called the uplift MSE-minimizing estimator*.*

Because the unknown values of β^T, β^C, σ^T, σ^C have been replaced with their estimators, we have no guarantee that $\hat{\beta}^U_{SMSE}$ minimizes the predictive mean squared error. However, under additional assumptions we are able to prove the following theorem.

Theorem 1 *Assume that the matrices X^T and X^C are orthogonal, i.e. ${X^T}'X^T = {X^C}'X^C = I$. Assume further, that the error vectors ε^T, ε^C are independent and normally distributed as $N(0, I)$, i.e. assume that $\sigma^T = \sigma^C = 1$. Then for $p \geqslant 6$*

$$\operatorname{E}\|X\beta^U - X\hat{\beta}^U_{SMSE}\|^2 \leqslant \operatorname{E}\|X\beta^U - X\hat{\beta}^U_d\|^2.$$

The proof of the theorem can be found in the Appendix. Additionally the supplementary material contains a symbolic computation script verifying the more technical sections of the proof.

The theorem says that under orthogonal design the uplift MSE-minimizing estimator given in Definition 1 has a lower expected prediction error than the double estimator given in Equation 3. The requirement for an orthogonal design is restrictive but we were not able to prove the theorem in a more general setting. Even with this assumption, the proof is long and fairly technical. For more general settings we resort to experimental verification in Section 4.

4 Experiments

In this section we present an experimental evaluation of the proposed shrinkage estimators. Before presenting the results we will describe two real life datasets used in the study, as well as the testing methodology we adopted.

4.1 Descriptions of datasets

The first dataset we consider is the well known Lalonde dataset [21] describing the effects of a job training program which addressed a population of low skilled adults. A randomly selected sample of the population was invited to take part in a job training program. Their income in the third year *after* randomization is the target variable. Our goal is to build a model predicting whether the program will be effective for a given individual. There are a total of 185 treatment records and 260 controls.

The second dataset we use is the IHDP dataset [16]. The dataset describes the results of a program whose target groups were low birth weight infants. A randomly selected subset of them received additional support such as home visits and access to a child development center. We want to identify infants for whose IQ (the target variable used in the study) increased *because* of the intervention program. There are 377 treatment and 608 control cases.

4.2 Methodology

The biggest problem in evaluating uplift models is that we never observe y_i^T and y_i^C simultaneously and, thus, do not know the true value of the quantity we want to predict $y_i^T - y_i^C$. Therefore we are forced to make the comparison on larger groups. Here we will estimate the so called Average Treatment Effect on the Treated (ATT) [5,7] using two methods: one based on predictions of a model with coefficients $\hat{\beta}^U$, the other based on true outcomes using a so called difference-in-means estimator [5]. Both quantities are given, respectively, by the following equations

$$ATT_{model}(\hat{\beta}^U) = \frac{1}{n^T} \sum_{i=1}^{n^T} X_i \hat{\beta}^U, \qquad ATT_{means} = \frac{1}{n^T} \sum_{i=1}^{n^T} y_i^T - \frac{1}{n^C} \sum_{i=1}^{n^C} y_i^C.$$

The difference-in-means estimator will play the role of ground truth. We define the absolute error in ATT estimation of a model with coefficients $\hat{\beta}^U$ as

$$ErrATT(\hat{\beta}^U) = |ATT_{model}(\hat{\beta}^U) - ATT_{means}|.$$

Model comparison will be based on their ErrATTs.

Each dataset was split into a training and testing part. Splitting was done separately in the treatment and control groups, 70% of cases assigned to the training set and the remaining cases to the test set. We repeat this procedure 1000 times and aggregate the results. To compare estimators $\hat{\beta}_1^U$ and $\hat{\beta}_2^U$ we will

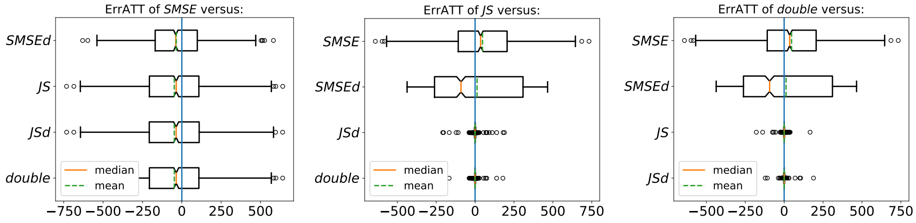

Fig. 1. Differences between errors in ATT estimation for pairs of models for the Lalonde dataset. Each boxplot summarizes the distribution of differences for a pair of models over 1000 train/test splits. For example, the first chart compares the proposed uplift MSE-minimizing shrinkage estimator $\hat{\beta}^U_{SMSE}$ against all other estimators. The mean/median lines on the negative side indicate the model in figure title performs better

compute the difference $ErrATT(\hat{\beta}^U_1) - ErrATT(\hat{\beta}^U_2)$ for each simulation and display the differences using box plots in order to better visualize how often an by what margin each model is better. We found this approach to give more meaningful results than simply comparing mean prediction errors.

4.3 Results

Our experiments involve five different estimators: the double regression estimator $\hat{\beta}^U_d$ given in Equation 3, two double shrinkage estimators: the double James-Stein estimator $\hat{\beta}^U_{JSd}$ and the double MSE minimizing shrinkage estimator $\hat{\beta}^U_{SMSE}$ given, respectively, in Equations 10 and 12, and finally the two direct uplift shrinkage estimators: $\hat{\beta}^U_{JS}$ given in Equation 11 and $\hat{\beta}^U_{SMSEd}$ given in Definition 1. In the figures the estimators are denoted with just their subscripts, e.g. $SMSEd$ instead of $\hat{\beta}^U_{SMSEd}$, except for $\hat{\beta}^U_d$ denoted by *double* for easier readability.

Results on the Lalonde dataset are shown in Figure 1. The first chart on the figure compares the proposed uplift MSE-minimizing estimator will all remaining estimators. It can be seen that the estimator outperforms all others: the original double regression and all three other shrinkage estimators. The improvement can be seen both in the mean and in the median of differences between $ErrATT$'s which are negative. The difference is not huge, but it is consistent, so there is little argument for not using the shrinkage estimator. Moreover, the results are statistically significant (notches in the box plot denote a confidence interval for the median).

The second chart shows the performance of another proposed estimator, the James-Stein version of uplift estimator $\hat{\beta}^U_{JS}$. Here, a different story can be seen. The performance is practically identical to that of the classical double regression and double James-Stein estimator; the boxplots have in fact collapsed at zero. There is an improvement in the median of error difference over $\hat{\beta}^U_{SMSEd}$ but it disappears when one looks at the mean: one cannot expect practical gains from using this estimator.

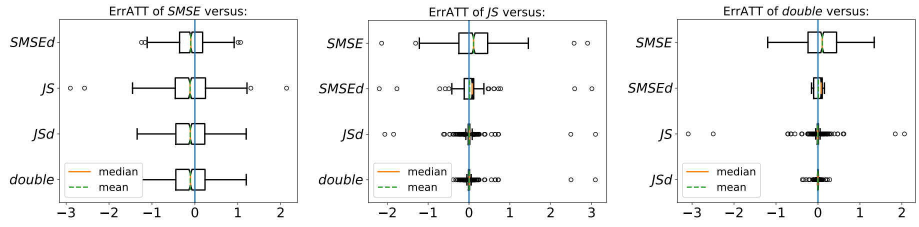

Fig. 2. Differences between errors in ATT estimation for pairs of models for the IHDP dataset. Each boxplot summarizes the distribution of differences for a pair of models over 1000 train/test splits. For example, the first chart compares the proposed uplift MSE-minimizing shrinkage estimator $\hat{\beta}^U_{SMSE}$ against all other estimators. The mean/median lines on the negative side indicate the model in figure title performs better

The two remaining double shrinked estimators performed similarly and charts comparing them to all other models are not shown. For completeness we compare the unshrinked double regression with all shrinkage estimator in the third chart of Figure 1. It can be seen that only the estimator given in Definition 1 dominates it.

The results for the IHDP dataset are shown in Figure 2. All conclusions drawn from the Lalonde dataset are essentially replicated also on IHDP, giving the results more credibility.

5 Conclusions and future work

We have proposed four different shrinkage estimators for uplift regression problem. One of them successfully and consistently reduced prediction error on two real life datasets. The estimator was different from others in that it jointly optimized the treatment and control shrinkage factors such that good uplift predictions are obtained.

The three other estimators did not bring improvement over the classical double regression model. One concludes, that simply applying shrinkage to treatment and control models separately is not enough to obtain a good uplift shrinkage estimator. Neither is applying the James-Stein approach directly the estimated uplift coefficients as is done in the $\hat{\beta}^U_{JS}$ estimator.

Future work will address the problem of adapting shrinkage methods to other uplift regression estimators such as those proposed in [9]. The task is challenging since the finite sample variance of those estimators is not known.

References

1. Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

2. Heumann C., Nittner T., Rao C.R., Scheid S., and Toutenburg H. *Linear Models: Least Squares and Alternatives*. Springer New York, 2013.
3. Bradley Efron and Trevor Hastie. *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Cambridge University Press, New York, NY, USA, 1st edition, 2016.
4. L. Guelman, M. Guillén, and A.M. Pérez-Marín. Random forests for uplift modeling: An insurance customer retention case. In *Modeling and Simulation in Engineering, Economics and Management*, volume 115 of *Lecture Notes in Business Information Processing (LNBIP)*, pages 123–133. Springer, 2012.
5. Imbens G.W. and Rubin D.B. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, New York, NY, USA, 2015.
6. Theil H. *Principles of Econometrics*. John Wiley, New York, 1971.
7. Robins J.M. Hernán M.A. *Causal Inference*. Boca Raton: Chapman & Hall/CRC, 2018. forthcoming.
8. Jennifer L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.
9. Rudas K. Jaroszewicz S. Linear regression for uplift modeling. *Data Mining and Knowledge Discovery*, 32(5):1275–1305, 2018.
10. M. Jaśkowski and S. Jaroszewicz. Uplift modeling for clinical trial data. In *ICML 2012 Workshop on Machine Learning for Clinical Data Analysis*, Edinburgh, June 2012.
11. Kane K., Lo V.S.Y., and Zheng J. Mining for the truly responsive customers and prospects using true-lift modeling: Comparison of new and existing methods. *Journal of Marketing Analytics*, 2(4):218–238, Dec 2014.
12. Ohtani K. Exact small sample properties of an operational variant of the minimum mean squared error estimator. *Communications in Statistics - Theory and Methods*, 25(6):1223–1231, 1996.
13. F. Kuusisto, V. Santos Costa, H. Nassif, E. Burnside, D. Page, and J. Shavlik. Support vector machines for differential prediction. In *ECML-PKDD*, 2014.
14. Zaniewicz Ł. and Jaroszewicz S. l_p-support vector machines for uplift modeling. *Knowledge and Information Systems*, 53(1):269–296, Oct 2017.
15. L.Y-T. Lai. Influential marketing: A new direct marketing strategy addressing the existence of voluntary buyers. Master's thesis, Simon Fraser University, 2006.
16. Brooks-Gunn J. Liaw F., Klebanov P. Effects of early intervention on cognitive function of low birth weight preterm infants. *Journal of Pediatrics,*, 120:350–359, 1991.
17. Aaron Meurer, Christopher P. Smith, Mateusz Paprocki, Ondřej Čertík, Sergey B. Kirpichev, Matthew Rocklin, AMiT Kumar, Sergiu Ivanov, Jason K. Moore, Sartaj Singh, Thilina Rathnayake, Sean Vig, Brian E. Granger, Richard P. Muller, Francesco Bonazzi, Harsh Gupta, Shivam Vats, Fredrik Johansson, Fabian Pedregosa, Matthew J. Curry, Andy R. Terrel, Štěpán Roučka, Ashutosh Saboo, Isuru Fernando, Sumith Kulal, Robert Cimrman, and Anthony Scopatz. Sympy: symbolic computing in python. *PeerJ Computer Science*, 3:e103, January 2017.
18. Efron B. Morris C. Stein's paradox in statistics. *Scientific American*, 236(5):119–127, 1977.
19. Namba A. Ohtani K. MSE performance of the weighted average estimators consisting of shrinkage estimators. *Communications in Statistics - Theory and Methods*, 47(5):1204–1214, 2018.
20. J. Pearl. *Causality*. Cambridge University Press, 2009.

21. Lalonde R. Evaluating the econometric evaluations of training programs. *American Economic Review*, 76:604–620, 1986.
22. N. J. Radcliffe and P. D. Surry. Real-world uplift modelling with significance-based uplift trees. Portrait Technical Report TR-2011-1, Stochastic Solutions, 2011.
23. Farebrother R.W. The minimum mean square error linear estimator and ridge regression. *Technometrics*, 17(1):127–128, 1975.
24. P. Rzepakowski and S. Jaroszewicz. Decision trees for uplift modeling with single and multiple treatments. *Knowledge and Information Systems*, 2011.
25. M. Sołtys, S. Jaroszewicz, and P. Rzepakowski. Ensemble methods for uplift modeling. *Data Mining and Knowledge Discovery*, pages 1–29, 2014. online first.
26. P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search.* MIT Press, 2001.
27. James W. Stein C. Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.*, 1:361–379, 1961.
28. Holland P. W. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, December 1986.
29. Ł. Zaniewicz and S. Jaroszewicz. Support vector machines for uplift modeling. In *The First IEEE ICDM Workshop on Causal Discovery (CD 2013)*, Dallas, December 2013.

A Proof of Theorem 1

Since large parts of the proof require lengthy derivations, the supplementary material contains a Python script which verifies certain equations symbolically using the Sympy [17] package.

From assumptions we have $X^{T'}X^T = X^{C'}X^C = I$, implying $X'X = 2I$. Taking this into account we can simplify the system of equations (14) to

$$\begin{bmatrix} \hat{\beta^T}'\hat{\beta^T} + p & -\hat{\beta^T}'\hat{\beta^C} \\ -\hat{\beta^C}'\hat{\beta^T} & \hat{\beta^C}'\hat{\beta^C} + p \end{bmatrix} \begin{bmatrix} \hat{\alpha}^T \\ \hat{\alpha}^C \end{bmatrix} = \begin{bmatrix} \hat{\beta^T}'\hat{\beta^U} \\ -\hat{\beta^C}'\hat{\beta^U} \end{bmatrix}. \tag{15}$$

Denote $b^{TT} = \hat{\beta^T}'\hat{\beta^T}$, $b^{CC} = \hat{\beta^C}'\hat{\beta^C}$ and $b^{TC} = \hat{\beta^T}'\hat{\beta^C}$. Denoting further $A = \begin{bmatrix} b^{TT} + p & -b^{TC} \\ -b^{TC} & b^{CC} + p \end{bmatrix}$ and $B = \begin{bmatrix} b^{TT} - b^{TC} \\ b^{CC} - b^{TC} \end{bmatrix}$ the system of equations simplifies further to:

$$A \begin{bmatrix} \hat{\alpha^T} \\ \hat{\alpha^C} \end{bmatrix} = B. \tag{16}$$

As a result, denoting $K = \frac{1}{(b^{TT}+p)(b^{CC}+p)-(b^{TC})^2}$, we obtain:

$$\begin{bmatrix} \hat{\alpha^T} \\ \hat{\alpha^C} \end{bmatrix} = A^{-1}B = K \begin{bmatrix} b^{CC} + p & b^{TC} \\ b^{TC} & b^{TT} + p \end{bmatrix} \begin{bmatrix} b^{TT} - b^{TC} \\ b^{CC} - b^{TC} \end{bmatrix}.$$

Equivalently we can write the parameters (verified in the supplementary material) as

$$\hat{\alpha^T} = K\left(b^{CC}b^{TT} - (b^{TC})^2 + pb^{TT} - pb^{TC}\right) = 1 - pK\left(b^{CC} + b^{TC} + p\right)$$
$$\hat{\alpha^C} = K\left(b^{CC}b^{TT} - (b^{TC})^2 + pb^{CC} - pb^{TC}\right) = 1 - pK\left(b^{TT} + b^{TC} + p\right) \quad (17)$$

The predictive MSE of the double model is

$$\mathrm{E}(\hat\beta^U_d - \beta^U)'(\hat\beta^U_d - \beta^U) = 2p. \quad (18)$$

To see this, note that the estimator is unbiased and, under the assumptions of the theorem, $\operatorname{Var}\hat\beta^T = \operatorname{Var}\hat\beta^C = I$. Apply the trace in (13) to get the result. Now we will calculate the predictive MSE of new shrinked and prove that it is less than $2p$. It is easy to see that

$$\mathrm{E}(\hat\beta^U_{SMSE} - \beta^U)'(\hat\beta^U_{SMSE} - \beta^U) = \mathrm{E}(\hat\beta^U_{SMSE} - \hat\beta^U_d)'(\hat\beta^U_{SMSE} - \hat\beta^U_d)$$
$$- \mathrm{E}(\hat\beta^U_d - \beta^U)'(\hat\beta^U_d - \beta^U) + 2\,\mathrm{E}(\hat\beta^U_{SMSE} - \beta^U)'(\hat\beta^U_d - \beta^U). \quad (19)$$

Denote:

$$\hat\beta^T_s = p\left(b^{CC} + b^{TC} + p\right)\hat\beta^T,$$
$$\hat\beta^C_s = p\left(b^{TT} + b^{TC} + p\right)\hat\beta^C.$$

Then the first term of (19) is (verified in the supplementary material)

$$\begin{aligned}
&\mathrm{E}(\hat\beta^U_{SMSE} - \hat\beta^U_d)'(\hat\beta^U_{SMSE} - \hat\beta^U_d)\\
&\quad = \mathrm{E}((\hat\alpha^T - 1)\hat\beta^T - (\hat\alpha^C - 1)\hat\beta^C)'((\hat\alpha^T - 1)\hat\beta^T - (\hat\alpha^C - 1)\hat\beta^C)\\
&\quad = K^2\,\mathrm{E}\left(\hat\beta^T_s - \hat\beta^C_s\right)'\left(\hat\beta^T_s - \hat\beta^C_s\right) \quad (20)\\
&\quad = K^2\,\mathrm{E}\left(p^4(b^{TT} - 2b^{TC} + b^{CC})\right.\\
&\qquad \left. + p^2\left(b^{TT}b^{CC} - b^{TC^2}\right)\left((b^{TT} + 2b^{TC} + b^{CC}) + 4p\right)\right). \quad (21)
\end{aligned}$$

Now we will concentrate on third term of (19):

$$\begin{aligned}
2\,\mathrm{E}(\hat\beta^U_{SMSE} - \beta^U)'(\hat\beta^U_d - \beta^U) &= 2\,\mathrm{E}(\hat{\alpha^T}\hat\beta^T - \beta^T)'(\hat\beta^T - \beta^T) \quad (22)\\
&+ 2\,\mathrm{E}(\hat{\alpha^C}\hat\beta^C - \beta^C)'(\hat\beta^C - \beta^C) \quad (23)\\
&- 2\,\mathrm{E}(\hat{\alpha^T}\hat\beta^T - \beta^T)'(\hat\beta^C - \beta^C) - 2\,\mathrm{E}(\hat\alpha^C\hat\beta^C - \beta^C)'(\hat\beta^T - \beta^T). \quad (24)
\end{aligned}$$

We will first look at $2\,\mathrm{E}(\hat{\alpha^T}\hat\beta^T - \beta^T)'(\hat\beta^T - \beta^T)$ following the classical proof for James-Stein estimator

$$2\,\mathrm{E}(\hat\beta^T - \beta^T)'(\hat\alpha^T\hat\beta^T - \beta^T) = 2\sum_{i=1}^{p}\mathrm{E}(\hat\beta^T_i - \beta^T_i)(\hat\alpha^T\hat\beta^T_i - \beta^T_i)$$
$$= 2\sum_{i=1}^{p}\int ..\int(\hat\beta^T_i - \beta^T_i)(\hat\alpha^T\hat\beta^T_i - \beta^T_i)f(\hat\beta^T)d\hat\beta^T_i, d\hat\beta^T_1 ... d\hat\beta^T_p$$

where $f(\hat{\beta}^T)$ is density function of distribution of $\hat{\beta}^T$, which, by assumptions is multivariate normal $N(\beta^T, I)$. Using integration by parts (derivatives calculated w.r.t. $\hat{\beta}_i$) with

$$u = \hat{\alpha}^T\hat{\beta}_i^T - \beta_i^T \qquad dv = (\hat{\beta}_i^T - \beta_i^T)\frac{\exp\left(-\frac{1}{2}\sum_{j=1}^p\left(\hat{\beta}_j^T - \beta_j^T\right)^2\right)}{2\pi^{\frac{p}{2}}}d\hat{\beta}_i^T$$

$$du = \frac{d}{d\hat{\beta}_i^T}(\hat{\alpha}^T\hat{\beta}_i^T - \beta_i^T)d\hat{\beta}_i^T \qquad v = -\frac{\exp\left(-\frac{1}{2}\sum_{j=1}^p\left(\hat{\beta}_j^T - \beta_j^T\right)^2\right)}{2\pi^{\frac{p}{2}}}$$

we obtain:

$$\int \ldots \int (\hat{\beta}_i^T - \beta_i^T)(\hat{\alpha}^T\hat{\beta}_i^T - \beta_i^T)f(\hat{\beta}^T)d\hat{\beta}_i^T d\hat{\beta}_1^T \ldots d\hat{\beta}_p^T$$

$$= \int \ldots \int (\hat{\beta}_i^T - \beta_i^T)(\hat{\alpha}^T\hat{\beta}_i^T - \beta_i^T)\frac{1}{2\pi^{\frac{p}{2}}}\exp\left(-\frac{1}{2}\sum_{j=1}^p\left(\hat{\beta}_j^T - \beta_j^T\right)^2\right)d\hat{\beta}_i^T d\hat{\beta}_1^T \ldots d\hat{\beta}_p^T$$

$$= \left[\int \ldots \int -\left(\hat{\alpha}^T\hat{\beta}_i^T - \beta_i^T\right)\frac{\exp\left(-\frac{1}{2}\sum_{j=1}^p\left(\hat{\beta}_j^T - \beta_j^T\right)^2\right)}{2\pi^{\frac{p}{2}}}\right]_{-\infty}^{\infty} d\hat{\beta}_1^T \ldots d\hat{\beta}_p^T$$

$$+ \int \ldots \int \frac{d}{d\hat{\beta}_i^T}(\hat{\alpha}^T\hat{\beta}_i^T - \beta_i^T)\frac{\exp\left(-\frac{1}{2}\sum_{j=1}^p\left(\hat{\beta}_j^T - \beta_j^T\right)^2\right)}{2\pi^{\frac{p}{2}}}d\hat{\beta}_i^T d\hat{\beta}_1^T \ldots d\hat{\beta}_p^T.$$

First term in last expression is 0, due to exponential decrease of normal density. Finally we obtain:

$$= \int \ldots \int f(\hat{\beta}^T)\frac{d}{d\hat{\beta}_i^T}(\hat{\alpha}^T\hat{\beta}_i^T - \beta_i^T)d\hat{\beta}_i^T d\hat{\beta}_1^T \ldots d\hat{\beta}_p^T.$$

Repeating the above process for $i = 1, \ldots, p$ we get

$$2\,\mathrm{E}(\hat{\alpha^T}\hat{\beta}^T - \beta^T)'(\hat{\beta}^T - \beta^T) = 2\,\mathrm{E}\sum_{i=1}^p\frac{d}{d\hat{\beta}_i^T}(\hat{\alpha}^T\hat{\beta}_i^T - \beta_i^T) = 2p\hat{\alpha}^T + 2\hat{\beta^T}'\frac{d\hat{\alpha}^T}{d\hat{\beta}^T}.$$

For the third term of (24) we obtain:

$$2\,\mathrm{E}(\hat{\alpha^T}\hat{\beta}^T - \beta^T)'(\hat{\beta}^C - \beta^C) = 2\,\mathrm{E}\sum_{i=1}^p\frac{d}{d\hat{\beta}_i^C}(\hat{\alpha}^T\hat{\beta}_i^T - \beta_i^T) = 2\hat{\beta^T}'\frac{d\hat{\alpha}^T}{d\hat{\beta}^C},$$

where the last factor is the vector derivative of a scalar. The remaining two terms are analogous. Combining the expressions and using the chain rule we obtain:

$$
\begin{aligned}
&\mathrm{E}(\hat{\alpha^T}\hat{\beta}^T-\beta^T)'(\hat{\beta}^T-\beta^T)-\mathrm{E}(\hat{\alpha}^C\hat{\beta}^C-\beta^C)'(\hat{\beta}^T-\beta^T)\\
&=\left[p\mid 0\right]\begin{bmatrix}\hat{\alpha}^T\\ \hat{\alpha}^C\end{bmatrix}+\left[\hat{\beta^T}'\mid -\hat{\beta^C}'\right]\begin{bmatrix}\frac{d\hat{\alpha}^T}{d\hat{\beta}^T}\\ \frac{d\hat{\alpha}^C}{d\hat{\beta}^T}\end{bmatrix}\\
&=\left[p\mid 0\right]\begin{bmatrix}\hat{\alpha}^T\\ \hat{\alpha}^C\end{bmatrix}+\left[\hat{\beta^T}'\mid -\hat{\beta^C}'\right]\begin{bmatrix}\frac{d\hat{\alpha}^T}{d\,b^{TT}}\frac{d\,b^{TT}}{d\hat{\beta^T}}+\frac{d\hat{\alpha}^T}{d\,b^{TC}}\frac{d\,b^{TC}}{d\hat{\beta^T}}+\frac{d\hat{\alpha}^T}{d\,b^{CC}}\frac{d\,b^{CC}}{d\hat{\beta^T}}\\ \frac{d\hat{\alpha}^C}{d\,b^{TT}}\frac{d\,b^{TT}}{d\hat{\beta^T}}+\frac{d\hat{\alpha}^C}{b^{TC}}\frac{d\,b^{TC}}{d\hat{\beta^T}}+\frac{d\hat{\alpha}^C}{d\,b^{CC}}\frac{d\,b^{CC}}{d\hat{\beta^T}}\end{bmatrix}\\
&=\left[p\mid 0\right]\begin{bmatrix}\hat{\alpha}^T\\ \hat{\alpha}^C\end{bmatrix}+2\left[b^{TT}\mid -b^{TC}\right]\begin{bmatrix}\frac{d\hat{\alpha}^T}{d\,b^{TT}}\\ \frac{d\hat{\alpha}^C}{d\,b^{TT}}\end{bmatrix}+\left[b^{TC}\mid -b^{CC}\right]\begin{bmatrix}\frac{d\hat{\alpha}^T}{d\,b^{TC}}\\ \frac{d\hat{\alpha}^C}{d\,b^{TC}}\end{bmatrix}\\
&=\left[p\mid 0\right]A^{-1}B+2\left[b^{TT}\mid -b^{TC}\right]\frac{dA^{-1}B}{d\,b^{TT}}+\left[b^{TC}\mid -b^{CC}\right]\frac{dA^{-1}B}{d\,b^{TC}}. \qquad (25)
\end{aligned}
$$

Analogously we obtain:

$$
\begin{aligned}
&\mathrm{E}(\hat{\alpha^C}\hat{\beta}^C-\beta^C)'(\hat{\beta}^C-\beta^C)-\mathrm{E}(\hat{\alpha}^T\hat{\beta}^T-\beta^T)'(\hat{\beta}^C-\beta^C)\\
&=\left[0\mid p\right]A^{-1}B+2\left[-b^{TC}\mid b^{CC}\right]\frac{dA^{-1}B}{d\,b^{CC}}+\left[-b^{TT}\mid b^{TC}\right]\frac{dA^{-1}B}{d\,b^{TC}}. \qquad (26)
\end{aligned}
$$

Combining (25) and (26) we obtain (verified in the supplementary material):

$$
\begin{aligned}
&2\,\mathrm{E}(\hat{\beta}^U_{SMSE}-\beta^U)'(\hat{\beta}^U_d-\beta^U)=2\left[p\mid p\right]A^{-1}B\\
&-2K\begin{bmatrix}3b^{TT}b^{CC}-3\left(b^{TC}\right)^2+2b^{TT}p+b^{CC}p-b^{TC}p\\ 3b^{TT}b^{CC}-3\left(b^{TC}\right)^2+2b^{CC}p+b^{TT}p-b^{TC}p\end{bmatrix}\left(A^{-1}B-\mathbf{1}\right) \qquad (27)\\
&=2p(\hat{\alpha^T}+\hat{\alpha^C})-2(2\hat{\alpha^T}+1)\hat{\alpha^T}-2(2\hat{\alpha^C}+1)\hat{\alpha^C}-4Kp^2\\
&=4p+2(p-3)(\hat{\alpha^T}-1+\hat{\alpha^C}-1)-4Kp^2-4(\hat{\alpha^T}-1)^2-4(\hat{\alpha^C}-1)^2, \qquad (28)
\end{aligned}
$$

where $\mathbf{1}=(1,1)'$. Now we have calculated each term of (19). Now we will combine (28) and (18).

$$
\begin{aligned}
&2\,\mathrm{E}(\hat{\beta}^U_{SMSE}-\beta^U)'(\hat{\beta}^U_d-\beta^U)-\mathrm{E}(\hat{\beta}^U_d-\beta^U)'(\hat{\beta}^U_d-\beta^U)\\
&\quad=2p-\mathrm{E}\,K^2\left(\frac{1}{K}\left(2p(p-3)\left(b^{TT}+2b^{TC}+b^{CC}\right)-4p^2(p-2)\right)\right.\\
&\qquad\left.-4\left(\left(pb^{TT}+pb^{TC}+p^2\right)^2+\left(pb^{CC}+pb^{TC}+p^2\right)^2\right)\right). \qquad (29)
\end{aligned}
$$

Combining further (29) with (21) we obtain the following expression for (19) (verified in the supplementary material):

$$\begin{aligned}
&\mathrm{E}(\hat{\beta}^U_{SMSE}-\beta^U)'(\hat{\beta}^U_{SMSE}-\beta^U) = 2p - && (30)\\
&\mathrm{E}\,K^2\Big(p(p-6)(b^{TT}+2b^{TC}+b^{CC})\left(b^{TT}b^{CC}-b^{TC^2}\right) && (31)\\
&\quad +2p^2(p-3)(b^{TT}+2b^{TC}+b^{CC})\left(b^{TT}+b^{CC}\right) && (32)\\
&\quad +2p^3(p-3)(b^{TT}+2b^{TC}+b^{CC}) && (33)\\
&\quad -8p^2\left(b^{TT}b^{CC}-b^{TC^2}\right) && (34)\\
&\quad +4p^3(p-2)\left(b^{TT}+b^{CC}+p\right) && (35)\\
&\quad +4p^2\left(b^{TT}+b^{TC}+p\right)^2+4p^2\left(b^{CC}+b^{TC}+p\right)^2 && (36)\\
&\quad -p^4(b^{TT}-2b^{TC}+b^{CC})\Big). && (37)
\end{aligned}$$

We see that (31) is greater than or equal to 0 when $p \geqslant 6$ and (33) when $p \geqslant 3$. Now combining (35) and (37) for $p \geqslant 4$ we get:

$$\begin{aligned}
&4p^3(p-2)\left(b^{TT}+b^{CC}+p\right)-p^4(b^{TT}-2b^{TC}+b^{CC})\\
&=2p^3(p-4)\left(b^{TT}+b^{CC}+p\right)+2p^4\left(b^{TT}+b^{CC}+p\right)-p^4(b^{TT}-2b^{TC}+b^{CC})\\
&\geqslant 2p^3(p-4)\left(b^{TT}+b^{CC}+p\right)+2p^4\left(b^{TT}+b^{CC}+p\right)-2p^4(b^{TT}+b^{CC})\\
&=2p^3(p-4)\left(b^{TT}+b^{CC}+p\right)+2p^5>0,
\end{aligned}$$

where the first inequality follows from $b^{TT}-2b^{TC}+b^{CC} \leqslant 2(b^{TT}+b^{CC})$.

Now, combining (32), (34) and (36), we obtain (verified in the supplementary material):

$$\begin{aligned}
&-K^2\Big(2p^2(p-3)(b^{TT}+2b^{TC}+b^{CC})\left(b^{TT}+b^{CC}\right)-8p^2(b^{TT}b^{CC}-b^{TC^2})\\
&\qquad +4p^2\left(b^{TT}+b^{TC}+p\right)^2+4p^2\left(b^{CC}+b^{TC}+p\right)^2\Big)\\
&=K^2\Big(-2p^2(p-5)(b^{TT}+2b^{TC}+b^{CC})\left(b^{TT}+b^{CC}\right) && (38)\\
&\qquad -\left(8p^2(b^{TT}+b^{TC})^2+8p^2(b^{CC}+b^{TC})^2\right) && (39)\\
&\qquad -\left(8p^3(b^{TT}+2b^{TC}+b^{CC})+p\right)\Big) && (40)
\end{aligned}$$

We can observe that (39) and (40) are always non positive and (38) is negative when $p > 5$. So the whole expression above is also negative when $p > 5$. So, the only positive term in (30)–(37) is $2p$ proving that predictive error of the shrinked estimator is lower than that of double regression (equal to $2p$) for $p \geqslant 6$.