Regularization for uplift regression

Krzysztof Rudaś[1,2][0000−0002−8309−9952] ✉ and Szymon Jaroszewicz[1,2][0000−0001−9327−5019]

[1] Institute of Computer Science, Polish Academy of Sciences
[2] Faculty of Mathematics and Information Science, Warsaw University of Technology

Abstract. We address the problem of regularization of linear regression models in uplift modeling and heterogeneous treatment effect estimation. We consider interaction models which are commonly used by statisticians in medicine and social sciences to estimate the causal effect of a treatment, and introduce a new type of such a model. We demonstrate the equivalence of all interaction models when no regularization is present, and that this is no longer the case when the model is regularized. Interaction terms introduce implicit correlations between treatment and control coefficients into the regularizer, a fact which has not been previously noted. The correlations depend on the type of interaction model, and by interpreting the regularizer as a prior distribution we were able to pinpoint cases when a given regularized interaction model is most appropriate. An interesting property of the proposed new interaction type is that it allows for smooth interpolation between two types of uplift regression models: the double model and the transformed target model. Our results are valid for both ridge (L_2) and Lasso (L_1) regularization. Experiments on synthetic data fully confirm our analyses. We also compare the usefulness of various regularization schemes on real data.

Keywords: uplift modeling · heterogeneous treatment effect · regularization · linear models · Lasso.

1 Introduction

Uplift modeling is a method of selecting targets for an action, such as a marketing campaign or a medical treatment. To clarify the problem, consider the following example. We administer a factory which produces a certain kind of product i.e. skis. In order to increase our income we send discounts to potential customers. Consider three kinds of customers. The first kind decides to buy skis, *because* they received a discount (without the discount they wouldn't have bought). The second kind decides to not buy skis and sending the discount had no effect. Customers of the third kind bought the skis but would have bought even without the discount. For us it is profitable to send the discount only to the customers of the first kind, but not to the second (no profits) and especially not to the third (lost income due to sale at a lower price). A typical approach to solving the problem of choosing appropriate targets for an action is to predict results after conducting a pilot campaign on a sample of the customers. If the predicted income is above a given threshold, the observation is classified as suitable for the action. However this approach in not correct because it doesn't take into consideration the counterfactual response in case the action would not have been taken. The three groups of customers cannot be distinguished.

In uplift modeling our goal is to predict, for the i-th observation, the difference of responses y_i^T when action was taken on it, and y_i^C when the action was not taken. Unfortunately we cannot directly compare those two outcomes, because we observe only one of them. This is known as the *Fundamental Problem of Causal Inference* [10].

Uplift modeling offers a solution of this problem. In this method we divide our population into two groups: treatment on which the action is taken, and control, which is not subjected to the action. Thanks to this we may decompose the effect observed in the treatment group into two parts. The first is the background (control) outcome and the second is the influence of the action which is only observed in the treatment group. Using this decomposition we may construct a model which will estimate the true effect of an action on an individual.

1.1 Related work

Uplift modeling is a part of a broader problem of causal discovery, concentrating not on predicting future responses, but on effects of interventions, which may be dependent on the values of other variables [23]. Causal discovery has two major branches. The first uses purely observational data [23, 30]. In the second, the action being analyzed has to be actively applied to a subgroup of individuals. Those methods have many applications in social science and medicine [12].

Methods presented in this paper are relevant to the second approach. Most research in this area focuses on cases when the treatment group is not selected randomly, i.e. the treatment assignment mechanism is biased [12, 7]. Those methods typically come under the name of Heterogeneous Treatment Effect estimation [1, 9]. Unfortunately those approaches (e.g. propensity score matching or weighting) are based on untestable assumptions like 'no unmeasured confounders'. The main focus of those methods is to correct the assignment bias not on the estimation problem itself. Uplift modeling, in contrast, concentrates on finding the best possible estimator under random assignment assumptions, which guarantee that the causal effect of the action is identified correctly.

Most of publications on uplift modeling concentrate on the classification problem. First works were based on decision trees [29, 25]. They modified splitting criteria in order to maximize difference in responses between two groups. Similar methods have been invented under the name of estimating heterogeneous treatment effect [1, 9]. Several publications use modified response variable [13, 14, 18] with linear models with such as logistic regression or Support Vector Machines [17, 32, 31]. Estimators for regression problem where analyzed in [27], where basic double regression approach is confronted with some new ideas. Another way of improving on double regression is using shrinkage estimators such as those proposed in [28].

Regression models with interaction terms have been used for causal prediction for decades, see [12, Chap. 7.6] or [7, Chap. 15]. The majority of works use treatment interaction models described in the next section.

There are currently few works devoted specifically to the problem of regularization in uplift modeling or heterogeneous treatment effect estimation. The main textbook on causal effect modeling [12] only discusses the Lasso method for variable selection on one page, and another [7] mentions the term 'regularization' only twice.

There are a few papers which introduce regularized uplift models but do not thoroughly analyze the problem. Imai et al. [11] proposed an SVM model for treatment effect estimation which used the Lasso penalty. This is in fact a variant of a regularized treatment interaction model, frequently used in literature. In [5] a Lasso style model for uplift regression has been introduced, inspired by multitask learning. The proposed model is similar to the models analyzed in this work but includes interaction terms for both treatment and control making it overparametrized which may lead to estimation problems. A similar approach called Shared Data Representation was presented in [3].

The problem of regularization has been addressed by several authors working on nonrandomized treatment assignment. In [22] fused lasso was applied to regularize propensity scores. Hahn et al. [6] discussed pitfalls of regularizing causal models under non-random treatment assignment. Chernozhukov [2] addressed the problem of variable selection for instrumental variables and confounding controls. The goals of those works are different from ours, since we focus on predictive accuracy in the case of randomized treatment assignment.

1.2 Notation

In the text, lowercase Latin and Greek letters denote vectors, uppercase letters: matrices. Let $'$ denote matrix transpose, I_p a $p \times p$ identity matrix, 0 the matrix of zeros of appropriate size, and $\otimes$ the Kronecker product of matrices. All vectors will be assumed to be column vectors, except the feature vectors, denoted with letter x, assumed to be row vectors.

We assume to have a training set of n samples, with i-th sample being a triple (x_i, y_i, t_i), where $x_i \in \mathbb{R}^p$ is a p-dimensional feature vector, $y_i \in \mathbb{R}$ the response, and $t_i \in \{0, 1\}$ the treatment indicator, where $t_i = 1$ means that the i-th case is in the experimental group (was subjected to the action) and $t_i = 0$ indicates a control case.

Quantities related to the treatment group will be denoted with superscript T and to the control group with superscript C. The superscript U will indicate quantities related to the estimated uplift, i.e. the effect of the action. For example, n^T (n^C) denotes the number of cases in the treatment (control) group. We will make the usual assumptions taken when working with linear models, namely, that the treatment and control responses are linear functions of the predictors [8]

$$y_i = \begin{cases} x_i\beta^C + \varepsilon_i & \text{if } t_i = 0 \\ x_i\beta^T + \varepsilon_i = x_i\beta^C + x_i\beta^U + \varepsilon_i & \text{if } t_i = 1, \end{cases} \tag{1}$$

where β^T and β^C are the true coefficient vectors for treatment and control cases, and ε_i are independent random error terms with equal variances and $\mathrm{E}\,\varepsilon_i = 0$.

Notice that for treatment cases the response is the sum of control response $x_i\beta^C$ and the effect of the action $x_i\beta^U$. Clearly, $\beta^U = \beta^T - \beta^C$ is the parameter of interest we want to estimate. We also introduce a vector $\beta^S = \beta^T + \beta^C$ such that $\frac{1}{2}x_i\beta^S$ is the average of treatment and control responses for a case x_i.

Finally, let us introduce a matrix $X^T \in \mathbb{R}^{n^T \times p}$ whose rows are feature vectors of treatment cases, and a vector $y^T \in \mathbb{R}^{n^T}$ of corresponding responses. For the control group, X^C and y^C are defined analogously. The pairs (X^T, y^T), (X^C, y^C) can be interpreted as two separate treatment and control training sets.

2 Linear models of causal influence

In this section we describe basic types of linear models used to estimate causal effects and demonstrate their equivalence. Here we assume that the models do not use regularization, which will be discussed in the next section.

2.1 The double model

The most common approach to uplift regression is the so called *double model* [27], denoted as D. The model is also known as the T-learner [16]. To estimate β^U, the model simply subtracts the coefficient vectors estimated separately on the treatment and control samples: $\hat{\beta^U} = \hat{\beta^T} - \hat{\beta^C}$, where both sub-estimators are obtained by minimizing some loss function ℓ, such as square loss:

$$\hat{\beta^T} = \arg\min_{\beta} \sum_{i=1}^{n^T} \ell(y_i^T, x_i^T\beta), \quad \hat{\beta^C} = \arg\min_{\beta} \sum_{i=1}^{n^C} \ell(y_i^C, x_i^C\beta). \tag{2}$$

Let us rewrite the double model as a single regression model. Define an $n \times 2p$ matrix $\tilde{X}$ and coefficient vector $\tilde{\beta} \in \mathbb{R}^{2p}$ as

$$\tilde{X} = \begin{bmatrix} X^T & 0 \\ 0 & X^C \end{bmatrix}, \qquad \tilde{\beta} = \begin{bmatrix} \beta^T \\ \beta^C \end{bmatrix}, \tag{3}$$

and let $\tilde{x}_i$ denote the i-th row of X. It is easy to see that estimating $\tilde{\beta}$ by minimizing

$$\sum_{i=1}^{n} \ell(y_i, \tilde{x}_i\tilde{\beta}) \tag{4}$$

is equivalent to Equation 2.

2.2 Interaction models

In medicine and social sciences casual effects are often estimated using so called *interaction models*. A single regression model is build on combined treatment and control data. The model includes a special interaction term which allows for estimating the causal effect's coefficients β^U. We now discuss several such models.

Treatment interaction model (TI). The most common approach [11] is to use an interaction between treatment indicator and all predictor variables, resulting in a model based on the following assumption

$$y_i = t_i x_i \beta^U + x_i \beta^C + \varepsilon_i. \tag{5}$$

The coefficient β^C describes the responses in the control group. Since $x_i\beta^C$ is also present in the treatment group, $x_i\beta^U$ has to represent the effect of the treatment. We call this model the *treatment interaction model* because the interaction involves the treatment indicator. Later in the text the model will be denoted with abbreviation TI.

It is easy to see that the model can be represented with a single regression model whose design matrix and coefficient vector are

$$\begin{bmatrix} X^T & X^T \\ 0 & X^C \end{bmatrix}, \quad \begin{bmatrix} \beta^U \\ \beta^C \end{bmatrix}, \tag{6}$$

respectively. While this is the most common interaction model, other approaches are also possible.

Symmetric interaction model (SI). Let us now introduce another interaction model which is one of the contributions of this paper

$$y_i = \left(t_i - \tfrac{1}{2}\right) x_i \beta^U + \tfrac{1}{2} x_i \beta^S + \varepsilon_i. \tag{7}$$

The model uses so called effect or deviation coding of the categorical treatment variable, see [8, Section 10.8] or [4, Section 2.3.2]. The interpretation is that $\frac{1}{2}x_i\beta^S$ is the average of treated and control outcomes for case x_i, and $\pm\frac{1}{2}x_i\beta^U$ is the difference from the mean for control/treatment response. The design matrix and coefficient vector for the corresponding single regression model are

$$\tfrac{1}{2}\begin{bmatrix} X^T & X^T \\ -X^C & X^C \end{bmatrix}, \quad \begin{bmatrix} \beta^U \\ \beta^S \end{bmatrix}. \tag{8}$$

Some advantages of this model, such as lack of correlations in the prior and a relationship with a model based on target variable transformation will be discussed in the following sections.

The model is called the *symmetric interaction model* since the indicators for treatment and control groups are treated in a symmetric fashion.

Control interaction model (CI). For completeness we also introduce a model with interaction between the control group indicator and x's, although we have never seen this model used in literature:

$$y_i = \beta^T x_i - (1 - t_i)\beta^U x_i + \varepsilon_i. \tag{9}$$

Here we estimate the treatment response for all cases, and correct for the strength of causal influence in the control group.

All proposed models are summarized in Table 1. The first row displays the models' names and abbreviations. The following rows provide the models' formulas, design matrices and coefficient vectors. The remaining rows will be explained in the next section.

Model	Double (D)	Treatment Interaction (TI)	Symmetric Interaction (SI)	Control Interaction (CI)
Form	$t_i\beta^T x_i + (1-t_i)\beta^C x_i$	$t_i\beta^U x_i + \beta^C x_i$	$\left(t_i - \frac{1}{2}\right)x_i\beta^U + \frac{1}{2}x_i\beta^S$	$-(1-t_i)\beta^U x + \beta^T x$
Design matrix	$\begin{bmatrix} X^T & 0 \\ 0 & X^C \end{bmatrix}, \begin{bmatrix} \beta^T \\ \beta^C \end{bmatrix}$	$\begin{bmatrix} X^T & X^T \\ 0 & X^C \end{bmatrix}, \begin{bmatrix} \beta^U \\ \beta^C \end{bmatrix}$	$\frac{1}{2}\begin{bmatrix} X^T & X^T \\ -X^C & X^C \end{bmatrix}, \begin{bmatrix} \beta^U \\ \beta^S \end{bmatrix}$	$\begin{bmatrix} 0 & X^T \\ -X^C & X^C \end{bmatrix}, \begin{bmatrix} \beta^U \\ \beta^T \end{bmatrix}$
Matrix $\tilde{A}$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$	$\sqrt{2}\begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{bmatrix}$	$\begin{bmatrix} 0 & 1 \\ -1 & 1 \end{bmatrix}$
Regularizer	$\lambda_1\lVert\beta^T\rVert + \lambda_2\lVert\beta^C\rVert$	$\lambda_1\lVert\beta^U\rVert + \lambda_2\lVert\beta^C\rVert$	$\lambda_1\lVert\beta^U\rVert + \lambda_2\lVert\beta^S\rVert$	$\lambda_1\lVert\beta^U\rVert + \lambda_2\lVert\beta^T\rVert$

Table 1: Summary of linear interaction models analyzed in the paper

2.3 Unified representation of interaction models

In this section we aim to unify all interaction models and demonstrate their equivalence. The theorem below shows that when no regularization is present all interaction models are in fact statistically equivalent with the double model and, as a consequence, with each other.

Theorem 1. *There is a one-to-one mapping between treatment interaction models and double models such that the corresponding models have identical values of the training set losses and provide the same estimates of β^U. An analogous result holds for symmetric and control interaction models.*

Proof. Recall that the double model (Equation 2) can be recast as a single model (Equation 2) trained on the matrix $\tilde{X}$ given in Equation 3, leading to an optimization problem given in Equation 4. For any nonsingular $2p \times 2p$ matrix A we have

$$\sum_{i=1}^{n} \ell(y_i, \tilde{x}_i\tilde{\beta}) = \sum_{i=1}^{n} \ell\left(y_i, (\tilde{x}_i A)(A^{-1}\tilde{\beta})\right). \tag{10}$$

So, for a given double model, multiplying the feature vectors and the coefficient vector respectively by A and A^{-1} does not change the predicted value and thus yields a model with the same empirical risk. This is a direct consequence of the so called affine equivariance of classic least squares linear models [26, p. 116].

Take A to be

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \otimes I_p = \begin{bmatrix} I_p & I_p \\ 0_p & I_p \end{bmatrix},$$

and apply Equation 10 to each row of the design matrix. We have

$$\tilde{X}A = \begin{bmatrix} X^T & 0 \\ 0 & X^C \end{bmatrix} \begin{bmatrix} I_p & I_p \\ 0_p & I_p \end{bmatrix} = \begin{bmatrix} X^T & X^T \\ 0 & X^C \end{bmatrix},$$

and after left-multiplying $\tilde{\beta}$ by A^{-1}

$$A^{-1}\tilde{\beta} = \left(\begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \otimes I_p\right) \begin{bmatrix} \beta^T \\ \beta^C \end{bmatrix} = \begin{bmatrix} \beta^T - \beta^C \\ \beta^C \end{bmatrix} = \begin{bmatrix} \beta^U \\ \beta^C \end{bmatrix},$$

which is the design matrix and coefficient vector defining the treatment interaction model (Equation 6). Thus, the matrix A defines a linear mapping between the double and treatment interaction models such that the corresponding models have the same empirical risk. The fact that the correspondence is one-to-one follows from nonsingularity of matrix A. As a result, both types of models lead to the same empirical risk minimizer and the same estimate of β^U.

To obtain an analogous mapping for symmetric interaction model use the matrix $A = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{bmatrix} \otimes I_p$, and for the control interaction model, the matrix $A = \begin{bmatrix} 0 & 1 \\ -1 & 1 \end{bmatrix} \otimes I_p$.

To show the equivalence between two types of interaction models, consider mapping models of the first type to the double model and than to the interaction model of the second type. Equivalence follows from the fact that the composition of one-to-one mappings is one-to-one. □

A consequence of the theorem is that, when no regularization is used, all interaction models and the double model are essentially equivalent from the statistical perspective: they provide identical estimates and the same predictions on future data. A generic estimation procedure for interaction models can thus be implemented conceptually as follows

1. Form the matrix $\tilde{X}$
2. Compute the matrix $\tilde{X}A$
3. Obtain an estimate $\hat{\beta}$ based on $\tilde{X}A$ and y
4. Compute $\hat{\beta^U} = [I_p| - I_p]A^{-1}\hat{\beta}$

In the last step above, we first transform $\hat{\beta}$ into $(\hat{\beta^T}, \hat{\beta^C})$ and then multiply it by $[I_p| - I_p]$ to obtain $\hat{\beta^U} = \hat{\beta^T} - \hat{\beta^C}$.

Notice that all transformation matrices used in the proof have the form

$$A = \tilde{A} \otimes I_p$$

for some 2×2 matrix $\tilde{A}$. The third row of Table 1 lists the matrices $\tilde{A}$ for all considered interaction models. The matrix for the symmetric interaction model has an additional $\sqrt{2}$ factor. This factor cancels out in Equation 10 and in step 4 of the above procedure, so it will not affect the final estimate. The reason for its introduction is explained in the next section.

3 Regularized interaction models

In the previous section we showed that all unregularized interaction models are equivalent. We will now show that when regularization is present, this will no longer be the case. Our analysis will be valid for all regularizers based on L_q norms raised to the power q, but later we will focus on L_1 and L_2 norms.

The most obvious way to regularize the double model is to separately regularize the estimators for β^T and β^C thus minimizing the following cost function

$$\sum_{i=1}^{n} \ell(y, \tilde{x}_i\tilde{\beta}) + \lambda_1\|\beta^T\|_q^q + \lambda_2\|\beta^C\|_q^q = \sum_{i=1}^{n} \ell(y, \tilde{x}_i\tilde{\beta}) + \left\| \begin{bmatrix} \sqrt[q]{\lambda_1} & 0 \\ 0 & \sqrt[q]{\lambda_2} \end{bmatrix} \tilde{\beta} \right\|_q^q . \quad (11)$$

For the interaction models, all approaches in literature apply regularization directly to coefficient vectors present in the model. The fourth row of Table 1 lists the form of the regularizer for each type of model considered in the paper.

For example, in the treatment interaction model we separately regularize β^U and β^C. This scheme looks appealing since we directly regularize the quantity of interest, which is β^U. However, this type of regularization introduces unexpected interactions between the two regularized vectors. For example, letting $\lambda_2 \to \infty$ does not just influence the estimate of β^C. The estimate of β^U is also affected: regularization will force $\beta^C \to 0$ and, as a result, β^U will tend towards β^T.

We will now analyze those issues further and provide guidelines on the scenarios where different types of regularized interaction models are most useful.

As we have seen above in Equation 10, every interaction model can be expressed as a linear transformation of the double model with an appropriately chosen nonsingular matrix A. Using this fact and Equation 11, every regularized interaction model can be expressed as

$$\sum_{i=1}^{n} \ell\left(y, (\tilde{x}_i A)(A^{-1}\tilde{\beta})\right) + \left\|\varLambda_q A^{-1}\tilde{\beta}\right\|_q^q , \quad (12)$$

where $\varLambda_q = \begin{bmatrix} \sqrt[q]{\lambda_1} & 0 \\ 0 & \sqrt[q]{\lambda_2} \end{bmatrix}$. Here the regularization is applied to the transformed coefficient vector $A^{-1}\tilde{\beta}$. Since A^{-1} does not cancel within the regularization term, the model is no longer invariant under linear transformations, and therefore different interaction models will lead to different regularization terms.

The equation demonstrates one of the main claims of the paper: regularized interaction models are equivalent to the double model regularized with a penalty based on a linear transformation of a unit sphere determined by the type of interaction model used.

Indeed, let us now analyze the generic regularizer by looking at the shape of the contours of its regularization regions. The contours are the sets of points

$$\left\{\tilde{\beta} : \left\|\varLambda_q A^{-1}\tilde{\beta}\right\|_q = r\right\}, \quad (13)$$

where $r > 0$ is a positive constant defining the contour. Substituting $\beta = \varLambda_q A^{-1}\tilde{\beta}$ the contour equation becomes

$$\left\{A\varLambda_q^{-1}\beta : \|\beta\|_q = r\right\}. \quad (14)$$

Therefore the contour is a linear transform of an L_q norm sphere of radius r. The shape of the contour will depend on the transformation matrix $A\varLambda_q^{-1}$.

Before providing a detailed analysis of the regularizers, let us first address the question of scaling of the regularization parameters λ_1, λ_2. It would be desirable if the same values of those parameters led to regularization regions of identical size regardless of the type of interaction model used. Here, we chose to measure the size of regularization contours by the volume they enclose. Let $V_q(r,p)$ be the volume of an L_q norm p-dimensional sphere of radius r. Since the regularization regions are linear transformations of such spheres their volume is

$$\left|\det\left(A\Lambda_q^{-1}\right)\right| V_q(r,p) = |\det(A)| \det\left(\Lambda_q^{-1}\right) V_q(r,p). \tag{15}$$

The equation follows since the Jacobian matrix of a linear transformation is constant. Notice that the type of interaction only affects the matrix A which has the form $\tilde{A} \otimes I_p$, and whose determinant is $\det^p(\tilde{A}) \det(I_p)^2 = \det^p(\tilde{A})$ [24]. Notice that $|\det(\tilde{A})| = 1$ for all matrices $\tilde{A}$ given in Table 1, so the volume of the regularization regions will not depend on the type of interaction model used, only on the values of λ_1, λ_2. To ensure this property, an additional $\sqrt{2}$ factor was added to the symmetric interaction model's design matrix.

3.1 Interpretation of regularized interaction models

In order to give an intuition and visualize those contours we restrict ourselves to the one variable case $p = 1$. The two coefficient vectors now become scalars which can be visualized on a two dimensional plot. Figure 1 shows regularization regions for $r = 1$ (unit sphere being transformed) and selected values of λ_1 and λ_2 parameters for the four types of models given in Table 1. The corresponding figure for the L_2 norm is given in the supplementary materials [3]: it gives the same overall picture with polygons replaced by ellipses. Supplementary materials also include an illustrative figure with superimposed regions for different methods.

The main axes of the plots correspond to coefficients β^T and β^C. Additionally we introduce two more diagonal axes corresponding to β^U and β^S respectively, such that it is possible to see how the parameter of interest β^U is regularized. It can be seen (supplementary material) the for the L_2 norm, the regularization regions are ellipsoids whose main axes do not necessarily align with the main axes of the plot. For the L_1 norm the shapes are analogues of ellipsoids in that norm.

Equivalently, we can view the regularizers from a Bayesian perspective as prior distributions. For the L_2 norm the prior will be Gaussian but with a non-spherical covariance matrix; that is we assume a-priori, that parameter vectors are correlated. In other words we assume some combinations of values of parameters vectors to be more likely than others. For the L_1 norm the prior is a form of multivariate Laplace distribution which, to the best of our knowledge, has not been analyzed in literature.[4] Nevertheless, correlation patterns are clearly visible. Let us now discuss the priors of the four types of regularized models.

[3] https://github.com/RudasKAP/ECML_PKDD_2023_supplementary

[4] The most popular definition of the multivariate Laplace distribution is based on the square root of a quadratic form, see e.g. [15].

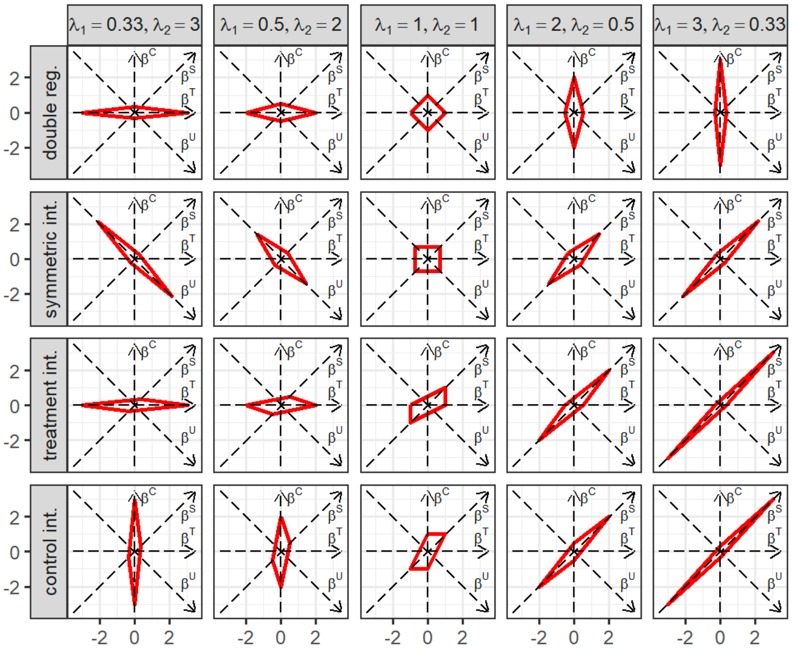

Fig. 1: Regularization regions in L_1 norm, for different types of estimators and different parameters λ_1, λ_2 for $p = 1$

First, it can be seen that the regularized double model does not assume a-priori correlation between β^T and β^C. More interestingly the symmetric interaction model does not assume correlation between β^U and β^S. We believe this property to be important in practice, since the parameter of interest β^U is not affected by the other regularization term. In other words the average response β^S can be regularized with arbitrary strength without affecting β^U. Out of all four models, this is the *only one* possessing this property.

On the other hand the TI and CI models assume a-priori correlations between the true uplift β^U and other coefficients. For example, in the TI model β^U is assumed to be positively correlated with β^T and β^S.

3.2 Applicability scenarios for regularized interaction models

Let us now examine scenarios in which various types of regularized interaction models are likely to yield the most accurate predictive models. We confirm those arguments experimentally in the next section.

We assume that the regularization (or equivalently prior distribution) gives the best results if it corresponds to the true values of the estimated parameters. Table 2 lists several such scenarios and indicates which models are appropriate for them. The third column gives the condition describing the region to which

			Estimator			
	Scenario	Condition	D	SI	TI	CI
1.	$\beta^S \approx 0$; β^T, β^C large	$\beta^T \approx -\beta^C$		✓		
2.	$\beta^T \approx 0$; β^U, β^C large	$\beta^U \approx -\beta^C$	✓			✓
3.	$\beta^C \approx 0$; β^U, β^T large	$\beta^U \approx \beta^T$	✓		✓	
4.	$\beta^U \approx 0$; β^T, β^C large	$\beta^T \approx \beta^C$		✓	✓	✓

Table 2: Scenarios in which different interaction models match the true coefficients

the parameters belong. For example, in Scenario 1, the average of treatment and control responses for a given feature vector x is close to zero, while treatment and control responses are relatively large. This implies $\beta^T \approx -\beta^C$ and the true parameters lie in the upper left and lower right corners in the plots in Figure 1. Looking at the figure it can be seen that only the symmetric interaction model (SI) is able to provide a prior matching those areas (the first chart in the second row in the figure). Other models can only achieve this by significantly decreasing the overall regularization strength.

Similar arguments can be used to pinpoint models most suitable in other scenarios in Table 2. From the practical point of view the most important are Scenarios 3 and 4, which correspond, respectively, to low control response and small effect of the action. The treatment interaction model is able to cover both those cases which may explain its popularity in literature.

Notice also that when both $\beta^T, \beta^C \approx 0$ all models should provide effective regularization.

3.3 Relationship between symmetric interaction model and transformed target variable regression

In [27] a different estimator for treatment effect coefficients has been proposed, which works by concatenating the treatment and control training sets and building a single regression model on a transformed target variable

$$\bar{y}_i = \begin{cases} 2y_i & \text{if } t_i = 1, \\ -2y_i & \text{if } t_i = 0. \end{cases}$$

Theorem 2. *When $n^T = n^C$, the square loss is used, and $\lambda_2 \to \infty$ with λ_1 held fixed, the symmetric interaction model (SI) tends to the variable transformation model regularized with $4\lambda_1 \|\beta^U\|_q^q$.*

The proof can be found in the supplementary material.

4 Experimental evaluation

In this section let n_{test} denote the number of test cases, x_{test_i} the feature vector of i-th test case and τ_i the true uplift for the i-th test case, i.e. the difference between potential outcome has case i been subjected to the action and the potential outcome has case i been a control. This value is available only for synthetic data since in real data only one of the outcomes is observed [10].

4.1 Evaluation of uplift regression models

We first need to discuss the issue of evaluation of uplift regression models. A natural choice is the Mean Squared Error $MSE(\hat{\beta}^U) = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \left(\tau_i - x_{test_i}\hat{\beta}^U\right)^2$. Unfortunately, τ_i is unknown for real data so there is a need for an alternative measure. We therefore propose a measure for evaluating uplift regression models which we call *by quantile MSE* or QMSE for short. This measure is similar to *expected uplift calibration error (EUCE)* proposed in [21], except that squared loss is used instead of absolute value. The measure is calculated as follows.

Let X_{test}^T and X_{test}^C be the treatment and control test sets. Compute the corresponding vectors of model predictions $X_{test}^T\hat{\beta}^U$, $X_{test}^C\hat{\beta}^U$. The vectors are sorted and split into J quantiles (10 in our case). Let Q_j^T and Q_j^C denote the indices of, respectively, treatment and control test records in the j-th treatment or, respectively, control quantile. Compute the MSE within j-th quantile as

$$MSE_j(\hat{\beta}^U) = \frac{1}{n_j^T} \sum_{i \in Q_j^T} \left(x_{test_i}\hat{\beta}^U - \left(\frac{1}{n_j^T} \sum_{i \in Q_j^T} y_i - \frac{1}{n_j^C} \sum_{i \in Q_j^C} y_i \right) \right)^2,$$

where n_j^T and n_j^C are the number of treatment and control records in the j-th quantile. The final QMSE measure is $QMSE(\hat{\beta}^U) = \frac{1}{J} \sum_{i=1}^{J} MSE_j(\hat{\beta}^U)$.

4.2 Synthetic data

In this section we evaluate regularized uplift regression estimators on synthetic data. We begin by describing the experimental procedure.

For a given number of columns ($p = 160$) we generated random predictor matrices X with increasing number of rows. For L_1 regularization we used $n \in \{60, 80, 100, 120\}$ and for L_2 regularization $n \in \{180, 200, 250, 500\}$. The reason was that L_1 regularization is supposed to work better when $p > n$ and L_2 regularization when $n > p$. Each row x_i of X is generated from the multivariate normal distribution with zero mean and unit covariance matrix. Each sample is assigned to the treatment or control group at random but with fixed group proportions $\frac{n^T}{n} = \frac{n^C}{n} = \frac{1}{2}$. The outcome variables are then generated based on Equation 1 with $\varepsilon_i \sim \mathcal{N}(0, 1)$. $n_{test} = 10\,000$ was used with identical data generation mechanism.

Regularized models require the choice of regularization parameter values. In our case we use 3-fold crossvalidation and select all regularization parameters from the set $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$ for both λ_1 and λ_2.

Since τ_i is known for simulated data, we use the classic MSE criterion to assess model performance. Parameter selection is still performed based on $QMSE$ for consistency with experiments on real data.

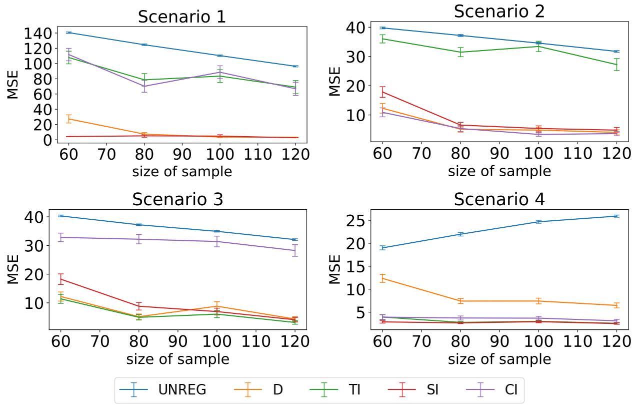

Fig. 2: MSE of estimators with L_1 penalty under different simulation scenarios

In our simulations we use vectors β^C and β^U corresponding to the four scenarios presented in Table 2. The actual coefficient vectors are given in supplementary materials. We include five estimators in the comparison: the nonregularized double model (UNREG), the regularized double model (D), and three regularized interaction models: treatment interaction (TI), symmetric interaction (SI) and control interaction (CI). Note that all unregularized models are equivalent so only one of them is included.

Results for L_1 regularization are presented in Figure 2. We observe that for the first scenario the best results are achieved by the symmetric interaction method. This is consistent with Table 2 and discussion in Section 3 which suggest the SI method is most suitable when values of β^S are small. Interestingly the double regularized model also performed well.

The second and third plots correspond to the situation when $\beta^T \approx 0$ and $\beta^C \approx 0$ respectively. In both cases double regularized method performs well. When $\beta^T \approx 0$ the control interaction model also attains good results, but treatment interaction model behaves badly. For $\beta^C \approx 0$ we have the opposite situation. Again, those results are in line with theoretical predictions.

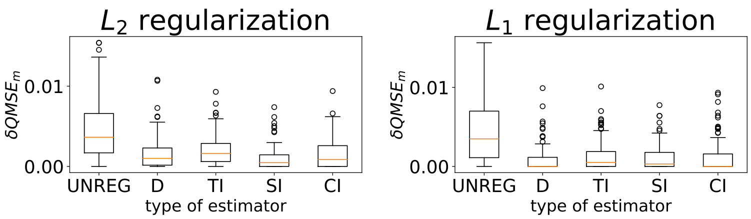

Fig. 3: Results for the IHDP dataset

The fourth plot presents the case when the action's impact (β^U) is small. All regularized methods, except the double regularized model achieve good and comparable results. This observation is again consistent with the fact that regularization regions with small values of uplift occur naturally in those methods.

Similar conclusions could be drawn from the results for L_2 regularization, which are shown in the supplementary material due to lack of space. Overall we conclude that experiments on synthetic data fully confirm theoretical analysis from Section 3 for both L_1 and L_2 regularized models.

4.3 Experiments on real data

Description of datasets. The first dataset we use is the IHDP dataset [20]. 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 whose IQ (the target variable) increased *because* of the intervention program. There are 377 treatment and 608 control cases. We also ran experiments on the well known Lalonde dataset [19], see supplementary materials.

Results. During experiments each dataset was split into training (70%) and test parts (30%), stratified by treatment. Models are built and tuned on the training part (λ_1 and λ_2 are chosen form the same set $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$) and their QMSE's are computed on the test part. To make the results easier to understand for each model we compute the difference $\delta\,\mathrm{QMSE}$ from the best model, i.e. for a model m

$$\delta\,\mathrm{QMSE}_m = |\,\mathrm{QMSE}_m - \min_i \mathrm{QMSE}_i\,|,$$

where QMSE_m is the QMSE of model m. The train/test split is repeated 100 times and box plots of $\delta\,\mathrm{QMSE}_m$ are shown for each model. This way we can visualize in a single plot how well, each model performed relative to others.

Results for the `IHDP` dataset are presented in Figure 3. All regularizers perform very well and beat unregularized models by a wide margin. We notice that symmetric interaction method achieves the best results out of all of L_2 regularizers. For L_1 regularization the smallest values of $\delta\,\mathrm{QMSE}_m$ were obtained by the regularized double model. While all methods perform well in general, it is worth trying different interaction models since there is a possibility that some of them may better match true coefficient vectors.

5 Conclusions

We have analyzed the problem of regularizing uplift regression models. We have shown that the type of interaction term used has a strong influence on the corresponding prior in unexpected ways. As a result, we were able to describe scenarios where each regularized model is most useful. Experiments on simulated data fully confirm our analyses, and experiments on real data demonstrate the usefulness of regularizing interaction models.

References

1. Athey, S., Imbens, G.: Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences **113**(27), 7353–7360 (2016)
2. Belloni, A., Chernozhukov, V., Hansen, C.: High-dimensional methods and inference on structural and treatment effects. Journal of Economic Perspectives **28**(2), 1–23 (2014)
3. Betlei, A., Diemert, E., Amini, M.R.: Uplift prediction with dependent feature representation in imbalanced treatment and control conditions. In: Proc. of the 25th Conf. on Neural Information Processing (ICONIP'18). pp. 47–57. Springer International Publishing (2018)
4. Chambers, J.M., Hastie, T.J.: Statistical Models in S. Chapman & Hall (1993)
5. Gross, S.M., Tibshirani, R.: Data shared lasso: A novel tool to discover uplift. Computational Statistics & Data Analysis **101**, 226–235 (2016)
6. Hahn, P.R., Murray, J.S., Carvalho, C.M.: Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects (with Discussion). Bayesian Analysis **15**(3), 965 – 2020 (2020)
7. Hernán, M., Robins, J.: Causal Inference. Boca Raton: Chapman & Hall/CRC (2018), forthcoming
8. Heumann, C., Nittner, T., Rao, C., Scheid, S., Toutenburg, H.: Linear Models: Least Squares and Alternatives. Springer New York (2013)
9. Hill, J.L.: Bayesian nonparametric modeling for causal inference. Journal of Computational and Graphical Statistics **20**(1), 217–240 (2011)
10. Holland, P.: Statistics and causal inference. Journal of the American Statistical Association **81**(396), 945–960 (Dec 1986)
11. Imai, K., Ratkovic, M.: The annals of applied statistics. Estimating treatment effect heterogeneity in randomized program evaluation **7**, 443–470 (2013)
12. Imbens, G., Rubin, D.: Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press, New York, NY, USA (2015)
13. Jaśkowski, M., Jaroszewicz, S.: Uplift modeling for clinical trial data. In: ICML 2012 Workshop on Machine Learning for Clinical Data Analysis. Edinburgh (Jun 2012)
14. Kane, K., Lo, V.S.Y., 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)
15. Kozubowski, T.J., Podgórski, K., Rychlik, I.: Multivariate generalized laplace distribution and related random fields. J. Multivar. Anal. **113**, 59–72 (2013)
16. Künzel, S.R., Sekhon, J.S., Bickel, P.J., Yu, B.: Metalearners for estimating heterogeneous treatment effects using machine learning. Proceedings of the National Academy of Sciences **116**(10), 4156–4165 (2019). https://doi.org/10.1073/pnas.1804597116
17. Kuusisto, F., Santos Costa, V., Nassif, H., Burnside, E., Page, D., Shavlik, J.: Support vector machines for differential prediction. In: ECML-PKDD (2014)
18. Lai, L.Y.T.: Influential Marketing: A New Direct Marketing Strategy Addressing The Existence Of Voluntary Buyers. Master's thesis, Simon Fraser University (2006)
19. Lalonde, R.: Evaluating the econometric evaluations of training programs. American Economic Review **76**, 604–620 (1986)
20. Liaw, F., Klebanov, P., Brooks-Gunn, J.: Effects of early intervention on cognitive function of low birth weight preterm infants. Journal of Pediatrics, **120**, 350–359 (1991)

21. Nyberg, O., Kuśmierczyk, T., Klami, A.: Uplift modeling with high class imbalance. In: Proceedings of the 13th Asian conference on machine learning. pp. 315–330. Bangkok (Nov 2021)
22. Padilla, O.H.M., Chen, Y., Ruiz, G.: A causal fused lasso for interpretable heterogeneous treatment effects estimation (2022)
23. Pearl, J.: Causality. Cambridge University Press (2009)
24. Petersen, K.B., Pedersen, M.S.: The Matrix Cookbook. Technical University of Denmark (nov 2012), version 20121115
25. Radcliffe, N.J., Surry, P.D.: Real-world uplift modelling with significance-based uplift trees. Portrait Technical Report TR-2011-1, Stochastic Solutions (2011)
26. Rousseeuw, P.J., Leroy, A.M.: Robust Regression and Outlier Detection. John Wiley & Sons (1987)
27. Rudas, K., Jaroszewicz, S.: Linear regression for uplift modeling. Data Mining and Knowledge Discovery **32**(5), 1275–1305 (2018)
28. Rudas, K., Jaroszewicz, S.: Shrinkage estimators for uplift regression. In: European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD'19). Würzburg, Germany (Sep 2019)
29. Rzepakowski, P., Jaroszewicz, S.: Decision trees for uplift modeling with single and multiple treatments. Knowledge and Information Systems (2011)
30. Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction, and Search. MIT Press (2001)
31. Zaniewicz, Ł., Jaroszewicz, S.: Support vector machines for uplift modeling. In: The First IEEE ICDM Workshop on Causal Discovery (CD 2013). Dallas (Dec 2013)
32. Zaniewicz, Ł., Jaroszewicz, S.: l_p-support vector machines for uplift modeling. Knowledge and Information Systems **53**(1), 269–296 (Oct 2017)