Comparing methods for variable selection in individual patient data meta-analysis

Michael Seo

# Introduction

We would like to explore how to best select covariates when we need to synthesize treatment effects from multiple clinical trials. Individual patient data (IPD) from all studies can be analyzed simultaneously by adopting a single model. IPD-MA model with random effects on the treatment effect is a common approach. However, selecting covariates in a IPD-MA framework is not clear. We compare different methods ranging from naive models that pool all clinical trials into one dataset and random effects models that properly take into account of the study level structure.

# Data and Simulation

## Motivating data: Drug-eluting or bare-metal stents for percutaneous coronary intervention

We have an individual patient data of randomised clinical trials to compare outcomes after implantation of new-generation DES or BMS among patients undergoing percutaneous coronoary intervention. Individual patient data were sought and obtained for all 20 studies. The primary outcome is the composite of cardiac death or myocardinal infarction at a 5 year landmark. Our motivating dataset is made of one continuous dependent variable, age, one count variable, the number of stented segments, and nine binary covariates: gender, diabetes, clinical presentation at the time of percutaneous coronary intervention, stent placement in the left anterior descending artery, overlapping stents, multivessel disease, mean stent diameter greater than 3, use of glycoprotein IIb/IIIa receptor inhibitors, and use of newer P2Y\_12 receptor inhibitors (ticagrelor or prasugrel).

Continuous variables are scaled. Binary variables are not scaled since in the imputation step, these variables are assumed to be bernoulli distributed, so the values need to be either 0 or 1. Count variables are assumed to be poisson distributed in the imputation stage. Bare-metal stent is chosen as the baseline treatment.

## Simulation

IPD-MA model is used to generate the simulations. The model is described below.

where refers to each patient, refers to baseline treatment, refers to study specific baseline risk, refers to average treatment effect, refers to coefficient of the prognostic factor, and refers to coefficient of the effect modifier.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Response | # of studies | # of patients | # of covariates | Prognostic factors | Effect Modifiers | Nuisance parameters |
| Sim 1 | Continuous | 5 | 30-300 | 10 | 2 cont 1 discrete | 1 cont 1 discrete | 3 cont 2 discrete |
| Sim 2 | Continuous | 10 | 30-300 | 20 | 4 cont 2 discrete | 2 cont 2 discrete | 6 cont 4 discrete |
| Sim 3 | Binary | 5 | 30-300 | 10 | 2 cont 1 discrete | 1 cont 1 discrete | 3 cont 2 discrete |
| Sim 4 | Binary | 10 | 30-300 | 20 | 4 cont 2 discrete | 2 cont 2 discrete | 6 cont 4 discrete |

From the chart above, first simulation has continuous response with 5 clinical trials. Each trial has around 50-100 patients and 10 covariates are measured. We generate 2 continuous and 1 discrete prognostic factors and 3 continuous and 2 discrete effect modifiers. Furthermore, 3 continuous and 2 discrete nuisance parameters are entered in the model fitting, but are not used to generate the simulation set. Our process of generating simulation can be described as follows:

1. we draw the rows of the predictor matrix X from , where has entry equal to for continuous covariates and for discrete covariates, treatment indicators from , number of patients for each study from ;
2. we draw treatment effect of each study from where , average treatment effect, is fixed at 1 and is drawn from ;
3. we draw baseline effect of the treatment , from and assume independent baseline effect;
4. we assume for the first study with 5 studies and 10 covariates that coefficients of prognostic factors are 0.1, -0.1, 0.2, 0.2, -0.2 and coefficients of effect modifiers are 0.2, 0.3. We assume for the second simulation with 10 studies and 20 coariates that coefficients of prognostic factors are 0.1, -0.1, 0.2, 0.2, -0.2, 0.3, 0.1, 0.1, -0.1, -0.2 and coefficients of effect modifiers are 0.2, 0.3, -0.1, -0.2. All the nuisance predictors are not used in the simulation;
5. we add in a random error component, denoted as , sampled from for normal and for binomial;
6. we draw the response vector Y from .

# Methods

We evaluate how different models compare in terms of model selection using the simulated dataset. To validate selection of each model, evaluation criteria such as correct model identified and number of false positive and false negative prognostic factors and effect modifiers are used. Moreover, primary goal of (individual) patient data meta-analysis is to obtain accurate estimation of treatment effect and effect modifiers. Thus, we calculate the sum of mean squared error of these parameter estimates from the true value.

When selecting important covariates, we might be tempted to pool all different studies and use naive models that ignore the study level structure. We compare these naive models with mixed effects models that account for study level structure. Following models are used: naive approach using step function (both direction), naive approach using LASSO, naive approach using best subset model, generalized linear mixed effects model using LASSO, and stochastic search variable selection (SSVS), which is a Bayesian variable selection model.

## Naive approach using step function

This method aggregates the studies into one and fits a simple linear model and uses stepwise function to select variables. This naive approach ignores the study level structure. Both direction is used for stepwise search.

## Naive approach using LASSO

This method aggregates the studies into one and fits a simple linear model and uses lasso penalization to select variables. All variables except treatment effect are penalized. When picking the optimal lambda value, we consider using lambda.1se instead of lambda.min. 1 SE rule selects the most parsimonious model whose error is no more than one standard error above the error of the best model. The main point of the 1 SE rule is to choose the simplest model whose accuracy is comparable with the best model. We used 10 fold cross validation of mean squared error to obtain the lambda.min and lambda.1se. Cross validation using mean squared error for continuous outcome and using misclassification rate for binary outcome are used.

## Best subset model

The best subsets regression is a model selection approach that consists of testing all possible combination of the predictor variables, and then selecting the best model according to some statistical criteria. One caveat with this approach is the high computational cost as all combination has to be considered. Recently, Bertsimas et al. (2016) showed that the best subset selection problem can be formulated as a mixed integer optimization (MIO) problem and demonstrated that this can be solved at even large problem sizes. We use cross validated mean squared error to select the best possible combination. We use R packages leaps to implement best subset models, but method from Bertsimas can be utilized using a R package bestsubset for more complicated problems.

## Generalized linear mixed effects model using LASSO

This model uses generalized linear mixed model and uses LASSO to identify important variables. Treatment effect is assumed to be a random effect and we incorporate study level structure by adding study level information as a baseline risk. Treatment effect is excluded from penalization. R package glmmLasso is used to implement this method. For a set of lambda values, we chose the lambda value that had the lowest BIC. Another way to find optimal lambda value would be through cross validation. We perform final fisher scoring re-estimation once the variables are selected.

## Bayesian variable selection model (SSVS)

This Bayesian model introduces indicator variable to select each covariate the model. A mixture prior for is used: , where the first density (the spike) is centred around zero and has a small variance. We let to be estimated in the model with own prior, and g fixed at 0.01. A natural alternative would be to fix , and estimate g, by placing a prior on the product .

# Simulation Results

We ran 100 simulations for the random effects models. Then, we calculated the following performance measure: correct model identified, false positive, and false negative. For the following simulations we have the following:

Scenario 1: continuous, few covariates

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | false em mse | true em mse | treatment mse | treatment\_sd |
| simple null | 0.0000000 | 0.0650000 | 0.1369622 | 0.0950777 |
| simple lm | 0.0175540 | 0.0215089 | 0.1557926 | 0.2079884 |
| naive step | 0.0092890 | 0.0323935 | 0.1451939 | 0.1271915 |
| naive lasso | 0.0000077 | 0.0603649 | 0.1368791 | NA |
| glmmLasso | 0.0019501 | 0.0468985 | 0.0906495 | NA |
| bayes lasso | 0.0010186 | 0.0419999 | 0.1090435 | 0.4819709 |
| SSVS | 0.0014421 | 0.0313511 | 0.1096103 | 0.4838256 |

Scenario 2: continuous, many covariates

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | false em mse | true em mse | treatment mse | treatment\_sd |
| simple null | 0.0000000 | 0.0450000 | 0.0800948 | 0.0725957 |
| simple lm | 0.0100855 | 0.0113586 | 0.0967060 | 0.1991320 |
| naive step | 0.0060647 | 0.0144095 | 0.0864682 | 0.1090227 |
| naive lasso | 0.0000698 | 0.0158049 | 0.0770593 | NA |
| glmmLasso | 0.0035954 | 0.0095506 | 0.1109269 | NA |
| bayes lasso | 0.0017630 | 0.0061731 | 0.0545835 | 0.2715918 |
| SSVS | 0.0014669 | 0.0066101 | 0.0535050 | 0.2710989 |

Scenario 3: binary, few covariates

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | false em mse | true em mse | treatment mse | treatment\_sd |
| simple null | 0.0000000 | 0.0650000 | 0.0992089 | 0.1543591 |
| simple lm | 0.0581641 | 0.0657322 | 0.2804521 | 0.3570444 |
| naive step | 0.0334639 | 0.0720479 | 0.2089268 | 0.2030755 |
| naive lasso | 0.0010561 | 0.0631050 | 0.1050510 | NA |
| glmmLasso | 0.0000175 | 0.0650104 | 0.0990628 | NA |
| bayes lasso | 0.0021531 | 0.0533504 | 0.1350061 | 0.5560701 |
| SSVS | 0.0026283 | 0.0505450 | 0.1415782 | 0.5630694 |

Scenario 4: binary, many covariates

# Case study - Stent data results

coefficients for simple lm, naive step, naive lasso, glmmLasso, SSVS, and Lykou.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | simple null | simple lm | naive step | naive lasso | glmmLasso | bayes lasso | SSVS | SSVS Ind |
| (Intercept) | -2.783 | -3.135 | -3.506 | -2.783 | -3.760 | NA | 0.000 | 0.000 |
| age | 0.000 | 0.838 | 0.806 | 0.000 | 0.048 | NA | 0.660 | 1.000 |
| gender | 0.000 | 0.012 | 0.000 | 0.000 | 0.000 | NA | -0.005 | 0.235 |
| diabetes | 0.000 | 0.540 | 0.506 | 0.000 | 0.000 | NA | 0.416 | 0.996 |
| stable\_cad | 0.000 | -0.532 | -0.472 | 0.000 | 0.000 | NA | -0.486 | 0.999 |
| multivessel | 0.000 | 0.250 | 0.155 | 0.000 | 0.000 | NA | 0.162 | 0.617 |
| ladtreated | 0.000 | 0.231 | 0.244 | 0.000 | 0.000 | NA | 0.055 | 0.320 |
| overlap | 0.000 | 0.481 | 0.485 | 0.000 | 0.000 | NA | 0.336 | 0.911 |
| m\_dia\_above\_3 | 0.000 | -0.433 | 0.000 | 0.000 | 0.000 | NA | -0.037 | 0.363 |
| num\_stent | 0.000 | 0.038 | 0.054 | 0.000 | 0.000 | NA | 0.017 | 0.166 |
| age:treat | 0.000 | -0.056 | 0.000 | 0.000 | 0.142 | NA | -0.037 | 0.260 |
| gender:treat | 0.000 | 0.031 | 0.000 | 0.000 | 0.000 | NA | 0.012 | 0.256 |
| diabetes:treat | 0.000 | -0.067 | 0.000 | 0.000 | 0.000 | NA | -0.009 | 0.291 |
| stable\_cad:treat | 0.000 | 0.113 | 0.000 | 0.000 | 0.000 | NA | 0.023 | 0.311 |
| multivessel:treat | 0.000 | -0.184 | 0.000 | 0.000 | 0.000 | NA | -0.050 | 0.365 |
| ladtreated:treat | 0.000 | -0.313 | -0.336 | 0.000 | 0.000 | NA | -0.165 | 0.576 |
| overlap:treat | 0.000 | 0.008 | 0.000 | 0.000 | 0.000 | NA | -0.029 | 0.326 |
| m\_dia\_above\_3:treat | 0.000 | 0.445 | 0.000 | 0.000 | 0.000 | NA | 0.087 | 0.430 |
| num\_stent:treat | 0.000 | -0.082 | -0.112 | 0.000 | 0.000 | NA | -0.041 | 0.254 |
| treat | -0.210 | -0.132 | 0.251 | -0.210 | 0.083 | NA | 0.023 | 1.000 |
| as.factor(studyid)1 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | NA | -4.417 | 1.000 |
| as.factor(studyid)2 | 0.000 | 0.000 | 0.000 | 0.000 | 2.324 | NA | -3.765 | 1.000 |
| as.factor(studyid)3 | 0.000 | 0.000 | 0.000 | 0.000 | -0.140 | NA | -3.541 | 1.000 |
| as.factor(studyid)4 | 0.000 | 0.000 | 0.000 | 0.000 | 0.627 | NA | -3.203 | 1.000 |
| as.factor(studyid)5 | 0.000 | 0.000 | 0.000 | 0.000 | -0.900 | NA | -3.530 | 1.000 |
| as.factor(studyid)6 | 0.000 | 0.000 | 0.000 | 0.000 | 0.195 | NA | -3.974 | 1.000 |
| as.factor(studyid)7 | 0.000 | 0.000 | 0.000 | 0.000 | 0.082 | NA | -2.953 | 1.000 |
| as.factor(studyid)8 | 0.000 | 0.000 | 0.000 | 0.000 | 0.493 | NA | -2.521 | 1.000 |
| heterogeneity | NA | NA | NA | 0.072 | 0.032 | NA | NA | NA |

Below are the estimates of standard error for these coefficients.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | simple null | simple lm | naive step | naive lasso | glmmLasso | bayes lasso | SSVS |
| (Intercept) | 0.061 | 0.308 | 0.141 | NA | NA | NA | 0.000 |
| age | 0.000 | 0.083 | 0.055 | NA | NA | NA | 0.080 |
| gender | 0.000 | 0.134 | 0.000 | NA | NA | NA | 0.065 |
| diabetes | 0.000 | 0.132 | 0.092 | NA | NA | NA | 0.104 |
| stable\_cad | 0.000 | 0.148 | 0.101 | NA | NA | NA | 0.115 |
| multivessel | 0.000 | 0.134 | 0.093 | NA | NA | NA | 0.144 |
| ladtreated | 0.000 | 0.129 | 0.128 | NA | NA | NA | 0.102 |
| overlap | 0.000 | 0.176 | 0.126 | NA | NA | NA | 0.152 |
| m\_dia\_above\_3 | 0.000 | 0.259 | 0.000 | NA | NA | NA | 0.148 |
| num\_stent | 0.000 | 0.064 | 0.057 | NA | NA | NA | 0.043 |
| age:treat | 0.000 | 0.114 | 0.000 | NA | NA | NA | 0.085 |
| gender:treat | 0.000 | 0.189 | 0.000 | NA | NA | NA | 0.087 |
| diabetes:treat | 0.000 | 0.185 | 0.000 | NA | NA | NA | 0.094 |
| stable\_cad:treat | 0.000 | 0.202 | 0.000 | NA | NA | NA | 0.102 |
| multivessel:treat | 0.000 | 0.186 | 0.000 | NA | NA | NA | 0.141 |
| ladtreated:treat | 0.000 | 0.179 | 0.178 | NA | NA | NA | 0.172 |
| overlap:treat | 0.000 | 0.252 | 0.000 | NA | NA | NA | 0.125 |
| m\_dia\_above\_3:treat | 0.000 | 0.405 | 0.000 | NA | NA | NA | 0.207 |
| num\_stent:treat | 0.000 | 0.095 | 0.074 | NA | NA | NA | 0.058 |
| treat | 0.085 | 0.467 | 0.178 | NA | NA | NA | 0.277 |
| as.factor(studyid)1 | 0.000 | 0.000 | 0.000 | NA | NA | NA | 0.406 |
| as.factor(studyid)2 | 0.000 | 0.000 | 0.000 | NA | NA | NA | 0.270 |
| as.factor(studyid)3 | 0.000 | 0.000 | 0.000 | NA | NA | NA | 0.890 |
| as.factor(studyid)4 | 0.000 | 0.000 | 0.000 | NA | NA | NA | 0.255 |
| as.factor(studyid)5 | 0.000 | 0.000 | 0.000 | NA | NA | NA | 0.286 |
| as.factor(studyid)6 | 0.000 | 0.000 | 0.000 | NA | NA | NA | 0.283 |
| as.factor(studyid)7 | 0.000 | 0.000 | 0.000 | NA | NA | NA | 0.228 |
| as.factor(studyid)8 | 0.000 | 0.000 | 0.000 | NA | NA | NA | 0.237 |

# Note on standard error

<https://stats.stackexchange.com/questions/91462/standard-errors-for-lasso-prediction-using-r>

It is a very natural question to ask for standard errors of regression coefficients or other estimated quantities. In principle such standard errors can easily be calculated, e.g. using the bootstrap. Still, this package deliberately does not provide them. The reason for this is that standard errors are not very meaningful for strongly biased estimates such as arise from penalized estimation methods. Penalized estimation is a procedure that reduces the variance of estimators by introducing substantial bias. The bias of each estimator is therefore a major component of its mean squared error, whereas its variance may contribute only a small part. Unfortunately, in most applications of penalized regression it is impossible to obtain a sufficiently precise estimate of the bias. Any bootstrap-based calculations can only give an assessment of the variance of the estimates. Reliable estimates of the bias are only available if reliable unbiased estimates are available, which is typically not the case in situations in which penalized estimates are used.