High Dimensional Metrics in Julia

VICTOR CHERNOZHUKOV, CHRISTIAN HANSEN, MARTIN SPINDLER

October - 13 - 2022

Contents

1 Introduction

Analysis of high-dimensional models, models in which the number of parameters to be estimated is large relative to the sample size, is becoming increasingly important. Such models arise naturally in modern data sets which have many measured characteristics available per individual observation as in, for example, population census data, scanner data, and text data. Such models also arise naturally even in data with a small number of measured characteristics in situations where the exact functional form with which the observed variables enter the model is unknown and we create many technical variables, a dictionary, from the raw characteristics. Examples covered by this scenario include semi-parametric models with non-parametric nuisance functions. More generally, models with many parameters relative to the sample size often arise when attempting to model complex phenomena. With increasing availability of such data sets in economics and other data science fields, new methods for analyzing those data have been developed. The `Julia` package `HDMjl` contains implementations of recently developed methods for high-dimensional approximately sparse models, mainly relying on forms of lasso and post-lasso as well as related estimation and inference methods. The methods are illustrated with econometric applications, but are also useful in other disciplines such as medicine, biology, sociology or psychology.

The methods which are implemented in this package are distinct from already available methods in other packages in the following four major ways:

1. First, we provide a version of Lasso regression that expressly handles and allows for non-Gaussian and heteroscedastic errors.
2. Second, we implement a theoretically grounded, data-driven choice of the penalty level λ in the Lasso regressions. To underscore this choice, we call the Lasso implementation in this package "rigorous"Lasso (=`rlasso`). The prefix `r` in function names should underscore this. In high dimensional settings cross-validation is very popular; but it lacks a theoretical justification for use in the present context and some theoretical proposals for the choice of λ are often not feasible. Moreover, the theoretically grounded, data-driven choice redundancies cross-validation which is time-consuming particularly in large data sets.
3. Third, we provide efficient estimators and uniformly valid confidence intervals for various low-dimensional causal/structural parameters appearing in high-dimensional approximately sparse models. For example, we provide efficient estimators and uniformly valid confidence intervals for a regression coefficient on a target variable (e.g., a treatment or policy variable) in a high-dimensional sparse regression model. Target variable in this context means the object not interest, e.g. a pre-specified regression coefficient. We also provide estimates and confidence intervals for average treatment effect (ATE) and average treatment effect for the treated (ATET), as well extensions of these parameters to the endogenous setting.
4. Fourth, joint/ simultaneous confidence intervals for estimated coefficients in a high-dimensional approximately sparse models are provided, based on the methods and theory developed in Belloni, Chernozhukov, and Kato (2014). They proposed uniformly valid confidence regions for regressions coefficients in a high-dimensional sparse Z-estimation problems, which include median, mean, and many other regression problems as special cases. In this article we apply this method to the coefficients of a Lasso regression and highlight this method with an empirical example.

2 How to get started

Julia is an open source software project and can be freely downloaded from the julialan g.org website along with its associated documentation. The Julia package `HDMjl` can be downloaded from juliahub. To install the `HDMjl` package from Julia we simply type.

```
# ] add HDMjl
```

The most current version of the package (development version) can installed by

```
using  Pkg; Pkg.add("HDMjl")
```

You may also install the dev version of the package by directly acquiring it from the repository by using

```
# ] add https://github.com/d2cml-ai/HDMjl.jl
```

or

```
import Pkg; Pkg.add(url = "https://github.com/d2cml-ai/HDMjl.jl")
```

Provided that your machine has a proper internet connection and you have write permission in the appropriate system directories, the installation of the package should proceed automatically. Once the `HDMjl` package is installed, it can be loaded to the current Julia session by the command

```
using HDMjl
```

3 Prediction using Approximate Sparsity

3.1 Prediction in Linear Models using Approximate Sparsity.

Consider high dimensional approximately sparse linear regression models. These models have a large number of regressors p, possibly much larger than the sample size n, but only a relatively small number $s = o(n)$ of these regressors are important for capturing accurately the main features of the regression function. The latter assumption makes it possible to estimate these models effectively by searching for approximately the right set of regressors.

The model reads

$$y_i = x_i'\beta_0 + \varepsilon_i, \quad E\left[\varepsilon_i x_i\right] = 0, \quad \beta_0 \in \mathbb{R}^p, \quad i = 1, \ldots, n$$

where y_i are observations of the response variable, $x_i = (x_{i,j}, \ldots, x_{i,p})$'s are observations of p-dimensional regressors, and ε_i's are centered disturbances, where possibly $p \gg n$. Assume that the data sequence is i.i.d. for the sake of exposition, although the framework covered is considerably more general. An important point is that the errors ε_i may be non-Gaussian or heteroskedastic (Belloni, Chen, Chernozhukov, and Hansen, 2012).

The model can be exactly sparse, namely

$$\|\beta_0\|_0 \leq s = o(n)$$

or approximately sparse, namely that the values of coefficients, sorted in decreasing order, $(|\beta_0|_{(j)})_{j=1}^p$ obey,

$$|\beta_0|_{(j)} \leq \mathrm{A} j^{-\mathrm{a}(\beta_0)}, \quad \mathrm{a}(\beta_0) > 1/2, \quad j = 1, \ldots, p$$

An approximately sparse model can be well-approximated by an exactly sparse model with sparsity index

$$s \propto n^{1/(2a(\beta_0))}.$$

In order to get theoretically justified performance guarantees, we consider the Lasso estimator with data-driven penalty loadings:

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \mathbb{E}_n \left[(y_i - x_i'\beta)^2 \right] + \frac{\lambda}{n} \|\hat{\Psi}\beta\|_1$$

where $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$, $\hat{\Psi} = \mathrm{diag}\left(\hat{\psi}_1, \ldots, \hat{\psi}_p\right)$ is a diagonal matrix consisting of penalty loadings, and En abbreviates the empirical average. The penalty loadings are chosen to insure basic equivariance of coefficient estimates to rescaling of $x_{i,j}$ and can also be chosen to address heteroskedasticity in model errors. We discuss the choice of λ and $\hat{\Psi}$ below.

Regularization by the **l** 1-norm naturally helps the Lasso estimator to avoid overfitting, but it also shrinks the fitted coefficients towards zero, causing a potentially significant bias. In order to remove some of this bias, consider the Post-Lasso estimator that applies ordinary least squares to the model $\hat{T}$ selected by Lasso, formally,

$$\hat{T} = \mathrm{support}(\hat{\beta}) = \{j \in \{1, \ldots, p\} : |\hat{\beta}| > 0\}.$$

The Post-Lasso estimate is then defined as

$$\tilde{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \mathbb{E}_n \left\{ y_i - \sum_{j=1}^p x_{i,j}\beta_j \right)^2 : \beta_j = 0 \quad \text{if } \hat{\beta}_j = 0, \quad \forall j.$$

In words, the estimator is ordinary least squares applied to the data after removing the regressors that were not selected by Lasso. The Post-Lasso estimator was introduced and analysed in Belloni and Chernozhukov (2013).

A crucial matter is the choice of the penalization parameter λ. With the right choice of the penalty level, Lasso and Post-Lasso estimators possess excellent performance guarantees: They both achieve the near-oracle rate for estimating the regression function, namely with probability $1 - \gamma - o(1)$,

$$\sqrt{\mathbb{E}_n \left[\left(x_i' \left(\hat{\beta} - \beta_0 \right) \right)^2 \right]} \lesssim \sqrt{(s/n) \log p}$$

In high-dimensions setting, cross-validation is very popular in practice but lacks theoretical justification and so may not provide such a performance guarantee. In sharp contrast, the

choice of the penalization parameter λ in the Lasso and Post-Lasso methods in this package is theoretical grounded and feasible. Therefore we call the resulting method the "rigorous"Lasso method and hence add a prefix **r** to the function names.

In the case of homoscedasticity, we set the penalty loadings $\hat{\psi}_j = \sqrt{\mathrm{E}_n x_{i,j}^2}$, which insures basic equivariance properties. There are two choices for penalty level λ : the X-independent choice and X dependent choice. In the X-independent choice we set the penalty level to

$$\lambda = 2c\sqrt{n}\hat{\sigma}\Phi^{-1}(1-\gamma/(2p)),$$

where Φ denotes the cumulative standard normal distribution, $\hat{\sigma}$ is a preliminary estimate of $\sigma = \sqrt{\mathbb{E}\varepsilon^2}$, and c is a theoretical constant, which is set to $c = 1.1$ by default for the Post-Lasso method and $c = .5$ for the Lasso method, and γ is the probability level, which is set to $\gamma = .1$ by default. The parameter γ can be interpreted as the probability of mistakenly not removing X 's when all of them have zero coefficients. In the X-dependent choice the penalty level is calculated as

$$\lambda = 2c\hat{\sigma}\Lambda(1-\gamma \mid X),$$

where

$$\Lambda(1-\gamma \mid X) = (1-\gamma) \text{ - quantile of } n\left\|\mathbb{E}_n\left[x_i e_i\right]\right\|_\infty \mid X,$$

where $X = [x_1, ..., x_n]'$ and e_i are iid $N(0,1)$, generated independently from X; this quantity is approximated by simulation. The X-independent penalty is more conservative than the X-dependent penalty. In particular the X-dependent penalty automatically adapts to highly correlated designs, using less aggressive penalization in this case Belloni, Chernozhukov, and Hansen (2010).

In the case of heteroskedasticity, the loadings are set to $\hat{\psi}_j = \sqrt{\mathbb{E}_n\left[x_{ij}^2\hat{\varepsilon}_i^2\right]}$, where $\hat{\varepsilon}_i$ are preliminary estimates of the errors. The penalty level can be X-independent (Belloni, Chen, Chernozhukov, and Hansen, 2012):

$$\lambda = 2c\sqrt{n}\Phi^{-1}(1-\gamma/(2p))$$

or it can be X-dependent and estimated by a multiplier bootstrap procedure (Chernozhukov, Chetverikov, and Kato, 2013)

$$\lambda = c \times c_W(1-\gamma),$$

where $c_W(1-\gamma)$ is the $1-\gamma$-quantile of the random variable W, conditional on the data, where

$$W := n \max_{1\le j\le p}\left|2\mathbb{E}_n\left[x_{ij}\hat{\varepsilon}_i e_i\right]\right|,$$

where e_i are iid standard normal variables distributed independently from the data, and $\hat{\varepsilon}_i$ denotes an estimate of the residuals.

Estimation proceeds by iteration. The estimates of residuals $\hat{\varepsilon}_i$ are initialized by running least squares of y_i on five regressors that are most correlated to y_i. This implies conservative starting values for λ and the penalty loadings, and leads to the initial Lasso and Post-Lasso estimates, which are then further updated by iteration. The resulting iterative procedure is fully justified in the theoretical literature.

3.2 A Joint Significance Test for Lasso Regression.

A basic question frequently arising in empirical work is whether the Lasso regression has explanatory power, comparable to a F-test for the classical linear regression model. The construction of a joint significance test follows (Chernozhukov, Chetverikov, and Kato, 2013) (Appendix M), and can be described as: Based on the model $y_i = a_0 + x_i' b_0 + \varepsilon_i$, the null hypothesis of joint statistical in-significance is $b_0 = 0$. The alternative is that of the joint statistical significance: $b_0 \neq 0$. The null hypothesis implies that

$$\mathrm{E}\left[(y_i - a_0)\, x_i\right] = 0$$

and restriction can be tested using the sup-score statistic:

$$S = \left\| \sqrt{n} \mathbb{E}_n \left[(y_i - \hat{a}_0)\, x_i\right] \right\|_\infty$$

where $\hat{a}_i = \mathbb{E}_n [y_i]$. The critical value for this statistic can be approximated by the multiplier bootstrap procedure, which simulates the statistic:

$$S^* = \left\| \sqrt{n} \mathbb{E}_n \left[(y_i - \hat{a}_0)\, x_i g_i\right] \right\|_\infty$$

where g_i 's are iid $N(0,1)$, conditional on the data. The $(1-\alpha)$-quantile of S^* serves as the critical value, $c(1-\alpha)$. We reject the null if $S > c(1-\alpha)$ in favor of statistical significant, and we keep the null of non-significance otherwise. This test procedure is implemented in the package when calling the `r_summary`-method of `rlasso-constructor`.

Julia Implementation The function `rlasso` implements Lasso and post-Lasso, where the prefix "r" signifies that these are theoretically rigorous versions of Lasso and post-Lasso. The default option is post-Lasso, `post=true`. This function returns an `constructor` of `rlasso` for which methods like `r_predict`, `r_print`, `r_summary` are provided.

`lassoShooting_fit` is the computational algorithm that underlies the estimation procedure, which implements a version of the Shooting Lasso Algorithm (Fu, 1998). The user has several options for choosing the non-default options. Specifically, the user can decide if an unpenalized intercept should be included (`true` by default). The option penalty of the function rlasso allows different choices for the penalization parameter and loadings. It allows for homoskedastic or heteroskedastic errors with default `homoscedastic = false`. Moreover, the dependence structure of the design matrix might be taken into consideration for calculation of the penalization parameter with `X_dependent_lambda = true`. In combination with these options, the option `lambda_start` allows the user to set a starting value for λ for the different algorithms. Moreover, the user can provide her own fixed value for the penalty level - instead of the data-driven methods discussed above - by setting `homoscedastic = "none"` and supplying the value via `lambda_start`.

The constants c and γ from above can be set in the option penalty. The quantities $\hat{\varepsilon}$, $\hat{\psi}$, $\hat{\sigma}$ are calculated in a iterative manner. The maximum number of iterations and the tolerance when the algorithms should stop can be set with control.

The method `r_summary` of rlasso-objects displays additionally for model diagnosis the R2 value, the adjusted R2 with degrees of freedom equal to the number of selected parameters, and the sup-score statistic for joint significance - described above - with corresponding p-value.

Example. (Prediction Using Lasso and Post-Lasso) Consider generated data from a sparse linear model:

```
using CodecXz, RData, DataFrames, StatsModels, LinearAlgebra
using Statistics, Distributions, PrettyTables, GLM, CSV

dta = get_data("seed_100")
n, p = size(dta);
Y = dta[:,1];
X = dta[:,2:end];
```

Next we estimate the model, print the results, and make in-sample and out-of sample predictions. We can use methods print and summarize to print the results, where the option all can be set to `false` to limit the print only to the non-zero coefficients.

We can estimate the models using Lasso

```
lasso_reg = rlasso(X, Y, post = false);
sum_lasso = r_summary(lasso_reg)
```

```
Post-Lasso Estimation: false
    Total number of variables: 100
    Number of selected variables: 11
    ---

============ ==============
  Variable    Estimate
============ ==============
  Intercept   0.056855
  X2          4.77121
  X3          4.69284
  X4          4.76568
  X14         -0.0453685
  X16         -0.0467382
  X17         -0.00499617
  X20         -0.0922336
  X23         -0.0272553
  X41         -0.0105032
  X62         0.113585
  X101        -0.0247296
============ ==============

    ----
```

```
    Multiple R-squared: 0.9912720815874809
    Adjusted R-squared: 0.9901810917859161
```

```
new_dta = get_data("seed_200")
Xnew = new_dta[:, Not(1)]
Ynew = new_dta[:, 1]
yhat_lasso_new = r_predict(lasso_reg, xnew = Matrix(Xnew))
post_lasso_reg = rlasso(X, Y, post = true)
y_hat_postlasso = r_predict(post_lasso_reg, xnew = Matrix(Xnew))
r_summary(post_lasso_reg)
```

```
Post-Lasso Estimation: true
    Total number of variables: 100
    Number of selected variables: 3
    ---

============ ============
  Variable    Estimate
============ ============
  Intercept   0.0341043
  X2          4.92413
  X3          4.85787
  X4          4.96442
============ ============

    ----
    Multiple R-squared: 0.9906284190077158
    Adjusted R-squared: 0.990335557101707
```

```
#in-sample prediction
yhat_postlasso = r_predict(post_lasso_reg)
#in-sample prediction;
yhat_postlasso_new = r_predict(post_lasso_reg, xnew = Matrix(Xnew));
```

```
MAE = mean(eachrow(hcat(
    abs.(Ynew - yhat_lasso_new), abs.(Ynew - yhat_postlasso_new)
    )))
MAE = DataFrame([[MAE[1]], [MAE[2]]], :auto)
MAE = rename!(MAE, ["lasso MAE", "Post-lasso MAE"])
pretty_table(MAE, tf = tf_simple, nosubheader = true)
```

```
============ =================
  lasso MAE   Post-lasso MAE
============ =================
   0.879583          0.78017
============ =================
```

4 Inference on Target Regression Coefficient

Here we consider inference on the target coefficient α in the model:

$$y_i = d_i\alpha_0 + x_i'\beta_0 + \epsilon_i, \quad \mathbb{E}_i\left(x_i', d_i'\right)' = 0$$

Here d_i is a target regressor such as treatment, policy or other variable whose regression coefficient α_0 we would like to learn (Belloni, Chernozhukov, and Hansen, 2014). If we are interested in coefficients of several or even many variables, we can simply write the model in the above form treating each variable of interest as d_i in turn and then applying the estimation and inference procedures described below.

We assume approximate sparsity for $x_i'\beta_0$ with sufficient speed of decay of the sorted components of β_0, namely $\mathrm{a}(\beta_0) > 1$. This condition translates into having a sparsity index $s \ll \sqrt{n}$. In general d_i is correlated to x_i, so α_0 cannot be consistently estimated by the regression of y_i on d_i. To keep track of the relationship of d_i to x_i, write

$$d_i = x_i'\pi_0^d + \rho_i^d, \quad \mathbb{E}\rho_i^d x_i = 0$$

To estimate α_0, we also impose approximate sparsity on the regression function $x_i'\pi_0^d$ with sufficient speed of decay of sorted components of π_0^d, namely a(^d_0) > 1$.

The Orthogonality Principle. Note that we can not use naive estimates of α_0 based simply on applying Lasso and Post-Lasso to the first equation. Such a strategy in general does not produce root-n consistent and asymptotically normal estimators of α, due to the possibility of large omitted variable bias resulting from estimating the nuisance function $x_i'\beta_0$ in high-dimensional setting. In order to overcome the omitted variable bias, we need to use orthogonalized estimating equations for α_0. Specifically we seek to find a score $\psi\left(w_i, \alpha, \eta\right)$, where $w_i = (y_i, x_i')'$ and η is the nuisance parameter, such that

$$\mathbb{E}\psi(w_i, \alpha_0, \eta_0) = 0, \quad \frac{\partial}{\partial\eta}\mathbb{E}\psi(w_i, \alpha_0, \eta_0) = 0$$

The second equation is the orthogonality condition, which states that the equations are not sensitive to the first-order perturbations of the nuisance parameter η near the true value. The latter property allows estimation of these nuisance parameters η_0 by regularized estimators $\hat{\eta}$, where regularization is done via penalization or selection. Without this property, regularization may have too much effect on the estimator of α_0 for regular inference to proceed. The estimators α of α_0 solve the empirical analog of the equation above,

$$\mathbb{E}_n\psi(w_i, \hat{\alpha}, \hat{\eta}) = 0,$$

where we have plugged in the estimator $\hat{\eta}$ for the nuisance parameter.

Due to the orthogonality property the estimator is first-order equivalent to the infeasible estimator $\tilde{\alpha}$ solving

$$\mathbb{E}_n\psi(w_i, \tilde{\alpha}, \eta_0) = 0,$$

where we use the true value of the nuisance parameter. The equivalence holds in a variety of models under plausible conditions. The systematic development of the orthogonality condition for inference on low-dimensional parameters in modern high-dimensional settings is given in Chernozhukov, Hansen, and Spindler (2015a).

In turns out that in the linear model the orthogonal equations are closely connected to the classical ideas of partialling out.

4.1 Intuition for the Orthogonality Principle in Linear Models via Partialling Out.

One way to think about estimation of α_0 is to think of the regression model:

$$\rho_i^y = \alpha_0 \rho_i^d + \epsilon_i$$

where ρ_i^y is the residual that is left after partialling out the linear effect of x_i from y_i and ρ_i^d is the residual that is left after partialling out the linear effect of x_i from d_i, both done in the population. Note that we have $\mathbb{E}\rho_i^y x_i = 0$, i.e. $\rho_i^y = y_i - x_i'\pi_0^y$ where $x_i'\pi_0^y$ is the linear projection of y_i on x_i. After partialling out, α_0 is the population regression coefficient in the univariate regression of ρ_i^y on ρ_i^d. This is the Frisch-Waugh-Lovell theorem. Thus, $\alpha = \alpha_0$ solves the population equation:

$$\mathbb{E}\left(\rho_i^y - \alpha\rho_i^d\right)\rho_i^d = 0$$

The score associated to this equation is:

$$\psi\left(w_i, \alpha, \eta\right) = \left(y_i - x_i'\pi^y\right) - \alpha\left(d_i - x_i'\pi^d\right)\left(d_i - x_i'\pi^d\right), \quad \eta = \left(\pi^{y'}, \pi^{d'}\right)',$$
$$\psi\left(w_i, \alpha_0, \eta_0\right) = \left(\rho_i^y - \alpha\rho_i^d\right)\rho_i^d, \quad \eta_0 = \left(\pi_0^{y'}, \pi_0^{d'}\right).$$

It is straightforward to check that this score obeys the orthogonality principle; moreover, this score is the semi-parametrically efficient score for estimating the regression coefficient α_0.

In low-dimensional settings, the empirical version of the partialling out approach is simply another way to do the least squares. Let's verify this in an example. First, we generate some data

```
dta = get_data("seed_300")
n, p = size(dta);
y = dta[:,"y"];
d = dta[:,"d"];
x = dta[:,3:end];
```

We can estimate α_0 by running full least squares:

```
full_fit = lm(hcat(ones(length(y)), Matrix(dta[:,2:end])), y);
DataFrame(
    Estimate = coef(full_fit)[2],
    Std_Error = stderror(full_fit)[2])
```

```
1×2 DataFrame
 Row   Estimate  Std_Error
       Float64   Float64
```

```
   1   0.978075   0.0137122
```

Another way to estimate $alpha_0$ is to first partial out the x-variables from y_i and d_i, and run least squares on the residuals:

```
rY_1 = lm(hcat(ones(length(y)), Matrix(dta[:,3:end])), y);
rY = y - predict(rY_1)
rD_1 = lm(hcat(ones(length(y)), Matrix(dta[:,3:end])), d);
rD = d - predict(rD_1);
partial_fit_ls = lm(hcat(ones(length(y)), rD), rY)
DataFrame(
    Estimate = coef(partial_fit_ls)[2],
    Std_Error = stderror(partial_fit_ls)[2]
    )
```

```
1×2 DataFrame
 Row   Estimate   Std_Error
       Float64    Float64

   1   0.978075   0.0136862
```

One can see that the estimates are identical, while standard errors are nearly identical. In fact, the standard errors are asymptotically equivalent due to the orthogonality property of the estimating equations associated with the partialling out approach.

In high-dimensional settings, we can no longer rely on the full least-squares and instead may rely on Lasso or Post-Lasso for partialling out. This amounts to using orthogonal estimating equations, where we estimate the nuisance parameters by Lasso or Post-Lasso. Let's try this in the above example, using Post-Lasso for partialling out:

```
rY_1 = rlasso(hcat(ones(length(y)), Matrix(dta[:,3:end])), y);
rY = rY_1["residuals"]
rD_1 = rlasso(hcat(ones(length(y)), Matrix(dta[:,3:end])), d);
rD = rD_1["residuals"]
partial_fit_postlasso = lm(hcat(ones(length(y)), rD), vec(rY))
DataFrame(
    Estimate = coef(partial_fit_postlasso)[2],
    Std_Error = stderror(partial_fit_postlasso)[2]
    )
```

```
1×2 DataFrame
 Row   Estimate   Std_Error
       Float64    Float64

   1   0.972739   0.0136868
```

We see that this estimate and standard errors are nearly identical to the previous estimates given above. In fact they are asymptotically equivalent to the previous estimates in the low-dimensional settings, with the advantage that, unlike the previous estimates, they will continue to be valid in the high-dimensional settings.

The orthogonal estimating equations method ' based on partialling out via Lasso or post-Lasso ' is implemented by the function `rlassoEffect`, using `method= "partialling out"`:

```
Eff = rlassoEffect(x, y, d, method = "partialling out");
r_summary(Eff);
```

```
Estimates and significance testing of the effect of target variables
  Row   Estimate.   Std. Error    t value   Pr(>|t|)

    1     0.97274      0.01369   71.05478    0.0 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Another orthogonal estimating equations method ' based on the double selection of covariates ' is implemented by the the function `rlassoEffect`, using `method= "double selection"`. Algorithmically the method is as follows:

1. Select controls x_{ij}'s that predict y_i by Lasso.
2. Select controls x_{ij}'s that predict d_i by Lasso.
3. Run OLS of y_i on d_i and the union of controls selected in steps 1 and 2.

```
Eff = rlassoEffect(x, y, d, method = "double selection");
r_summary(Eff);
```

```
Estimates and significance testing of the effect of target variables
  Row   Estimate.   Std. Error    t value   Pr(>|t|)

    1     0.97807      0.01416   69.07274    0.0 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4.2 Inference: Confidence Intervals and Significance Testing.

The function `rlassoEffects` does inference ' namely construction of confidence intervals and significance testing ' for target variables. Those can be specified either by the variable names, an integer valued vector giving their position in x or by a logical vector indicating the variables for which inference should be conducted. It returns an object of S3 class `rlassoEffects` for which the methods `r_summary`, `r_print`, and `r_confint` are provided. `rlassoEffects` is a wrap function for the function `rlassoEffect` which does inference for a single target regressor. The theoretical underpinning is given in Belloni, Chernozhukov, and Hansen (2014) and for a more general class of models in Belloni, Chernozhukov, and Kato (2014). The function `rlassoEffects` might either be used in the form `rlassoEffects(x, y, index)` where `x` is a matrix, `y` denotes the outcome variable and index specifies the variables of `x` for which inference is conducted. This can done by an integer vector (postion of the variables), a logical vector or the name of the variables. An alternative usage is as `rlassoEffects(formula, data, I)` where I is a one-sided formula which specifies the variables for which is inference is conducted. For further details we refer to the help page of the function and the following examples where both methods for usage are shown.

Here is an example of the use.

```
data = get_data("seed_400")
n, p = size(data);
```

```
y = data[:,1];
#d = dta[:,"d"];
x = data[:,2:end];
```

We can do inference on a set of variables of interest, e.g. the first, second, third, and the fiftieth:

```
lasso_effects = rlassoEffects(x, y, index = [1,2,3,50]);
```

```
r_print(lasso_effects, digits = 4)
```

```
Coefficients:

     X1          X2          X3          X50

   2.9445      3.0413      2.9754       0.072
```

```
r_summary(lasso_effects);
```

```
Estimates and significance testing of the effect of target variables
                    Estimate.   Std. Error    t value   Pr(>|t|)

   X1     2.94448     2.94448      0.08815   33.40306    0.0 ***
   X2     3.04127     3.04127      0.08389   36.25307    0.0 ***
   X3      2.9754      2.9754      0.07804    38.1266    0.0 ***
  X50   0.0719553     0.07196      0.07765    0.92672   0.35407
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
r_confint(lasso_effects);
```

```
2.5%      97.5%

   X1      2.77171    3.11724
   X2      2.87685     3.2057
   X3      2.82245    3.12836
  X50   -0.0802271   0.224138
```

The two methods are first-order equivalent in both low-dimensional and high-dimensional settings under regularity conditions. Not surprisingly we see that in the numerical example of this section, the estimates and standard errors produced by the two methods are very close to each other.

It is also possible to estimate joint confidence intervals. The method relies on a multiplier bootstrap method as described in Belloni, Chernozhukov, and Kato (2014). Joint confidence intervals can be invoked by setting the option joint to `true` in the method `confint` for objects of class `rlassoEffects`.

```
r_confint(lasso_effects, joint = true);
```

```
2.5%     97.5%
```

```
   X1      2.72818   3.16077
   X2      2.83734   3.24521
   X3      2.78352   3.16728
  X50   -0.115249   0.25916
```

We will also demonstrate the application of joint confidence intervals in an empirical application in the next section.

For logistic regression we provide the functions `rlassologit` and `rlassologitEffects`.

4.3 Application: the effect of gender on wage.

In Labour Economics an important question is how the wage is related to the gender of the employed. We use US census data from the year 2012 to analyse the effect of gender and interaction effects of other variables with gender on wage jointly. The dependent variable is the logarithm of the wage, the target variable is `female` (in combination with other variables). All other variables denote some other socio-economic characteristics, e.g. marital status, education, and experience. For a detailed description of the variables we refer to the help page.

First, we load and prepare the data.

```
cps2012 = get_data("cps2012")
n, p = size(cps2012);
size(cps2012)

## Two parts X <- model.matrix(~-1 + female + female:(widowed + divorced + separated


x_formula1 = @formula(
    lnw ~ female + female &
        (widowed + divorced + separated + nevermarried +
        hsd08 + hsd911 + hsg + cg + ad + mw + so + we +
        exp1 + exp2 + exp3)
    )
y, x1 = data_formula(x_formula1, cps2012);

# (widowed + divorced + separated + nevermarried + hsd08 + hsd911 + hsg + cg + ad + m
# we + exp1 + exp2 + exp3)^2

x0 = data_formula(x_names = ["widowed", "divorced",  "separated",
    "nevermarried", "hsd08", "hsd911", "hsg", "cg", "ad", "mw", "so",
    "we", "exp1", "exp2", "exp3"],
    y_name = "lnw", Data = cps2012)


x = hcat(x1, x0);

index_gender = []
female = findfirst.("female", names(x))
for i in eachindex(female)
```

```
    if  isnothing(female[i])
        continue
    else
        append!(index_gender, i)
    end
end
y = cps2012.lnw;
```

```
["female", "female & widowed", "female & divorced", "female & separated", "
female & nevermarried", "female & hsd08", "female & hsd911", "female & hsg"
, "female & cg", "female & ad", "female & mw", "female & so", "female & we"
, "female & exp1", "female & exp2", "female & exp3"]
```

The parameter estimates for the target parameters, i.e. all coefficients related to gender (i.e. by interaction with other variables) are calculated and summarized by the following commands

```
effects_female = rlassoEffects(x, y, index = index_gender);
r_summary(effects_female);
```

```
Estimates and significance testing of the effect of target variables
                                    Estimate.   Std. Error    t value
    Pr(>|t|)

                female     -0.177119    -0.17712       0.0363   -4.87934
     0.0 ***
       female & widowed      0.136095      0.1361      0.09066    1.50121
     0.1333
        female & divorced      0.136939     0.13694      0.02218    6.17403
     0.0 ***
     female & separated     0.0233028      0.0233      0.05321    0.43789
    0.66147
  female & nevermarried      0.186853     0.18685      0.01994    9.37061
     0.0 ***
         female & hsd08     0.0278103     0.02781      0.12091    0.23001
    0.81808
        female & hsd911     -0.114916    -0.11492      0.05151   -2.23102
   0.02568 *
           female & hsg     -0.012445    -0.01245      0.01922   -0.64776
    0.51714
            female & cg     0.0101386     0.01014      0.01833    0.55319
    0.58013
            female & ad    -0.0304637    -0.03046      0.02181   -1.39661
    0.16253
            female & mw   -0.00214701    -0.00215      0.01918    -0.1121
    0.91074
            female & so   -0.00818334    -0.00818      0.01936   -0.42252
    0.67265
            female & we   -0.00422613    -0.00423      0.02117   -0.19981
    0.84163
```

```
          female & exp1    0.00493526     0.00494       0.0078    0.63333
    0.52652
          female & exp2     -0.159519    -0.15952       0.0453   -3.52141
 0.00043 ***
          female & exp3     0.0384506     0.03845      0.00786    4.89186
     0.0 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Finally, we estimate and plot confident intervals, first "pointwise" and then the joint confidence intervals.

```
joint_CI = r_confint(effects_female, 0.95, joint = true);
joint_CI;
```

```
2.5%        97.5%

                 female    -0.280849   -0.0733889
       female & widowed    -0.141442     0.413633
      female & divorced    0.0730069     0.200872
     female & separated     -0.12075     0.167356
  female & nevermarried     0.127599     0.246108
         female & hsd08    -0.386091     0.441712
        female & hsd911    -0.268431    0.0385989
           female & hsg   -0.0662278    0.0413378
            female & cg   -0.0431745    0.0634516
            female & ad   -0.0977201    0.0367926
            female & mw   -0.0572478    0.0529538
            female & so   -0.0642917    0.0479251
            female & we   -0.0673975    0.0589452
          female & exp1   -0.0171127    0.0269832
          female & exp2    -0.287735   -0.0313041
          female & exp3    0.0162255    0.0606757
```

This analysis allows a closer look how discrimination according to gender is related to other socio- economic variables.

4.4 Application: Estimation of the treatment effect in a linear model with many confounding factors.

A part of empirical growth literature has focused on estimating the effect of an initial (lagged) level of GDP (Gross Domestic Product) per capita on the growth rates of GDP per capita. In particular, a key prediction from the classical Solow-Swan-Ramsey growth model is the hypothesis of convergence, which states that poorer countries should typically grow faster and therefore should tend to catch up with the richer countries, conditional on a set of institutional and societal characteristics. Covariates that describe such characteristics include variables measuring education and science policies, strength of market institutions, trade openness, savings rates and others.

Thus, we are interested in a specification of the form:

$$y_i = \alpha_0 d_i + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i,$$

where y_i is the growth rate of GDP over a specified decade in country i, d_i is the log of the initial level of GDP at the beginning of the specified period, and the $x_{ij}'s$ form a long list of country i's characteristics at the beginning of the specified period. We are interested in testing the hypothesis of convergence, namely that $\alpha_1 < 0$. Given that in the Barro and Lee (1994) data, the number of covariates p is large relative to the sample size n, covariate selection becomes a crucial issue in this analysis. We employ here the partialling out approach (as well as the double selection version) introduced in the previous section.

First, we load and prepare the data

```
GrowthData = get_data("GrowthData")
y = GrowthData[:, 1];
d = GrowthData[:, 3:3];
X = Matrix(GrowthData[:, Not(1, 2, 3)]);
X_1 = Matrix(GrowthData[:, Not(1, 2)]);
```

Now we can estimate the effect of the initial GDP level. First, we estimate by OLS:

```
Q, R = qr(hcat(ones(length(y)), X_1))
beta_i = pinv(hcat(ones(length(y)), X_1)) * y

res = y - hcat(ones(length(y)), X_1) * beta_i;
n = size(hcat(ones(length(y)), X_1))[1]
k = size(hcat(ones(length(y)), X_1))[2]

sigma2_hat = (res' * res) / (n - k)
vcov_beta_hat = sigma2_hat .* inv(
    hcat(ones(length(y)), X_1)' * hcat(ones(length(y)), X_1)
    );
se = sqrt.(diag(vcov_beta_hat))

ls_effect = DataFrame(Estimate = beta_i, stderror = se);
```

Second, we estimate the effect by the partialling out by Post-Lasso:

```
lasso_effect = rlassoEffect(X, y, d, method = "partialling out");
r_summary(lasso_effect);
```

```
Estimates and significance testing of the effect of target variables
  Row   Estimate.   Std. Error    t value      Pr(>|t|)

    1    -0.04981      0.01394   -3.57317   0.00035 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Third, we estimate the effect by the double selection method:

```
double_effect = rlassoEffect(X, y, d, method = "double selection");
r_summary(double_effect);
```

```
Estimates and significance testing of the effect of target variables
             Estimate.   Std. Error    t value     Pr(>|t|)

  gdpsh465    -0.05001      0.01579   -3.16719   0.00154 **
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We then collect results in a nice table:

```
table = zeros(3,2)
table[1,:] = [
    round(Matrix(ls_effect)[2,1], digits = 2),
    round.(Matrix(ls_effect)[2,2], digits = 5)
    ]
table[2,:] = [
    round(lasso_effect.coefficients, digits =2),
    round(lasso_effect.se, digits = 5)
    ]
table[3,:] = [
    round(double_effect.coefficients, digits =2),
    round(double_effect.se, digits = 5)
    ];
index = ["full reg via ols", "partial reg
via post-lasso ", "partial reg via double selection"]
pretty_table(
    hcat(index, table), show_row_number = false,
    header = [" ", "Estimate", "Std. Error"],
    tf = tf_simple, nosubheader = true
    )
```

```
================================== ========== =============
                                     Estimate   Std. Error
================================== ========== =============
                 full reg via ols      -0.01      0.02989
      partial reg\nvia post-lasso      -0.05      0.01394
  partial reg via double selection     -0.05      0.01579
================================== ========== =============
```

We see that the OLS estimates are noisy, which is not surprising given that p is comparable to n. The partial regression approaches, based on Lasso selection of covariates in the two projection equations, in contrast yields much more precise estimates, which does support the hypothesis of conditional convergence. We note that the partial regression approaches represent a special case of the orthogonal estimating equations approach.

5 Instrumental Variable Estimation in a High-Dimensional Setting

In many applied settings the researcher is interested in estimating the (structural) effect of a variable (treatment variable), but this variable is endogenous, i.e. correlated with the error term. In this case, instrumental variables (IV) methods are used for identification of the causal parameters.

We consider the linear instrumental variables model:

$$y_i = \alpha_0 d_i + \gamma_0 x_i' + \epsilon_i,$$

$$d_i = z_i' \prod + \beta_0 x_i' + \nu_i,$$

where $\mathbb{E}\left[\epsilon_i(x_i', z_i')\right] = 0$, $\mathbb{E}\left[\nu(x_i', z_i')\right] = 0$, but $\mathbb{E}[\epsilon_i, \nu_i] \neq 0$ leading to endogeneity. In this setting d_i is a scalar endogenous variable of interest, z_i is a pz-dimensional vector of instruments and x_i is a px-dimensional vector of control variables.

In this section we present methods to estimate the effect α_0 in a setting where either x is high-dimensional or z is high-dimensional. Instrumental variables estimation with very many instruments was analysed in Belloni, Chen, Chernozhukov, and Hansen (2012), the extension with many instruments and many controls in Chernozhukov, Hansen, and Spindler (2015b).

5.1 Estimation and Inference.

To get efficient estimators and uniformly valid confidence intervals for the structural parameters there are different strategies which are asymptotically equivalent where again orthogonalization (via partialling out) is a key concept.

In the case of the high-dimensional instrument z_i and low-dimensional x_i. We predict the endogenous variable d_i using (Post-)Lasso regression of d_i on the instruments z_i and x_i, forcing the inclusion of x_i. Then we compute the IV estimator (2SLS) $\hat{\alpha}$ of the parameter α_0 using the predicted value $\hat{d}_i$ as instrument and using x_i's as controls. We then perform inference on α_0 using $\hat{\alpha}$ and conventional heteroskedasticity robust standard errors.

In the case of the low-dimensional instrument z_i and high-dimensional x_i, we first partial out the effect of x_i from d_i, y_i, and z_i by (Post-)Lasso. Second, we then use the residuals to compute the IV estimator (2SLS) $\hat{\alpha}$ of the parameter 0. We then perform inference on α_0 using $\hat{\alpha}$ and conventional heteroskedasticity robust standard errors.

In the case of the high-dimensional instrument z_i and high-dimensional x_i the algorithm described in Chernozhukov, Hansen, and Spindler (2015b) is adopted.

Julia Implementation. The wrap function `rlassoIV` handles all of the previous cases. It has the options **select.X** and **select.Z** which implement selection of either covariates or instruments, both with default values set to **true**. The class of the return object depends on the chosen options, but the methods summary, print and confint are available for all. The functions **rlassoSelectX** and **rlassoSelectZ** do selection on x-variables only and z-variables only. Selection on both is done in **rlassoIV**. All functions employ the option **post=true** as default, which uses post-Lasso for partialling out. By setting **post=true** we can employ

Lasso instead of Post-Lasso. Finally, the package provides the standard function **tsls**, which implements the "classical" two-stage least squares estimation.

Function usage both the family of **rlassoIV**-functions and the family of the functions for treatment effects , which are introduced in the next section, allow use with both formula-interface and by handing over the prepard model matrices. Hence the general pattern for use with formula is function(formula, data, ...) where formula consists of two-parts and is a member of the class Formula. These formulas are of the pattern y d + x | x + z where y is the outcome variable, x are exogenous variables, d endogenous varialbes (if several ones are allowed depends on the concrete function), and z denote the instrumental variables. A more primitive use of the functions is by simply hand over the required group of variables as matrices: function(x=x, d=d, y=y, z=z). In some of the following examples both alternatives are displayed.

5.2 Application: Economic Development and Institutions.

Estimating the causal effect of institutions on output is complicated by the simultaneity between institutions and output: specifically, better institutions may lead to higher incomes, but higher incomes may also lead to the development of better institutions. To help overcome this simultaneity, Acemoglu, Johnson, and Robinson (2001) use mortality rates for early European settlers as an instrument for institution quality. The validity of this instrument hinges on the argument that settlers set up better institutions in places where they are more likely to establish long-term settlements, that where they are likely to settle for the long term is related to settler mortality at the time of initial colonization, and that institutions are highly persistent. The exclusion restriction for the instrumental variable is then motivated by the argument that GDP, while persistent, is unlikely to be strongly influenced by mortality in the previous century, or earlier, except through institutions.

In this application, we consider the problem of selecting controls. The input raw controls are the Latitude and the continental dummies. The technical controls can include various polynomial transformations of the Latitude, possibly interacted with the continental dummies. Such flexible specifications of raw controls results in the high-dimensional x, with dimension comparable to the sample size.

First, we process the data

```
AJR = get_data("AJR")
y = AJR[!,"GDP"]
d = AJR[!, 2]
z = AJR[!,"logMort"];
x_formula = @formula(GDP ~ -1 + Latitude + Latitude2 + Africa + Asia +
    Namer + Samer + Latitude*Latitude2 + Latitude*Africa +
    Latitude*Asia + Latitude*Namer + Latitude*Samer + Latitude2*Africa +
    Latitude2*Asia + Latitude2*Namer + Latitude2*Samer + Africa*Asia
    + Africa*Namer + Africa*Samer + Asia*Namer + Asia*Samer
    + Namer*Samer
    )
y, x = data_formula(x_formula, AJR)
size(x)
```

```
Any["Latitude", "Latitude2", "Africa", "Asia", "Namer", "Samer", "Latitude
& Latitude2", "Latitude & Africa", "Latitude & Asia", "Latitude & Namer", "
```

```
Latitude & Samer", "Latitude2 & Africa", "Latitude2 & Asia", "Latitude2 & N
amer", "Latitude2 & Samer", "Africa & Asia", "Africa & Namer", "Africa & Sa
mer", "Asia & Namer", "Asia & Samer", "Namer & Samer"]
(64, 21)
```

Then we estimate an IV model with selection on the X

```
AJR_Xselect  = rlassoIV(x, d, y, z, select_X=true, select_Z=false);
r_summary(AJR_Xselect);
```

```
Estimates and Significance Testing of the effect of target variables in the
 IV regression model
       coeff.       se.   t-value      p-value

  d1   0.84503   0.26993   3.13055   0.00174 **
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
r_confint(AJR_Xselect);
```

```
2.5%     97.5%

  d1   0.315981   1.37407
```

It is interesting to understand what the procedure above is doing. In essence, it partials out x_i from y_i, d_i and z_i using Post-Lasso and applies the 2SLS to the residual quantities. Let us investigate partialling out in more detail in this example. We can first try to use OLS for partialling out:

```
rY_1 = lm(@formula(
    GDP ~ Latitude + Latitude2 + Africa + Asia + Namer + Samer +
    Latitude*Latitude2 + Latitude*Africa + Latitude*Asia +
    Latitude*Namer + Latitude*Samer + Latitude2*Africa +
    Latitude2*Asia + Latitude2*Namer + Latitude2*Samer +
    Africa*Asia + Africa*Namer + Africa*Samer + Asia*Namer +
    Asia*Samer + Namer*Samer), AJR
    )
rY = y - predict(rY_1)

rD_1 = lm(@formula(
    Exprop ~ Latitude + Latitude2 + Africa + Asia + Namer + Samer +
    Latitude*Latitude2 + Latitude*Africa + Latitude*Asia +
    Latitude*Namer + Latitude*Samer + Latitude2*Africa +
    Latitude2*Asia + Latitude2*Namer + Latitude2*Samer +
    Africa*Asia + Africa*Namer + Africa*Samer + Asia*Namer +
    Asia*Samer + Namer*Samer), AJR
    )
rD = d - predict(rD_1)

rZ_1 = lm(@formula(
    logMort ~ Latitude + Latitude2 + Africa + Asia + Namer + Samer +
```

```
    Latitude*Latitude2 + Latitude*Africa + Latitude*Asia +
    Latitude*Namer + Latitude*Samer + Latitude2*Africa +
    Latitude2*Asia + Latitude2*Namer + Latitude2*Samer +
    Africa*Asia + Africa*Namer + Africa*Samer + Asia*Namer +
    Asia*Samer + Namer*Samer), AJR
    )
rZ = z - predict(rZ_1);
```

```
ivfit_lm = tsls(rD, rY, rZ, nothing, intercept=false)
DataFrame(Estimate = ivfit_lm["coefficients"][1,2],
    Std_Error = ivfit_lm["se"])
```

```
1×2 DataFrame
 Row   Estimate  Std_Error
       Float64   Float64

   1    1.26721    1.73054
```

We see that the estimates exhibit large standard errors. The imprecision is expected because dimension of x is quite large, comparable to the sample size.

Next, we replace the OLS operator by post-Lasso for partialling out

```
x_formula1 = @formula(GDP ~ Latitude + Latitude2 + Africa + Asia +
    Namer + Samer + Latitude*Latitude2 + Latitude*Africa +
    Latitude*Asia + Latitude*Namer + Latitude*Samer +
    Latitude2*Africa + Latitude2*Asia + Latitude2*Namer +
    Latitude2*Samer + Africa*Asia + Africa*Namer + Africa*Samer
    + Asia*Namer + Asia*Samer + Namer*Samer
    )


y, xx = data_formula(x_formula1, AJR)

rY_1 = rlasso(xx, y);
rY = rY_1["residuals"]
rD_1 = rlasso(xx, d);
rD = rD_1["residuals"]
rZ_1 = rlasso(xx, z);
rZ = rZ_1["residuals"]

ivfit_lasso = tsls(rD, rY, rZ)
DataFrame(
    Estimate = ivfit_lasso["coefficients"][1,2],
    Std_Error = ivfit_lasso["se"][1]
    )
```

```
["Latitude", "Latitude2", "Africa", "Asia", "Namer", "Samer", "Latitude & L
atitude2", "Latitude & Africa", "Latitude & Asia", "Latitude & Namer", "Lat
itude & Samer", "Latitude2 & Africa", "Latitude2 & Asia", "Latitude2 & Name
r", "Latitude2 & Samer", "Africa & Asia", "Africa & Namer", "Africa & Samer
```

```
", "Asia & Namer", "Asia & Samer", "Namer & Samer"]
1×2 DataFrame
 Row   Estimate  Std_Error
       Float64   Float64

   1   0.845027   0.272094
```

This is exactly what command `rlassoIV` is doing by calling the command `rlassoSelectX`, so the numbers we see are identical to those reported first. In comparison to OLS results, we see precision is improved by doing selection on the exogenous variables.

5.3 Application: Impact of Eminent Domain Decisions on Economic Outcomes.

Here we investigate the effect of pro-plaintiff decisions in cases of eminent domain (government's takings of private property) on economic outcomes. The analysis of the effects of such decisions is complicated by the possible endogeneity between judicial decisions and potential economic outcomes. To address the potential endogeneity, we employ an instrumental variables strategy based on the random assignment of judges to the federal appellate panels that make the decisions. Because judges are randomly assigned to three-judge panels, the exact identity of the judges and their demographics are randomly assigned conditional on the distribution of characteristics of federal circuit court judges in a given circuit-year.

Under this random assignment, the characteristics of judges serving on federal appellate panels can only be related to property prices through the judges' decisions; thus the judge's characteristics will plausibly satisfy the instrumental variable exclusion restriction. For further information on this application and the data set we refer to Chen and Yeh (2010) and Belloni, Chen, Chernozhukov, and Hansen (2012).

First, we load the data a construct the matrices with the controls (x), instruments (z), outcome (y),and treatment variables (d). Here we consider regional GDP as the outcome variable.

```
EminentDomain = get_data("EminentDomain")
z = EminentDomain["logGDP"]["z"];
x = EminentDomain["logGDP"]["x"];
d = EminentDomain["logGDP"]["d"];
y = EminentDomain["logGDP"]["y"];
x = x[:, (mean(x, dims = 1) .> 0.05)'];
z = z[:, (mean(z, dims = 1) .> 0.05)'];
```

As mentioned above, y is the economic outcome, the logarithm of the GDP, d the number of pro plaintiff appellate takings decisions in federal circuit court c and year t, x is a matrix with control variables, and z is the matrix with instruments. Here we consider socio-economic and demographic characteristics of the judges as instruments.

First, we estimate the effect of the treatment variable by simple OLS and 2SLS using two instruments:

```
ED_ols = lm(hcat(ones(length(vec(y))), hcat(d, x)), vec(y));
ED_2sls = tsls(d, y, z[:,1:2], x, intercept = false);
```

Next, we estimate the model with selection on the instruments

```
lasso_IV_Z = rlassoIV(x, d, y, z, select_X = false, select_Z = true);
r_summary(lasso_IV_Z);
```

```
Estimates and Significance Testing of the effect of target variables in the
 IV regression model
      coeff.       se.   t-value    p-value

  d1   0.4146   0.29025   1.42842   0.15317
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
r_confint(lasso_IV_Z);
```

```
2.5%     97.5%

  d1   -0.154276   0.98348
```

Finally, we do selection on both the x and z variables.

```
lasso_IV_XZ = rlassoIV(x, d, y, z, select_X = true, select_Z = true);
r_summary(lasso_IV_XZ)
```

```
Estimates and Significance Testing of the effect of target variables in the
 IV regression model
        coeff.       se.    t-value    p-value

  d1   -0.02383   0.12851   -0.18543   0.85289
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
1×5 DataFrame
 Row        coeff.    se.       t-value   p-value
       Any  Any       Any       Any       Any

   1   d1   -0.02383  0.12851  -0.18543  0.85289
```

```
r_confint(lasso_IV_XZ);
```

```
2.5%      97.5%

  d1   -0.275703   0.228033
```

Comparing the results we see, that the OLS estimates indicate that the influence of pro plaintiff appellate takings decisions in federal circuit court is significantly positive, but the 2SLS estimates which account for the potential endogeneity render the results insignificant. Employing selection on the instruments yields similar results. Doing selection on both the x- and z-variables results in extremely imprecise estimates.

Finally, we compare all results

```
table = zeros(4,2)
table[1,:] = [coef(ED_ols)[2], stderror(ED_ols)[2]];
```

```
table[2,:] = [ED_2sls["coefficients"][1,2], ED_2sls["se"][1]];
table[3,:] = Matrix(r_summary(lasso_IV_Z)[:,2:3]);
table[4, :] = Matrix(r_summary(lasso_IV_XZ)[:, 2:3]);
index = [
    "ols regression", "IV estimation ",
    "selection on Z", "selection on X and Z"
    ]
pretty_table(
    hcat(index, table),
    show_row_number = false, header = [" ", "Estimate", "Std. Error"],
    tf = tf_simple, nosubheader = true)
```

```
Estimates and Significance Testing of the effect of target variables in the
 IV regression model
       coeff.       se.    t-value    p-value

  d1   0.4146   0.29025   1.42842   0.15317
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Estimates and Significance Testing of the effect of target variables in the
 IV regression model
         coeff.       se.     t-value    p-value

  d1   -0.02383   0.12851   -0.18543   0.85289
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
======================== ============ =============
                           Estimate   Std. Error
======================== ============ =============
        ols regression   0.00786473   0.00986593
        IV estimation    -0.0107097    0.0337652
        selection on Z       0.4146      0.29025
  selection on X and Z     -0.02383      0.12851
======================== ============ =============
```

6 Inference on Treatment Effects in a High-Dimensional Setting

In this section, we consider estimation and inference on treatment effects when the treatment variable d enters non-separably in determination of the outcomes. This case is more complicated than the additive case, which is covered as a special case of Section 3. However, the same underlying principle ' the orthogonality principle ' applies for the estimation and inference on the treatment effect parameters. Estimation and inference of treatment effects in a high-dimensional setting is analysed in Belloni, Chernozhukov, Fern´andez-Val, and Hansen (2013).

6.1 Treatment Effects Parameters ' a short Introduction.

In many situations researchers are asked to evaluate the effect of a policy intervention. Examples are the effectiveness of a job related training program or the effect of a newly developed drug. We consider n units or individuals, $i = 1, ..., n$. For each individual we observe the treatment status. The treatment variable D_i takes the value 1, if the unit received (active) treatment, and 0, if it received the control treatment. For each individual we observe the outcome for only one of the two potential treatment states. Hence, the observed outcome depends on the treatment status and is denoted by $Y_i(D_i)$. One important parameter of interest is the average treatment effect (ATE):

$$\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)].$$

This quantity can be interpreted as the average effect of the policy intervention.

Researchers might also be interested in the average treatment effect on the treated (ATET) given by

$$\mathbb{E}[Y(1) - Y(0)|D = 1] = \mathbb{E}[Y(1)|D = 1] - \mathbb{E}[Y(0)|D = 1].$$

This is the average treatment effect restricted to the population the treated individuals. When treatment D is randomly assigned conditional on confounding factors X, the ATE and ATET can be identified by regression or propensity score weighting methods. Our identification and estimation method rely on the combination of two methods to create orthogonal estimating equations for these parameters.

In observational studies, the potential treatments are endogenous, i.e. jointly determined with the outcome variable. In such cases, causal effects may be identified with the use of a binary instrument Z that affects the treatment but is independent of the potential outcomes. An important parameter in this case is the local average treatment effect (LATE):

$$E[Y(1) - Y(0)|D(1) > D(0)].$$

The random variables $D(1)$ and $D(0)$ indicate the potential participation decisions under the instrument states 1 and 0. LATE is the average treatment effect for the subpopulation of compliers ' those units that respond to the change in the instrument. Another parameter of interest is the local average treatment effect of the treated (LATET):

$$\mathbb{E}\left[Y(1) - Y(0)|D(1) > D(0), D = 1\right],$$

which is the average effect for the subpopulation of the treated compliers.

When the instrument Z is randomly assigned conditional on confounding factors X, the LATE and LATET can be identified by instrumental variables regression or propensity score weighting methods. Our identification and estimation method rely on the combination of two methods to create orthogonal estimating equations for these parameters.[1]

[1]It turns out that the orthogonal estimating equations are the same as doubly robust estimating equations, but emphasizing the name "doubly robust" could be confusing in the present context.

6.2 Estimation and Inference of Treatment effects.

We consider the estimation of the effect of an endogenous binary treatment, D, on an outcome variable, Y , in a setting with very many potential control variables. In the case of endogeneity, the presence of a binary instrumental variable, Z, is required for the estimation of the LATE and LATET.

When trying to estimate treatment effects, the researcher has to decide what conditioning variables to include. In the case of a non-randomly assigned treatment or instrumental variable, the researcher must select the conditioning variables so that the instrument or treatment is plausibly exogenous. Even in the case of random assignment, for a precise estimation of the policy variable selection of control variables is necessary to absorb residual variation, but overfitting should be avoided. For uniformly valid post-selection inference, "orthogonal"estimating equations as described above are they key to the efficient estimation and valid inference. We refer to Belloni, Chernozhukov, Fernandez-Val, and Hansen (2013) for details.

6.3 Application: 401(k) plan participation.

Though it is clear that 401(k) plans are widely used as vehicles for retirement saving, their effect on assets is less clear. The key problem in determining the effect of participation in 401(k) plans on accumulated assets is saver heterogeneity coupled with nonrandom selection into participation states. In particular, it is generally recognized that some people have a higher preference for saving than others. Thus, it seems likely that those individuals with the highest unobserved preference for saving would be most likely to choose to participate in tax-advantaged retirement savings plans and would also have higher savings in other assets than individuals with lower unobserved saving propensity. This implies that conventional estimates that do not allow for saver heterogeneity and selection of the participation state will be biased upward, tending to overstate the actual savings effects of 401(k) and IRA participation.

Again, we start first with the data preparation:

```
pension = get_data("pension")
y = pension[:, "tw"];
d = pension[:, "p401"];
z = pension[:, "e401"];
X = pension[:,
        ["i2", "i3", "i4", "i5", "i6", "i7", "a2", "a3",
        "a4", "a5", "fsize", "hs", "smcol", "col", "marr",
        "twoearn", "db", "pira", "hown"]];
```

Now we can compute the estimates of the target treatment effect parameters. For ATE and ATET we report the the effect of eligibility for 401(k)

```
pension_ate = rlassoATE(X, d, y);
r_summary(pension_ate);
```

```
Estimation and significance tesing of the treatment effect
    Type: ATE
    Bootstrap: none
```

```
========== ========= ==========
  Coeff      SE         t.value
========== ========= ==========
  10180.1    1930.68    5.2728
========== ========= ==========
```

```
pension_atet = rlassoATET(X, d, y);
r_summary(pension_atet);
```

```
Estimation and significance tesing of the treatment effect
    Type: ATET
    Bootstrap: none

========== ========= ==========
  Coeff      SE         t.value
========== ========= ==========
  12628.5    2944.43    4.28893
========== ========= ==========
```

For LATE and LATET we estimate the effect of 401(k) participation (d) with plan eligibility (z) as instrument.

```
pension_late = rlassoLATE(X, d, y, z);
r_summary(pension_late);
```

```
Estimation and significance tesing of the treatment effect
    Type: LATE
    Bootstrap: none

========== ======== ==========
  Coeff      SE       t.value
========== ======== ==========
  12992.1    2326.9   5.58344
========== ======== ==========
```

```
pension_latet = rlassoLATET(X, d, y, z);
r_summary(pension_latet);
```

```
Estimation and significance tesing of the treatment effect
    Type: LATET
    Bootstrap: none

========== ========= ==========
  Coeff      SE         t.value
========== ========= ==========
  15323.2    3645.28    4.20357
========== ========= ==========
```

The results are summarized into a table:

```
using PrettyTables
table = zeros(4,2)
table[1,:] = round.(vec(r_summary(pension_ate)[:, 1:2]), digits = 2);
```

```
table[2,:] = round.(vec(r_summary(pension_atet)[:, 1:2]), digits = 2);
table[3,:] = round.(vec(r_summary(pension_late)[:, 1:2]), digits = 2);
table[4,:] = round.(vec(r_summary(pension_latet)[:, 1:2]), digits = 2);
index = ["ATE", "ATET ", "LATE", "LATET"];
```

```
Estimation and significance tesing of the treatment effect
    Type: ATE
    Bootstrap: none

========== ========= ==========
  Coeff      SE         t.value
========== ========= ==========
  10180.1    1930.68    5.2728
========== ========= ==========

    Estimation and significance tesing of the treatment effect
    Type: ATET
    Bootstrap: none

========== ========= ==========
  Coeff      SE         t.value
========== ========= ==========
  12628.5    2944.43    4.28893
========== ========= ==========

    Estimation and significance tesing of the treatment effect
    Type: LATE
    Bootstrap: none

========== ======== ==========
  Coeff      SE        t.value
========== ======== ==========
  12992.1    2326.9    5.58344
========== ======== ==========

    Estimation and significance tesing of the treatment effect
    Type: LATET
    Bootstrap: none

========== ========= ==========
  Coeff      SE         t.value
========== ========= ==========
  15323.2    3645.28    4.20357
========== ========= ==========
```

```
pretty_table(
    hcat(index, table), show_row_number = false,
    header = [" ", "Estimate", "Std. Error"],
    tf = tf_simple, nosubheader = true
```

```
    )
```

```
======== ========== =============
          Estimate   Std. Error
======== ========== =============
    ATE    10180.1      1930.68
  ATET     12628.5      2944.43
   LATE    12992.1       2326.9
  LATET    15323.2      3645.28
======== ========== =============
```

Finally, we estimate a model including all interaction effects:

```
pension_ate = rlassoATE(X, z, y);
pension_atet = rlassoATET(X, z, y);
pension_late = rlassoLATE(X, d, y, z);
pension_latet = rlassoLATET(X, d, y, z);
```

```
table = zeros(4, 2)
table[1,:] = r_summary(pension_ate)[:, 1:2];
table[2,:] = r_summary(pension_atet)[:, 1:2];
table[3,:] = r_summary(pension_late)[:, 1:2];
table[4,:] = r_summary(pension_latet)[:, 1:2];
index = ["ATE", "ATET ", "LATE", "LATET"]
pretty_table(
    hcat(index, table), show_row_number = false,
    header = [" ", "Estimate", "Std. Error"],
    tf = tf_simple, nosubheader = true
    )
```

```
Estimation and significance tesing of the treatment effect
    Type: ATE
    Bootstrap: none

========== ========= ==========
  Coeff      SE         t.value
========== ========= ==========
  8491.99    1902.92    4.4626
========== ========= ==========

    Estimation and significance tesing of the treatment effect
    Type: ATET
    Bootstrap: none

========== ========= ==========
  Coeff      SE         t.value
========== ========= ==========
  10795.3    2568.13    4.20357
========== ========= ==========

    Estimation and significance tesing of the treatment effect
```

```
    Type: LATE
    Bootstrap: none

========== ======== ==========
  Coeff     SE       t.value
========== ======== ==========
  12992.1   2326.9   5.58344
========== ======== ==========

    Estimation and significance tesing of the treatment effect
    Type: LATET
    Bootstrap: none

========== ========= ==========
  Coeff     SE        t.value
========== ========= ==========
  15323.2   3645.28   4.20357
========== ========= ==========
======== ========== =============
          Estimate   Std. Error
======== ========== =============
    ATE    8491.99      1902.92
  ATET     10795.3      2568.13
   LATE    12992.1       2326.9
  LATET    15323.2      3645.28
======== ========== =============
```

7 The Lasso Methods for Discovery of Significant Causes amongst Many Potential Causes, with Many Controls

Here we consider the model

$$\underbrace{Y_i}_{\text{Outcome}} = \underbrace{\sum_{l=1}^{p_1} D_{il}\alpha_\ell}_{\text{Causes}} + \underbrace{\sum_{j=1}^{p_2} W_{ij}\beta_j}_{\text{Controls}} + \underbrace{\epsilon_i}_{\text{Noise}}$$

where the number of potential causes p_1 could be very large and the number of controls p_2 could also be very large. The causes are randomly assigned conditional on controls. Under approximate sparsity of $\alpha = (\alpha_l)_{l_1}^{p1}$ and $\beta = (\beta_l)_{l=1}^{p2}$, we can use Lasso-based method of Belloni, Chernozhukov, and Kato (2014) for estimating $(\alpha_l)_{l=1}^{p1}$ and constructing a joint confidence band on $(\alpha_l)_{l=1}^{p1}$ and then checking which αl's are significantly different from zero. The approach is based on buliding orthogonal estimating equations for each of $(\alpha_l)_{l=1}^{p_1}$, and can be interpreted as doing Frisch-Waugh procedure for each coefficient of interest, where we do partialling out via Lasso or OLS-after-Lasso.

This procedure is implemented in the `Julia` package `hdm`. Here is an example in which

$n = 100$, $p_1 = 20$, and $p_2 = 20$, so that total number of regressors is $p = p_1 + p_2 = 40$. In this example $\alpha_1 = 5$ and $\beta_1 = 5$, i.e. there is only one true cause D_{i1}, among the large number of causes, $D_{i1}, \dots, D_{i20}$, and only one true control W_{i1}. This example is made super-simple for clarity sake. The Belloni, Chernozhukov, and Kato (2014) procedure, implemented by rlassoEffects function in Julia package HDMjl.

```
data = get_data("seed_500")
n, p = size(data);
p1 = 20;
X = data[:,2:end]
Y = data[:,1];

r_confint(rlassoEffects(X, Y, index = [1:p1;]), joint = true);
```

```
2.5%        97.5%

   V2      4.50295     5.22594
   V3    -0.324586    0.315242
   V4    -0.361376    0.195754
   V5    -0.263248    0.296394
   V6    -0.285772    0.285464
   V7    -0.331704    0.304463
   V8    -0.235016    0.309707
   V9   -0.0560166    0.482326
  V10    -0.196153    0.399825
  V11    -0.245571    0.272447
  V12    -0.323424    0.218174
  V13    -0.318749     0.27528
  V14    -0.183316    0.385985
  V15    -0.335361     0.39722
  V16    -0.332536    0.323684
  V17    -0.274859    0.340915
  V18     -0.18912    0.426874
  V19    -0.376246   0.0538804
  V20    -0.115641    0.402017
  V21    -0.223552    0.263272
```

As you can see the procedure correctly tells that only the first cause $D_{i1}, among the large number of causes, $Di1, \dots D_{i20}$, is a statistically significant cause of Y (see the confidence interval for variable V1).

8 References

Acemoglu, D., Johnson, S., & Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation. American economic review, 91(5), 1369-1401.

Barro, R. J., & Lee, J. W. (1994). Data set for a panel of 138 countries.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012): "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain," Econometrica, 80, 2369-2429, Arxiv, 2010.

Belloni, A., and V. Chernozhukov (2013): "Least Squares After Model Selection in High-dimensional Sparse Models," Bernoulli, 19(2), 521-547, ArXiv, 2009

Belloni, A., V. Chernozhukov, I. Fernández-Val, and C. Hansen (2013): "Program Evaluation with High-Dimensional Data," arXiv:1311.2645, ArXiv, 2013

Belloni, A., V. Chernozhukov, and C. Hansen (2010): "Inference for High-Dimensional Sparse Econometric Models," Advances in Economics and Econometrics. 10th World Congress of Econometric Society. August 2010, III, 245-295, ArXiv, 2011.

Belloni, A., V. Chernozhukov, and C. Hansen (2014): "Inference on Treatment Effects After Selection Amongst High-Dimensional Controls," Review of Economic Studies, 81, 608'650, ArXiv, 2011

Belloni, A., V. Chernozhukov, and K. Kato (2014): "Uniform post-selection inference for least absolute deviation regression and other Z-estimation problems," Biometrika.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile Prices in Market Equilibrium," Econometrica, 63, 841'890

Chen, D. L., & Yeh, S. (2010). The economic impacts of eminent domain. Unpublished Manuscript.[2373, 2403-2407].

Chernozhukov, V., Chetverikov, D., & Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. The Annals of Statistics, 41(6), 2786-2819.

Chernozhukov, V., & Hansen, C. (2004). The impact of 401 (k) participation on the wealth distribution: An instrumental quantile regression analysis. Review of Economics and statistics, 86(3), 735-751.

Chernozhukov, V., Hansen, C., & Spindler, M. (2015). Valid post-selection and post-regularization inference: An elementary, general approach. Annu. Rev. Econ., 7(1), 649-688.

Chernozhukov, V., Hansen, C., & Spindler, M. (2015). Post-selection and post-regularization inference in linear models with many controls and instruments. American Economic Review, 105(5), 486-90.

Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. Journal of computational and graphical statistics, 7(3), 397-416.