High-dimensional doubly robust inference for regression parameters

Oliver Dukes, Vahe Avagyan and Stijn Vansteelandt
Department of Applied Mathematics, Computer Science and Statistics
Ghent University, Belgium

1 Introduction

The aim of many economic studies is to infer a cause-effect relationship between an exposure A and outcome Y. The standard analytical approach is to fit a regression model for Y, adjusted for the exposure and any confounders which may distort the $A - Y$ association. The estimated coefficient for A is then used to obtain inference on the (conditional) exposure effect. In practice, there is often little prior knowledge on which variables in a given data set are confounders and furthermore how one should model their association with Y.

Data-adaptive procedures are therefore routinely employed in order to select the variables to adjust for and/or choose a model for their dependence on Y. Such procedures are essential when p, the dimension of the covariates, is close to or greater than n, the number of observations. Popular methods include stepwise variable selection strategies and the Lasso.

However, current data adaptive techniques are subject to several pitfalls:

1. **Model misspecification**: the series of models considered may not contain the truth.
2. **Optimal prediction methods may be suboptimal for exposure effect estimation**: variable selection methods developed to deliver predictions with minimal error do not necessarily lead to exposure effect estimates with minimal bias or variance.
3. **Lack of uniformity**: there may be no finite sample size at which a given procedure is guaranteed to attain its nominal coverage/size. In particular, the exposure effect estimator may have a complex, non-normal distribution (due to jumping back and forth between different selected models), even when the sample size is large.

Hence post-selection p-values and intervals are at best overly optimistic and at worst invalid.

We will describe how to obtain confidence intervals and p-values for an exposure effect in a high-dimensional model, which are *uniformly valid* over the parameter space. Compared to competing approaches, the proposed estimators are less sensitive to model misspecification and the choice of selection strategy; the confidence intervals are straightforward to calculate, whilst reflecting the uncertainty about the exposure effect after model selection.

2 The proposal

2.1 Doubly robust estimation of the exposure effect

Consider a study design which collects i.i.d. data on an outcome Y, a binary exposure A and a vector of covariates L. Furthermore, we consider the high dimensional regression model

$$E(Y|A, L; \theta, \beta) = g(\theta A + \beta' L)$$

Here, $g()$ is a known link function and θ is the parameter of interest.

We will base estimation of the exposure effect θ on the estimating function

$$\psi(W;\theta;\eta) \equiv d(A,L;\theta,\eta)\{Y - g(\theta A + \beta' L)\}$$

where $W = (Y, A, L)$ and η is a vector of nuisance parameters. The function $d(A, L; \theta, \eta)$ will be chosen such that $\psi(W;\theta;\eta)$ is *doubly robust*; it has mean zero (and hence is an unbiased estimating function for θ) if either a model for the outcome or the exposure is correct.

For example, for a continuous outcome we may choose $g()$ to be the identity link, so $E(Y|A, L; \theta, \beta) = \theta A + \beta' L$. We will also postulate a model for the conditional mean of the exposure $E(A|L) = E(A|L;\gamma)$ where $E(A|L;\gamma)$ is a known function smooth in an unknown parameter γ. One typically uses a logistic model e.g. $E(A|L;\gamma) = \text{expit}(\gamma' L)$. Then $d(A, L; \theta, \eta)$ can be chosen to equal $\{A - E(A|L;\gamma)\}$, in which case

$$E[\psi(W;\theta;\eta)] = E[\{A - E(A|L;\gamma)\}\{Y - \theta A - E(Y|A = 0, L;\beta)\}] = 0$$

if either model $E(A|L;\gamma)$ or model $E(Y|A = 0, L;\beta)$ is correct. If $g()$ is the log link, then $d(A, L; \theta, \eta) = \{A - E(A|L;\gamma)\}\exp(-\theta A)$ (Robins et al., 1992). An alternative exposure model is required if $g()$ is the logit link. Estimating θ therefore requires estimation of the nuisance parameter $\eta = (\gamma, \beta)$; the validity of our proposal rests on the procedure described below.

2.2 Estimation of the nuisance parameter η

The score $\psi(W;\theta,\eta)$ is doubly robust, in the sense of having expectation zero if either of the working models are correctly specified, but not necessarily both. However, plugging in an arbitrary sparse estimator $\tilde{\eta}$ of η will not generally deliver uniformly valid, doubly robust inference. Consider performing a score test of $\theta = 0$ based on the score $\psi(W, 0, \tilde{\eta})$, when one of the models is misspecified. By a Taylor expansion, we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(W_i;0,\tilde{\eta}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(W_i;0,\eta) + \frac{1}{n}\sum_{i=1}^{n}\frac{\partial\psi(W_i;0,\tilde{\eta})}{\partial\eta}\sqrt{n}(\tilde{\eta}-\eta) + \text{Remainder}$$

A sparse estimator of η is required in high-dimensional settings to make the 'Remainder' term sufficiently small. But in general this does not prevent the existence of converging sequences η_n for which $\sqrt{n}(\tilde{\eta} - \eta_n)$ (and thus the test statistic) has a complex non-normal distribution.

For a given function $\psi(W;\theta,\eta)$, we therefore propose to to use the gradient $\partial\psi(W;\theta,\eta)/\partial\eta$ as an estimating function for η, so as to ensure that $0 = \sum_{i=1}^{n}\partial\psi(W_i,\theta,\eta)/\partial\eta$ at the nuisance parameter estimator $\hat{\eta}$. This leaves (aside from the remainder) only a term involving $\psi(W;\theta,\eta)$ which is uniformly asymptotically normal. Specifically, one can estimate η by solving the following penalized estimating equations with a bridge penalty:

$$0 = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\gamma}\psi(W_i,\theta,\eta) + \lambda_\gamma\delta|\hat{\gamma}|^{\delta-1}\circ\text{sign}(\gamma) \tag{1}$$

$$0 = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\beta}\psi(W_i,\theta,\eta) + \lambda_\beta\delta|\hat{\beta}|^{\delta-1}\circ\text{sign}(\beta) \tag{2}$$

Here, $\lambda_\gamma > 0$ and $\lambda_\beta > 0$ are penalty parameters, $\delta \geq 1$ and $\circ$ is the Hadamard product operator. We will use $\hat{\gamma}$ and $\hat{\beta}$ to refer to the resulting estimators of γ and β respectively.

Note that by letting $\delta \to 1+$, the penalty terms correspond to the sub-gradient of the ℓ_1 or Lasso norm penalty $||\eta||_1$ with respect to η. Therefore the estimates $\hat{\gamma}$ and $\hat{\beta}$ can be obtained using standard software for (weighted) Lasso regression (e.g. 'glmnet'). Then θ can be estimated in closed form and confidence intervals can be constructed by inverting a score test.

3 Theoretical properties

For any vector $a \in \mathbb{R}^p$, define its support as $\text{support}(a) = \{j \in \{1, .., p\} : a_j \neq 0\}$. Let us define the active sets of variables as $S_\gamma = \text{support}(\gamma_n)$ and $S_\beta = \text{support}(\beta_n)$. Furthermore, let s_γ denote the cardinality $|S_\gamma|$ and likewise $s_\beta = |S_\beta|$. We consider the behavior of our estimators under two scenarios:

***Either* the exposure model *or* $E(Y|A = 0, L; \beta)$ is correct**: assuming no $A - L$ interaction on the relevant scale, then tests and confidence intervals for our estimators can be shown to be uniformly valid if $s_\gamma^2 \log^2(p \vee n) = o(n)$ and $s_\beta^2 \log^2(p \vee n) = o(n)$ hold. These so-called 'ultra-sparsity' conditions on γ and β are common in the literature on high-dimensional inference e.g. Belloni et al. (2014). However, unlike other proposals, we do not assume that both considered series of working models contain the truth.

All models are correct: we can get uniformly valid inference under the weaker conditions $s_\gamma \log(p \vee n) = o(n)$, $s_\beta \log(p \vee n) = o(n)$ and $s_\gamma s_\beta \log^2(p \vee n) = o(n)$, without requiring ultra-sparsity. Hence we can allow for more complexity in one model as long as the other model is easier to estimate. These sharpened rates are due to the test (and estimators) being doubly robust; unlike Chernozhukov et al. (2017), we do not require sample splitting to obtain them.

Our procedure is related to the bias-reduced doubly robust estimation methodology of Vermeulen and Vansteelandt (2015), and additionally incorporates regularization. Thus estimating η in this way is also expected to prevent the inflation of bias upon minor misspecification of one or both working models. Our proposal is further illustrated through simulations and a data analysis.

References

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on Treatment Effects after Selection among High-Dimensional Controls. *The Review of Economic Studies*, 81(2):608–650.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2017). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*.

Robins, J. M., Mark, S. D., and Newey, W. K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, 48(2):479–495.

Vermeulen, K. and Vansteelandt, S. (2015). Bias-Reduced Doubly Robust Estimation. *Journal of the American Statistical Association*, 110(511):1024–1036.