Slice Inverse Regression with Score Functions

Dmitry Babichev
INRIA - École normale supérieure
Paris, France
Email: dmitry.babichev@inria.fr

Francis Bach
INRIA - École normale supérieure
Paris, France
Email: francis.bach@inria.fr

***Abstract*—We consider non-linear regression problems where we assume that the response depends non-linearly on a linear projection of the covariates. We propose score function extensions to sliced inverse regression (SIR) problems, both for the first- order and second-order score functions. We show that they provably improve estimation in the population case over the non-sliced versions and we study finite sample estimators and their consistency given the exact score functions. We also propose to learn the score function as well, in two steps, i.e., first learning the score function and then learning the effective dimension reduction space, or directly, by solving a convex optimization problem regularized by the nuclear norm. We illustrate our results on a series of experiments.**

I. Introduction

In this work, we consider a random vector $x \in \mathbb{R}^d$, a random response $y \in \mathbb{R}$, and a regression model of the form $y = f(x) + \varepsilon$, which we want to estimate from n independent and identically distributed (i.i.d.) observations (x_i, y_i), $i = 1, \dots, n$. Our goal is to estimate the function f from these data. To avoid the curse of dimensionality, we make the following assumption:

(A1) For all $x \in \mathbb{R}^d$, we have $f(x) = g(w^\top x)$ for a certain $w \in \mathbb{R}^{d\times k}$ and a function $g : \mathbb{R}^k \to \mathbb{R}$. Moreover, $y = f(x)+\varepsilon$ with ε independent of x with zero mean and finite variance.

We consider a specific instantiation of the method of moments, which partially circumvents the curse of dimensionality by estimating w (which is called effective dimension reduction or e.d.r. space) directly without the knowledge of g. The starting point for this method is the work by Brillinger [1], which shows, as a simple consequence of Steins lemma [2], that if the distribution of x is Gaussian, and **(A1)** is satisfied with $k = 1$, then the expectation $\mathbb{E}(yx)$ is proportional to w. The use of Stein's lemma with a Gaussian random variable can be directly extended using the **score function** $\mathcal{S}_1(x)$ defined as $\mathcal{S}_1(x) = -\nabla \log p(x) = \frac{1}{p(x)}\nabla p(x)$, where $p(x)$ is the probability density of x. Known extensions: SIR [3] - works only for Gaussian x, uses moment $\mathbb{E}(x|y)$), PHD (Principal Hessian Directions) [4] - works only for Gaussian, uses moment $\mathbb{E}(y \cdot xx^T)$ and PHD+ [5] - works for any smooth distribution, doesn't use slices and uses 2-nd order score $\mathbb{E}(y\cdot\mathcal{S}_2(x))$ (where $\mathcal{S}_2(x) = \frac{1}{p(x)}\nabla^2 p(x)$), which is hard to learn in real problems.

II. Estimation with infinite sample size

Here we focus on the population situation and propose two new methods, SADE: Sliced average derivative estimation and SPHD: Sliced principal Hessian directions. Under regularity assumptions:

Lemma 1 (SADE moment)**.** *Assume **(A1)**. Then, $\mathbb{E}(\mathcal{S}_1(x)|y)$ is in the e.d.r. space almost surely (in y).*

The key difference is now that by conditioning on different values of y, we have access to *several* vectors $\mathbb{E}(\mathcal{S}_1(x)|y)$. In the population case, we will consider the matrix $\mathcal{V}_1 = \mathbb{E}\big[\mathbb{E}(\mathcal{S}_1(x)|y)\mathbb{E}(\mathcal{S}_1(x)|y)^\top\big]$, which will take eigenvectors of.

Similar results can be obtained for the second score function: $\mathcal{S}_2(x) = \frac{1}{p(x)}\nabla^2 p(x)$:

Lemma 2 (SPHD moment)**.** *Assume **(A1)**. Then, $\mathbb{E}(\mathcal{S}_2(x)|y)$ has a column space within the e.d.r. space almost surely.*

III. Estimation from finite sample and algorithm

The moments from Section II can be easily estimated from finite date. Here we provide an estimator for $\mathcal{V}_1$ for SADE (the estimator for SPHD is straightforward). This leads to the following algorithm:

- Divide range of $y_1, \dots, y_n$ into H *slices* $I_1, \dots, I_H$. Let $\hat{p}_h > 0$ be the proportion of y_i, $i = 1, \dots, n$, that fall in slice I_h.
- For each slice I_h, compute the sample mean $(\hat{\mathcal{S}}_1)_h$ of $\mathcal{S}_1(x)$: $(\hat{\mathcal{S}}_1)_h = \frac{1}{n\hat{p}_h}\sum\limits_{i=1}^{n} 1_{y_i\in I_h}\mathcal{S}_1(x_i)$.
- Compute the weighted covariance matrix $\hat{\mathcal{V}}_1 = \sum\limits_{h=1}^{H} \hat{p}_h(\hat{\mathcal{S}}_1)_h(\hat{\mathcal{S}}_1)_h^\top$.
- Find the k largest eigenvalues and let $\hat{w}_1, \dots, \hat{w}_k$ be eigenvectors in $\mathbb{R}^d$ corresponding to these eigenvalues.

With some extra regularity assumptions, $\hat{\mathcal{V}}_1$ is a $\sqrt{n}$-consistent estimator of $\mathcal{V}_1$, and this leads to a $\sqrt{n}$-consistent estimator of the e.d.r. subspace when the score function is known.

IV. Learning score functions

In practice we have to learn score functions from sample data. Under the parametric assumption:

(A4) The score function $\ell(x)$ is a linear combination of known basis functions $\psi^j(x)$, $j = 1, \dots, m$, where $\psi^j : \mathbb{R}^d \to \mathbb{R}^d$,

Using notation: $\mathbf{\Psi}$ is $m \times d$ matrix with rows equal to $\phi^j(x)$ and θ is m-dimensional coefficient vector, the empirical *score matching* cost function of [6] may be written as:

$$\hat{\mathcal{R}}_{\text{score}}(\theta) = \frac{1}{2}\theta^\top\big(\frac{1}{n}\sum_{i=1}^{n}\Psi(x_i)\Psi(x_i)^\top\big)\theta - \theta^\top\big(\frac{1}{n}\sum_{i=1}^{n}(\nabla\cdot\Psi)(x_i)\big).$$

This direct estimation of score function is more robust than estimation of $p(x)$ and then evaluating score function.

Introduce the notation

$$\hat{\mathbf{\Psi}}_h = \frac{1}{|I_h|}\sum_{i=1}^{n} 1_{y_i\in I_h}\Psi(x_i) \in \mathbb{R}^{m\times d}\,;\ \hat{\mathcal{V}}_1 = \sum_{h=1}^{H}\hat{p}_h\hat{\mathbf{\Psi}}_h^\top\theta\theta^\top\hat{\mathbf{\Psi}}_h \in \mathbb{R}^{d\times d}.$$

- Two-step approach: solve the score matching optimization problem to obtain the optimal parameters of θ and then use them to get the k largest eigenvectors of matrix $\hat{\mathcal{V}}_1$.
- Direct approach: minimize
$$\hat{\mathcal{R}}(\theta) = \hat{\mathcal{R}}_{\text{score}}(\theta) + \lambda\cdot\operatorname{tr}\big[\sum_{h=1}^{H}\hat{p}_h\hat{\mathbf{\Psi}}_h^\top\theta\theta^\top\hat{\mathbf{\Psi}}_h\big]^{1/2},$$
where $\operatorname{tr}[\hat{\mathcal{V}}_1]^{1/2}$ is the nuclear norm of the matrix $\hat{\mathcal{V}}_1$

By enforcing the low-rank constraint, the direct approach will circumvent a potential poor estimation of the score function, which could be enough for the task of estimating the e.d.r. space.

Experiments

For experiments we consider a Gaussian mixture model with 2 components in $\mathbb{R}^d$: The error ε has a standard normal distribution. To estimate the effectiveness of an estimated e.d.r. subspace, we use the square trace error $R^2(w, \hat{w})$ [7]

$$R^2(w,\hat{w}) = 1 - \frac{1}{k}\operatorname{tr}\big[(w^\top w)^{-1}w^\top\hat{w}(\hat{w}^\top\hat{w})^{-1}\hat{w}^\top w\big] = 1 - \frac{1}{k}\operatorname{tr}\big[P\cdot\hat{P}\big].$$

On graphs 1 and 2 we compare methods with known scores, and on graphs 3 and 4 we consider unknown score case, where we choose basis functions as Gaussian kernels centered in the sample points, matrix $w = [I_2, 0] \in \mathbb{R}^{d\times 2}$. We can see from graphs 1 and 2, that direct approach works better and has wider range of applicable bandwidths. Graph 3 shows, that usually first-orders method work better, than second-order. Graph 4 shows failure mode of SADE, where only one direction from e.d.r. can be recovered.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Unions H2020 Framework Programme (H2020-MSCA-ITN-2014) under grant agreement n° 642685 MacSeNet.

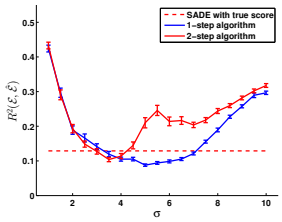

Fig. 1. Rational function $f(x) = x_1/(0.5 + (x_2 + 2)^2)$, $d = 10$, $n = 1000$; error dependence of bandwidth of Gaussian kernels

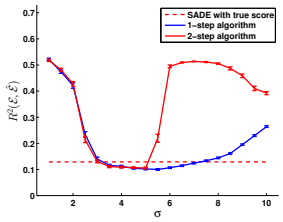

Fig. 2. Rational function $f(x) = x_1/(0.5 + (x_2 + 2)^2)$, $d = 20$, $n = 2000$; error dependence of bandwidth of Gaussian kernels

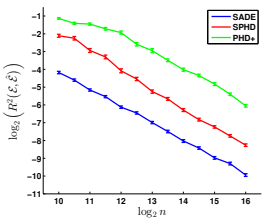

Fig. 3. Mean and standard deviation of $R^2(\mathcal{E}, \hat{\mathcal{E}})$ for the function $f(x) = x_1/(0.5 + (x_2 + 2)^2)$

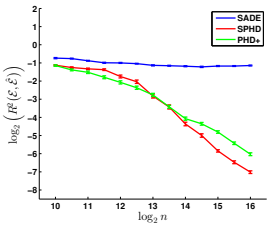

Fig. 4. Mean and standard deviation of $R^2(\mathcal{E}, \hat{\mathcal{E}})$ for the function $f(x) = \mathcal{I}(x_1^2 + 2x_2^2 > 4)$

REFERENCES

[1] D. R. Brillinger. *A Generalized Linear Model with Gaussian Regressor Variables*. In K.A. Doksum P.J. Bickel and J.L. Hodges, editors, A Festschrift for Erich L. Lehmann. Woodsworth International Group, Belmont, California, 1982. California, 1982.

[2] T. Stoker, *Consistent estimation of scaled coefficients*, Econometrica, 54 (1986), p. 1461-1481.

[3] K.-C. Li, *Sliced Inverse Regression for Dimension Reductions*, Journal of the American Statistical Association, 86 (1991), pp 316-327

[4] K.-C. Li, *On Principal Hessian Directions for Data Visualization and Dimension Reduction: Another Application of Stein's Lemma*, Journal of the American Statistical Association, 87 (1992), pp. 1025-1039.

[5] M. Janzamin, H. Sedghi, and A. Anandkumar. *Score function features for discriminative learning: Matrix and tensor framework*. CoRR, abs/1412.2863, 2014.

[6] A. Hyvarinen. *Estimation of non-normalized statistical models by score matching*. Journal of Machine Learning Research, 6:695-709, 2005.

[7] J. Hooper. *Simultaneous Equations and Canonical Correlation Theory.* Econometrica, 27:245256, 1959.