Slice inverse regression with score functions

Dmitry Babichev    Francis Bach
INRIA - École normale supérieure Paris

Abstract

We consider the problem of dimension reduction in non-linear regression, which is often formulated as a non-convex optimization problem. We propose score function extensions to sliced inverse regression problems, both for the first-order and second-order score functions, which provably improve estimation in the population case over the non-sliced versions; we study finite sample estimators and study their consistency given the exact score functions. We propose also 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.

1 Introduction

Non-linear regression and related problems such as non-linear classification are core important tasks in machine learning and statistics. In this paper, 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 $(x_i, y_i)$, $i = 1, \ldots, n$. Our goal is to estimate the function $f$ from these data. A traditional key difficulty in this general regression problem is the lack of parametric assumptions regarding the functional form of $f$, leading to a problem of non-parametric regression. This is often tackled by searching implicitly or explicitly a function $f$ within an infinite-dimensional vector space.

While several techniques exist to estimate such a function, e.g., kernel methods, local-averaging, or neural networks (see, e.g., [1]), they also suffer from the curse of dimensionality, that is, the rate of convergence of the estimated function to the true function (with any relevant performance measure) can only decrease as a small power of $n$, and this power cannot be larger than a constant divided by $d$. In other words, the number $n$ of observations for any level of precision is exponential in dimension.

A classical way of by-passing the curse of dimensionality is to make extra assumption regarding the function to estimate, such as the dependence on a lower unknown low-dimensional subspace, such as done by projection pursuit or neural networks. More precisely, throughout the paper, we make the following assumption:

(A1) For all $x \in \mathbb{R}^d$, we have $f(x) = g(w^T x)$ 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 $\varepsilon$ independent of $x$ with zero mean and finite variance.

The subspace of $\mathbb{R}^d$ spanned by the $k$ columns $w_1, \ldots, w_k \in \mathbb{R}^d$ of $w$ has dimension less than or equal to $k$, and is often called the effective dimension reduction (e.d.r.) space. We will always make the assumption that the e.d.r. space has exactly rank $k$, that is the matrix $w$ has rank $k$.

Our goal is to estimate both the non-linear function $g$ and the matrix $w$, with the hope of obtaining a convergence rate where the inverse power of $n$ will now be proportional to $k$ and not $d$. Note that the matrix $w$ is only identifiable up to a (right) linear transform, since only the subspace spanned by its column is characteristic.

Given \(w\), estimating \(g\) may be done by any technique in non-parametric regression, with a convergence rate which requires a number of observations \(n\) to be exponential in \(k\). Given the non-linear function \(g\), estimating \(w\) is computationally difficult because the resulting optimization problem may not be convex and thus leads to several local minima. Our difficulty is even stronger since we want to estimate both the function \(g\) and the matrix \(w\).

In this paper, we consider a specific instantiation of the method of moments, which partially circumvents this difficulty by estimating \(w\) directly without the knowledge of \(g\). The starting point for this method is the work by Brillinger [2], which shows, as a simple consequence of Stein’s lemma [3], that if the distribution of \(x\) is Gaussian, and (A1) is satisfied with \(k = 1\) (e.d.r. of dimension one), then the expectation \(\mathbb{E}(yx)\) is proportional to \(w\). Thus, a certain expectation, which can be easily approximated given i.i.d. observations, simultaneously eliminates \(g\) and reveals \(w\).

While the result above provides a very simple algorithm to recover \(w\), it has several strong limitations: (a) it only applies to normally distributed data \(x\), or more generally to elliptically symmetric distributions [4], (b) it only applies to \(k = 1\), and (c) in many situations with symmetries, the proportionality constant is equal to zero and thus we cannot recover the vector \(w\). This has led to several extensions in the statistical literature which we now present.

**Using score functions.** The use of Stein’s lemma with a Gaussian random variable can be directly extended using the score function \(S_1(x)\) defined as the negative gradient of the log-density, that is, \(S_1(x) = -\nabla \log p(x) = \frac{1}{p(x)} \nabla p(x)\), which leads to the following assumption:

\[
\text{(A2) The distribution of } x \text{ has a strictly positive density } p(x) \text{ which is differentiable with respect to the Lebesgue measure, and such that } p(x) \rightarrow 0 \text{ when } ||x|| \rightarrow +\infty.
\]

Given Assumption (A2), then [5] showed that, for \(k = 1\) and if Assumption (A1) is satisfied and the input data have zero mean and identity covariance matrix, then \(\mathbb{E}(yS_1(x))\) is proportional to \(w\), for all differentiable functions \(g\), with a proportionality constant that depends on \(w\) and \(\nabla g\). This leads to the “average derivative method” (ADE) and thus replaces the Gaussian assumption by the existence of a differentiable log-density, which is much weaker. This however does not remove the restriction \(k = 1\), which can be done in two ways which we now present.

**Sliced inverse regression.** Given a Gaussian distribution for \(x\) (or any elliptically symmetric distribution), then, if (A1) is satisfied, almost surely in \(y\), the conditional expectation \(\mathbb{E}(x|y)\) happens to belong to the e.d.r. subspace. Given several distinct values of \(y\), the vectors \(\mathbb{E}(x|y)\) or any estimate thereof, will hopefully span the entire e.d.r. space and we can recover the entire matrix \(w\), leading to “slice inverse regression” (SIR), originally proposed by Li and Duan [6]. Duan and Li [7] and Li [8]. This allows the estimation with \(k > 1\), but this is still restricted to Gaussian data. In this paper, we propose to extend SIR by the use of score function to go beyond elliptically symmetric distributions, and we show that the new method combining SIR and score functions is formally better than the plain ADE method.

**From first-order to second-order moments.** Another line of extension of the simple method of Brillinger [2] is to consider higher-order moments, namely the matrix \(\mathbb{E}(yxx^\top) \in \mathbb{R}^{d \times d}\), which, with normally distributed input data \(x\) and if (A1) is satisfied, will be proportional to the Hessian of the function \(g\), leading to the method of “principal Hessian directions” (PHD) from Li [9]. Again, \(k > 1\) is allowed (more than a single projection), but thus is limited to elliptically symmetric data. Janzamin, Sedghi and Anandkumar [10] proposed to used second-order score functions to go beyond this assumption. In order to define this new method, we consider the following assumption:

\[
\text{(A3) The distribution of } x \text{ has a strictly positive density } p(x) \text{ which is twice differentiable with respect to the Lebesgue measure, and such that } ||\nabla p(x)|| \rightarrow 0 \text{ when } ||x|| \rightarrow +\infty.
\]

Given (A1) and (A3), then [10] shows that \(\mathbb{E}(yS_2(x))\) will be proportional to the Hessian of the function \(g\), where \(S_2(x) = \nabla^2 \log p(x) + S_1(x)S_1(x)^\top = \frac{1}{p(x)} \nabla^2 p(x)\), thus extending the Gaussian situation above where \(S_1\) was a linear function and \(S_2\), up to linear terms, proportional to \(xx^\top\).

In this paper, we extend the method above to allow an SIR estimator for the second-order score functions, where we condition on \(y\), and we show that the new method is formally better than [10].
Learning score functions through score matching. Relying on score functions immediately raises the following question: is estimating the score function really simpler than our original problem of non-parametric regression? Fortunately, a recent line of work [11] has considered this exact problem, and formulated the task of density estimation directly on score functions, which is particularly useful in our context. We may then use the data, first to learn the score, and then to use the novel score-based moments to estimate $w$. We will also consider a direct approach that jointly estimates the score function and the e.d.r. subspace.

Fighting the curse of dimensionality. Learning the score function is still a non-parametric problem, with the associated curse of dimensionality. If we first learn the score function (through score matching) and then learn the matrix $w$, we will not escape that curse, while our direct approach is empirically more robust.

2 Estimation with infinite sample size

In this section, we focus on the population situation, where we can compute expectations and conditional expectations exactly. We propose two new methods: SADE: Sliced average derivative estimation and SPHD: Sliced principal Hessian directions.

We can now present our first new lemma (see proof in Appendix):

**Lemma 1** (SADE moment). Assume (A1) and (A2). Then, $\mathbb{E}(S_1(x)|y)$ is in the e.d.r. subspace almost surely (in $y$).

The key difference is now that by conditioning on different values of $y$, we have access to several vectors $\mathbb{E}(S_1(x)|y)$. In the population case, we will consider the matrix $V_1 = \mathbb{E}[\mathbb{E}(S_1(x)|y)\mathbb{E}(S_1(x)|y)^\top]$, which will take eigenvectors of.

We also present our second new lemma (see proof in Appendix):

**Lemma 2** (SPHD moment). Assume (A1) and (A3). Then, $\mathbb{E}(S_2(x)|y)$ has a column space within the e.d.r. subspace almost surely.

We will estimate $V_2 = \mathbb{E}\left( y^2|\mathbb{E}(S_2(x)|y)\right)^2$, which is the expectation with respect to $y$ of the square of the conditional expectation from Lemma 2.

These moments are superior to previous moments, as they allow correct estimation in more situations:

**Proposition 1.** Assume (A1) and (A2), with $k = 1$. The vector $w$ may be recovered from the ADE moment (up to scale) if and only if $\mathbb{E} g'(w^\top x) \neq 0$. If this condition is satisfied, then SADE also recovers $w$ up to scale (i.e., $V_1$ is different from zero).

**Proposition 2.** Assume (A1) and (A3). The matrix $w$ may be recovered from the moment $\mathbb{E}(\nabla y^2 g(w^\top x))$ (up to right linear transform) if and only if $\mathbb{E}[\nabla^2 y^2 g(w^\top x)]$ has full rank. If this condition is satisfied, then SPHD also recovers $w$ up to scale.

3 Estimation from finite sample and algorithm

The moments from Section 2 can be easily estimated from finite data. Here we provide an estimator for $V_1$ for SADE (the estimator for SPHD is straightforward). This leads to the following algorithm:

- Divide range of $y_1, \ldots, y_n$ into $H$ “slices” (i.e., contiguous intervals) $I_1, \ldots, I_H$. Let $\hat{p}_h > 0$ be the proportion of $y_i$, $i = 1, \ldots, n$, that fall in slice $I_h$.
- For each slice $I_h$, compute the sample mean $(\hat{S}_1)_h$ of $S_1(x)$: $(\hat{S}_1)_h = \frac{1}{n_{I_h}} \sum_{i=1}^{n_{I_h}} 1_{y_i \in I_h} S_1(x_i)$.
- Compute the weighted covariance matrix $\hat{V}_1 = \sum_{h=1}^{H} \hat{p}_h (\hat{S}_1)_h (\hat{S}_1)_h^\top$.
- Find the $k$ largest eigenvalues and let $\hat{w}_1, \ldots, \hat{w}_k$ be eigenvectors in $\mathbb{R}^d$ corresponding to these eigenvalues.

As shown in Appendix B, with some extra regularity assumptions, $\hat{V}_1$ is a $\sqrt{n}$-consistent estimator of $V_1$, and this leads to a $\sqrt{n}$-consistent estimator of the e.d.r. subspace when the score function is known.
4 Learning score functions

All previous methods can work only if we know the score function of first or second order. In practice, we do not have such information, and we have to learn score functions from sample data. In this section, we only consider the first-order score function $\ell(x) = S_1(x) = -\nabla \log p(x)$.

Under the parametric assumption

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

The empirical score matching cost function of (11) may then be written as:

$$\hat{R}_{\text{score}}(\theta) = \frac{1}{2} \theta^\top \left( \frac{1}{n} \sum_{i=1}^{n} \psi(x_i) \psi(x_i)^\top \right) \theta - \theta^\top \left( \frac{1}{n} \sum_{i=1}^{n} (\nabla \cdot \psi)(x_i) \right),$$

and thus leads to a simple least-squares problem.

4.1 Two-steps approach

We can now combine our linear parametrization of the score with the SIR approach outlined in Section 3. We first estimate the score function from minimizing Eq. (2), and run the SADE algorithm with this estimated score. Introducing the notation

$$\hat{\Psi} = \frac{1}{n} \sum_{i=1}^{n} \psi(x_i),$$

and thus leads to a simple least-squares problem.

4.2 Direct approach

We can also try to combine these two steps to avoid “the curse of dimensionality”. Our estimation of the score, i.e., of the parameter $\theta$ is done only to be used within the SIR approach where we expect the matrix $\hat{\Psi}_1$ to have rank $k$. Thus when estimating $\theta$ by minimizing $\hat{R}_{\text{score}}(\theta)$, we may add a regularization that penalizes large ranks for $\hat{\Psi}_1(\theta) = \sum_{h=1}^{H} \hat{\psi}_h \hat{\Psi}^h \theta \theta^\top \hat{\Psi}_h$, where we highlight the dependence on $\theta \in \mathbb{R}^m$.

We may thus also consider the $H$ vectors $\hat{\Psi}^h \theta \in \mathbb{R}^d$, $h \in \{1, \ldots, H\}$ and put them in a matrix $\mathcal{A}(\theta) = (\sqrt{\hat{\rho}_1} \hat{\Psi}^1 \theta, \ldots, \sqrt{\hat{\rho}_H} \hat{\Psi}^H \theta) \in \mathbb{R}^{d \times H}$.

We may then penalize the nuclear norm of $\mathcal{A}(\theta)$, which happens to be equal to $[\hat{\Psi}_1(\theta)^{1/2}]$, or potentially consider norms that take into account that we look for a rank $k$ (e.g., the $k$-support norm on the spectrum of $\mathcal{A}(\theta)$). Combining two penalties, we have a convex optimization task:

$$\hat{R}(\theta) = \hat{R}_{\text{score}}(\theta) + \lambda \text{tr}[\hat{\Psi}_1(\theta)^{1/2}].$$

This is much more robust in practice than the two-step approach, as we present in a set of experiments in Appendix C.

5 Conclusion

In this paper we consider a general non-linear regression model and the dependence on a unknown $k$-dimensional subspace assumption. Our goal was direct estimation of this unknown $k$-dimensional space, which is often called effective dimension reduction or e.d.r. space. We proposed new approaches (SADE and SPHD), combining two existing techniques (sliced inverse regression (SIR) and score function-based estimation). We obtained consistent estimation for $k > 1$ only using the first-order score and proposed explicit approaches to learn the score from data.

It would be interesting to extend our sliced extensions to learning neural networks: indeed, our work focused on the subspace spanned by $w$ and cannot identify individual columns, while [12] showed that by using proper tensor decomposition algorithms and second-order score functions, columns of $w$ can be consistently estimated (in polynomial time).
References


A  Proofs

A.1  Appendix. Proof of Lemma\[1\]

Proof. We consider any vector \( b \in \mathbb{R}^d \) in the orthogonal complement of \( \text{Span}\{w_1, \ldots, w_k\} \). We need to show, that \( b^\top \mathbb{E}(S_1(x)|y) = 0 \) with probability 1. We have by the law of total expectation

\[
b^\top \mathbb{E}(S_1(x)|y) = \mathbb{E}\left( b^\top S_1(x)|w_1^\top x, \ldots, w_k^\top x, y \right) \cdot \mathbb{E}(y | y).
\]

Because of Assumption (A1), we have \( y = g(w^\top x) + \varepsilon \) with \( \varepsilon \) independent of \( x \), and thus

\[
\mathbb{E}(b^\top S_1(x)|w_1^\top x, y) = \mathbb{E}(b^\top S_1(x)|w_1^\top x, \varepsilon) = \mathbb{E}(b^\top S_1(x)|w_1^\top x) \text{ almost surely.}
\]

This leads to

\[
b^\top \mathbb{E}(S_1(x)|y) = \mathbb{E}[b^\top S_1(x)|w_1^\top x].
\]

We now prove that almost surely \( \mathbb{E}(b^\top S_1(x)|w_1^\top x) = 0 \), which will be sufficient to prove Lemma\[1\]. We consider the linear transform of coordinates: \( \tilde{x} = \tilde{w}^\top x \in \mathbb{R}^d \), where \( \tilde{w} = (w_1, \ldots, w_k, w_{k+1}, \ldots, w_d) \) is a square matrix with full rank obtained by adding a basis of the subspace orthogonal to the span of the \( k \) columns of \( w \). Then \( \nabla p(x) = \tilde{w}^\top \cdot \nabla \tilde{p}(\tilde{x}) \) and \( \tilde{b} = \tilde{w}^\top b = \{0, \ldots, 0, \tilde{b}_{k+1}, \ldots, \tilde{b}_{d}\} \) (because \( b \perp \text{Span}\{w_1, \ldots, w_k\} \)) and the desired conditional expectation equals

\[
J \cdot \mathbb{E}(\tilde{b}^\top \tilde{S}_1(\tilde{x})|\tilde{x}_1, \ldots, \tilde{x}_k),
\]

where \( J \) is the Jacobian and \( \tilde{w}^\top \tilde{S}_1(\tilde{x}) = \tilde{w}^\top \nabla \tilde{p}(\tilde{x}) = \nabla p(x) \tilde{p}(x) = S_1(x) \).

It is thus sufficient to show that

\[
\int_{\mathbb{R}^{d-k}} \tilde{b}^\top \tilde{S}_1(\tilde{x})\tilde{p}(\tilde{x}_1, \ldots, \tilde{x}_d) d\tilde{x}_{k+1} \ldots d\tilde{x}_d = 0, \text{ for all } \tilde{x}_1, \ldots, \tilde{x}_k.
\]

We have

\[
= \sum_{j=k+1}^d \tilde{b}_j \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{\partial \tilde{p}(\tilde{x})}{\partial \tilde{x}_j} d\tilde{x}_j \cdot \prod_{k+1 \leq t \leq d \atop t \neq j} \tilde{x}_t = 0,
\]

because for any \( j \in \{k+1, \ldots, n\} \), \( \int_{-\infty}^{\infty} \frac{\partial \tilde{p}(\tilde{x})}{\partial \tilde{x}_j} d\tilde{x}_j = 0 \) by Assumption (A2). This leads to the desired result. \( \square \)

A.2  Appendix. Proof of Lemma\[2\]

Proof. We consider any \( a \in \mathbb{R}^d \) and \( b \in \mathbb{R}^d \) orthogonal to the e.d.r. subspace, and prove, that \( a^\top \mathbb{E}(S_2(x)|y)b = 0 \). We use the same transform of coordinates as in the proof of Lemma\[1\]

\( \tilde{x} = \tilde{w}^\top x \in \mathbb{R}^d \). Then \( \nabla^2 p(x) = \tilde{w} \nabla^2 \tilde{p}(\tilde{x}) \tilde{w}^\top \), \( \tilde{b} = \tilde{w}^\top b = \{0, \ldots, 0, \tilde{b}_{k+1}, \ldots, \tilde{b}_{d}\} \) and we will prove, that \( J \cdot \mathbb{E}(a^\top S_2(\tilde{x})|w_1^\top x) = 0 \) almost surely, where \( J \) is the Jacobian and \( \tilde{w}^\top \tilde{S}_2(\tilde{x}) \tilde{w} = \frac{\tilde{w}^\top \nabla^2 \tilde{p}(\tilde{x}) \tilde{w}}{\tilde{p}(\tilde{x})} = \nabla^2 p(x) \tilde{p}(x) = S_2(x) \). It is sufficient to show, that for all \( \tilde{x}_1, \ldots, \tilde{x}_k \):

\[
\int_{\mathbb{R}^{d-k}} \tilde{a}^\top \cdot \tilde{S}_2(\tilde{x}) \cdot b \cdot \tilde{p}(\tilde{x}_1, \ldots, \tilde{x}_d) d\tilde{x}_{k+1} \ldots d\tilde{x}_d = 0.
\]

We have:

\[
\int_{\mathbb{R}^{d-k}} \tilde{a}^\top \cdot \tilde{S}_2(\tilde{x}) \cdot b \cdot \tilde{p}(\tilde{x}_1, \ldots, \tilde{x}_d) d\tilde{x}_{k+1} \ldots d\tilde{x}_d = \tilde{a}^\top \left[ \int_{\mathbb{R}^{d-k}} \nabla^2 \tilde{p}(\tilde{x}) \cdot d\tilde{x}_{k+1} \ldots d\tilde{x}_d \right] \cdot \tilde{b}
\]
We consider the case when the score function is known. Following Hsing and Carroll [14], the nuclear norm in future work. The key novelty compared to [15]. In this section, we focus on the simplest set of assumptions to pave the way to the analysis for the √ algorithm.

**Proof of consistency**

We have:

\[
\sum_{1 \leq i \leq d} \hat{a}_i b_j \int_{-\infty}^{\infty} \cdots \left[ \int_{-\infty}^{\infty} \frac{\partial^2 \hat{p}(\tilde{x})}{\partial \tilde{x}_i \partial \tilde{x}_j} \cdot d\tilde{x}_j \right] \prod_{k+1 \leq i \leq d} d\tilde{x}_i = 0,
\]

because for any \( j \in \{k+1, \ldots, n\} \):

\[
\int_{-\infty}^{\infty} \frac{\partial^2 \hat{p}(\tilde{x})}{\partial \tilde{x}_i \partial \tilde{x}_j} \cdot d\tilde{x}_j = 0 \quad \text{due to Assumption (A3)}.
\]

**A.3 Appendix. Proof of Proposition 1**

**Proof.** The first statement is a consequence of the proof for ADE moment [5]. If SADE fails, that is, for almost all \( y \), \( \mathbb{E}(S_1(x)|y) = 0 \), then \( \mathbb{E}(S_1(x)|y) = 0 \) which implies that \( \mathbb{E}(S_1(x)|y) = 0 \) and thus ADE fails. Moreover, we have:

\[
\mathcal{V}_1 = \mathbb{E}[(yS_1(x)|y)E(yS_1(x)|y)^\top] \succ [\mathbb{E}(yS_1(x)|y)]\mathbb{E}(yS_1(x)|y)\mathbb{E}(yS_1(x)|y)\mathbb{E}(yS_1(x)|y)^\top - [\mathbb{E}(yS_1(x))]\mathbb{E}(yS_1(x))\mathbb{E}(yS_1(x))\mathbb{E}(yS_1(x)|y)],
\]

showing that the new moment is dominating the ADE moment, which provides an alternative proof. □

**A.4 Appendix. Proof of Proposition 2**

**Proof.** Since \( y = f(x) + \varepsilon \), and \( \varepsilon \) is independent of \( x \), we have

\[
\mathbb{E}(yS_2(x)) = \mathbb{E}(f(x))S_2(x) = \int \nabla^2 p(x) (p(x)f(x)dx = \int \nabla^2 p(x) \cdot f(x)dx = \int p(x) \cdot \nabla^2 f(x)dx = \mathbb{E}[\nabla^2 f(x)],
\]

using integration by parts and the decay of \( p(x) \) and \( \nabla p(x) \) for \( \|x\| \to \infty \). This leads to the desired result since \( \nabla^2 f(x) = w\nabla^2 g(w^\top x)w^\top \). Moreover, using Lowner-Heinz theorem about operator convexity [13]:

\[
\mathcal{V}_2 = \mathbb{E}[\mathbb{E}(yS_2(x)|y)^2] \succeq [\mathbb{E}[\mathbb{E}(yS_2(x)|y)]^2 = [\mathbb{E}(yS_2(x))]^2,
\]

showing that the new moment is dominating the PHD moment, thus implying that

\[
\text{rank}[\mathcal{V}_2] \succeq \text{rank}[\mathbb{E}(yS_2(x))].
\]

□

**B Proof of consistency**

In this section, we prove the consistency of the SADE moment estimator and the resulting algorithm, when the score function is known. Following Hsing and Carroll [14] and Xhu and Ng [15], we can get \( \sqrt{n} \) consistency for the SADE algorithm with very broad assumptions regarding the problem.

In this section, we focus on the simplest set of assumptions to pave the way to the analysis for the nuclear norm in future work. The key novelty compared to [14] [15] is a precise non-asymptotic analysis with precise constants.

We consider the case \( c = 2 \), that is every slice contains 2 points. We make the following assumptions:

**(L1)** The function \( m : \mathbb{R} \to \mathbb{R}^d \) such that \( \mathbb{E}(\ell(x)|y) = m(y) \) is \( L \)-Lipschitz-continuous.

**(L2)** The score function has limited growth, that is \( \|\ell(x)\| \leq M\|x\| \) for all \( x \in \mathbb{R}^d \).

**(L3)** The random variable \( y \in \mathbb{R} \) is sub-Gaussian, i.e., such that \( \mathbb{E}e^{t(y-E_y)} \leq e^{\tau_y t^2/2} \), for some \( \tau_y > 0 \).

**(L4)** The random variables \( \eta_j = \ell_j(x) - m_j(y) \in \mathbb{R} \) are sub-Gaussian, i.e., such that \( \mathbb{E}e^{\tau \eta} \leq e^{\tau \eta^2/2} \) for each component \( j \in \{1, \ldots, d\} \), for some \( \tau > 0 \).
Let us first formulate and proof auxiliary lemma:

**Lemma 3.** Let $X$ be a non-negative random variable such that for some positive constants $A$ and $B$, and all $p$:

\[ \mathbb{E}X^p \leq (A\sqrt{p} + Bn^{1/p}p^2)^p. \]

Then, for $t \geq (\log n)/2$:

\[ \mathbb{P}(X \geq 3A\sqrt{t} + 3Bn^{1/t}t^2) \leq 3e^{-t}. \]

**Proof.** By Markov’s inequality, for every non-negative integer $p$:

\[ \mathbb{P}(X \geq 3A\sqrt{t} + 3Bn^{1/t}t^2) = \mathbb{P}(X^p \geq [3A\sqrt{\mathbb{E}(X)} + Bn^{1/t}t^2]^p) \leq \frac{[3A\sqrt{\mathbb{E}(X)} + Bn^{1/t}t^2]^p}{\mathbb{E}(X^p)} \leq e^{-p}. \]

Consider any $t > (\log n)/2$ and $p = \lfloor t \rfloor$, then:

\[ \mathbb{P}(X \geq 3A\sqrt{t} + 3Bn^{1/t}t^2) \leq \mathbb{P}(X \geq 3A\sqrt{t} + 3Bn^{1/t}t^2) \leq e^{-p} \leq 3e^{-t}, \]

because function $f(t) = 3A\sqrt{t} + 3Bn^{1/t}t^2$ increases on $[(\log n)/2; +\infty)$.

Now we can formulate and proof the main theorem:

**Theorem 1.** Under assumptions (L1) - (L4) we get the following bound on $\|\hat{V}_1 - V_1\|_{\text{nuclear}}$:

for any $\delta < 1/n$, with probability not less than $1 - \delta$:

\[ \|\hat{V}_1 - V_1\|_{\text{nuclear}} \leq \frac{195d^2\eta^2}{\sqrt{n}} \sqrt{\log \frac{12d^2}{\delta}} + \frac{155d^2\eta^2}{n} \log^2 \left( \frac{12d^2}{\delta} \right) + \frac{8L^2\tau^2 + 16\eta \tau_L \sqrt{d} + 2\tau^2_0}{n} \log \frac{16d^2}{\delta}. \]

**Proof.** Introduce notations: $\ell(h,1) = \ell(x(h-1)), \ell(h,2) = \ell(x(2h)), y(h,1) = y(2h-1)$ and $y(h,2) = y(2h)$. We have then, for $c = 2$, and $H = n/c = n/2$:

\[ \hat{V}_1 = \frac{1}{n} \sum_{h=1}^{H} \left( \ell(h,1) - \ell(h,2) \right) \left( \ell(h,1) - \ell(h,2) \right) \top, \]

Thus, the deviation from the population version, may be split into four terms as follows:

\[ \hat{V}_1 - V_1 = \frac{1}{n} \sum_{h=1}^{H} \left( \ell(h,1) - \ell(h,2) \right) \left( \ell(h,1) - \ell(h,2) \right) \top - \mathbb{E}\eta \top \]

\[ = \frac{1}{n} \sum_{h=1}^{H} \left( \eta(h,1) - \eta(h,2) \right) \left( \eta(h,1) - \eta(h,2) \right) \top - \mathbb{E}\eta \top \]

\[ + \frac{1}{n} \sum_{h=1}^{H} \left( m(y(h,1)) - m(y(h,2)) \right) \left( \eta(h,1) - \eta(h,2) \right) \top \]

\[ + \frac{1}{n} \sum_{h=1}^{H} \left( \eta(h,1) - \eta(h,2) \right) \left( m(y(h,1)) - m(y(h,2)) \right) \top \]

\[ + \frac{1}{n} \sum_{h=1}^{H} \left( m(y(h,1)) - m(y(h,2)) \right) \left( m(y(h,1)) - m(y(h,2)) \right) \top \]

\[ = T_4 + T_3 + T_2 + T_1. \]

We now bound each term separately.
Bounding $T_1$. We have

$$T_1 \leq \frac{1}{n} \sum_{h=1}^{H} L^2 (y(h, 1) - y(h, 2))(y(h, 1) - y(h, 2))^\top$$

$$\text{tr}T_1 = \|T_1\|_{\text{nuclear}} \leq \frac{1}{n} \sum_{h=1}^{H} L^2 \text{diameter}(y_1, \ldots, y_n)|y(h, 1) - y(h, 2)|$$

$$\leq \frac{1}{n} L^2 \text{diameter}(y_1, \ldots, y_n)^2.$$  

The range cannot grow too much, i.e., as $\log n$. Indeed, assuming without loss of generality that $\mathbb{E}y = 0$, we have $\max\{y_1, \ldots, y_n\} \leq u/2$ and $\min\{y_1, \ldots, y_n\} \geq -u/2$ implies that the range is less than $u$, and thus, $\mathbb{P}(\text{diameter}(y_1, \ldots, y_n) \geq u) \leq \mathbb{P}(\max\{y_1, \ldots, y_n\} \geq u/2) + \mathbb{P}(\min\{y_1, \ldots, y_n\} \leq -u/2) \leq n\mathbb{P}(y > u/2) + n\mathbb{P}(y < -u/2) \leq 2n \exp(-u^2/8\tau_y^2)$ by using sub-Gaussianity. Then, by selecting $u^2/8\tau_y^2 = \log(2n) + \log(8/\delta)$, we get with probability greater than $1 - \delta/8$ that diameter$(y_1, \ldots, y_n) \leq 2\sqrt{2}\tau_y\sqrt{\log(2n) + \log(8/\delta)}$.

Bounding $T_2$ and $T_3$. We also have

$$\max\{\|T_2\|_{\text{nuclear}}, \|T_3\|_{\text{nuclear}}\} \leq \frac{1}{n} \sum_{h=1}^{H} L|y(h, 1)y(h, 2)| \text{diameter}(\eta_1, \ldots, \eta_n)$$

$$\leq \frac{1}{n} L \text{diameter}(y_1, \ldots, y_n) \text{diameter}(\eta_1, \ldots, \eta_n).$$

Like for $T_1$, the ranges cannot grow too much, i.e., as $\log n$. Similarly $\mathbb{P}(\text{diameter}(\eta_1, \ldots, \eta_n) \geq u) \leq n\mathbb{P}(\eta_1 > u/2) + n\mathbb{P}(\eta_1 < -u/2) \leq 2n \exp(-u^2/8\tau_y^2)$. We thus with get with probability greater than $1 - \delta/(8d)$ that max$_{j \in \{1, \ldots, d\}} \text{diameter}(\eta_1, \ldots, \eta_n) \leq 2\sqrt{2}\tau_y\sqrt{\log(2n) + \log(8d/\delta)}$.

Thus combining the two terms above, with probability greater than $1 - \delta/4$,

$$\|T_1\|_{\text{nuclear}} + \|T_2\|_{\text{nuclear}} + \|T_3\|_{\text{nuclear}} \leq \frac{8L(L\tau_y^2 + 2\tau_y\sqrt{d})}{n}\left(\log(2n) + \log(8d) + \log(1/\delta)\right).$$

Note the term in $\sqrt{d}$, which corresponds to the definition of the diameter $\text{diameter}(\eta_1, \ldots, \eta_n)$ in terms of the $\ell_2$-norm.

Bounding $T_4$. We have:

$$T_4 = \frac{1}{n} \sum_{h=1}^{H} \left\{ \eta(h, 1)^\top \eta(h, 1) + \eta(h, 2)^\top \eta(h, 2) - \eta(h, 2)^\top \eta(h, 1) - \eta(h, 1)^\top \eta(h, 2) \right\} - \mathbb{E}\eta^\top$$

$$= \frac{1}{n} \sum_{h=1}^{H} \left\{ -\eta(h, 2)^\top \eta(h, 1) - \eta(h, 1)^\top \eta(h, 2) \right\} + \frac{1}{n} \sum_{i=1}^{n} \eta_i^\top - \mathbb{E}\eta^\top = T_{4,1} + T_{4,2}.$$  

For the second term $T_{4,2}$ above, if we select any element indexed by $a, b$, then

$$\frac{1}{n} \sum_{i=1}^{n} (\eta_i)_a (\eta_i)_b - \mathbb{E}(\eta)_a (\eta)_b.$$  

Using [16] Theorem 2.1, we get

$$\mathbb{E}[(\eta)_a (\eta)_b] \leq \sqrt{\mathbb{E}(\eta)_a^2 \mathbb{E}(\eta)_b^2} \leq 4(2\tau_y^2)^2 = 16\tau_y^4,$$

and

$$\mathbb{E}|(\eta)_a (\eta)_b|^q \leq \sqrt{\mathbb{E}(\eta)_a^{2q} \mathbb{E}(\eta)_b^{2q}} \leq 2q! (2\tau_y^2)^q = \frac{q!}{2} (2\tau_y^2)^q - 16\tau_y^4.$$
We can then use Bernstein’s inequality [16] Theorem 2.10, to get that with probability less than $e^{-t}$

\[
\frac{1}{n} \sum_{i=1}^{n} (\eta_i)_a(\eta_i)_b - E\eta_a\eta_b \geq 2\frac{\tau_2}{n} t + \sqrt{32\tau_4^2 \sqrt{t}/\sqrt{n}}
\]

Thus, with $t = \log \frac{8\pi^2}{\delta}$, we get that all $d(d+1)/2$ absolute deviations are less than $2\tau_2^2 \left( \frac{\log \frac{8\pi^2}{\delta}}{n} + \sqrt{2 \log \frac{8\pi^2}{\delta}} \right)$, with probability greater than $1 - \delta/4$. This implies that the nuclear norm of the second term is less than $2d\tau_2^2 \left( \frac{\log \frac{8\pi^2}{\delta}}{n} + \sqrt{2 \log \frac{8\pi^2}{\delta}} \right)$, because for any matrix $K \in \mathbb{R}^{d \times d}$:

\[
\|K\|_{\text{nuclear}} \leq d\|K\|_{\infty}.
\]

For the first term, we consider $Z = \sum_{h=1}^{H} (\eta_{(h,2)})_a(\eta_{(h,1)})_b$, and considering conditioning on $Y = (y_1, \ldots, y_n)$. A key result from the theory of order statistics is that the $n$ random variables $\eta_{(h,2)}$, $\eta_{(h,1)}$, $h \in \{1, \ldots, n/2\}$ are independent given $Y$ [17]. This allows us to compute expectations.

Using Rosenthal’s inequality [16] Theorem 15.11 conditioned on $Y$, for which we have $E((\eta_{(h,2)})_a(\eta_{(h,1)})_b | Y) = 0$, we get:

\[
E(|Z|^p | Y) \leq 2^{p-1} \sqrt{8p} \left[ \sum_{h} E\left[\left((\eta_{(h,2)})^2_a(\eta_{(h,1)})^2_b\right) | Y\right]\right]^{p/2} + 2^{p-1} p \cdot 2^p \sum_{h} E\left[\left((\eta_{(h,2)})^p_a(\eta_{(h,1)})^p_b\right) | Y\right].
\]

By taking the $p$-th power, we get:

\[
E(|Z|^p) \leq 2^{p-1} \sqrt{8p} \left[ \sum_{h} E\left[\left((\eta_{(h,2)})^2_a(\eta_{(h,1)})^2_b\right) | Y\right]\right]^{p/2} + 2^{p-1} p \cdot 2^p \sum_{h} E\left[\left((\eta_{(h,2)})^p_a(\eta_{(h,1)})^p_b\right) | Y\right].
\]

By now taking expectations with respect to $Y$, we get, using Jensen’s inequality:

\[
E|Z|^p \leq 2^{p-1} \sqrt{8p} E\left[ \left( \sum_{h} (\eta_{(h,2)})^4_a + (\eta_{(h,1)})^4_b \right)^{p/2} \right] + 2^{p-1} p \cdot 2^p \sum_{h} E\left[\left((\eta_{(h,2)})^p_a + (\eta_{(h,1)})^p_b\right) | Y\right].
\]

Because summing over all order statistics is equivalent to summing over all elements. Thus, using the bound on moments of $(\eta_i)_a$, we get:

\[
E|Z|^p \leq 2^{p-1} \sqrt{8p} 2^{p/2-1} E\left( \left( \sum_{i} (\eta_i)_a^4 \right)^{p/2} \right) + 2^{p-1} \sqrt{8p} 2^{p/2-1} E\left( \left( \sum_{i} (\eta_i)_b^4 \right)^{p/2} \right) + 2^{p-1} p \cdot 2^p n \cdot 4p!(2\tau_4^2)^p
\]

We can now use [16] Theorem 15.10, to get

\[
E\left( \left( \sum_{i} (\eta_i)_a^4 \right)^{p/2} \right)^2 / p \leq 2E[\sum_{i} (\eta_i)_a^4] + \frac{p}{2} E\left[ \max_{i}(\eta_i)_a^4 \right]^{2/p}
\]

\[
\leq 2E[\sum_{i} (\eta_i)_a^4] + \frac{p}{2} E\left[ \sum_{i} (\eta_i)_a^4 \right]^{2/p}
\]

\[
\leq 2n \times 4(2\tau_4^2)^2 + \frac{p}{2} \gamma_{\eta}^2 p\tilde{E}_\eta^2
\]

\[
\leq (32n + n^2/2P(2p)^{2/p}) \tau_4^4
\]

\[
\leq (32n + n^2/2P^3) \tau_4^4
\]
Thus
\[ \mathbb{E}[Z]^p \leq 2^p \sqrt{8p} \frac{2p}{2^p - 1} (32n + n^2/p^3)^{1/2} \tau^2 \eta + 2^p - 1 p^2 \cdot 2p \cdot 4p!(2\tau^2_{\eta})^p \]
\[ \leq 2^{p-1} p^2 \cdot 2^{p-1} \frac{1}{2^{p-1}} (32n)^{p/2} + n \cdot 2^{p/2} \tau^2 \eta + 2^{p+1} \tau^2 \eta \cdot n p^2 \eta \]
\[ \leq (2^{p-2} p^2 n p^2 + n p^2 (2^{p/2} - 2 + 2^{p+1}) \tau^2 \eta \]
\[ \leq (64p \cdot p^2 n p^2 + 19p \cdot n p^2) \tau^2 \eta \]
Thus
\[ (\mathbb{E}[Z]^p)^{1/p} \leq \left( 64 \cdot \sqrt{\frac{n}{p}} + 19 \cdot n^{1/p} p^2 \right)^{2} \tau^2 \eta \]
Thus, for any \( \delta \leq 1/n \), using Lemma 3 for random variable \( Z/n \) with \( t = \log(\frac{12d^2}{\delta}) > (\log n)/2 \) and we obtain:
\[ \mathbb{P}\left[ \left| \frac{Z}{n} \right| \geq 192 \tau^2_{\eta} \sqrt{\frac{1}{n} + \frac{57 \tau^2_{\eta}}{n^{1/4} t^2}} \right] \leq 3 e^{-t} \Rightarrow \]
\[ \mathbb{P}\left[ \left| \frac{Z}{n} \right| \geq 192 \tau^2_{\eta} \sqrt{\log(\frac{12d^2}{\delta}) + \frac{57 \tau^2_{\eta}}{n \log(\frac{12d^2}{\delta})} \log^2\left(\frac{12d^2}{\delta}\right)} \right] \leq \frac{\delta}{4d^2} \Rightarrow \]
\[ \mathbb{P}\left[ \left| \frac{Z}{n} \right| \geq 192 \tau^2_{\eta} \sqrt{\log(\frac{12d^2}{\delta}) + \frac{155 \tau^2_{\eta}}{n \log(\frac{12d^2}{\delta})} \log^2\left(\frac{12d^2}{\delta}\right)} \right] \leq \frac{\delta}{4d^2} \]
Combining all terms \( T_1, T_2, T_3, T_{4,1} \) and \( T_{4,2} \) we get with probability not less than \( 1 - \delta \):
\[ \| \hat{V}_1 - V_1 \|_{\text{nuclear}} \leq \frac{1}{n} \left[ 8L(L \tau^2_{\eta} + 2\tau_{\eta} \sqrt{d}) \cdot \log(\frac{12d^2}{\delta}) + 2d \tau^2_{\eta} \log \frac{12d^2}{\delta} + 155 \tau^2_{\eta} d \log^2\left(\frac{12d^2}{\delta}\right) \right] \]
\[ + \frac{1}{\sqrt{n}} \left[ 2d \tau^2_{\eta} \sqrt{2 \log \frac{12d^2}{\delta} + 192 \tau^2_{\eta} \sqrt{\log(\frac{12d^2}{\delta})} \log^2\left(\frac{12d^2}{\delta}\right)} \right] . \]
Finally, rearranging terms, with probability not less than \( 1 - \delta \):
\[ \| \hat{V}_1 - V_1 \|_{\text{nuclear}} \leq \frac{195d \tau^2_{\eta}}{\sqrt{n}} \sqrt{\log\left(\frac{12d^2}{\delta}\right) + \frac{155d \tau^2_{\eta}}{n} \log^2\left(\frac{12d^2}{\delta}\right)} + \frac{8L^2 \tau^2_{\eta} + 16\tau_{\eta} \sqrt{d} L + 2 \tau^2_{\eta} d}{n} \log\frac{16d^2 n}{\delta} \]
\[ \square \]

## C Appendix. Experiments

In this section we provide numerical experiments of SADE, PHD+ and SPHD on different functions \( f \). We denote the true and estimated e.d.r. as \( \hat{E} \) and \( \hat{E} \) respectively.

Consider a Gaussian mixture model with 2 components in \( \mathbb{R}^d \):
\[ p(x) = \sum_{i=1}^{2} \theta_i \cdot \frac{1}{(2\pi)^{d/2} \cdot |\Sigma_i|^{1/2}} \cdot \exp \left\{ -\frac{1}{2} (X - \mu_i)^\top \Sigma_i^{-1} (X - \mu_i) \right\}, \tag{4} \]
where \( \theta = (6/10, 4/10), \mu_1 = \frac{1}{d} (d-1, \ldots, -1), \mu_2 = \frac{1}{d} (1, \ldots, 1), \Sigma_1 = I_d, \Sigma_2 = 2 \cdot I_d. \)

The error \( \varepsilon \) is a standard normal distribution. To estimate the effectiveness of an estimated e.d.r. subspace, we use the square trace error \( R^2(w, \hat{w}) \) \[ \mathbb{R}^2(w, \hat{w}) = 1 - \frac{1}{k} \text{tr} \left[ (w^\top w)^{-1} w^\top \hat{w} (\hat{w}^\top \hat{w})^{-1} \hat{w}^\top w \right] = 1 - \frac{1}{k} \text{tr} [ P \cdot \hat{P} ]. \]
We consider two models:

1) rational model \( d = 10; \quad y = \frac{x_1}{1/2 + (x_2 + 2)^2} + \sigma \cdot \varepsilon \) with \( \sigma = 1/4 \).

Results are shown of Figure 1 and we can see, that SPHD works better, than PHD+, that is slicing make method more robust.

2) classification problem in form \( y = I(x_1^2 + 2x_2^2 > 4) + \varepsilon / 4, \quad \varepsilon \in \mathcal{N}(0, 1) \)

We can see, that SADE is close to 0.5 (Figure 2). It means that the method finds only one direction in the e.d.r. space. Moreover, SPHD gave better results than PHD+ (using the moments of 10).

![Figure 1: Mean and standard deviation of \( R^2(\mathcal{E}, \hat{\mathcal{E}}) \) for the rational function](image1)

![Figure 2: Mean and standard deviation of \( R^2(\mathcal{E}, \hat{\mathcal{E}}) \) for the the classification problem](image2)

Also we conduct experiments for unknown score functions in two ways: 1-step and 2-step algorithms. We choose basis functions as Gaussian kernels centered in the sample points. Result are shown on Figures 3 and 4.

![Figure 3: Rational model, \( d = 10, n = 1000 \)](image3)

![Figure 4: Rational model, \( d = 20, n = 2000 \)](image4)