ESTIMATING ACTIVE SUBSPACES WITH RANDOMIZED GRADIENT SAMPLING

Farhad Pourkamali-Anaraki, Stephen Becker

SIAM Workshop on Dimension Reduction 2017
July 9-10 · Pittsburg

Summary

In this work, we present an efficient method for estimating active subspaces using only random observations of gradient vectors. Our method is based on the bi-linear representation of low-rank gradient matrices with a novel initialization step for alternating minimization.

Active Subspaces

In modern computer simulations, scientists and engineers seek to study the relationships between high-dimensional spaces of input parameters and quantities of interest. Due to the large number of input parameters and high cost of simulations, many methods have been proposed to reduce the dimension of the input parameter space. These methods often find small subsets or linear combinations of the input parameters that approximately preserve input-output relationships. This low-dimensional characterization of complex problems with hundreds or thousands of input parameters is a crucial tool for modern computer simulations.

Active subspaces are powerful tools for identifying important directions in the high-dimensional space of input parameters [1, 2]. Let $\mathbf{x} \in \mathbb{R}^m$ be a vector of simulation inputs and assume that $f(\mathbf{x}) : \mathbb{R}^m \mapsto \mathbb{R}$ is the mapping between $\mathbf{x}$ and a quantity of interest. The active subspace is defined by the top $n < m$ eigenvectors of the following $m \times m$ symmetric positive semidefinite matrix

$$\mathbf{C} = \int \nabla f(\mathbf{x}) \nabla f(\mathbf{x})^T \rho(\mathbf{x}) d\mathbf{x}, \qquad (1)$$

where $\nabla f(\mathbf{x}) \in \mathbb{R}^m$ is the gradient vector and ρ is a user-specified probability density function. Consider the eigenvalue decomposition of $\mathbf{C} = \mathbf{W}\mathbf{\Lambda}\mathbf{W}^T$, where $\mathbf{\Lambda} \in \mathbb{R}^{m\times m}$ is the diagonal matrix of eigenvalues, listed in decreasing order, and $\mathbf{W} \in \mathbb{R}^{m\times m}$ contains m orthonormal eigenvectors. The matrix $\mathbf{W}$ can then be partitioned: $\mathbf{W} = [\mathbf{W}_1, \mathbf{W}_2]$. The column space of $\mathbf{W}_1 \in \mathbb{R}^{m\times n}$ is the n-dimensional active subspace, where n is usually chosen so the first n eigenvalues are much larger than the remaining $m - n$ eigenvalues.

In high-dimensional settings, computing the integral in (1) for constructing the matrix $\mathbf{C}$ is impractical. Moreover, in some applications, the gradient vector $\nabla f(\mathbf{x})$ may not have a closed-form expression. In such cases, gradients can be approximated by the first-order finite difference with $m + 1$ function evaluations

$$\mathbf{e}_j^T \nabla f(\mathbf{x}) \approx \left(f(\mathbf{x} + h\mathbf{e}_j) - f(\mathbf{x})\right)/h, \quad j = 1, \dots, m, \qquad (2)$$

where $\mathbf{e}_j$ is the j-th canonical basis vector in $\mathbb{R}^m$ and $h > 0$ is the finite difference parameter.

In [4], it is shown that the matrix $\mathbf{C}$ and its eigenpairs can be approximated using the following Monte Carlo method. First, M samples $\mathbf{x}_1, \dots, \mathbf{x}_M \in \mathbb{R}^m$ are drawn i.i.d. according to ρ. Then, $\nabla f_i := \nabla f(\mathbf{x}_i)$ are estimated via (2) using $M(m + 1)$ function evaluations to form

$$\widehat{\mathbf{C}} = \frac{1}{M} \sum_{i=1}^{M} \nabla f_i \nabla f_i^T = \widehat{\mathbf{W}}\widehat{\mathbf{\Lambda}}\widehat{\mathbf{W}}^T. \qquad (3)$$

The leading n eigenvectors of $\widehat{\mathbf{C}}$ provide an accurate estimate of $\mathbf{W}_1$ when M is sufficiently large [4]. These eigenvectors are equivalent to computing the top left singular vectors of the gradient matrix

$$\mathbf{G} := [\nabla f_1, \dots, \nabla f_M] \in \mathbb{R}^{m\times M}. \qquad (4)$$

In this work, we show that the *low-rank* structure of the gradient matrix $\mathbf{G}$ allows us to find accurate estimates of active subspace using fewer function evaluations. In particular, we consider a scheme where only k entries of each gradient vector $\nabla f_i \in \mathbb{R}^m$ are computed uniformly at random, thus the total number of function evaluations required would be $M(k + 1)$. To estimate the active subspace, $\mathbf{G}$ is written in a bi-linear form $\mathbf{G} = \mathbf{A}\mathbf{B}$ and then alternating minimization [5] is used to find $\mathbf{A}$ and $\mathbf{B}$ that best fit the observed entries of gradient matrix. To further improve the performance for small values of k, we use an unbiased estimate of the left singular vectors of $\mathbf{G}$ as the initial point for alternating minimization [6].

Estimating Active Subspaces with Gradient Sampling

Let us define the linear measurement operator $\mathcal{L}(\cdot)$ as

$$\mathcal{L}(\mathbf{G}) := \left[\mathbf{R}_1^T \nabla f_1, \dots, \mathbf{R}_M^T \nabla f_M\right] \in \mathbb{R}^{k\times M}, \qquad (5)$$

where each sampling matrix $\mathbf{R}_i \in \mathbb{R}^{m\times k}$ contains k canonical basis vectors in $\mathbb{R}^m$ chosen uniformly at random without replacement. Given the incomplete observations and the bi-linear parameterization of $\mathbf{G}$, our goal is to minimize

$$\min_{\mathbf{A}\in\mathbb{R}^{m\times n},\mathbf{B}\in\mathbb{R}^{n\times M}} \|\mathcal{L}(\mathbf{G}) - \mathcal{L}(\mathbf{AB})\|_F. \quad (6)$$

This problem can be reformulated by using the vector operator and kronecker product

$$\begin{aligned}\|\mathcal{L}(\mathbf{G}) - \mathcal{L}(\mathbf{AB})\|_F &= \|\mathbf{y} - \mathbf{R}(\mathbf{I}_{M\times M}\otimes \mathbf{A})\mathrm{vec}(\mathbf{B})\|_2 \\ &= \|\mathbf{y} - \mathbf{R}(\mathbf{B}^T\otimes \mathbf{I}_{m\times m})\mathrm{vec}(\mathbf{A})\|_2, \quad (7)\end{aligned}$$

where $\mathbf{y} = \mathbf{R}\mathrm{vec}(\mathbf{G})$ and $\mathbf{R} \in \mathbb{R}^{kM\times mM}$ contains all sampling matrices $\mathbf{R}_i^T$, $i = 1,\ldots,M$, on its main diagonal. Thus, our method iteratively keep one of $\mathbf{A},\mathbf{B}$ fixed and optimize over the other. Each subproblem is convex and can be solved efficiently

$$\begin{aligned}\mathrm{vec}(\mathbf{B}) &\leftarrow [\mathbf{R}(\mathbf{I}_{M\times M}\otimes\mathbf{A})]^\dagger\,\mathbf{y},\\ \mathrm{vec}(\mathbf{A}) &\leftarrow \left[\mathbf{R}(\mathbf{B}^T\otimes\mathbf{I}_{m\times m})\right]^\dagger\mathbf{y}. \quad (8)\end{aligned}$$

In [3], estimation of active subspaces is considered in a similar framework where each $\mathbf{R}_i \in \mathbb{R}^{m\times k}$ has independent standard Gaussian entries. As we see from the update rule in (8), our method is more efficient in terms of computation and memory. The proposed method in [3] must store Mmk nonzero entries of matrix $\mathbf{R}$, whereas our method requires only Mk nonzero entries to be stored. Similarly, our method reduces the cost of matrix-matrix multiplications. We expect our proposed sampling method to perform as well as Gaussian samples when the active subspaces are incoherent with the standard basis.

Another contribution of our work is the novel initialization step based on [6]. The initial iterate for alternating minimization is preferred to be chosen based on a *good* estimate of $\mathbf{A}$, rather than random initializations, to guarantee the convergence [5]. The recent work [6] presents an unbiased estimator for the matrix $\widehat{\mathbf{C}}$

$$\widetilde{\mathbf{\Sigma}} := \widetilde{\mathbf{C}} - \eta\,\mathrm{diag}(\widetilde{\mathbf{C}}), \quad \eta = \frac{m-k}{m-1}, \quad (9)$$

where $\mathrm{diag}(\widetilde{\mathbf{C}})$ represents the matrix formed by zeroing all but the diagonal elements of $\widetilde{\mathbf{C}}$, which is defined as

$$\widetilde{\mathbf{C}} := \frac{m(m-1)}{k(k-1)}\frac{1}{M}\sum_{i=1}^{M}(\mathbf{R}_i\mathbf{R}_i^T\nabla f_i)(\mathbf{R}_i\mathbf{R}_i^T\nabla f_i)^T. \quad (10)$$

Numerical Experiments

Let $\mathbf{H} \in \mathbb{R}^{m\times m}$ be symmetric positive semidefinite and $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x}$, defined on the domain $\mathbf{x} \in [-1,1]^m$ with a uniform density ρ. Thus, the gradient is $\nabla f(\mathbf{x}) = \mathbf{H}\mathbf{x}$. The eigenvalues of $\mathbf{C}$ in are the eigenvalues of $\mathbf{H}$, squared and divided by 3. Moreover, the eigenvectors of $\mathbf{C}$ and $\mathbf{H}$ are identical. The matrix $\mathbf{H}$ is constructed so that its eigenvalues decay at a slow rate, except for a large gap between the fifth and sixth eigenvalues. We set parameters $m = 100$, $M = 2000$, and $n = 5$. The subspace estimation error is defined as $\mathcal{E} := \|\widehat{\mathbf{W}}_1\widehat{\mathbf{W}}_1^T - \widetilde{\mathbf{W}}_1\widetilde{\mathbf{W}}_1^T\|_2$, where $\widetilde{\mathbf{W}}_1$ is the active subspace estimate using incomplete gradients. In Fig. 1, the mean estimation error over 100 trials is reported for various values of measurements k. Alternating minimization with 20 iterations is used in two cases: (1) our proposed initial point based on (9), and (2) random initialization based on a Gaussian matrix.

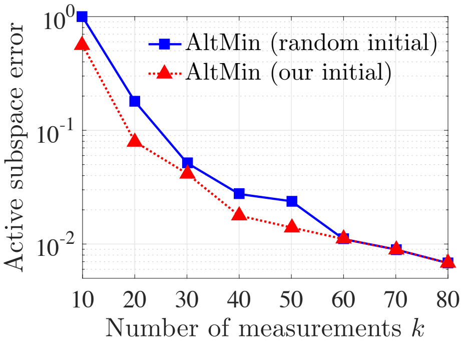

Figure 1: Active subspace estimation error for varying number of measurements k and fixed dimension $m = 100$.

A good initialization point becomes more important for small values of k, which are crucial for large-scale problems. For example, at $k = 20$, the mean estimation errors for our proposed initial point and random initialization are 0.08 and 0.18, respectively. Thus, our initialization procedure reduces the error by almost a factor of 2 in this case.

References

[1] P. Constantine. *Active subspaces: Emerging ideas for dimension reduction in parameter studies.* SIAM, 2015.

[2] P. Constantine, E. Dow, and Q. Wang. Active subspace methods in theory and practice: applications to kriging surfaces. *SIAM Journal on Scientific Computing*, 36(4):A1500–A1524, 2014.

[3] P. Constantine, A. Eftekhari, and M. Wakin. Computing active subspaces efficiently with gradient sketching. In *CAMSAP*, pages 353–356, 2015.

[4] P. Constantine and D. Gleich. Computing active subspaces with Monte Carlo. *arXiv preprint arXiv:1408.0545*, 2014.

[5] P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In *ACM symposium on Theory of computing*, pages 665–674, 2013.

[6] F. Pourkamali-Anaraki and S. Becker. Preconditioned data sparsification for big data with applications to PCA and K-means. *IEEE Transactions on Information Theory*, 2017.