Computing Active Subspaces Efficiently with Gradient Sketching

Paul G. Constantine

Applied Mathematics and Statistics

Colorado School of Mines

Golden, Colorado 80401

Email: pconstan@mines.edu

Armin Eftekhari and Michael B. Wakin

Electrical Engineering and Computer Science

Colorado School of Mines

Golden, Colorado 80401

Email: aeftekha,mwakin@mines.edu

Abstract—Active subspaces are an emerging set of tools for identifying and exploiting the most important directions in the space of a computer simulation's input parameters; these directions depend on the simulation's quantity of interest, which we treat as a function from inputs to outputs. To identify a function's active subspace, one must compute the eigenpairs of a matrix derived from the function's gradient, which presents challenges when the gradient is not available as a subroutine. We numerically study two methods for estimating the necessary eigenpairs using only linear measurements of the function's gradient. In practice, these measurements can be estimated by finite differences using only two function evaluations, regardless of the dimension of the function's input space.

I. ACTIVE SUBSPACES

Modern physics and engineering simulations take several inputs—e.g., boundary conditions, material properties, and forcings—and output several quantities of interests. The scientist uses these simulations to study the relationship between inputs and outputs. Uncertainty quantification seeks precise characterization of the simulation's quantities of interest subject to variability in the inputs. These characterizations often reduce to parameter studies—such as optimization, integration, or response surface modeling—that treat the simulation as a mapping between inputs $\mathbf{x}$ and a quantity of interest $f(\mathbf{x})$. However, thorough parameter studies quickly become infeasible as the dimension of $\mathbf{x}$ grows, particularly if evaluating $f(\mathbf{x})$ (i.e., running the physical simulation) is computationally expensive. To combat this curse of dimensionality, one may seek a low-dimensional parameterization of $f(\mathbf{x})$ that (i) maintains the input/output representation and (ii) enables otherwise infeasible parameter studies. One idea is to identify the least important input parameters and fix them at nominal values, thus reducing the dimension of the parameter study. Such identification is the domain of *sensitivity analysis*, and several techniques exist that use a few simulation runs to screen the inputs' importance—such as local perturbations, elementary effects, or sensitivity indices [1]. A more general approach is to identify important linear combinations of the inputs $\mathbf{x}$ and focus parameter studies along the associated directions. *Active subspaces* are defined by important directions in the high-dimensional space of inputs; once identified, the scientist can exploit the active subspace to enable otherwise infeasible parameter studies for expensive simulations [2].

Assume that $\mathbf{x} \in \mathbb{R}^m$ is a vector of simulation inputs, and let the input space be equipped with a probability density function $\rho(\mathbf{x})$ that is strictly positive in the domain of f and zero outside the domain. In practice, ρ identifies the set of inputs of interest and quantifies the variability. We assume that the independent inputs have been shifted and scaled to be centered at the origin and have equal variances. Assume that $f : \mathbb{R}^m \to \mathbb{R}$ is continuous, square-integrable with respect to ρ, and differentiable with gradient vector $\nabla f \in \mathbb{R}^m$; also assume that f's gradient is square-integrable with respect to ρ. The active subspace is defined by the first $n < m$ eigenvectors of the following $m \times m$ symmetric positive semi-definite matrix,

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

where the non-negative eigenvalues are ordered in descending order. The eigenvalue λ_i measures the average change in f subject to perturbations in $\mathbf{x}$ along the corresponding eigenvector $\mathbf{w}_i$,

$$\lambda_i = \int (\nabla f^T \mathbf{w}_i)^2 \, \rho \, d\mathbf{x}. \qquad (2)$$

For example, if $\lambda_i = 0$, then f is constant along the direction $\mathbf{w}_i$. If f is constant along a direction, then one can ignore this direction when studying the behavior of f under changes in $\mathbf{x}$. Suppose that the first $n < m$ eigenvalues are much larger than the trailing $m - n$, and let $\mathbf{W}_1$ be the first n columns of the orthogonal eigenvector matrix $\mathbf{W}$. Then a reasonable approximation for f is

$$f(\mathbf{x}) \approx g(\mathbf{W}_1^T \mathbf{x}), \qquad (3)$$

where g is a properly constructed map from $\mathbb{R}^n$ to $\mathbb{R}$. In [3], we study a particular construction of g, its approximation properties, and how those properties change when $\mathbf{W}_1$ is estimated.

There are several recent works that study models similar to (3) in uncertainty quantification [4], [5], computational engineering [6]–[9], and approximation theory [10], [11]. In statistics, subspace-based dimension reduction has wide use in regression modeling under the monikers *sufficient dimension reduction* [12] and *efficient dimension reduction* [13]. Many in that community have recognized the connection between the dimension reduction space and the matrix $\mathbf{C}$ in (1); see [14], [15]. However, in contrast to our case where the gradient ∇f is given, the gradient of the regression function with respect to predictors must be learned. Additionally, there is no random noise in computer simulations in contrast to regression modeling.

The tremendous potential benefits of dimension reduction drive us to pursue methods to estimate the eigenvalues Λ and eigenvectors $\boldsymbol{W}$ from (1). In [16], we analyze the following Monte Carlo method. First, draw M samples $\{\mathbf{x}_i\}$ independently according to ρ. For each $\mathbf{x}_i$, compute $\nabla f_i = \nabla f(\mathbf{x}_i)$. Then compute

$$\boldsymbol{C} \approx \widehat{\boldsymbol{C}} = \frac{1}{M}\sum_{i=1}^{M} \nabla f_i \nabla f_i^T = \widehat{\boldsymbol{W}}\widehat{\Lambda}\widehat{\boldsymbol{W}}^T. \qquad (4)$$

In [16], we use tools from nonasymptotic random matrix theory to study the approximation error in the estimated eigenvalues and subspaces [17], [18]. The approach assumes that one has access to the gradient $\nabla f(\mathbf{x})$ as a black box, which is not unreasonable in modern simulation codes due to adjoint methods or algorithmic differentiation [19]. However, many legacy simulation codes lack gradient routines, and one must estimate the eigenpairs with only evaluations of $f(\mathbf{x})$.

II. ESTIMATING ACTIVE SUBSPACES WITHOUT GRADIENTS

When the gradient is not available as a black box, one might estimate the gradient through evaluations of f, e.g., with finite differences or related techniques [20]. A first-order finite difference approximation to the gradient vector requires $m+1$ evaluations of f, so estimating $\widehat{\boldsymbol{C}}$ in (4) takes $M(m+1)$ evaluations, which may be prohibitively expensive for large-scale simulations. To surmount this challenge, we take advantage of a finite difference approximation of directional derivatives. For sufficiently small $h > 0$ and $\mathbf{a} \in \mathbb{R}^m$,

$$\nabla f(\mathbf{x})^T \mathbf{a} \approx (f(\mathbf{x} + h\mathbf{a}) - f(\mathbf{x}))/h. \qquad (5)$$

The left side of (5) can be interpreted as a linear measurement of the gradient vector, and the right side takes only two evaluations of f—regardless of the dimension m. We study two techniques that exploit (5) to estimate $\boldsymbol{W}_1$ from (1) using only evaluations of f. The relationship (5) is also exploited in the works [10], [11] to reconstruct functions of the form $f(\mathbf{x}) = g(\boldsymbol{A}\mathbf{x})$, where $\boldsymbol{A} \in \mathbb{R}^{n \times m}$. Note the subtle difference between this reconstruction and the approximation model in (3). Our goal in this paper is to estimate $\boldsymbol{W}_1$, not propose a choice for g in (3). Another important difference is in the linear measurement operator; [10], [11] use the same operator for each ∇f_i in (4), while we use an independent measurement operator for each ∇f_i.

A. Eigenvectors from projections

Our first approach is based on work by Qi and Hughes [21] for estimating principal components from linear measurements of a collection of vectors. Suppose that $\mathbf{z}_i \in \mathbb{R}^m$, $i = 1, \ldots, M$, are independent, zero-mean random vectors of the form

$$\mathbf{z}_i = \sum_{j=1}^{d} w_{ij}\, \sigma_j \,\mathbf{v}_j, \qquad w_{ij} \sim \mathcal{N}(0,1), \qquad (6)$$

where the $\mathbf{v}_j$ are orthonormal vectors. Let $\boldsymbol{E}_i \in \mathbb{R}^{m \times k}$, $k < m$, have independent standard Gaussian entries, and define the measurements $\mathbf{m}_i = \boldsymbol{E}_i^T \mathbf{z}_i$. Then the orthogonal projection of $\mathbf{z}_i$ onto the column space of $\boldsymbol{E}_i$ is

$$\mathcal{P}_i \mathbf{z}_i := \boldsymbol{E}_i(\boldsymbol{E}_i^T \boldsymbol{E}_i)^{-1}\mathbf{m}_i = \boldsymbol{E}_i(\boldsymbol{E}_i^T \boldsymbol{E}_i)^{-1}\boldsymbol{E}_i^T \mathbf{z}_i. \qquad (7)$$

Qi and Hughes [21, Theorem 2] show that the first d eigenvectors of the matrix

$$\frac{1}{M}\sum_{i=1}^{M} \mathcal{P}_i \mathbf{z}_i \,(\mathcal{P}_i \mathbf{z}_i)^T \qquad (8)$$

converge to the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_d$ as M goes to infinity. The eigenvalues converge to quantities related to σ_i from (6). In fact, their results show how to estimate the mean and principal components even when the model (6) has a nonzero mean and random noise.

There is no reason to suspect that the gradient vectors ∇f_i from (4) satisfy the model (6). Nonetheless, we can numerically check if the eigenvectors of the matrix

$$\widehat{\boldsymbol{C}}_{\mathcal{P}} = \frac{1}{M}\sum_{i=1}^{M} \mathcal{P}_i \nabla f_i \,(\mathcal{P}_i \nabla f_i)^T, \qquad (9)$$

where $\mathcal{P}_i$ is defined as in (7), are close to the eigenvectors of (4) for chosen test problems. If so, we can exploit the finite difference relationship (5) to efficiently estimate each element of the k-vector $\boldsymbol{E}_i^T \nabla f_i$, which is the analogue of $\mathbf{m}_i$ in (7). In this case, we can estimate the eigenvectors $\widehat{\boldsymbol{W}}$ from (4) using $M(k+1) < M(m+1)$ evaluations of f. Note that, by the analysis in [21], we do not expect the eigenvalues of $\widehat{\boldsymbol{C}}_{\mathcal{P}}$ to converge to those of $\widehat{\boldsymbol{C}}$.

B. Low-rank approximation from linear measurements

Our second approach is based on low-rank approximation of the matrix of gradients $\widehat{\boldsymbol{G}} = [\nabla f_1, \ldots, \nabla f_M] \in \mathbb{R}^{m \times M}$ using linear measurements of the gradients. Define the linear measurement operator $\mathcal{M}(\cdot)$ as

$$\mathcal{M}(\widehat{\boldsymbol{G}}) := \begin{bmatrix} \boldsymbol{E}_1^T \nabla f_1 & \cdots & \boldsymbol{E}_M^T \nabla f_M \end{bmatrix}, \qquad (10)$$

where $\boldsymbol{E}_i \in \mathbb{R}^{m \times k}$ has independent standard Gaussian entries as in the previous section. Let r be the rank of the low-rank approximation. We seek matrices $\boldsymbol{A} \in \mathbb{R}^{m \times r}$ and $\boldsymbol{B} \in \mathbb{R}^{M \times r}$ that solve

$$\underset{\boldsymbol{A},\,\boldsymbol{B}}{\text{minimize}} \; \|\mathcal{M}(\widehat{\boldsymbol{G}}) - \mathcal{M}(\boldsymbol{A}\boldsymbol{B}^T)\|_F, \qquad (11)$$

where $\|\cdot\|_F$ is the Frobenius norm. We estimate the minimizers with alternating least-squares. Given $\boldsymbol{A}$, (11) is a linear least-squares problem for $\boldsymbol{B}$. Similarly, given $\boldsymbol{B}$, (11) is a linear least-squares problem for $\boldsymbol{A}$. We choose a starting value for $\boldsymbol{A}$ as the first r eigenvectors, scaled by the square-roots of the first r eigenvalues, of $\widehat{\boldsymbol{C}}_{\mathcal{P}}$ in (9). In this sense, the alternating least-squares can be considered an iterative refinement on the estimates from the first method. Once $\boldsymbol{A}$ and $\boldsymbol{B}$ are estimated, we compute the left singular vectors of $\boldsymbol{A}\boldsymbol{B}^T$ to estimate $\boldsymbol{W}_1$ in (3).

The low-rank model $\boldsymbol{A}\boldsymbol{B}^T$ requires the user to choose the rank r. To solve the least-squares subproblem for $\boldsymbol{A}$ without additional regularization, we need r less than the number k of linear measurements (i.e., the number of columns in $\boldsymbol{E}_i$). Additionally, we are guided by the ultimate goal of the dimension reduction, which is to construct g in (3). Without prior knowledge of the relationship between f and $\mathbf{x}$, constructing a response surface is generally too expensive in more than a handful of dimensions. Thus, we may reasonably keep r less than 8 or 9. If there is no gap within the first 8 or 9 eigenvalue estimates, then (3) may be inappropriate.

III. Experiments

We test the two methods on two functions: (i) a quadratic polynomial in $m = 10$ dimensions and (ii) a quantity of interest from the solution of a PDE whose operator coefficients depend on $m = 100$ variables. In each experiment, we compute $\widehat{C}$ from (4) with a fixed number M of gradient samples, and we consider the eigenpairs $\widehat{W}$ and $\widehat{\Lambda}$ to be the *true* values. We compute the error in the first six eigenvalue estimates and the error in the subspaces defined by the eigenvectors estimates as

$$\left(\frac{\sum_{i=1}^{6} (\hat{\lambda}_i - \tilde{\lambda}_i)^2}{\sum_{i=1}^{6} \hat{\lambda}_i^2} \right)^{1/2} \tag{12}$$

and

$$\|\widehat{W}_1 \widehat{W}_1^T - \tilde{W}_1 \tilde{W}_1^T\|_2, \tag{13}$$

respectively, where $\tilde{\lambda}_i$ and $\tilde{W}$ are the eigenvalue and eigenvector estimates, respectively, from the linear measurement-based approaches. The subscript 1 on the matrices $\widehat{W}$ and $\tilde{W}$ indicates that they contain only the first n columns; we note n when needed. We do not expect the eigenvalue estimates from the projection-based method in Section II-A to converge, but we report them to compare with the errors from the alternating least-squares approach. We repeat the study 20 times with independently drawn Gaussian measurement matrices E_i. The resulting errors are averaged over these 20 trials. We do not study the approximation properties of the finite differences in (5), e.g., by varying the finite difference parameter h. Instead, we examine cases where the gradient is available, thus focusing on the performance of the linear measurement-based methods when the answer is known.

A. Quadratic function

Let $H \in \mathbb{R}^{10 \times 10}$ be symmetric and positive semidefinite, and let $f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T H \mathbf{x}$, defined on the domain $\mathbf{x} \in [-1, 1]^{10}$ with a uniform density ρ. The gradient is $\nabla f(\mathbf{x}) = H\mathbf{x}$. The eigenvectors of C from (1) are the eigenvectors of H, and the eigenvalues of C are the eigenvalues of H, squared and divided by 3. We construct H so that its eigenvalues decay at a slow spectral rate, except for a large gap between the third and fourth eigenvalues; this indicates that the active subspace is three-dimensional. We compute $\widehat{C}$ from (4) using $M = 200$ gradient samples. We study the quality of the estimates as the number k of measurements goes from 4 to 9. Note that $k = 10$ measurements would produce a perfect reconstruction, almost surely. For the alternating least-squares method, we choose the rank $r = 4$. The caption in Figure 1 describes the contents of the subfigures showing the results. In general, we see the least-squares method outperform the projection-based method once the number of measurements exceeds the rank $r = 4$.

B. PDE model

Let $u = u(\mathbf{s}, \mathbf{x})$ solve the Poisson equation in two spatial dimensions,

$$-\nabla_\mathbf{s} \cdot (a \, \nabla_\mathbf{s} u) = 1, \qquad \mathbf{s} \in [0, 1]^2, \tag{14}$$

with homogeneous Dirichlet boundary conditions on the left, top, and bottom of the domain and homogeneous Neumann conditions on the right side of the domain. The spatially varying operator coefficients $a = a(\mathbf{s}, \mathbf{x})$ are parameterized

by $m = 100$ independent parameters $\mathbf{x} \in \mathbb{R}^{100}$ via a truncated Karhunen-Loeve expansion of a Gaussian random field with a relatively long correlation length. This implies that ρ is a standard Gaussian density in 100 dimensions. The quantity of interest $f(\mathbf{x})$ is the average of the solution u over the right side of the spatial domain. The solution is approximated with a well-resolved finite element method, and the gradient $\nabla f(\mathbf{x})$ is computed with an adjoint scheme. More information on this problem can be found in [3].

The matrix $\widehat{C}$ from (4) is estimated with $M = 300$ gradient samples. The particular quantity of interest has a dominant one-dimensional active subspace; in other words, there is a large gap between the first and second eigenvalues of $\widehat{C}$. We study the approximations as the number k of measurements increases by 20 from 10 to 90. For the least-squares method, we choose the rank $r = 8$. The caption in Figure 2 describes the contents of the subfigures showing the results. In general, the alternating least-squares method outperforms the projection-based method. Note the sharp decrease in the error once the number k of measurements exceeds 50.

IV. Conclusion

We have studied the approximation properties of two methods for estimating the eigenvectors and eigenvalues that define a multivariate function's active subspace. These methods use linear measurements of the gradient instead of the full gradient, which can be efficiently computed with finite differences. The first method, based on work by Qi and Hughes [21], uses random projections of the gradient. The second uses a low-rank approximation of the matrix of gradient measurements fit with an alternating least-squares method. The low-rank approximation performs better than the projection-based method in two numerical tests. Future work will (i) analyze the improvement with the low-rank approximation and (ii) study the effect of finite difference approximations of the linear measurements. These initial results are sufficiently promising to pursue such studies.

References

[1] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola, *Global Sensitivity Analysis: The Primer*. John Wiley & Sons, Hoboken, 2008.

[2] P. G. Constantine, *Active Subspaces: Emerging Ideas for Dimension Reduction in Parameter Studies*. SIAM, Philadelphia, 2015.

[3] P. Constantine, E. Dow, and Q. Wang, "Active subspace methods in theory and practice: Applications to kriging surfaces," *SIAM Journal on Scientific Computing*, vol. 36, no. 4, pp. A1500–A1524, 2014.

[4] R. Tipireddy and R. Ghanem, "Basis adaptation in homogeneous chaos spaces," *Journal of Computational Physics*, vol. 259, pp. 304 – 317, 2014.
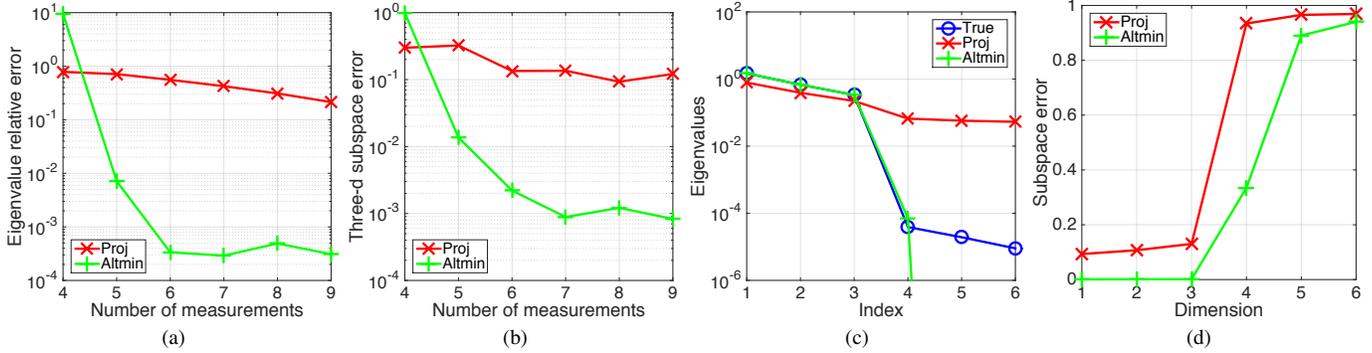
Fig. 1. In the legends, "Proj" indicates the projection-based method from Qi and Hughes [21] described in Section II-A, and "Altmin" indicates the alternating least-squares minimization from Section II-B. Figure 1a shows the error (12) in the first six eigenvalue estimates as a function of the number k of measurements; note that we do not expect the eigenvalues from the projection-based method to converge. Figure 1b shows the error in the estimate of the three-dimensional active subspace (13) as a function of the number of measurements. Figure 1c shows the first six eigenvalues using $k = 7$ measurements. Figure 1d shows the errors in the subspace estimates for dimension $n = 1$ to $n = 6$ using $k = 7$ measurements.
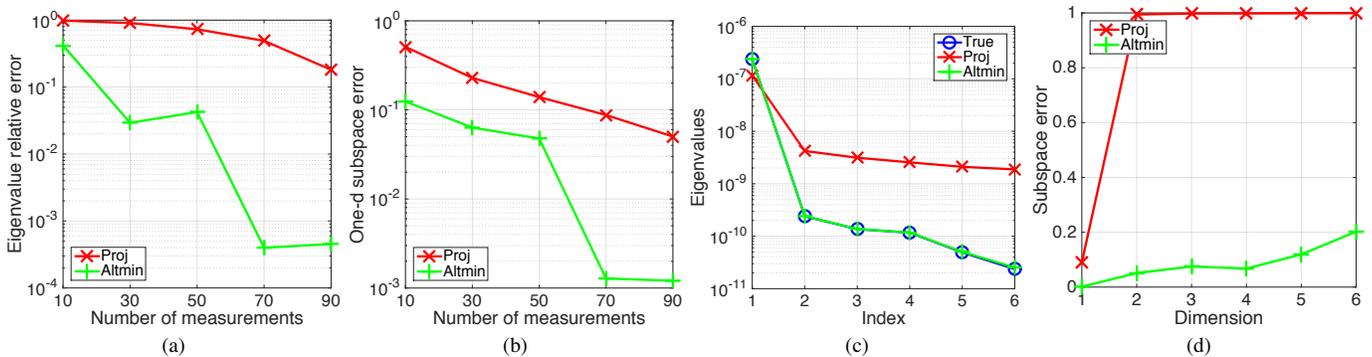

Fig. 2. In the legends, "Proj" indicates the projection-based method from Qi and Hughes [21] described in Section II-A, and "Altmin" indicates the alternating least-squares minimization from Section II-B. Figure 2a shows the error (12) in the first six eigenvalue estimates as a function of the number k of measurements; note that we do not expect the eigenvalues from the projection-based method to converge. Figure 2b shows the error in the estimate of the one-dimensional active subspace (13) as a function of the number k of measurements. Figure 2c shows the first six eigenvalues using $k = 70$ measurements. Figure 2d shows the errors in the subspace estimates for dimension $n = 1$ to $n = 6$ using $k = 70$ measurements.

[5] M. Stoyanov and C. G. Webster, "A gradient-based sampling approach for dimension reduction of partial differential equations with stochastic coefficients," *International Journal for Uncertainty Quantification*, vol. 5, no. 1, pp. 49–72, 2015.

[6] Y. Bang, H. S. Abdel-Khalik, and J. M. Hite, "Hybrid reduced order modeling applied to nonlinear models," *International Journal for Numerical Methods in Engineering*, vol. 91, no. 9, pp. 929–949, 2012.

[7] S. H. Berguin and D. N. Mavris, "Dimensionality reduction using principal component analysis applied to the gradient," *AIAA Journal*, vol. 53, no. 4, pp. 1078–1090, 2014.

[8] P. Constantine, M. Emory, J. Larsson, and G. Iaccarino, "Exploiting active subspaces to quantify uncertainty in the numerical simulation of the HyShot II scramjet," *Journal of Computational Physics*, vol. 302, pp. 1 – 20, 2015.

[9] T. M. Russi, "Uncertainty quantification with experimental data and complex system models," Ph.D. dissertation, UC Berkeley, 2010.

[10] M. Fornasier, K. Schnass, and J. Vybiral, "Learning functions of few arbitrary linear parameters in high dimensions," *Foundations of Computational Mathematics*, vol. 12, pp. 229–262, 2012.

[11] H. Tyagi and V. Cevher, "Learning non-parametric basis independent models from point queries via low-rank methods," *Applied and Computational Harmonic Analysis*, vol. 37, no. 3, pp. 389 – 412, 2014.

[12] R. D. Cook, *Regression Graphics: Ideas for Studying Regressions through Graphics*. John Wiley & Sons, Hoboken, 2009.

[13] 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*, vol. 87, no. 420, pp. 1025–1039, 1992.

[14] A. Samarov, "Exploring regression structure using nonparametric functional estimation," *Journal of the American Statistical Association*, vol. 88, no. 423, pp. 836–847, 1993.

[15] M. Hristache, A. Juditsky, J. Polzehl, and V. Spokoiny, "Structure adaptive approach for dimension reduction," *The Annals of Statistics*, vol. 29, no. 6, pp. 1537–1566, 12 2001.

[16] P. Constantine and D. Gleich, "Computing active subspaces with Monte Carlo," *arXiv preprint arXiv:1408.0545v2*, 2015.

[17] A. Gittens and J. A. Tropp, "Tail bounds for all eigenvalues of a sum of random matrices," *arXiv preprint arXiv:1104.4513*, 2011.

[18] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Foundations of Computational Mathematics*, vol. 12, no. 4, pp. 389–434, 2012.

[19] A. Griewank, *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*. SIAM, Philadelphia, 2000.

[20] A. R. Conn, K. Scheinberg, and L. N. Vicente, *Introduction to Derivative-Free Optimization*. SIAM, Philadelphia, 2009.

[21] H. Qi and S. Hughes, "Invariance of principal components under low-dimensional random projection of the data," in *2012 19th IEEE International Conference on Image Processing (ICIP)*, Sept 2012, pp. 937–940.