OVERCOMING NOISE, AVOIDING CURVATURE: OPTIMAL SCALE SELECTION FOR TANGENT PLANE RECOVERY

Daniel N. Kaslovsky and François G. Meyer

Department of Applied Mathematics, University of Colorado, Boulder, CO 80309
{kaslovsky, fmeyer}@colorado.edu

ABSTRACT

Constructing an efficient parametrization of a large, noisy data set of points lying close to a smooth manifold in high dimension remains a fundamental problem. One approach consists in recovering a local parametrization using the local tangent plane. Principal component analysis (PCA) is often the tool of choice, as it returns an optimal basis in the case of noise-free samples from a linear subspace. To process noisy data, PCA must be applied locally, at a scale small enough such that the manifold is approximately linear, but at a scale large enough such that structure may be discerned from noise. Using eigenspace perturbation theory, we adaptively select the scale that minimizes the angle between the subspace estimated by PCA and the true tangent space, revealing the optimal scale for local tangent plane recovery.

Index Terms— Manifold-valued data, tangent plane, principal component analysis, local linear models, curvature, noise

1. INTRODUCTION

Large data sets of points in high-dimension often lie close to a smooth low-dimensional manifold. A fundamental problem in processing such data sets is the construction of an efficient parameterization that allows for the data to be well represented in fewer dimensions. Such a parameterization may be realized by exploiting the inherent manifold structure of the data. However, discovering the geometry of an underlying manifold from only noisy samples remains an open topic of research.

The optimal parameterization for data sampled from a linear subspace is given by principal component analysis (PCA). However, most interesting manifold-valued data organize on or near a nonlinear manifold. PCA, by projecting data points onto the linear subspace of best fit, is not optimal in this case, as curvature may only be accommodated by choosing a subspace of dimension higher than that of the manifold. Such data are typically processed either via a global, nonlinear embedding, or in a piecewise-linear fashion allowing local application of linear methods (PCA). The latter is the subject of this work.

There have been several versions of localized PCA for tangent plane recovery proposed in the literature. While the need for locality has been acknowledged, a precise treatment of the size of the local neighborhood is often not addressed. The appropriate neighborhood size must be a function of intrinsic (manifold) dimensionality, curvature, and noise level. Despite the fact that these properties may change as different regions of the manifold are explored, locality is often defined via an *a priori* fixed number of neighbors or as the output of an algorithm (e.g., [1, 2]). Other methods (e.g., [3]) adaptively estimate local neighborhood size but are not equipped with optimality guarantees.

The selection of the optimal scale, or neighborhood size, for local tangent plane recovery is the key contribution of this paper. We use eigenspace perturbation theory to study the stability of the tangent plane as the size of the neighborhood varies. We bound, with high probability, the angle between the recovered linear subspace and the true tangent plane. In doing so, we are able to adaptively select the neighborhood that minimizes this bound, yielding the best approximate tangent plane.

Our approach is similar to the analysis presented by Nadler [4], who studies the finite-sample properties of the PCA spectrum using a linear data model. The present work recovers the results of [4] in the curvature-free setting and therefore generalizes the study of Nadler to noisy samples from a nonlinear manifold model. Other recent related works include that of Singer and Wu [5], who use local PCA to build a tangent plane basis in the absence of noise, that of Zhang, *et al.* [6], who use Jones' β-number to recover local "flats" (affine subspaces) from which data are assumed to be sampled, and that of Maggioni and coauthors [7], in which multiscale PCA is used to discover the intrinsic dimensionality of a data set.

2. PROBLEM FORMULATION

Our goal is to recover the best approximation to a local tangent space of a nonlinear d-dimensional Riemannian manifold $\mathcal{M}$ from noisy samples presented in dimension $D > d$. Working about a reference point x_0, an approximation to the linear tangent space of $\mathcal{M}$ at x_0 is given by the span of the d singular vectors associated with the largest d singular values of the centered data matrix. The question becomes: at which scale (number of neighbors or radius about x_0) may we recover the best approximation? The quality of PCA approximation to a noisy linear subspace increases as more points are included, but the curvature of $\mathcal{M}$ prevents the inclusion of a large number of points. Similarly, while a very small radius about x_0 avoids curvature ($\mathcal{M}$ locally resembles Euclidean space), the sample points are indistinguishable from noise at small scales.

2.1. Geometric Data Model

A d-dimensional manifold of codimension 1 may be described locally by the surface $y = f(\ell_1, \ldots, \ell_d)$, where ℓ_i is a coordinate in the tangent plane. After translating the origin, a rotation of the coordinate system can align the coordinate axes with the principal directions associated with the principal curvatures at the given reference point x_0. Using this choice of coordinates, the manifold may

This work was supported by NSF Project DMS-0941476, NSF-IGERT in Computational Optical Sensing and Imaging, and DOE Award DE-SC0004096.

978-1-4673-0183-1/12/$31.00 ©2012 IEEE

be described locally [8] by the Taylor series of f at the origin x_0:

$$y = f(\ell_1, \ldots, \ell_d) = \frac{1}{2}(\kappa_1 \ell_1^2 + \cdots + \kappa_d \ell_d^2) + o\left(\ell_1^2 + \cdots + \ell_d^2\right) \quad (1)$$

where $\kappa_1, \ldots, \kappa_d$ are the principal curvatures of $\mathcal{M}$ at x_0. Generalizing to a d-dimensional manifold of arbitrary codimension in $\mathbb{R}^D$ and truncating the Taylor expansion to retain only the quadratic approximation, there exist $(D-d)$ functions

$$f_i(\ell) = \frac{1}{2}(\kappa_1^{(i)} \ell_1^2 + \cdots + \kappa_d^{(i)} \ell_d^2) \quad (2)$$

for $i = (d+1), \ldots, D$, with $\kappa_1^{(i)}, \ldots, \kappa_d^{(i)}$ representing the principal curvatures in codimension i at x_0. Then, given the coordinate system aligned with the principal directions, a point in a neighborhood of x_0 has coordinates $[\ell_1, \ldots, \ell_d, f_{d+1}, \ldots, f_D]$.

Consider now discrete samples from $\mathcal{M}$ that are contaminated with an additive Gaussian noise vector e drawn from the $\mathcal{N}\left(0, \sigma^2 I_D\right)$ distribution. Each sample x is a D-dimensional vector and N such samples may be stored as columns of a matrix $X \in \mathbb{R}^{D \times N}$. The coordinate system above allows the decomposition of x into its linear (tangent plane) component ℓ, its quadratic (curvature) component c, and noise e, three D-dimensional vectors

$$\ell = [\ell_1 \; \ell_2 \; \cdots \; \ell_d \; 0 \; \cdots \; 0]^T \quad (3)$$

$$c = [0 \; \cdots \; 0 \; c_{d+1} \cdots c_D]^T \quad (4)$$

$$e = [e_1 \; e_2 \; \cdots \; e_D]^T \quad (5)$$

such that the last $(D-d)$ entries of c are of the form $c_i = f_i(\ell)$, $i = (d+1), \ldots, D$. We may store the N samples of ℓ, c, and e as columns of matrices L, C, E, respectively, such that our data matrix is decomposed as

$$X = L + C + E. \quad (6)$$

REMARK. We will ultimately quantify the norm of each matrix using the unitarily-invariant Frobenius norm. Thus the rotation needed to realize this coordinate system will not affect our analysis. We therefore proceed by assuming that the coordinate axes align with the principal directions.

2.2. Perturbation of Invariant Subspaces

The tangent plane we wish to recover is given by the PCA of L. Because we do not have direct access to L, we work with X as a proxy. The problem may be posed as a perturbation analysis of invariant subspaces. Let $\widetilde{M}$ denote the centered version of a matrix M as required by PCA. Then, given the decomposition of the data (6), we have

$$\frac{1}{N}\widetilde{X}\widetilde{X}^T = \frac{1}{N}\widetilde{L}\widetilde{L}^T + \Delta, \quad (7)$$

where

$$\Delta = \frac{1}{N}(\widetilde{C}\widetilde{C}^T + \widetilde{E}\widetilde{E}^T + \widetilde{L}\widetilde{C}^T + \widetilde{C}\widetilde{L}^T + \widetilde{L}\widetilde{E}^T + \widetilde{E}\widetilde{L}^T + \widetilde{C}\widetilde{E}^T + \widetilde{E}\widetilde{C}^T) \quad (8)$$

is the perturbation that prevents us from working directly with $\widetilde{L}\widetilde{L}^T$. The dominant eigenspace of $\widetilde{X}\widetilde{X}^T$ is therefore a perturbed version of the dominant eigenspace of $\widetilde{L}\widetilde{L}^T$. Seeking to minimize the effect of this perturbation, we look for the scale N^* (number of neighbors) at which the dominant eigenspace of $\widetilde{X}\widetilde{X}^T$ is closest to that of $\widetilde{L}\widetilde{L}^T$. We may define N^* to be the scale that minimizes the angle between these subspaces:

$$N^* = \arg\min_N \|P - \widehat{P}\|_F, \quad (9)$$

where P and $\widehat{P}$ are the orthogonal projectors onto the subspaces computed from L and X, respectively. Note that the Frobenius norm replaces the spectral norm to allow a simplification of Theorem 1. The solution to (9) is the main goal of this work.

The distance $\|P - \widehat{P}\|_F$ may be bounded by the classic $\sin\Theta$ theorem of Davis and Kahan [9]. We will use a version of this theorem presented by Stewart (Theorem V.2.7 of [10]), as it most efficiently facilitates our analysis. First, we establish some notation, following closely that found in [10]. Consider the eigendecompositions

$$\frac{1}{N}\widetilde{L}\widetilde{L}^T = U\Lambda U^T = [U_1 \; U_2] \, \Lambda \, [U_1 \; U_2]^T, \quad (10)$$

$$\frac{1}{N}\widetilde{X}\widetilde{X}^T = \widehat{U}\widehat{\Lambda}\widehat{U}^T = [\widehat{U}_1 \; \widehat{U}_2] \, \widehat{\Lambda} \, [\widehat{U}_1 \; \widehat{U}_2]^T. \quad (11)$$

The columns of U_1 are those eigenvectors associated with the d largest eigenvalues in Λ arranged in descending order, U_2 holds the remaining eigenvectors, and $\widehat{U}$ is similarly partitioned. The subspace we recover is spanned by the columns of $\widehat{U}_1$ and we wish to have this subspace as close as possible to the tangent space spanned by the columns of U_1. The orthogonal projectors onto the tangent and computed subspaces, P and $\widehat{P}$ respectively, are given by

$$P = U_1 U_1^T \quad \text{and} \quad \widehat{P} = \widehat{U}_1 \widehat{U}_1^T.$$

Define λ_d to be the dth largest eigenvalue of $\frac{1}{N}\widetilde{L}\widetilde{L}^T$, or the last entry on the diagonal of Λ_1. We are now in position to state the theorem.

Theorem 1. (Davis & Kahan [9], Stewart [10])
Let $\delta = \lambda_d - \left\|U_1^T \Delta U_1\right\|_F - \left\|U_2^T \Delta U_2\right\|_F$ *and consider*

- *(Condition 1)* $\quad \delta > 0$
- *(Condition 2)* $\quad \left\|U_1^T \Delta U_2\right\|_F \left\|U_2^T \Delta U_1\right\|_F < \frac{1}{4}\delta^2$.

Then, provided that conditions 1 and 2 hold,

$$\left\|P - \widehat{P}\right\|_F \leq 2\sqrt{2}\,\frac{\left\|U_1^T \Delta U_2\right\|_F}{\delta}. \quad (12)$$

The solution to (9) is impractical to compute. However, (12) is a tight bound, as will be demonstrated in Section 4. Thus, a solution may be approximated by minimizing the right-hand side of (12). To do so, and to give the conditions of the theorem a geometric interpretation, we must first understand the behavior of the perturbation Δ as a function of the scale parameter N.

3. ANALYSIS OF PERTURBATION TERMS

Theorem 1 requires an analysis of perturbation terms with the form $\|U_p^T \Delta U_q\|_F$ for $p, q = \{1, 2\}$. By (8) and the triangle inequality, each such norm is bounded by the sum of terms with the form

$$\left\|U_p^T \frac{1}{N}\widetilde{A}\widetilde{B} U_q\right\|_F, \quad (13)$$

where A and B represent the matrices L, C, and E. In [11] we present two approaches for bounding (13) with high probability. Here we summarize the approaches.

The key observation of the first approach is that $\frac{1}{N}\widetilde{A}\widetilde{B}^T$ is a sample mean of N outer products of vectors a and b, each sampled from a given distribution:

$$\frac{1}{N}\widetilde{A}\widetilde{B}^T = \widehat{\mathbb{E}}[(a - \widehat{\mathbb{E}}[a])(b - \widehat{\mathbb{E}}[b])^T], \tag{14}$$

where $\widehat{\mathbb{E}}[Y]$ is the sample mean of N realizations of random variable Y. We therefore expect that $\frac{1}{N}\widetilde{A}\widetilde{B}^T$ will converge toward the centered outer product of a and b. We use the following result to bound, with high probability, the norm of the difference between this sample mean and its expectation

$$\Big\| \mathbb{E}[U_p^T(a - \mathbb{E}[a])(b - \mathbb{E}[b])^T U_q] - \\ \widehat{\mathbb{E}}[U_p^T(a - \widehat{\mathbb{E}}[a])(b - \widehat{\mathbb{E}}[b])^T U_q] \Big\|_F$$

where $\mathbb{E}[Y]$ is the expectation of the random variable $Y \in \mathcal{Y}$.

Theorem 2. (Shawe-Taylor & Cristianini, [12]). *Given N samples of a random variable Y generated independently at random from $\mathcal{Y}$ according to the distribution P_Y, with probability at least $1 - e^{-\eta^2}$ over the choice of the samples, we have*

$$\Big\|\mathbb{E}[Y] - \widehat{\mathbb{E}}[Y]\Big\|_F \leq \frac{R}{\sqrt{N}}\left(2 + \eta\sqrt{2}\right) \tag{15}$$

where $R = \sup_{supp(P_Y)} \|Y\|_F$ and $supp(P_Y)$ is the support of distribution P_Y.

Theorem 2 may be used to bound (13) by letting a and b represent the vectors ℓ, c, and e, and the full calculations are detailed in [11]. Note that we must condition on e belonging to a set in which its supremum is bounded and that this set can be shown to have large measure.

This approach captures leading order behavior with high probability, but does so at the cost of attaching large constants to each term. Theorem 2 introduces constants based on suprema of functions of random variables taken over the support of their distributions. Accordingly, each perturbation term is shown to deviate from its expectation by factors larger than constant multiples of its variance. In [11] we demonstrate that tighter constants may be achieved and use the Central Limit Theorem (CLT) to show that the variance of the perturbation terms controls the deviation from their expectations. The CLT and Gaussian tail bound yield a confidence interval for each entry of the matrix $\frac{1}{N}AB^T$. Using a union bound to simultaneously control all of the entries of this matrix, this second approach yields an overall confidence interval for the Frobenius norm of the matrix. While such an approach yields a tighter result than that using Theorem 2, it holds with lower probability due to the use of many union bounds.

REMARK. While the CLT holds only as $N \to \infty$, this analysis may be rigorously applied to our finite-sample setting through use of Bernstein-type inequalities and concentration of measure, yielding only slightly larger constants. We appeal to the CLT for the purpose of demonstrating the tightest possible constants.

To leading order, both approaches yield the same results. Let r be the radius of the d-dimensional ball in the tangent plane from which we uniformly sample N points to populate the L matrix (noting that $N = \mathcal{O}(r^d)$). The curvature K of the local model is quantified by

$$K = \left(\sum_{i=d+1}^{D}\left(\sum_{n=1}^{d}\kappa_n^{(i)}\right)^2\right)^{\frac{1}{2}} \tag{16}$$

and is a natural result of our use of the Frobenius norm. Then each vector c has norm roughly of size Kr^2 and we expect the norm of the curvature matrix $\frac{1}{N}CC^T$ to grow as K^2r^4. Concentration of Gaussian measure indicates that the norm of the noise matrix $\frac{1}{N}EE^T$ will have size that depends on the variance σ and the square root of the projection dimension (as given by U_p). All other terms are zero in expectation and thus we expect $1/\sqrt{N}$ decay. The linear-curvature (LC^T), linear-noise (LE^T), and curvature-noise (CE^T) matrices should have norm $Kr^3/\sqrt{N}$, $\sigma r/\sqrt{N}$, and $K\sigma r^2/\sqrt{N}$, respectively. Finally, λ_d may be shown to have size r^2. The reader is referred to [11] for the full expression of each term and formal calculations.

4. OPTIMAL SCALE SELECTION

Our main result, a bound on the angle between the recovered and true tangent planes, is formulated by applying the triangle inequality to bound the norms appearing in Theorem 1 and then injecting the perturbation norm bounds described in the previous section. For ease of interpretation, we present the main result showing only leading order terms and neglecting probability-dependent constants (see [11] for the full result).

Theorem 3. (Main Result). *Let the following conditions hold:*

- *(Condition 1)* $\delta = \lambda_d - \left\|U_1^T\Delta U_1\right\|_F - \left\|U_2^T\Delta U_2\right\|_F > 0$,
- *(Condition 2)* $\|U_1^T\Delta U_2\|_F < \frac{1}{2}\delta$.

Then we have

$$\left\|P - \widehat{P}\right\|_F \leq \frac{2\sqrt{2}\,\frac{1}{\sqrt{N}}\left[\frac{K}{2}r^3 \; + \; \sigma^2\sqrt{d(D-d)}\right]}{\frac{r^2}{d+2} - \frac{K^2r^4(d+1)}{2(d+2)^2(d+4)} - \sigma^2\left(\sqrt{d} + \sqrt{D-d}\right)}. \tag{17}$$

Recalling that $N = \mathcal{O}(r^d)$, the optimal scale N^* may be selected as the N for which (17) is minimized. The behavior of the bound as a function of scale demonstrates the noise-curvature tradeoff. The linear and curvature contributions are small for small values of N. Thus for N small, the denominator δ of (17) is either negative or ill conditioned for most values of σ. This makes intuitive sense as we have not yet encountered much curvature but the linear structure has also not been explored. Therefore the noise dominates the early behavior of this bound and an approximating subspace may not be recovered from noise. As N increases, the conditioning of the denominator improves, and the bound is controlled by the $1/\sqrt{N}$ behavior of the numerator. This again corresponds with our intuition, as the addition of more points serves to overcome the effects of noise as the linear structure is explored. Thus, the bound becomes tighter. Eventually, N becomes large enough such that the curvature contribution approaches the size of the linear contribution, and δ^{-1} becomes large. The $1/\sqrt{N}$ term is overtaken by the ill conditioning of the denominator and the bound is forced to become large.

The reciprocal of the denominator, δ^{-1}, may therefore be interpreted as the condition number for the subspace recovery problem. When δ^{-1} is small we may recover a tight approximation to the true tangent space. The notion of an angle loses meaning as δ^{-1} tends to infinity, and we are unable to recover an approximating subspace. Condition 1 requires that we at least have $\delta > 0$ to approximate the appropriate linear subspace. This condition imposes that the spectrum corresponding to the linear subspace ($\lambda_d \sim \mathcal{O}(r^2)$) must be well separated from the spectra of the noise ($\|U_1\Delta U_1\|_F \sim \mathcal{O}(\sigma^2)$) and curvature ($\|U_2\Delta U_2\|_F \sim \mathcal{O}(K^2r^4)$) perturbations and the quality of approximation improves for larger δ. When the spectra are not well separated, the approximating subspace contains an eigenvector corresponding to a direction orthogonal to the true tangent plane. Condition 2 may then be interpreted as a control on sampling, as it may be satisfied by increasing the sampling density whenever δ^{-1} is well conditioned.

Imposing condition 1 yields a geometric uncertainty principle quantifying the limits of curvature and noise perturbation we may tolerate. To recover an approximating subspace, we must have that:

Geometric Uncertainty Principle.

$$K\sigma \;<\; \sqrt{\frac{(d+4)}{2(d+1)(\sqrt{d}+\sqrt{D-d})}}. \tag{18}$$

By preventing curvature and noise from simultaneously becoming large, this principle ensures that the geometry of the data is not destroyed by noise and expresses the intuitive requirement that the curvature of the manifold be less than the curvature of the perturbing noise-ball (see [11] for further analysis).

Figure 1 demonstrates that the bound in our main result accurately tracks the true subspace recovery error and may therefore be used to obtain the optimal scale for tangent plane recovery. We generate a data set sampled from a 3-dimensional manifold embedded in $\mathbb{R}^{20}$ according to the local model (2) and $N = 1.25 \times 10^6$ points are uniformly sampled from the tangent plane. The principal curvatures are set as follows: $\kappa_1^{(j)} = 3, \kappa_2^{(j)} = 1.5, \kappa_3^{(j)} = 1.5$ for $j = 4, 5, 6$; and $\kappa_1^{(j)} = 1.6351, \kappa_2^{(j)} = 0.1351, \kappa_3^{(j)} = 0.1351$ for $j = 7, \ldots, 20$. Gaussian noise ($\sigma = 0.01$) is added. The tangent plane at reference point x_0 is computed at each scale N via PCA of the N nearest neighbors of x_0. The true tangent plane recovery error $\|P - \widehat{P}\|_F$ is then computed at each scale. A "true bound" is computed by applying Theorem 1 after measuring each perturbation norm directly from the data. This true bound utilizes information that is not practically available, and therefore represents the best possible bound that we can hope to achieve. We compare the mean of the true error and mean of the true bound over 10 trials (with error bars indicating one standard deviation) to the following three curves: (1) Main Result using Theorem 2 holding with probability 0.5 (magenta); (2) Main Result using the CLT holding with probability 0.5 (black); and (3) Main Result using the CLT with all probability constants set to 1 (green). The third curve abandons any guarantee of providing an upper bound in favor of capturing the trend of the true error.

The bounds accurately track the behavior of the true error. Note that the true error is large at small scales due to noise and the bounds are accordingly ill conditioned. Eventually, a scale is reached at which there is too much curvature and the bounds blow up to infinity. This corresponds exactly to where the true error plateaus at its maximum value, representing the fact that the computed subspace is now orthogonal to the true tangent plane. In between, the bounds are well conditioned and track the true error. In fact, the curves are shown to be parallel on a logarithmic scale, indicating that they differ only by multiplicative constants. Note also that the true bound (red) tightly tracks the true error (blue), providing evidence that the triangle inequalities used in computing the bounds are reasonably tight. As no matrix decompositions are needed to compute our bounds, we have efficiently tracked the tangent plane recovery error. The black dots indicate the minimum of each curve and we see agreement of the location at which the minima occur, indicating the scale that will yield the optimal tangent plane approximation.

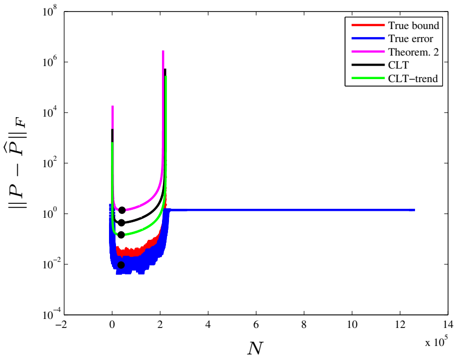

Fig. 1. Norm of the perturbation and bounds. Black dots indicate minima of the curves. Note the logarithmic scale on the Y-axis.

5. REFERENCES

[1] S.T. Roweis and L.K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323–2326, 2000.

[2] Z. Zhang and H. Zha, "Principal manifolds and nonlinear dimensionality reduction via tangent space alignment," *SIAM Journal on Scientific Computing*, vol. 26, pp. 313–338, 2004.

[3] M. Brand, "Charting a manifold," in *Advances in Neural Information Processing Systems 15*. 2003, pp. 961–968, MIT Press.

[4] B. Nadler, "Finite sample approximation results for principal component analysis: A matrix perturbation approach," *Annals of Statistics*, vol. 36, pp. 2792–2817, 2008.

[5] A. Singer and H.-T. Wu, "Vector diffusion maps and the connection Laplacian," *Communications on Pure and Applied Mathematics (to appear)*, 2012.

[6] T. Zhang, A. Szlam, Y. Wang, and G. Lerman, "Randomized hybrid linear modeling by local best-fit flats," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 1927–1934.

[7] G. Chen, A.V. Little, M. Maggioni, and L. Rosasco, "Some recent advances in multiscale geometric analysis of point clouds," in *Wavelets and Multiscale Analysis: Theory and Applications*, J. Cohen and A.I. Zayed, Eds., pp. 199–225. Springer, 2011.

[8] M. Giaquinta and G. Modica, *Mathematical Analysis: An Introduction to Functions of Several Variables*, Springer, 2009.

[9] C. Davis and W.M. Kahan, "The rotation of eigenvectors by a perturbation III," *SIAM Journal on Numerical Analysis*, vol. 7, pp. 1–46, 1970.

[10] G.W. Stewart and J. Sun, *Matrix Perturbation Theory*, Academic Press, 1990.

[11] D.N. Kaslovsky and F.G. Meyer, "Optimal tangent plane recovery from noisy manifold samples," *Submitted to Annals of Statistics*, pp. 1–57, 2011.

[12] J. Shawe-Taylor and N. Cristianini, "Estimating the moments of a random vector with applications," in *Proceedings of GRETSI 2003 Conference*, 2003, pp. 47–52.