Scale Selection for the Analysis of Point-Sampled Curves: Extended Report

Ranjith Unnikrishnan Jean-François Lalonde
Nicolas Vandapel Martial Hebert

CMU-RI-TR-06-25

June 2006

Robotics Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

Abstract

An important task in the analysis and reconstruction of curvilinear structures from unorganized 3-D point samples is the estimation of tangent information at each data point. Its main challenges are in (1) the selection of an appropriate scale of analysis to accommodate noise, density variation and sparsity in the data, and in (2) the formulation of a model and associated objective function that correctly expresses their effects. We pose this problem as one of estimating the neighborhood size for which the principal eigenvector of the data scatter matrix is best aligned with the true tangent of the curve, in a probabilistic sense. We analyze the perturbation on the direction of the eigenvector due to finite samples and noise using the expected statistics of the scatter matrix estimators, and employ a simple iterative procedure to choose the optimal neighborhood size. Experiments on synthetic and real data validate the behavior predicted by the model, and show competitive performance and improved stability over leading polynomial-fitting alternatives that require a preset scale.

Contents

1 Introduction

The inference of smooth geometric curves from a set of unorganized points is a challenging problem in several fields including computer vision, computational geometry and computer graphics. Its applications include feature extraction for indexing and geometric modeling, 3-D reconstruction of fine structures in medical imaging, skeletonization operations of handwritten character templates for recognition, and shape outline encoding for iso-contour predictive compression algorithms [5].

A popular first step in the analysis for many of these problems is the computation of first-order or tangent information at each point [3, 4, 6, 10]. One objective of this step is to reasonably capture the underlying curvilinear geometry of the point set, so as to be amenable to shape reconstruction and feature extraction. A closely related, but much more studied problem is that of computing surface normals from a set of unorganized points. There are several approaches in the literature, both non-parametric (tensor voting [6], radial basis functions, etc.) and parametric [5] (moving least-squares approximations [4], implicit parabolic fitting, B-splines, etc.).

Most practical algorithms achieve robustness by computing the relevant quantities in a local neighborhood [5, 6, 10]. Such algorithms crucially depend on knowledge of the radius of the neighborhood to be considered. The chosen radius at each point determines the *scale* of analysis – too small a radius would compromise the quality of the estimate due to the use of smaller number of noisy data points, while using too large a radius would permit a large number of potentially dissimilar points in the neighborhood to adversely influence the estimate. Hence the choice of a scale that reflects the underlying geometry is crucial to its analysis from finite samples.This requires the modeling of the observed data and the design of an objective function that faithfully matches the criteria of interest.

In this paper, we exploit the property of local linearity in the curve through principal component analysis (PCA) by choosing the tangent vector at a point as the principal eigenvector of a scatter matrix with local support [6, 3, 7]. We propose that for spatial curves, the neighborhood size should be chosen such that the principal eigenvalue of the scatter matrix is most closely aligned with the true tangent to the curve in an expected sense. We derive an upper bound on the angular error induced by finite sampling and sample noise as a function of neighborhood radius, and experimentally demonstrate that it matches the trend in variation of the optimal radius with the perturbing parameters.

1.1 Related work

There has been a fair amount of theoretical interest in curve reconstruction in the computational geometry community. As summarized in [2], the problem has largely been studied in the noise-free case of a single curve in the scene, and with the objective of inferring a geometric graph with an edge connecting points only if they are adjacent on the curve. The various approaches promise

differing extents of theoretical guarantees varying with assumptions on uniformity in sampling density, smoothness and presence of curve boundaries. However, their applicability in the presence of noise is largely unclear.

Most practical curve reconstruction algorithms are based on local polynomial fitting and its variants. Recent work by Lewiner et al. [5] computed the coefficients of an arc-length parameterized 3rd-order approximation to the curve by solving a weighted-least squares problem at each point using only the points in its local neighborhood. This procedure gave robust estimates of curvature and torsion, as well as the tangent as a by-product. They also reported favorable experimental comparison on a variety of synthetic curves with a large family of other fitting approaches based on Gaussian smoothing, Fourier transforms, circle and parabola fitting and others. The implicit parameter in the algorithm was the considered neighborhood radius, which was preset by fixing the number of neighbors considered at each point.

In the computer vision community, much work has been done on geometric reconstruction using non-parametric tensor voting [6, 10]. A key step in this framework is a voting procedure used to aggregate local information at each point or voxel of interest. The vote is in the form of a $d \times d$ tensor, where d is the data dimensionality, indicating preferred direction of normal/tangent, and the eigen decomposition of the aggregate tensor at a point gives the desired result. Again, a crucial parameter is the choice of the size of the support region for vote collection, usually chosen heuristically. Work in [10] proposed a fine-to-coarse approach in which points likely to form curves are linked together at fine scale to form fragments, and then linked together incrementally as the scale is increased using a heuristic inspired by perceptual grouping. Work in this paper focuses on sparser point sets than used in [10], necessitating a study of the small sample behavior of the tangent estimator. In the sections that follow, we show how the optimal neighborhood size may be estimated to minimize the expected angular error using basic matrix perturbation theory.

We also mention the related theoretical work by Mitra et al. [7] on optimal neighborhood size for normal estimation in surfaces using PCA. They obtained a closed form expression for a bound on the angular error between the estimated normal and true normal, and proposed the optimal radius as the value that minimized that bound. An iterative procedure was suggested that first estimated the local density and curvature, then computed the optimal radius for those values, and repeated the procedure until convergence. However, the closed form expression involved two parameters that relied on knowledge of the observed data distribution and had to be fixed a priori.

1.2 Approach and Paper Outline

In our proposed method, we exploit the property of local linearity in the curve through local principal component analysis using an adaptive neighborhood size. Our estimate of the tangent at a point is the principal eigenvector of the scatter matrix computed in its local neighborhood [3, 6, 10]. We propose that, for spatial curves, the neighborhood size should be chosen such that the

principal eigenvalue of the scatter matrix is most closely aligned with the true tangent to the curve. To make this choice, we derive an upper bound on the expected angular error induced by finite sampling and sample noise as a function of neighborhood radius. The optimal radius is then chosen as the value that minimizes this upper bound on angular error.

The main steps in our derivation are as follows:

1. We first model and state our assumptions of the underlying geometry (Section 2) and adopt a convenient reference frame.

2. We then compute the expected statistics of the estimators (Section 3) under the chosen geometry model. The scatter matrix (which we will also refer to as the covariance matrix) computed by the choice of estimators is then trivially expressed as the sum of the expected value in the limiting case and a random zero-mean perturbation matrix.

3. Using bounds on the Frobenius norm of the random matrix, we can then compute an upper bound on the angular deviation the eigenvector from the perturbed matrix as a function of sample noise, sample size, radius as well as the curvature and torsion of the curve. (Section 4)

4. We then analyze the cases of 2-D and 3-D curves (Section 5), make observations from the analytic behavior of their bounds, and derive the conditions under which the dominant eigenvector of their corresponding expected matrices align with the tangent to the true curve in the canonical frame.

5. The proposed scale-adaptive PCA algorithm then proceeds iteratively through fixed-point descent. First we estimate the curvature ($\kappa^{(t)}$) and torsion ($\tau^{(t)}$) for a starting neighborhood size $r^{(t)}$ and use a sensor model to obtain the value of sample noise. Then we perform line-estimation on r to obtain a value $r^{(t+1)}$ that minimizes the above bounds. We then re-estimate $\kappa^{(t+1)}$ and $\tau^{(t+1)}$ using the new neighborhood radius $r^{(t+1)}$ and iterate till convergence.

   Section 6 compares the stability and accuracy of the adaptive PCA algorithm with the local polynomial fitting algorithm of [5] through experiments on synthetic 2-D and 3-D curves.

2 Curve model

We assume the existence of a smooth curve (for both 2-D and 3-D cases) parameterized by distance s. Without loss of generality, we assume a Frenet reference frame (Figure 1) with origin located at the point of interest such that the tangent to the curve is aligned with the x-axis, the curvature vector in the plane of the osculating circle containing the point of interest is aligned with the y-axis and the normal to the osculating plane is aligned with the z-axis.

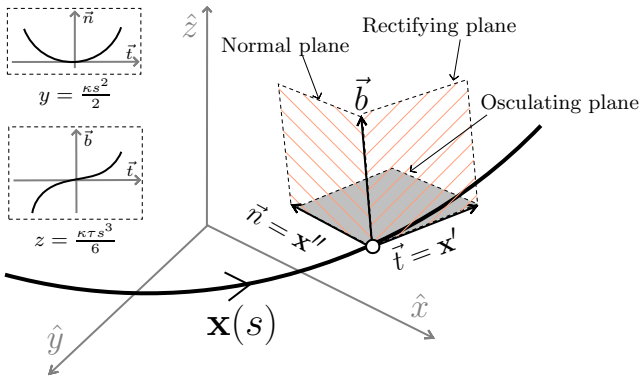


Figure 1: Model of local curve geometry

Our available data is a set of n unordered points $\{x_i, y_i, z_i\}$. Each such point may be thought of as a noisy observation of a true point lying on a spatial curve, parameterized by arc-length, at the (unknown) location s_i along the curve. The points are assumed to lie within a candidate distance r from a point of interest. We may then adopt the generative model of a set of n samples from $S \sim \text{Uniform}(-r, r)$ with additive Gaussian noise $\eta \sim \mathcal{N}(0, \sigma_0^2)$ as:

$$\begin{aligned} x_i &= s_i + \eta_{x,i} \\ y_i &= \frac{\kappa}{2} s_i^2 + \eta_{y,i} \\ z_i &= \frac{\kappa\tau}{6} s_i^3 + \eta_{z,i} \end{aligned} \tag{1}$$

which is valid for moderate slowly changing values of curvature (κ) and torsion (τ). We assume iid sensor noise that is zero-mean normally distributed with variance σ_0^2 affecting all three coordinates.

In summary, our assumptions are that :

1. In a local neighborhood around the point of interest, curvature κ and torsion τ are bounded and near constant, i.e. $\kappa'(s), \tau'(s) \approx 0$

2. The quantities x, y, z are observed with iid zero-mean Gaussian noise of standard deviation σ_0. The noise is assumed independent of the position on the curve. In practice, we allow the value of σ_0 to differ across the scene to account for variation in noise level with distance from the laser sensor.

3. There is a minimum density of points in the scene. In the neighborhood of radius r around the point of interest, there is a minimum point density ρ_0, so that $n \geq 2\rho_0 r$

3 The covariance matrix for curves

One technique to estimate the direction of the local tangent at a given sample point on a curve is to look at the shape of a scatter matrix computed using points in its neighborhood [3, 6, 10]. If the curve is smooth, it is reasonable to expect that the scatter matrix will be elongated and that its major axis, or principal eigenvector, will approximate the direction of the local tangent for some appropriate (and unknown) range of neighborhood sizes. In this and the following subsection, we will derive and analyze the conditions under which this assumption will hold for both 2-D and 3-D curves.

The random variables X, Y and Z (denoted in capitals to distinguish them from the data) are noisy functions of the random variable S whose distribution is assumed to be locally uniform. Hence the distribution of X, Y and Z, as well as estimators of their 1st and 2nd order statistics will depend on the coefficients (κ, τ) and order of the functions (given in (1)) as well as properties of the uniform (for S) and Gaussian (for η) distributions.

We start by computing the mean and variance of the estimators used to construct the sample covariance matrix $\hat{M}_n$. We will denote the true means of random variables by μ (e.g. μ_X for X) and standard deviation by σ (e.g. σ_X^2 for variance of X). Then:

$$\hat{M}_n = \begin{bmatrix} M_{11} & M_{12} & M_{13} \\ M_{12} & M_{22} & M_{23} \\ M_{13} & M_{23} & M_{33} \end{bmatrix} \tag{2}$$

where

$$M_{11} = \frac{\sum_i (x_i - \bar{X}_n)^2}{n-1} \qquad M_{12} = \frac{\sum_i (x_i - \bar{X}_n)(y_i - \bar{Y}_n)}{n-1} \tag{3}$$

$$M_{22} = \frac{\sum_i (y_i - \bar{Y}_n)^2}{n-1} \qquad M_{13} = \frac{\sum_i (x_i - \bar{X}_n)(z_i - \bar{Z}_n)}{n-1} \tag{4}$$

$$M_{33} = \frac{\sum_i (z_i - \bar{Z}_n)^2}{n-1} \qquad M_{23} = \frac{\sum_i (y_i - \bar{Y}_n)(z_i - \bar{Z}_n)}{n-1} \tag{5}$$

and $\bar{X}_n = \frac{1}{n}\sum_i x_i$ is the sample mean estimator for X, and similarly for $\bar{Y}_n$ and $\bar{Z}_n$.

Note that, as is well known from introductory statistics [11], the diagonal elements in M are unbiased estimators for variance (e.g. M_{11} is the estimator for variance σ_X^2 of X) and the off-diagonal elements are unbiased estimators of covariance (e.g. M_{13} is the estimator for covariance $\mathrm{cov}(X, Z)$ of X and Z).

In the following sections, we will compute the expected values and variances for each element in the covariance matrix under the distribution of $S \sim \mathrm{Uniform}(-r, r)$. First we will derive some useful identities of the random variables X, Y and Z as defined by our geometry model in (1).

3.1 Moments and Covariances

In this section we compute some useful moment and covariance identities relating the random variables X, Y and Z, which represent the observed point samples as a noisy function of the position S on the curve. We assume $S \sim$ Uniform$(-r, r)$ implying the distribution $f_S(s) = \frac{1}{2r}$ in $[-r, r]$ along the curve, and zero elsewhere. As stated earlier, we consider the observed quantities to be corrupted by iid noise terms η_X, η_Y and η_Z normally distributed with variance σ_0^2. In what follows, we make use of some statistics of the uniform distribution derived in Appendix A.
First we consider $X = S + \eta_X$

$$\mu_X = \mathbb{E}(S + \eta_X) = \int_{-r}^{r} s \frac{1}{2r} ds = 0 \tag{6}$$

$$\boxed{d_2(X)} = \sigma_X^2 = \int_{-r}^{r} s^2 \frac{1}{2r} ds + \sigma_0^2 = \frac{r^2}{3} + \sigma_0^2 \tag{7}$$

$$\tag{8}$$

$$\begin{aligned} \boxed{d_4(X)} &= \mathbb{E}(X + \eta_X - \mu_X)^4 = \mathbb{E}(S^4) + 6\mathbb{E}(S^2)\mathbb{E}(\eta_X^2) + \mathbb{E}(\eta_X^4) \\ &= \frac{r^4}{5} + 2\sigma_o^2 r^2 + 3\sigma_0^4 \end{aligned} \tag{9}$$

Next we consider $Y = \frac{\kappa}{2} S^2 + \eta_Y$

$$\mu_Y = \frac{\kappa}{2} \mathbb{E}(S^2) = \frac{\kappa r^2}{6} \tag{10}$$

$$\begin{aligned} \boxed{d_2(Y)} = \sigma_Y^2 &= \left(\frac{\kappa}{2}\right)^2 d_2(S^2) + \sigma_0^2 \\ &= \frac{\kappa^2 r^4}{45} + \sigma_0^2 \end{aligned} \tag{11}$$

$$\begin{aligned} \boxed{d_4(Y)} &= \left(\frac{\kappa}{2}\right)^4 d_4(S^2) + 6\sigma_0^2 \left(\frac{\kappa}{2}\right)^2 d_2(S^2) + \mathbb{E}(\eta_Y^4) \\ &= \frac{\kappa^4 r^8}{945} + \frac{2}{15}\sigma_0^2 \kappa^2 r^4 + 3\sigma_0^4 \end{aligned} \tag{12}$$

We can also compute the first and second order covariance between X and Y:

$$\boxed{c_1(X, Y)} = \text{cov}(X, Y) = \frac{\kappa}{2} \text{cov}(S, S^2) = 0 \tag{13}$$

$$\begin{aligned}\boxed{c_2(X,Y)} &= c_2(S,\frac{\kappa}{2}S^2) + \sigma_0^2\left(d_2(S) + \left(\frac{\kappa}{2}\right)^2 d_2(S^2)\right) + \sigma_0^4 \\ &= \frac{11}{945}\kappa^2 r^6 + \sigma_0^2\left(\frac{r^2}{3} + \left(\frac{\kappa}{2}\right)^2 \frac{4}{45}r^4\right) + \sigma_0^4 \\ &= \frac{11}{945}\kappa^2 r^6 + \sigma_0^2\left(\frac{r^2}{3} + \frac{\kappa^2}{45}r^4\right) + \sigma_0^4\end{aligned} \tag{14}$$

Next we consider $Z = \frac{\kappa\tau}{6}S^3 + \eta_Z$

$$\mu_Z = \mathbb{E}(\frac{\kappa\tau}{6}S^3 + \eta_Z) = 0 \tag{15}$$

$$\boxed{d_2(Z)} = \sigma_Z^2 = \left(\frac{\kappa\tau}{6}\right)^2 \int_{-r}^{r} s^6 \frac{1}{2r} ds + \sigma_0^2 = \left(\frac{\kappa\tau}{6}\right)^2 \frac{r^6}{7} + \sigma_0^2 \tag{16}$$

$$\begin{aligned}\boxed{d_4(Z)} &= \left(\frac{\kappa\tau}{6}\right)^4 d_4(S^3) + 6\left(\frac{\kappa\tau}{6}\right)^2 d_2(S^3) + \mathbb{E}(\eta_Y^4) \\ &= \left(\frac{\kappa\tau}{6}\right)^4 \frac{r^{12}}{13} + 6\left(\frac{\kappa\tau}{6}\right)^2 \frac{r^6}{7} + 3\sigma_0^4 \\ &= \left(\frac{\kappa\tau}{6}\right)^4 \frac{r^{12}}{13} + \frac{(\kappa\tau)^2}{42}r^6 + 3\sigma_0^4\end{aligned} \tag{17}$$

We can also compute the first and second order covariance between Z and X and Y.

$$\boxed{c_1(X,Z)} = cov(X,Z) = \frac{\kappa\tau}{6}cov(S,S^3) = \frac{\kappa\tau}{30}r^4 \tag{18}$$

$$\boxed{c_1(Y,Z)} = cov(Y,Z) = \frac{\kappa^2\tau}{12}cov(S^2,S^3) = 0 \tag{19}$$

$$\begin{aligned}\boxed{c_2(X,Z)} &= c_2(S,\frac{\kappa\tau}{6}S^3) + \sigma_0^2\left(d_2(S) + d_2(\frac{\kappa\tau}{6}S^3)\right) + \sigma_0^4 \\ &= \left(\frac{\kappa\tau}{6}\right)^2 \mathbb{E}(S^8) + \sigma_0^2\left(\frac{r^2}{3} + \left(\frac{\kappa\tau}{6}\right)^2 \frac{r^6}{7}\right) + \sigma_0^4 \\ &= \left(\frac{\kappa\tau}{6}\right)^2 \frac{r^8}{9} + \sigma_0^2\left(\frac{r^2}{3} + \left(\frac{\kappa\tau}{6}\right)^2 \frac{r^6}{7}\right) + \sigma_0^4\end{aligned} \tag{20}$$

$$\begin{aligned}\boxed{c_2(Y,Z)} &= c_2(\frac{\kappa}{2}S^2,\frac{\kappa\tau}{6}S^3) + \sigma_0^2\left(d_2(\frac{\kappa}{2}S^2) + d_2(\frac{\kappa\tau}{6}S^3)\right) + \sigma_0^4 \\ &= \left(\frac{\kappa^2\tau}{12}\right)^2 c_2(S^2,S^3) + \sigma_0^2\left(\frac{\kappa^2}{4}d_2(S^2) + \left(\frac{\kappa\tau}{6}\right)^2 d_2(S^3)\right) + \sigma_0^4 \\ &= \left(\frac{\kappa^2\tau}{12}\right)^2 \frac{68}{2079}r^{10} + \sigma_0^2\left(\frac{\kappa^2 r^4}{45} + \left(\frac{\kappa\tau}{6}\right)^2 \frac{r^6}{7}\right) + \sigma_0^4\end{aligned} \tag{21}$$

3.2 Estimators of the covariance matrix

In this section, we compute the expected values for each entry in the covariance matrix using the results from Section 3.1 as well as the identities for the variances of sample variance and sample covariance estimators derived in Appendix B.

The estimate of the covariance matrix is then given by:

$$\begin{aligned}\hat{M}_n &= \begin{bmatrix} \sum_i (x_i - \bar{X}_n)^2 & \sum_i (x_i - \bar{X}_n)(y_i - \bar{Y}_n) & \sum_i (x_i - \bar{X}_n)(z_i - \bar{Z}_n) \\ \sum_i (x_i - \bar{X}_n)(y_i - \bar{Y}_n) & \sum_i (y_i - \bar{Y}_n)^2 & \sum_i (y_i - \bar{Y}_n)(z_i - \bar{Z}_n) \\ \sum_i (x_i - \bar{X}_n)(z_i - \bar{Z}_n) & \sum_i (y_i - \bar{Y}_n)(z_i - \bar{Z}_n) & \sum_i (z_i - \bar{Z}_n)^2 \end{bmatrix} \\ &= \begin{bmatrix} M_{11} & M_{12} & M_{13} \\ M_{12} & M_{22} & M_{23} \\ M_{13} & M_{23} & M_{33} \end{bmatrix}\end{aligned} \tag{22}$$

The first and second moments of these quantities follow as:

1. M_{11} :

$$\mathbb{E}(M_{11}) = \mathbb{V}(X_i) = \mathbb{V}(S_i) + \sigma_0^2 = \frac{r^2}{3} + \sigma_0^2 \tag{23}$$

$$\begin{aligned}\mathbb{V}(M_{11}) &= \frac{d_4(X_i)}{n} - \frac{(n-3)}{n(n-1)}\sigma_X^4 \\ &= \frac{1}{5n}\left(r^4 + 10\sigma_0^2 r^2 + 15\sigma_0^4\right) + \frac{(n-3)}{9n(n-1)}\left(r^2 + 3\sigma_0^2\right)^2\end{aligned} \tag{24}$$

2. M_{12} :

$$\mathbb{E}(M_{12}) = \mathrm{cov}(X, Y) = \frac{\kappa}{2}\mathbb{E}(S^3) = 0 \tag{25}$$

and

$$\mathbb{V}(M_{12}) = \frac{1}{n}\left(\frac{11}{945}\kappa^2 r^6 + \sigma_0^4 + \sigma_0^2(\frac{r^2}{3} + \frac{\kappa^2 r^4}{45})\right) + \frac{\left(\frac{r^2}{3} + \sigma_0^2\right)\left(\frac{\kappa^2 r^4}{45} + \sigma_0^2\right)}{n(n-1)} \tag{26}$$

3. M_{13} :

$$\mathbb{E}(M_{13}) = \mathrm{cov}(X, Z) = \frac{\kappa\tau}{6}\mathbb{E}(S^4) = \frac{\kappa\tau}{30}r^4 \tag{27}$$

and

$$\begin{aligned}\mathbb{V}(M_{13}) =& \frac{1}{n}\left(\left(\frac{\kappa\tau}{6}\right)^2 \frac{r^8}{9} + \sigma_0^2\left(\frac{r^2}{3} + \left(\frac{\kappa\tau}{6}\right)^2 \frac{r^6}{7}\right) + \sigma_0^4\right) \\ &+ \frac{1}{n(n-1)}\left(\frac{r^2}{3} + \sigma_0^2\right)\left(\left(\frac{\kappa\tau}{6}\right)^2 \frac{r^6}{7} + \sigma_0^2\right) \\ &+ \frac{(n-2)}{n(n-1)}\left(\frac{\kappa\tau}{30}\right)^2 r^8\end{aligned} \tag{28}$$

4. M_{22} :

$$\mathbb{E}(M_{22}) = \mathbb{V}(Y_i) = \frac{\kappa^2}{4}\mathbb{V}(S^2) + \sigma_0^2 = \frac{\kappa^2}{45}r^4 + \sigma_0^2 \tag{29}$$

and

$$\begin{aligned}\mathbb{V}(M_{22}) &= \frac{d_4(Y)}{n} - \frac{(n-3)}{n(n-1)}\sigma_Y^4 \\ &= \frac{1}{n}\left(\frac{\kappa^4 r^8}{945} + \frac{2}{15}\sigma_0^2\kappa^2 r^4 + 3\sigma_0^4\right) + \frac{(n-3)}{n(n-1)}\left(\frac{\kappa^2 r^4}{45} + \sigma_0^2\right)^2\end{aligned} \tag{30}$$

5. M_{23} :

$$\mathbb{E}(M_{23}) = \text{cov}(Y, Z) = \frac{\kappa^2\tau}{18}(\mathbb{E}(S^5) - \mathbb{E}(S^2)\mathbb{E}(S^3)) = 0 \tag{31}$$

and

$$\begin{aligned}\mathbb{V}(M_{23}) =& \frac{1}{n}\left(\left(\frac{\kappa^2\tau}{12}\right)^2 \frac{68}{2079}r^{10} + \sigma_0^2\left(\frac{\kappa^2 r^4}{45} + \left(\frac{\kappa\tau}{6}\right)^2\frac{r^6}{7}\right) + \sigma_0^4\right) \\ &+ \frac{1}{n(n-1)}\left(\frac{\kappa^2 r^4}{45} + \sigma_0^2\right)\left(\left(\frac{\kappa\tau}{6}\right)^2\frac{r^6}{7} + \sigma_0^2\right)\end{aligned} \tag{32}$$

6. M_{33} :

$$\mathbb{E}(M_{33}) = \mathbb{V}(Z_i) = \left(\frac{\kappa\tau}{6}\right)^2\frac{r^6}{7} + \sigma_0^2 \tag{33}$$

and

$$\mathbb{V}(M_{33}) = \frac{1}{n}\left(\left(\frac{\kappa\tau}{6}\right)^4\frac{r^{12}}{13} + \frac{(\kappa\tau)^2}{42}r^6 + 3\sigma_0^4\right) + \frac{(n-3)}{n(n-1)}\left(\left(\frac{\kappa\tau}{6}\right)^2\frac{r^6}{7} + \sigma_0^2\right)^2 \tag{34}$$

Observe that the estimator for sample covariance matrix may be expressed as the sum of the matrix of its expected value and a matrix of random variables as:

$$\hat{M} = \tilde{M} + Q \tag{35}$$

where $\tilde{M} = \mathbb{E}(M)$ is a symmetric matrix with elements given by the equations (23), (25), (27), (29), (31) and (33), and Q is a symmetric *perturbation matrix* of random variables each with mean 0 and variance given by the equations (24) (26), (28), (30), (32), and (34).

4 Perturbation model

In the previous section, we were able to express the scatter matrix ($\hat{M}$) computed in a local neighborhood as a sum of an uncorrupted intrinsic quantity ($\tilde{M}$) and a random matrix (Q) existing due to finite sampling and noise. In this

section we compute the effect of the perturbation Q on the principal eigenvector of $\hat{M}$.

We denote the eigenvalues of $\tilde{M} = \mathbb{E}(M)$ by $\lambda_1 \geq \lambda_2 \geq \lambda_3$. Let the eigenvector corresponding to λ_1 be e_1. Let $\hat{e}_1$ be the eigenvector corresponding to the largest eigenvalue of the estimated $\hat{M}$. If Q is the symmetric perturbation to the positive semidefinite matrix $\tilde{M}$, then the application of the matrix perturbation theorem V.2.8 from [9] yields [8]:

$$||\hat{e}_1 - e_1|| \leq \frac{4||Q||_F}{\delta - \sqrt{2}||Q||_F} \tag{36}$$

where $\delta = \lambda_1 - \lambda_2$ is the spectral gap of the matrix $\mathbb{E}(M)$, and $||Q||_F$ represents Frobenius norm[1].

Since the matrix Q consists of random variables, we are confined to making probabilistic statements about $||Q||_F$. Using Chebyshev's inequality, the square of the value attained by each element Q_{ij} can be upper bounded by:

$$Q_{ij}^2 \leq \frac{\mathbb{V}(M_{ij})}{n\epsilon}$$

with probability $1 - \epsilon$, where $\mathbb{V}(M_{ij})$ is the variance of corresponding finite sample estimator of covariance (or variance if $i = j$). Note that minimizing the RHS of (36) is equivalent to minimizing the ratio:

$$B \triangleq ||Q||_F/\delta \tag{37}$$

5 Angular bounds and their behavior

We first analyze the behavior of the perturbation bound to variation in sampling density, noise and curvature by looking at the slightly simpler case of 2-D curves.

5.1 2-D curves

We analyze the 2-D case by working with the same assumptions as stated earlier except that we discard the z coordinate (or equivalently nullify torsion). The scatter matrix in this case is obtained as the top left 2×2 sub-matrix of Q, which we will refer to as Q_2. From our perturbation model in Section 4, we know that Frobenius norm of Q_2 is upper bounded with probability $1 - \epsilon$ by:

$$\begin{aligned} ||Q_2||_F^2 &\leq \frac{1}{n\epsilon} \sum_{i=1}^{2} \sum_{j=1}^{2} \mathbb{V}(M_{ij}) \\ &= \frac{1}{n\epsilon} \left[\mathbb{V}(M_{11}) + \mathbb{V}(M_{22}) + 2\mathbb{V}(M_{12})\right] \end{aligned} \tag{38}$$

[1]Forbenius norm is defined as $||A||_F^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2$ for a $m \times n$ matrix A with elements a_{ij}

Substituting the expressions from Section 3.2 gives:

$$\begin{aligned}||Q_2||_F^2 \leq & \frac{1}{n\epsilon}\Bigg[\frac{1}{5n}\left(r^4+10\sigma_0^2r^2+15\sigma_0^4\right)+\frac{(n-3)}{9n(n-1)}\left(r^2+3\sigma_0^2\right)^2 \\ & +\frac{1}{n}\left(\frac{\kappa^4r^8}{945}+\frac{2}{15}\sigma_0^2\kappa^2r^4+3\sigma_0^4\right)+\frac{(n-3)}{n(n-1)}\left(\frac{\kappa^2r^4}{45}+\sigma_0^2\right)^2 \\ & +\frac{2}{n}\left(\frac{11}{945}\kappa^2r^6+\sigma_0^4+\sigma_0^2(\frac{r^2}{3}+\frac{\kappa^2r^4}{45})\right)+2\frac{\left(\frac{r^2}{3}+\sigma_0^2\right)\left(\frac{\kappa^2r^4}{45}+\sigma_0^2\right)}{n(n-1)}\Bigg]\end{aligned} \tag{39}$$

The spectral gap of the corresponding top-left 2×2 sub-matrix $\tilde{M}_2$ of $\mathbb{E}(M)$ given by:

$$\tilde{M}_2 = \begin{bmatrix} \frac{r^2}{3}+\sigma_0^2 & 0 \\ 0 & \frac{\kappa^2}{45}r^4+\sigma_0^2 \end{bmatrix} \tag{40}$$

is obtained easily by inspection as

$$\delta_2 = \frac{r^2}{3}-\frac{\kappa^2r^4}{45} \tag{41}$$

This implies that for the dominant eigenvector of $\tilde{M}$ to be $\begin{bmatrix}1 & 0\end{bmatrix}$, the value of radius r must satisfy

$$0 < r < \sqrt{15}/\kappa \tag{42}$$

The bound to be minimized is then

$$B_2(r) \triangleq \frac{||Q_2||_F}{\delta_2} \tag{43}$$

To study the analytical behavior of this bound, we need to replace the discrete parameter n by a continuous function of radius r, and explicitly express their dependency. To do this, we use the assumption of minimum local point density ρ and substitute $n = 2\rho r$ to form the analytical plots that follow.

Note, however, that in the implementation of the proposed algorithm we directly set n in (43) to equal the number of points observed in the neighborhood of candidate radius r and do not ever need to estimate ρ. The assumption of an underlying ρ is used *only* for studying the expected behavior of the analytic bound in synthetic data and is *not* used at runtime.

Before proceeding, we point out that there are two expected limitations in the functional analysis of the derived expressions that will be relevant in their experimental validation.

1. The bounds are high order polynomial expressions having multiple discontinuities over r. However, we restrict our analysis to the regime of r where the constraints for dominant eigenvectors hold. In this regime, the function is smooth and has only one minima.

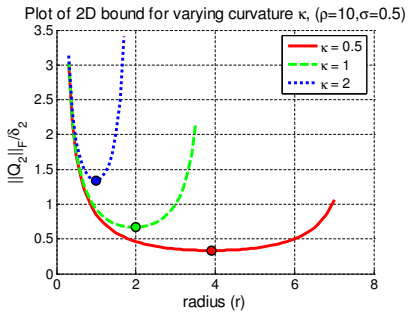


Figure 2: Plot of analytic 2D bound for varying curvature

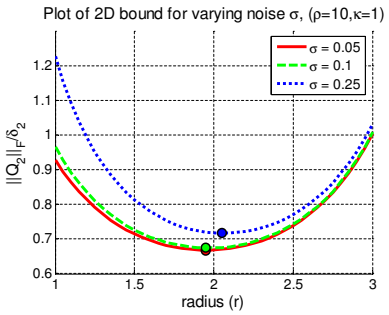


Figure 3: Plot of analytic 2D bound for varying sampling noise

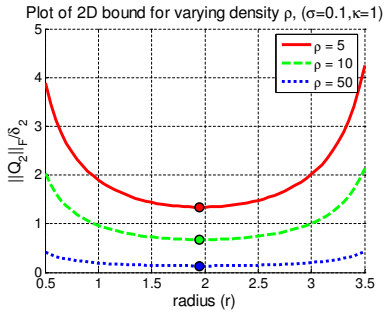


Figure 4: Plot of analytic 2D bound for varying sample density

2. As also observed experimentally in [7, 5], the predicted error tends to 0 as $r \to 0$ for noise-free data. But for $\sigma_0 > 0$, the error tends to sharply increase for the same condition. This behavior is not reflected by our current model for two related reasons. First, the continuous relaxation of n in terms of ρ is invalid for small number of points. Secondly, the effective n for small r becomes a non-sensical fractional number. So while the trend in shift of minima is correct, the analytical plot cannot be expected to fully match the experimental plot.

In short, we advocate the interpretation of the function in terms of the behavior of its minima in the meaningful regimes of interest, rather than throughout the domain of the function.

Based on the analytical plots of $B_2(r)$ in Figures 3–4, we make the following qualitative observations:

1. Complexity: Like most closed form expressions encountered in real-world problems, the formula in (43) is not short and has many terms. However, it can be easily shown that the terms with coefficients $(n(n-1))^{-1}$ in the numerator of $B_2(r)$ are dominated by the others for integer values

of $n \geq 2$, reducing the expression to the ratio of the root of a 6^{th} degree polynomial and a 4^{th} degree polynomial of r, both only containing even powers of r.

2. Variation with curvature κ: Figure 2 plots the function B_2 for multiple values of κ and fixed values of noise and sampling density. As one would expect, the optimal radius r tends to increase with decreasing curvature in order to compensate for noise and sparsity, without exceeding the bounds in (42) when the eigenvector more closely aligned to the x-axis is no longer dominant.

3. Variation with sampling noise σ_0 : Figure 3 plots the function B_2 for multiple values of κ and fixed values of noise and sampling density. It can be seen that as the noise increases, the point of minima of B_2 increases but only approaching the required bounds for eigenvector dominance in (42).

4. Variation with sampling density ρ : Figure 4 plots the function B_2 for multiple values of sampling density and fixed values of noise and curvature. It is interesting to note that although the value of the bound decreases as expected with increased number of points, the location of the extremum hardly changes. This is in contrast with the observations in [7] for surfaces which varies r with $\rho^{-0.5}$. We validate this later in Section 6.2.

5.2 3-D curves

The derivation and behavior of the angular bound for 3-D curves is fairly similar to the 2-D case. From Section 4, the $||Q||_F$ is upper bounded with probability $1-\epsilon$ by:

$$\begin{aligned} ||Q||_F^2 &\leq \frac{1}{n\epsilon}\sum_{i=1}^{3}\sum_{j=1}^{3}\mathbb{V}(M_{ij}) \\ &= \frac{1}{n\epsilon}\Big[\mathbb{V}(M_{11})+\mathbb{V}(M_{22})+\mathbb{V}(M_{33})+2\left(\mathbb{V}(M_{12})+\mathbb{V}(M_{13})+\mathbb{V}(M_{23})\right)\Big] \end{aligned} \tag{44}$$

Substituting the results from Section 3 gives the required final expression.

The matrix of expected values can be written as:

$$\tilde{M} = \mathbb{E}(M) = \begin{bmatrix} \frac{r^2}{3}+\sigma_0^2 & 0 & \frac{\kappa\tau}{30}r^4 \\ 0 & \frac{\kappa^2}{45}r^4+\sigma_0^2 & 0 \\ \frac{\kappa\tau}{30}r^4 & 0 & \left(\frac{\kappa\tau}{6}\right)^2\frac{r^6}{7}+\sigma_0^2 \end{bmatrix} \tag{45}$$

We denote the eigenvalues of $\tilde{M}$ as $\lambda_1 \geq \lambda_2 \geq \lambda_3$. The spectral gap of $\tilde{M}$ is not as straightforward due to its off-diagonal terms. However, we can lower

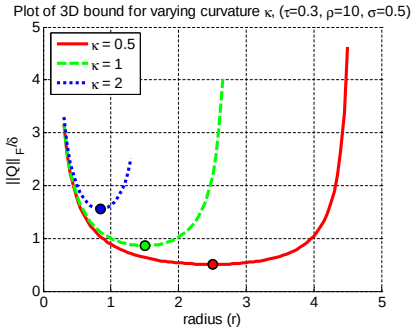


Figure 5: Plot of analytic 3-D bound for varying curvature

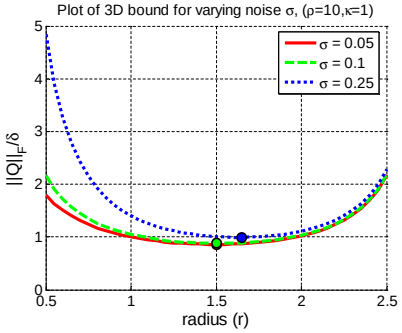


Figure 6: Plot of analytic 3-D bound for varying sampling noise

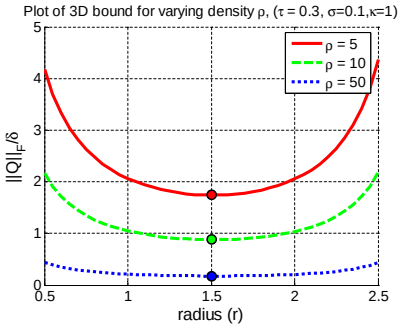


Figure 7: Plot of analytic 3-D bound for varying density

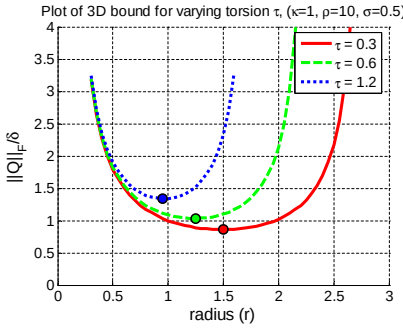


Figure 8: Plot of analytic 3-D bound for varying torsion

bound the spectral gap using the Gershgorin circle theorem (GCT). This gives the system of inequalities:

$$|\lambda_1 - \frac{r^2}{3} + \sigma_0^2| \leq \frac{\kappa\tau}{30} r^4 \tag{46}$$

$$\lambda_2 = \frac{\kappa^2}{45} r^4 + \sigma_0^2 \tag{47}$$

$$|\lambda_3 - \left(\frac{\kappa\tau}{6}\right)^2 \frac{r^6}{7} + \sigma_0^2| \leq \frac{\kappa\tau}{30} r^4 \tag{48}$$

This gives the additional constraint on r for principal eigenvector dominance as:

$$r \leq \sqrt{28/5\tau} \tag{49}$$

together with a bound on spectral gap as:

$$\delta_3 \geq \frac{r^2}{3} - \frac{\kappa^2 r^4}{45} - \kappa\tau\frac{r^4}{15} \tag{50}$$

Combining (44) and (50) with the continuous relaxation $n = 2\rho r$ in (37) gives the desired analytical result.

The observations we make on the analytic behavior of $B(r)$ are analogous to those in the 2-D case. The main effect of torsion is that due to its presence as an off-diagonal term in $\mathbb{E}(M)$, it always induces a finite angular offset of the dominant eigenvector in the rectifying plane (see Figure 1).

However as the radius is decreased, the off-diagonal term tends to 0 with r^4 while the leading eigen-vector decays with r^2. Thus in moving from the 2-D to 3-D analysis, the overall effect of torsion is to decrease the optimal scale of analysis with increasing τ. This shift can be verified in Figure 5 which has the same parameters as the 2-D curve of Figure 2 but with a non-zero torsion $\tau = 0.3$. Figures 6 and 7 show the shift in error bound with change in noise and density, and the plots are similar to the 2-D case. Figure 8 plots the effect of variation in torsion on the error bound, and it can be seen to be similar but less pronounced than the effect of change in curvature.

6 Experimental Results

In this section, we first outline the proposed algorithm used in later experimental results. We experimentally study the behavior of the choice of neighborhood radius on the error in tangent estimation. We then establish the utility of the proposed method for suggesting a radius that minimizes this error. We also study the effect of different starting points of radius for the proposed algorithm and different choices of the radius parameter on polynomial fitting based methods.

6.1 Algorithm and implementation

The analytical behavior of the bound derived in Section 5 suggests the following iterative algorithm:

At $t = 0$, for a starting neighborhood size $r^{(t)}$, we estimate the curvature ($\kappa^{(t)}$) and torsion ($\tau^{(t)}$) using [5] and use a sensor model to obtain the value of sample noise. Then we perform line-estimation on r to obtain the $r^{(t+1)}$ minimizing (43), subject to (42) using values at time t. We then re-estimate $\kappa^{(t+1)}$ and $\tau^{(t+1)}$ corresponding to the new value of radius r and iterate till convergence. To prevent large changes in estimates of r between iterations, we use a damping factor $\alpha = 0.5$, although no significant difference in results was observed without it.

To estimate κ and τ at each iteration, we use the procedure from [5] setting its scale parameter to the current estimate of r. Both the technique in [5] and our method for scale selection approximates distances between points along the underlying curve by a sum of edge distances in a graph constructed on the points.

We chose to construct the graph as the sum of disjoint minimum spanning trees (DMST) as suggested in [1], followed by a post-processing step of reject-

ing edges with length greater than that determined by our assumed minimum global density (ρ_0). Figure 9 shows an example of a construction for points acquired from a concertina wire. The range sensor used is a SICK LMS-291 attached to a custom made scanning mount. The angular separation between laser beams is $\frac{1}{4}^\circ$ over a 100° field of view. The angular separation between laser sweeps is $\frac{2}{3}^\circ$ over a range of 115°.

The construction using DMSTs has some desirable properties over traditional k-nearest neighbor or ϵ-ball schemes. In practice, it produces connected graphs without undesirable gaps and does not induce edges to clump together in noisy regions having relatively higher point density. Several authors have also noted its good empirical performance in problems that require good approximations of geodesic distances using graphs. The only parameter to be chosen is the number of spanning trees (in our case, $= 2$) and it has been observed to be robust to changes in the dataset for our choice.

6.2 Validation

As a first step, we test our model by attempting to validate the behavior predicted by the analytical bounds of Section 5 for the 2-D case. The test curve is a 2-D parabola and the error in tangent direction is evaluated at the apex for various values of curvature and point density. The estimation is done using PCA for various values of neighborhood radius. The reader is encouraged to compare Figures 10-12 with the analytic curves of Figures 3-4.

Figure 10 shows the observed angular error with varying curvature κ of the parabola. It can be seen to show the predicted systematic decrease in scale for increased curvature. The variation of estimation error with sample noise σ_0^2 in Figure 11 shows the increase in optimal scale for increased noise. Figure 12 shows the relatively small change in choice of optimal scale except at a low density. It also shows the expected decrease in error with increasing sample density.

6.3 Performance and Stability

We choose to compare the proposed method with the polynomial fitting algorithm of [5], as the latter performed nearly uniformly better experimentally on a variety of synthetic curves against a large family of other fitting approaches based on Gaussian smoothing, Fourier transforms and others.

Figure 13 presents results on 100 samples from two synthetic curves, a 2-D hypocycloid and a 3-D conical helix (as used in [5]). The hypocycloid has the parametric form $(4\cos(t) - \cos(2t), 4\sin(t) + \sin(2t))$ with $t \in [0, 2\pi]$ and the helix has the form $(t\cos(t), t\sin(t), t)$ with $t \in [\pi/2, 5\pi/2]$. These two are presented as their constantly varying curvature violates the assumptions made in both algorithms, and PCA is intuitively not expected to perform well on them under its simplistic assumption of local linearity. The algorithms were run for 30 datasets each for varied sample noise (σ). A range of values for

radius r_0 were used to fix the scale for polynomial fitting and correspondingly serve as the starting point of the proposed PCA algorithm.

As seen in Figure 13, the scale-adaptive PCA performs surprisingly well in terms of error rate, and is much more stable to varying values of r_0. Similar results were observed on comparison with other 2-D and 3-D curves from [5].

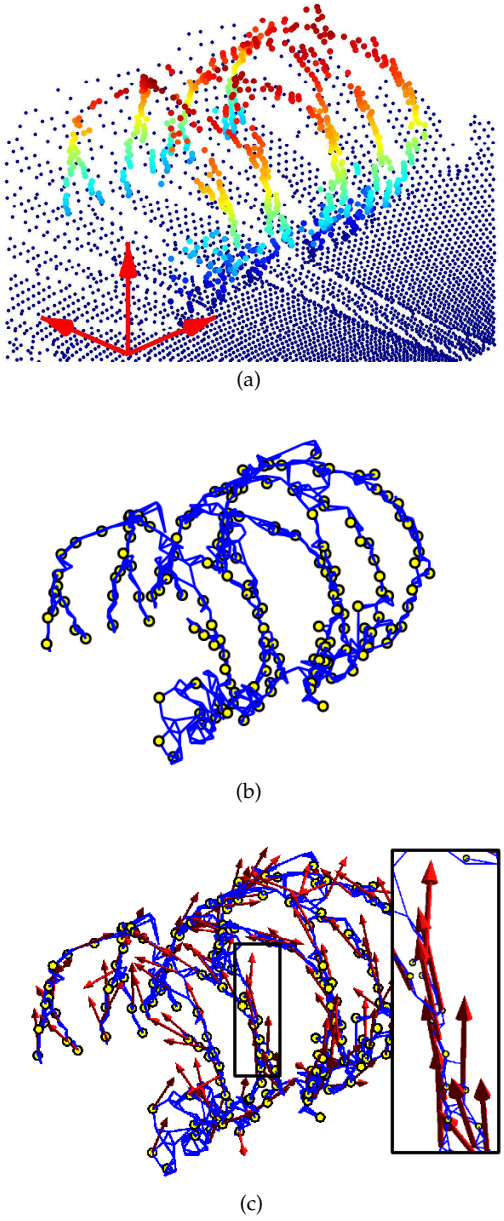
7 Conclusions

Selection of an appropriate scale of analysis is a challenging problem in several domains. This paper presented a technique for adaptive scale selection in estimating tangents of point-sampled curves. We derived analytical bounds for the perturbation of the leading eigenvector in PCA due to the influence of finite samples and noise. The predicted behavior of the change in optimal choice of scale with the perturbing parameters was validated on synthetic data. We also demonstrated experimentally that with principled scale selection, the error in tangent estimation using naïve local PCA is comparable, somewhat counter-intuitively, to the best fixed-scale alternative based on local polynomial fitting.

For future work, it would be interesting to study the effect of differing graph construction on the result of both algorithms. The same theoretical analysis could also be performed for the more robust variant of weighted PCA, for some fixed family of weighting functions (e.g. Gaussian). This would make the proposed algorithm more robust overall to outliers as well as to poor graph construction.

8 Acknowledgements

This document was prepared through collaborative participation in the Robotics Consortium sponsored by the U.S Army Research Laboratory under the Collaborative Technology Alliance Program, Cooperative Agreement DAAD19-01-209912.



(a)

(b)

(c)

Figure 9: Processing the laser scan of a concertina wire having the geometry of two oppositely wound helices of equal diameter: (a) Raw 3-D points color-coded by elevation [axis length = 0.5m], (b) DMST graph constructed on manually extracted non-ground points, and (c) Estimated tangents using scale-adaptive PCA.

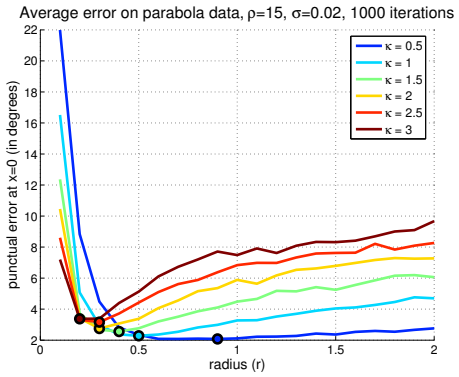


Figure 10: Plot of observed angular error for 2D parabola for varying curvature.

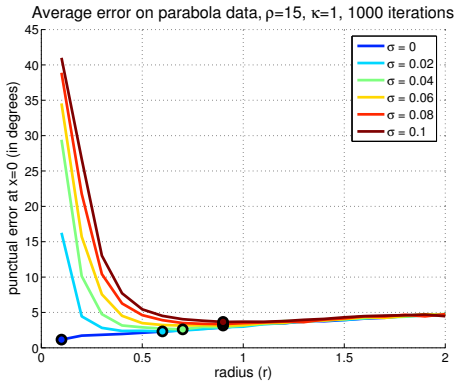


Figure 11: Plot of observed angular error for 2D parabola for varying sample noise.

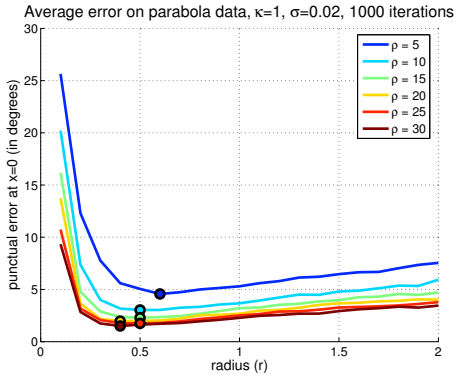
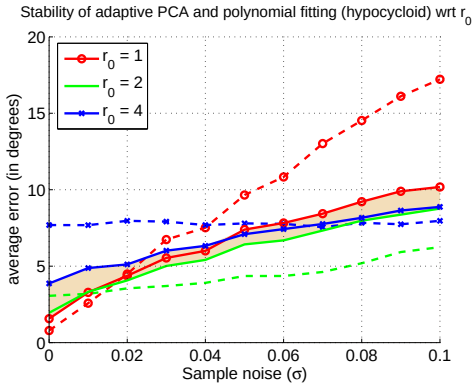
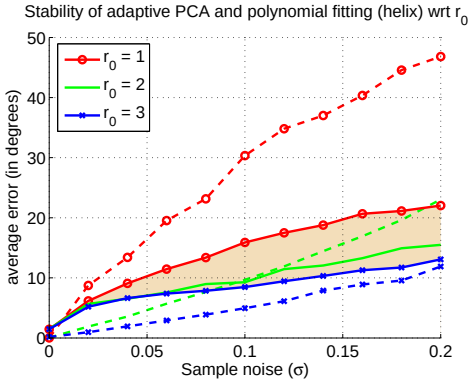


Figure 12: Plot of observed angular error for 2D parabola for varying sample density.



(a)



(b)

Figure 13: Plot of observed error on (a) 2-D hypocycloid and (b) 3-D conical helix dataset. The solid and dashed lines correspond to the proposed scale-adaptive PCA and the method of [5] respectively. The shaded region encloses the lower variation in the more stable PCA method.

References

[1] M. A. Carreira-Perpinán and R. S. Zemel. Proximity graphs for clustering and manifold learning. In *Neural Information Processing Systems (NIPS)*, pages 225–232, 2004.

[2] J. Goodman and J. O. Rourke, editors. *Handbook of Discrete and Computational Geometry*, chapter 30. CRC press, 2004.

[3] B. Kégl, A. Kryzak, T. Linder, and K. Zeger. Learning and design of principal curves. *IEEE Trans. Pattern Anal. Machine Intell.*, 22(3):281–297, 2000.

[4] I.-K. Lee. Curve reconstruction from unorganized points. *Computer Aided Geometric Design*, 17(2):161–177, 2000.

[5] T. Lewiner, J. D. Gomez, H. Lopes, and M. Craizer. Curvature and torsion estimators based on parametric curve fitting. *Computers & Graphics*, 29(5):641–655, 2005.

[6] G. Medioni and C. K. Tang. Inference of integrated surface, curve, and junction descriptions from sparse 3-d data. *IEEE Trans. Pattern Anal. Machine Intell.*, 20(11):1206–1223, 1998.

[7] N. J. Mitra, A. Nguyen, and L. Guibas. Estimating surface normals in noisy point cloud data. *Special issue of Int. Journal of Computational Geometry and Applications*, 14(4):261–276, 2004.

[8] A. Y. Ng, A. X. Zheng, and M. Jordan. Link analysis, eigenvectors, and stability. In *Proc. of the Intl. Joint Conference on Artificial Intelligence (IJCAI)*, pages 903–910, 2001.

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

[10] C. K. Tang, G. Medioni, P. Mordohai, and W. S. Tong. First order augmentations to tensor voting for boundary inference and multiscale analysis in 3-d. *IEEE Trans. Pattern Anal. Machine Intell.*, 26(5):594–611, 2004.

[11] L. Wasserman. *All of Statistics*. Springer, 2004.

Appendices

Appendix A Moments of the uniform distribution

In this section we list some useful statistics of the random variable $S \sim \text{Uniform}(-r, r)$ implying the distribution $f_S(s) = \frac{1}{2r}$ in $[-r, r]$ and zero elsewhere. It can be easily verified that:

$$\mu_S \triangleq \mathbb{E}(S) = \int_{-r}^{r} s \frac{1}{2r} ds = 0 \tag{51}$$

$$d_2(S) = \mathbb{V}(S) = \int_{-r}^{r} s^2 \frac{1}{2r} ds = \frac{r^2}{3} \tag{52}$$

$$\mathbb{E}(S^3) = \int_{-r}^{r} s^3 \frac{1}{2r} ds = 0 \tag{53}$$

$$d_4(S) = \mathbb{E}(S^4) = \int_{-r}^{r} s^4 \frac{1}{2r} ds = \frac{r^4}{5} \tag{54}$$

Let a new random variable $U = S^2$. Then it follows that:

$$\mu_U \triangleq \mathbb{E}(U) = \mathbb{E}(S^2) = \frac{r^2}{3} \tag{55}$$

$$\mathbb{E}(U^2) = \mathbb{E}(S^4) = \frac{r^4}{5} \tag{56}$$

$$d_2(S^2) = \mathbb{V}(U) = \mathbb{E}(U^2) - (\mathbb{E}(U))^2 = \frac{4}{45} r^4 \tag{57}$$

$$\mathbb{E}(U^3) = \int_{-r}^{r} s^6 \frac{1}{2r} ds = \frac{r^6}{7} \tag{58}$$

$$\mathbb{E}(U^4) = \mathbb{E}(S^8) = \frac{r^8}{9} \tag{59}$$

$$d_4(S^2) = \mathbb{E}((U - \mu_U)^4) = \frac{16}{945} r^8 \tag{60}$$

Let a new random variable $V = S^3$. It similarly follows that:

$$\mu_V \triangleq \mathbb{E}(V) = \mathbb{E}(S^3) = 0 \tag{61}$$

$$d_2(S^3) = \mathbb{E}(V^2) = \mathbb{E}(S^6) = \frac{r^6}{7} \tag{62}$$

$$\mathbb{V}(V) = \mathbb{E}(V^2) - (\mathbb{E}(U))^2 = \frac{r^6}{7} \tag{63}$$

$$\mathbb{E}(V^3) = \mathbb{E}(S^9) = 0 \tag{64}$$

$$d_4(S^3) = d_4(V) = \mathbb{E}(V^4) = \mathbb{E}(S^{12}) = \frac{r^{12}}{13} \tag{65}$$

Note also that:

1. Effect of Scaling:
$$\mathbb{E}((\alpha X)^n) = \alpha^n \mathbb{E}(X^n) \tag{66}$$
for a constant α.

2. Effect of Gaussian noise on variance:
$$d_4(X + \eta) = d_4(X) + 6\sigma^2 d_2(X) + 3\sigma^4 \tag{67}$$
for normally distributed noise $\eta \sim \mathcal{N}(o, \sigma^2)$ using the fact the odd moments of η are zero and that $\mathbb{E}(\eta^4) = 3\sigma^4$.

3. Effect of Gaussian noise on covariance:
$$\text{cov}(X + \eta_X, Y + \eta_Y) = \text{cov}(X, Y) \tag{68}$$
for zero-mean independent noise η_X and η_Y.

4. Effect of Gaussian noise on $c_2(X, Y)$:
$$c_2(X + \eta_X, Y + \eta_Y) = c_2(X, Y) + \sigma_0^2(\sigma_X^2 + \sigma_Y^2) + \sigma_0^4 \tag{69}$$
for zero mean iid noise η_X and η_Y that are independent of X, Y and have variance σ_0^2.

Appendix B Estimators and their properties

B.1 Sample mean

Consider a random variable X and the estimator of its mean μ_X from n samples $\{x_i\}$:
$$\bar{X}_n = \frac{1}{n} \sum_i x_i \tag{70}$$

It follows that the sample mean is a random variable with the following properties:
$$\mathbb{E}(\bar{X}_n) = \frac{1}{n} n \mathbb{E}(X_i) = \mu_X \tag{71}$$
$$\mathbb{V}(\bar{X}_n) = \frac{1}{n^2} n \mathbb{V}(X_i) = \frac{\mathbb{V}(X_i)}{n} \tag{72}$$

B.2 Sample Variance

Consider the estimator of variance of X from its n samples $\{x_i\}$:
$$\hat{\sigma}_X^2 = \tfrac{1}{n-1} \sum_i (x_i - \bar{x}_n)^2 \tag{73}$$

The expected value of this estimator (also a random variable) is given by

$$\begin{aligned}
\mathbb{E}(\hat{\sigma}_n^2) &= \tfrac{1}{n-1}\mathbb{E}\left[\sum_i (X_i - \mu_X) - (\bar{X}_n - \mu_X)^2\right] \\
&= \tfrac{1}{n-1}\sum_i \left[\mathbb{E}(X-\mu_X)^2 + \mathbb{E}(\bar{X}_n - \mu_X)^2 - 2\,\mathbb{E}(X_i - \mu_X)(\bar{X}_n - \mu_X)\right] \\
&= \tfrac{1}{n-1}\left[n\mathbb{V}(X) + n\frac{\mathbb{V}(X)}{n} - 2\,\mathbb{E}(\bar{X}_n - \mu_X)(\bar{X}_n - \mu_X)\right] \\
&= \tfrac{1}{n-1}(n+1-2)\mathbb{V}(X) = \mathbb{V}(X)
\end{aligned}$$

Hence $\hat{\sigma}_X^2$ is an unbiased estimator.

The variance of the estimator $\hat{\sigma}_n^2$ is a little less straightforward. We first prove some useful identities.

Identity 1.

$$\hat{\sigma}_n^2 = \tfrac{1}{2n(n-1)}\sum_{i,j}(x_i - x_j)^2$$

Proof.

$$\begin{aligned}
\text{RHS} &= \tfrac{1}{2n(n-1)}\sum_{i,j}\left[(x_i - \bar{x}_n) - (x_j - \bar{x}_n)\right]^2 \\
&= \tfrac{1}{2n(n-1)}\left[n\sum_i (x_i - \bar{x}_n)^2 + n\sum_j (x_j - \bar{x}_n)^2 - 2\sum_{i,j}(x_i - \bar{x}_n)(x_j - \bar{x}_n)\right] \\
&= \tfrac{1}{2n(n-1)}\left[2n\sum_i (x_i - \bar{x}_n)^2 - 2\left[\sum_i (x_i - \bar{x}_n)\right]^2\right] \\
&= \tfrac{1}{n-1}\sum_i (x_i - \bar{x}_n)^2
\end{aligned}$$

■

We will adopt a notation to conveniently represent the higher order centered moments of a random variable X having mean μ_X and variance σ_X^2 by

$$d_n(X) \triangleq \mathbb{E}(X - \mu_X)^n \tag{74}$$

Note that $d_0(X) = 1$, $d_1(X) = 0$ and $\sigma_X^2 = d_2(X)$ under the above definition.

Identity 2.

$$\mathbb{E}\left[(X_i - X_j)^m\right] = \sum_{k=0}^{m}\binom{m}{k} d_k d_{m-k} \tag{75}$$

Proof.

$$\begin{aligned}
(X_i - X_j)^m &= \left((X_i - \mu_X) - (X_j - \mu_X)\right)^m \\
&= \sum_{k=0}^{m}\binom{m}{k}(X_i - \mu_X)^k (X_j - \mu_X)^{m-k}
\end{aligned}$$

Using independence of $X_i - \mu_X$ and $X_j - \mu_X$, and taking expectations on both sides gives the desired result. ■

Lemma 1 (Variance of the sample variance estimator).

$$\mathbb{V}(\hat{\sigma}_X^2) = \frac{d_4(X)}{n} - \frac{(n-3)}{n(n-1)}\sigma_X^4 \tag{76}$$

Proof. By the property of covariance of sums of variables

$$\mathbb{V}(\hat{\sigma}_X^2) = \tfrac{1}{(4n(n-1))^2} \sum_{i,j,k,l} \text{cov}\left((x_i - x_j)^2, (x_k - x_l)^2\right) \tag{77}$$

There are 3 cases for the right hand side:

1. Distinct i, j, k, l :In this case $\text{cov}\left((x_i - x_j)^2, (x_k - x_l)^2\right) = 0$ by independence.
2. Distinct i, j in $\text{cov}\left((x_i - x_j)^2, (x_i - x_j)^2\right)$: There are $2n(n-1)$ terms fulfilling this case.

$$\begin{aligned}
&\text{cov}\left((x_i - x_j)^2, (x_i - x_j)^2\right) \\
&\quad = \mathbb{E}(x_i - x_j)^4 - [\mathbb{E}(x_i - x_j)^2]^2 \\
&\quad = d_4(X) + 8d_1(X)d_3(X) + 6d_2^2(X) + d_4(X) - (2d_2(X) + 2d_1^2(X))^2 \\
&\quad = 2d_4(X) + 2d_2(X)^2
\end{aligned}$$

3. Distinct i, j, k in $\text{cov}\left((x_i - x_j)^2, (x_k - x_j)^2\right)$: There are $4n(n-1)(n-2)$ terms fulfilling this case.

$$\begin{aligned}
\text{cov}\left((x_i - x_j)^2, (x_k - x_j)^2\right) &= \mathbb{E}[(x_i - x_j)^2(x_j - x_k)^2] \\
&= (d_4(X) + 3d_2^2(X)) - 4d_2^2(X) \\
&= d_4(X) - d_2^2(X)
\end{aligned}$$

Summing up the weighted contributions of each case gives the desired result. ■

B.3 Sample Covariance

Consider the sample covariance estimator between random variables X and Y from n samples $\{x_i\}$ and $\{y_i\}$:

$$\hat{S}_{XY} = \frac{1}{n-1}\sum_i (x_i - \bar{X}_n)(y_i - \bar{Y}_n) \tag{78}$$

Let the true value of the covariance be denoted by σ_{XY}.

The expected value of the estimator is given by:

$$\begin{aligned}
\mathbb{E}(\hat{S}_{XY}) &= \frac{1}{n-1}\sum_i \mathbb{E}(x_i - \bar{X}_n)(y_i - \bar{Y}_n) \\
&= \frac{1}{n-1}\sum_i \Big[\mathbb{E}(x_i - \mu_X)(y_i - \mu_Y) - \mathbb{E}(x_i - \mu_X)(\bar{Y}_n - \mu_Y) \\
&\qquad - \mathbb{E}(\bar{X}_n - \mu_X)(y_i - \mu_Y) + \mathbb{E}(\bar{X}_n - \mu_X)(\bar{Y}_n - \mu_Y)\Big] \\
&= \frac{n}{n-1}\left[\sigma_{XY} - \frac{\sigma_{XY}}{n} - \frac{\sigma_{XY}}{n} + n\frac{\sigma_{XY}}{n^2}\right] \\
&= \sigma_{XY}
\end{aligned}$$

Hence the estimator is unbiased.

We proceed to derive variance of $\hat{S}_{XY}$ in a manner similar to that of $\hat{\sigma}_X^2$ by first proving some identities.

Identity 3.

$$\hat{S}_{XY} = \frac{1}{2n(n-1)}\sum_{i,j}(x_i - x_j)(y_i - y_j) \tag{79}$$

Proof.

$$\begin{aligned}
\text{RHS} &= \frac{1}{2n(n-1)}\sum_{i,j}\left[(x_i - \bar{X}_n) - (x_j - \bar{X}_n)\right]\left[(y_i - \bar{Y}_n) - (y_j - \bar{Y}_n)\right] \\
&= \frac{1}{2n(n-1)}\sum_{i,j}\Big[(x_i - \bar{X}_n)(y_i - \bar{Y}_n) - (x_i - \bar{X}_n)(y_j - \bar{Y}_n) \\
&\qquad - (x_j - \bar{X}_n)(y_i - \bar{Y}_n) + (x_j - \bar{X}_n)(y_j - \bar{Y}_n)\Big] \\
&= \frac{1}{2n(n-1)}\left[n\sum_i (x_i - \bar{X}_n)(y_i - \bar{Y}_n) + n\sum_j (x_j - \bar{X}_n)(y_j - \bar{Y}_n)\right] \\
&= \frac{1}{(n-1)}\sum_i (x_i - \bar{X}_n)(y_i - \bar{Y}_n)
\end{aligned}$$

■

We introduce some new symbols for convenience. Let

$$c_m(X,Y) \triangleq \mathbb{E}\left[(X - \mu_X)(Y - \mu_Y)\right]^m \tag{80}$$

Note that $c_1(X,Y) = \sigma_{XY}$ and $c_0(X,Y) = 1$. To avoid cluttering the equations, we will drop the explicit arguments (X,Y) when it is clear from the context.

Identity 4.

$$\mathbb{E}[(X_i - X_j)(Y_i - Y_j)] = 2c_1 \tag{81}$$

Proof.

$$\begin{aligned}\text{LHS} &= \mathbb{E}(X_i - \mu_X)(Y_i - \mu_Y) + \mathbb{E}(X_j - \mu_X)(Y_j - \mu_Y) \\ &\quad + \mathbb{E}(X_i - \mu_X)(Y_j - \mu_Y) + \mathbb{E}(X_j - \mu_X)(Y_i - \mu_Y) \\ &= 2c_1\end{aligned}$$ ■

Identity 5.

$$\mathbb{E}[(X_i - X_j)^2(Y_i - Y_j)^2] = 2(c_2 + \sigma_X^2\sigma_Y^2 + 2c_1^2) \tag{82}$$

Proof.

$$\begin{aligned}\text{LHS} &= \mathbb{E}\Big(\left[(X_i - \mu_X)^2 + (X_j - \mu_X)^2 + 2(X_i - \mu_X)(X_j - \mu_X)\right] \\ &\quad \left[(Y_i - \mu_Y)^2 + (Y_j - \mu_Y)^2 + 2(Y_i - \mu_Y)(Y_j - \mu_Y)\right]\Big) \\ &= \mathbb{E}\left[2(X_i - \mu_X)^2(Y_i - \mu_Y)^2 + 2(X_i - \mu_X)^2(Y_j - \mu_Y)^2 + 4c_1.c_1\right] \\ &= 2c_2 + 2\sigma_X^2\sigma_Y^2 + 4c_1^2\end{aligned}$$ ■

Lemma 2 (Variance of the sample covariance estimator).

$$\boxed{\mathbb{V}(\hat{S}_{XY}) = \frac{c_2(X,Y)}{n} + \frac{\sigma_X^2\sigma_Y^2}{n(n-1)} - \frac{(n-2)}{n(n-1)}c_1^2(X,Y)} \tag{83}$$

Proof. By the property of variance of a sum of variables:

$$\mathbb{V}(\hat{\sigma}_{XY}) = \tfrac{1}{(2n(n-1))^2}\sum_{i,j,k,l}\text{cov}\left((x_i - x_j)(y_i - y_j), (x_k - x_l)(y_k - y_l)\right) \tag{84}$$

It is clear that the case of $i = j$ and $k = l$ reduces the corresponding covariance term to 0. Besides those, there are 3 cases to consider:

1. Distinct i, j, k, l :In this case $\text{cov}\left((x_i - x_j)(y_i - y_j), (x_k - x_l)(y_k - y_l)\right) = 0$ by independence.
2. Distinct i, j in $\text{cov}\left((x_i - x_j)(y_i - y_j), (x_i - x_j)(y_i - y_j)\right)$:
   There are $2n(n-1)$ terms fulfilling this case.

   $$\begin{aligned}&\text{cov}\left((x_i - x_j)(y_i - y_j), (x_i - x_j)(y_i - y_j)\right) \\ &\quad = \mathbb{E}\left((x_i - x_j)^2(y_i - y_j)^2\right) - \left(\mathbb{E}(x_i - x_j)(y_i - y_j)\right)^2 \\ &\quad = 2(c_2 + \sigma_X^2\sigma_Y^2 + 2c_1^2) - 4c_1^2 \\ &\quad = 2c_2 + 2\sigma_X^2\sigma_Y^2\end{aligned}$$

3. Distinct i, j, k in $\text{cov}\left((x_i - x_j)(y_i - y_j), (x_k - x_j)(y_k - y_j)\right)$: There are $4n(n-1)(n-2)$ terms fulfilling this case. After expressing each term wrt difference from the mean, expanding out the products and some manipulation, we get
   $$\mathbb{E}\left((x_i - x_j)(y_i - y_j)(x_k - x_j)(y_k - y_j)\right) = 3c_1^2 + c_2$$

which yields

$$\begin{aligned}\operatorname{cov}\left((x_i - x_j)(y_i - y_j), (x_k - x_j)(y_k - y_j)\right) &= (3c_1^2 + c_2) - 4c_1^2 \\ &= c_2 - c_1^2\end{aligned}$$

Summing up the weighted contributions of each case gives the desired result. ■