Scale Selection for the Analysis of Point-Sampled Curves

Ranjith Unnikrishnan   Jean-François Lalonde   Nicolas Vandapel   Martial Hebert

Robotics Institute, Carnegie Mellon University

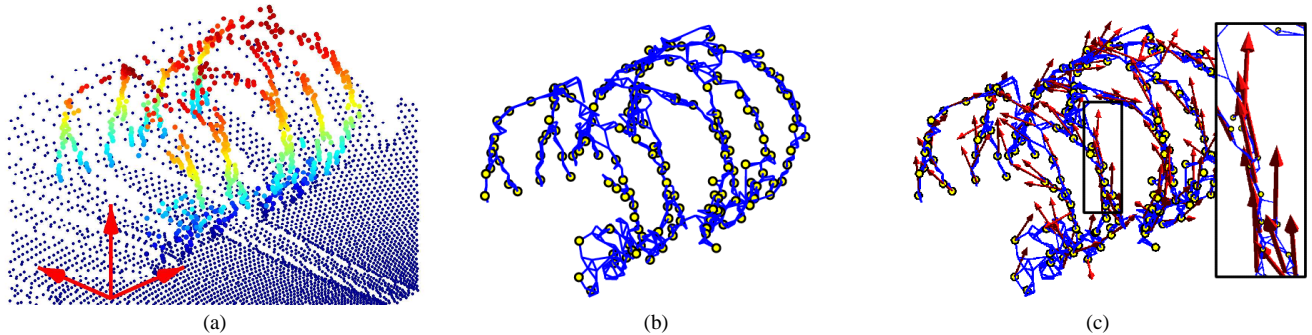{ranjith, jlalonde, vandapel, hebert}@cs.cmu.edu

Figure 1: 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, (c) Estimated tangents using scale-adaptive PCA.

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. [1]

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 ra-

dius 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.

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. 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 of the 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.

We also mention the related theoretical work by Mitra et al. [7] on optimal neighborhood size for normal estimation in surfaces using PCA. They derived 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.

In the next section we describe our proposed method and derive the required analytical results for the case of 2-D and 3-D curves. In Section 3, we present experimental evidence to validate our solution and demonstrate that it matches the predicted variation of the optimal radius with the perturbing parameters. We also compare its stability and performance to the local polynomial fitting algorithm of [5]. Section 4 concludes with discussion and directions for future work.

2. Approach

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. Our derivation proceeds as follows:

We first model and state our assumptions of the underlying geometry (Section 2.1) and adopt a convenient reference frame.

We then compute the expected statistics of the estimators (Section 2.2) 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. The effect of the perturbation matrix is the combination of finite sample size as well as noise in the data.

Using bounds on the Frobenius norm of the random matrix, we then compute (Section 2.3) an upper bound on the angular deviation of the eigenvector of the perturbed matrix as a function of sample noise, sample size, radius as well as the curvature and torsion of the curve.

We then analyze the cases of 2-D and 3-D curves (Section 2.4), make observations from the analytic behavior of their bounds, and derive the conditions under which the
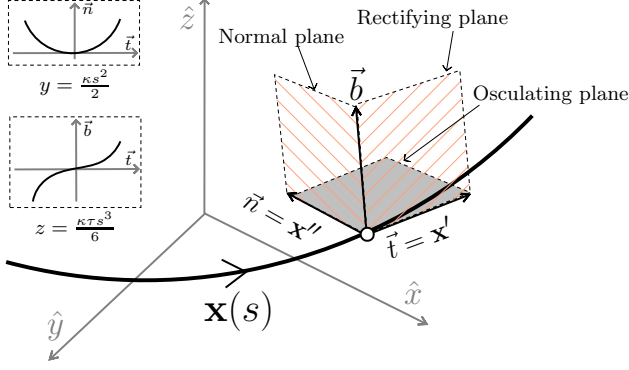
Figure 2: Model of local curve geometry

dominant eigenvector of their corresponding expected matrices align with the tangent to the true curve in the canonical frame.

2.1 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 2) 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.

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 $S \sim$ 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 $n \geq 2\rho_0 r$

2.2 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

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

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

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

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 the diagonal elements 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 $\text{cov}(X, Z)$ of X and Z).

From the distribution of $S \sim \text{Uniform}(-r, r)$, we can then compute the expected values of each of the above quantities.

For example, using $X = S + \eta_X$

$$\mathbb{E}(M_{11}) = \mathbb{V}(X_i) = \mathbb{V}(S + \eta) = \mathbb{V}(S_i) + \sigma_0^2$$
$$= \sigma_X^2 + \sigma_0^2 = \int_{-r}^{r} s^2 \frac{1}{2r} ds + \sigma_0^2 = \frac{r^2}{3} + \sigma_0^2 \quad (6)$$

Using a similar procedure, we can derive the following identities under the model defined in (1).

$$\mathbb{E}(M_{12}) = \text{cov}(X, Y) = \frac{\kappa}{2} \mathbb{E}(S^3) = 0 \quad (7)$$

$$\mathbb{E}(M_{13}) = \text{cov}(X, Z) = \frac{\kappa\tau}{6} \mathbb{E}(S^4) = \frac{\kappa\tau}{30} r^4 \quad (8)$$

$$\mathbb{E}(M_{22}) = \mathbb{V}(Y) = \frac{\kappa^2}{4} \mathbb{V}(S^2) + \sigma_0^2 = \frac{\kappa^2}{45} r^4 + \sigma_0^2 \quad (9)$$

$$\mathbb{E}(M_{23}) = \text{cov}(Y, Z) = \frac{\kappa^2\tau}{18} (\mathbb{E}(S^5) - \mathbb{E}(S^2)\mathbb{E}(S^3)) = 0 \quad (10)$$

$$\mathbb{E}(M_{33}) = \mathbb{V}(Z) = \left(\frac{\kappa\tau}{6}\right)^2 \frac{r^6}{7} + \sigma_0^2 \quad (11)$$

To proceed from here, we must use results on the variance of the sample variance and sample covariance estimators. We state them without proof [11] below.

Identity 1 (Variance of the *sample variance* estimator).

$$\mathbb{V}(\hat{\sigma}_X^2) = \frac{d_4(X)}{n} - \frac{(n-3)}{n(n-1)} \sigma_X^4 \quad (12)$$

for a random variable X, where

$$d_n(X) \triangleq \mathbb{E}(X - \mu_X)^n \quad (13)$$

Identity 2 (Variance of the *sample covariance* estimator).

$$\mathbb{V}(\hat{S}_{XY}) = \frac{c_2(X, Y)}{n} + \frac{\sigma_X^2 \sigma_Y^2}{n(n-1)} - \frac{(n-2)}{n(n-1)} c_1^2(X, Y) \quad (14)$$

for random variables X and Y, where

$$c_m(X, Y) \triangleq \mathbb{E}\left[(X - \mu_X)(Y - \mu_Y)\right]^m \quad (15)$$

Note that we use the hat symbol (ˆ) to distinguish the estimator from the true quantity.

Under the curve model defined in (1), we can derive the expressions for $d_4(X)$, $d_4(Y)$ and $d_4(Z)$ in a manner similar to that used for (7)–(11), using the identity:

$$d_4(X + \eta) = d_4(X) + 6\sigma^2 d_2(X) + 3\sigma_0^4 \quad (16)$$

for any random variable X affected by normally distributed independent noise $\eta \sim \mathcal{N}(o, \sigma_0^2)$. Note that the simplification is because the odd moments of η vanish and $\mathbb{E}(\eta^4) = 3\sigma_0^4$. We may also similarly derive the expressions for c_1 and c_2 for all pairs of X, Y and Z.

Once we have the required values for c_i and d_i, we can then substitute them back in (12) and (14) to get the variance of the individual estimators, which we denote as $\mathbb{V}(M_{ij})$ with $i, j = \{1, 2\}$. The final expressions for $\mathbb{V}(M_{ij})$ were obtained using Mathematica™ and are listed in [11] due to space limitations.

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:

$$\hat{M} = \tilde{M} + Q \quad (17)$$

Here $\tilde{M} = \mathbb{E}(M)$ is a symmetric matrix with elements given by (7)–(11), and Q is a symmetric *perturbation matrix* of random variables each with mean 0 and variance given by the expressions listed in [11].

2.3 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]:

$$\|\hat{e}_1 - e_1\| \leq \frac{4\|Q\|_F}{\delta - \sqrt{2}\|Q\|_F} \quad (18)$$

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

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 (18) is equivalent to minimizing the ratio:
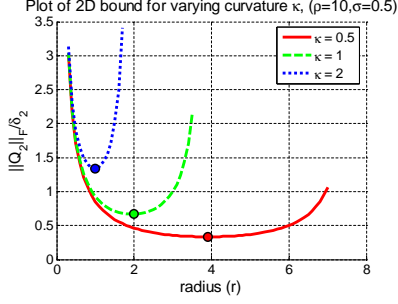
$$B \triangleq \|Q\|_F / \delta \quad (19)$$

4

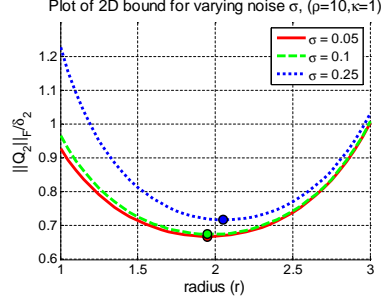Figure 3: Plot of analytic 2D bound for varying curvature

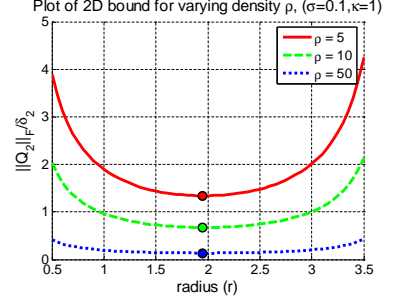Figure 4: Plot of analytic 2D bound for varying sampling noise

Figure 5: Plot of analytic 2D bound for varying sample density

We will analyze the function B for both 2-D and 3-D curves in the next section.

2.4 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.

2.4.1 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 2.3, we know that Frobenius norm of Q_2 is upper bounded with probability $1 - \epsilon$ by:

$$
\begin{aligned}
||Q_2||_F^2 &\leq \frac{1}{n\epsilon} \sum_{i=1}^{2} \sum_{j=1}^{2} \mathbb{V}(M_{ij}) \\
&= \frac{1}{n\epsilon} \left[\mathbb{V}(M_{11}) + \mathbb{V}(M_{22}) + 2\mathbb{V}(M_{12}) \right]
\end{aligned}
\tag{20}
$$

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

$$
\tilde{M}_2 = \begin{bmatrix} \frac{r^2}{3} + \sigma_0^2 & 0 \\ 0 & \frac{\kappa^2}{45}r^4 + \sigma_0^2 \end{bmatrix}
\tag{21}
$$

is obtained easily by inspection as

$$
\delta_2 = \frac{r^2}{3} - \frac{\kappa^2 r^4}{45}
\tag{22}
$$

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

$$
0 < r < \sqrt{15}/\kappa
\tag{23}
$$

The bound to be minimized then is

$$
B_2(r) \triangleq \frac{||Q_2||_F}{\delta_2}
\tag{24}
$$

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 (24) 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. Firstly, although the bounds are discontinuous functions of high order polynomials in r, our analysis is restricted to the regime where the constraints (23) required for eigenvector dominance are satisfied. In this regime, the bound is convex with a unique minimum.

Secondly, and as also observed experimentally in [5, 7], 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 in our model as our continuous relaxation of n as $2\rho r$ is invalid for small r. Hence, we advocate the interpretation of the function only 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 4–5, we make the following qualitative observations:

1. Complexity: Like most closed form expressions encountered in real-world problems, the formula in (24)

is messy. 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 6th degree polynomial and a 4th degree polynomial of r, both only containing even powers of r.

2. Variation with curvature κ: Figure 3 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 (23) when the eigenvector more closely aligned to the x-axis is no longer dominant.

3. Variation with sampling noise σ_0 : Figure 4 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 (23).

4. Variation with sampling density ρ : Figure 5 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 3.2.

2.4.2   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 2.3, the $||Q||_F$ is upper bounded with probability $1 - \epsilon$ by:

$$
\begin{aligned}
||Q||_F^2 &\leq \frac{1}{n\epsilon} \sum_{i=1}^{3} \sum_{j=1}^{3} \mathbb{V}(M_{ij}) \\
&= \frac{1}{n\epsilon} \Big[\mathbb{V}(M_{11}) + \mathbb{V}(M_{22}) + \mathbb{V}(M_{33}) \\
&\qquad + 2 \left(\mathbb{V}(M_{12}) + \mathbb{V}(M_{13}) + \mathbb{V}(M_{23}) \right) \Big]
\end{aligned}
\tag{25}
$$

Substituting the results from Section 2.2 gives the required final expression [11].

The matrix of expected values can be written as:

$$
\tilde{M} = \mathbb{E}(M) = \begin{bmatrix} \frac{r^2}{3} + \sigma_0^2 & 0 & \frac{\kappa\tau}{30}r^4 \\ 0 & \frac{\kappa^2}{45}r^4 + \sigma_0^2 & 0 \\ \frac{\kappa\tau}{30}r^4 & 0 & \left(\frac{\kappa\tau}{6}\right)^2 \frac{r^6}{7} + \sigma_0^2 \end{bmatrix}
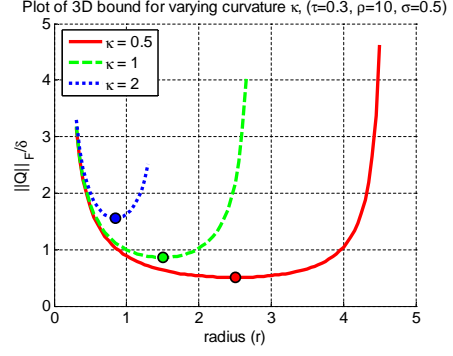\tag{26}
$$

Figure 6: Plot of analytic 3D bound for varying curvature

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

$$
\left| \lambda_1 - \frac{r^2}{3} + \sigma_0^2 \right| \leq \frac{\kappa\tau}{30}r^4
\tag{27}
$$

$$
\lambda_2 = \frac{\kappa^2}{45}r^4 + \sigma_0^2
\tag{28}
$$

$$
\left| \lambda_3 - \left(\frac{\kappa\tau}{6}\right)^2 \frac{r^6}{7} + \sigma_0^2 \right| \leq \frac{\kappa\tau}{30}r^4
\tag{29}
$$

Under the conditions of (23), this gives the additional constraint on radius as:

$$
r \leq \sqrt{28/5\tau}
\tag{30}
$$

and a bound on spectral gap as:

$$
\delta_3 \geq \frac{r^2}{3} - \frac{\kappa^2 r^4}{45} - \kappa\tau \frac{r^4}{15}
\tag{31}
$$

Combining (25) and (31) with the continuous relaxation $n = 2\rho r$ in (19) gives the desired 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 2).

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 6 which has the same parameters as the 2-D curve of Figure 3 but with a non-zero torsion $\tau = 0.3$.

3. Experimental Results

In this section, we first outline the proposed algorithm used in later experimental results. We experimentally study the
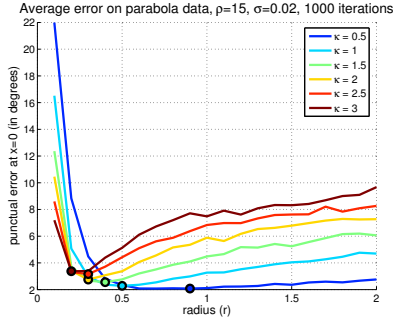
Figure 7: Plot of observed angular error for 2D parabola for varying curvature.
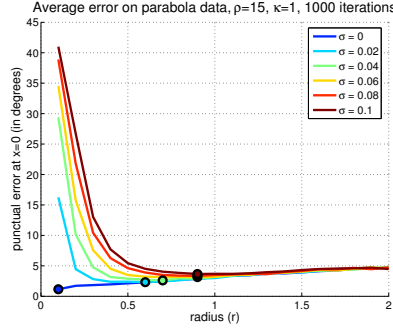
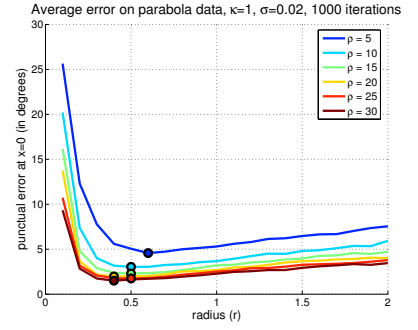Figure 8: Plot of observed angular error for 2D parabola for varying sample noise.

Figure 9: Plot of observed angular error for 2D parabola for varying sample density.

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.

3.1 Algorithm and implementation

The analytical behavior of the bound derived in Section 2.4 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 (24), subject to (23) 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 rejecting edges with length greater than that determined by our assumed minimum global density (ρ_0). Figure 1 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. 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.

3.2 Validation

As a first step, we test our model by attempting to validate the behavior predicted by the analytical bounds of Section 2.4 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 7-9 with the analytic curves of Figures 4-5.

Figure 7 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 8 shows the increase in optimal scale for increased noise. Figure 9 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.

3.3 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.

4. 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.

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] R. Unnikrishnan, J.-F. Lalonde, N. Vandapel, and M. Hebert. Scale selection for the analysis of point sampled curves. Technical report, Robotics Institute, CMU, 2006.
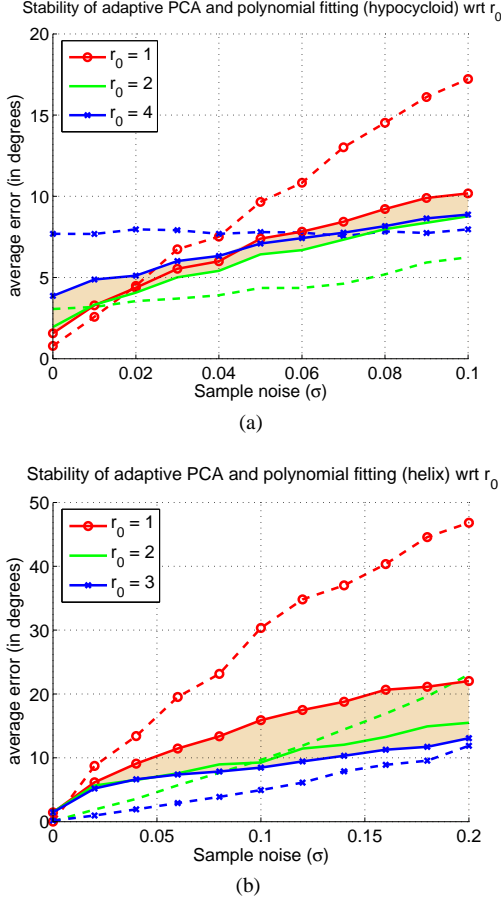
Figure 10: Plot of observed error on (a) 2-D hypocycloid and (b) 3-D conical helix dataset. The solid and dashed lines correspond to the proposed scale-adaptive PCA and the method of [5] respectively. The shaded region encloses the lower variation in the more stable PCA method.

Figure 10 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 10, 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].