Sankhyā : The Indian Journal of Statistics
2018, Volume 80-A, Supplement 1, pp. S8-S32

Dedicated to the memory of P. C. Mahalanobis

Statistical Shape Methodology for the Analysis of Helices

Mai F. Alfahad, John T. Kent and Kanti V. Mardia
University of Leeds, Leeds, UK
Kanti V. Mardia
University of Oxford, Oxford, UK

Abstract

Consider a helix in three-dimensional space along which a sequence of equally spaced points is observed, subject to statistical noise. For data coming from a single helix, a two-stage algorithm based on a profile likelihood is developed to compute the maximum likelihood estimate of the helix parameters. Statistical properties of the estimator are studied and comparisons are made to other estimators found in the literature. Next a likelihood ratio test is developed to test if there is a change point in the helix, splitting the data into two sub-helices. The shapes of protein α-helices are used to illustrate the methodology.

AMS (2000) *subject classification.* Primary 62H11; Secondary 62P10, 92C40.
Keywords and phrases. Change point, Helix axis, Kinked helix, Principal component analysis, Procrustes analysis, Shape analysis

1 Introduction

Proteins form a major part of all living organisms, with their shape being specific to their function. The most common shape is the α-helix, which is a right-handed helix (see for examples, Campbell and Farrell (2014), Wilman (2014a), and Section 2 below). Many helices have a *kink*, i.e. a point where the helix axis locally changes direction (Wilman et al., 2014b; Mardia, 2014). Various statistical methods in the literature have been developed to analyze kinks as local change points, such as *Helanal* by Bansal et al. (2000), *Kinkfinder* by Wilman (2014a), and *Kink-Detector* by Mardia et al. (2018). Kinkfinder uses all the atoms on the helix, whereas the Helanal and Kink-Detector use just the C_α atoms. These methods estimate the kink position by looking for a local change point in the direction of the helix axis. For further review of the topic, see Mardia (2013) and Mardia et al. (2018).

Blundell et al. (1983) and Barlow and Thornton (1988) initiated the work on helix structure based on curvature; they used main classifications:

straight, kinked and curved. There are several ways to define these classifications but it is better to recall how Wilman et al. (2014b) formulated these for their experiments.

- *Kinked:* There is a clear location where the direction of the helix changes. Only a small part of the helix is involved in this.
- *Curved:* There is a slow but steady change of the direction of the helix. This can happen over a large part or even all of the helix.
- *Straight:* There is no change in the overall direction of the helix.

Mardia et al. (1999) was another early paper on the estimation of curvature and torsion for a helix.

The study of helix structure falls within the scope of statistical shape analysis. Shape analysis deals with geometric objects in Euclidean space and is concerned with the properties that remain invariant under a group of transformations. There is now a rich collection of statistical tools for shape data (e.g. Dryden and Mardia, 2016). The simplest type of object consists of a collection of points or *landmarks*, and the most important groups are the similarity transformations (location, scale and rotation) and the rigid body transformations (location and rotation). For this paper an *object* consists of a set of points in $\mathbb{R}^3$ that lie on or near a helix and the relevant group is the rigid body transformations.

A *helix* in three dimensions is defined by the function

$$\boldsymbol{f}(t) = \begin{bmatrix} r\cos t \\ r\sin t \\ ct \end{bmatrix} \tag{1.1}$$

where r and $2\pi c$ denote the *radius* and *pitch*. In this representation the helix is traversed at constant speed as a function of an independent variable t, which can be regarded either as arc length after projecting the helix onto the plane of the first two coordinates, or as a sort of *time*, even though the helix exists as a static object. In addition, the position of the helix in $\mathbb{R}^3$ can be altered by a rigid body motion. In a statistical helix, regular measurements are available along the helix subject to statistical error.

There are two main purposes for the present paper. First we revisit the estimation problem for a statistical helix and show that maximum likelihood estimation can be reduced to an optimized least squares problem under certain assumptions. Secondly, we present a new global approach to the change point problem and investigate the use of feature statistics to highlight the character of a change point.

In more detail, the paper is organized as follows. Section 2 introduces the statistical helix model (similar in spirit to Mardia et al. (2018), but more geometrically explicit). Section 3 gives the details behind the optimized least squares algorithm. This algorithm needs an initial estimate of the helix axis to start the iterations, and there are a large number of methods in the literature such as *Rotfit* described by Christopher et al. (1996) and *HELFIT* by Enkhbayar et al. (2008). Two methods to estimate the initial axis are compared in Section 4: Rotfit and a new method based on modified principal components. Section 6 presents a new likelihood ratio test for the presence of a global change point, both where the time index of the potential change point is known and where it needs to be estimated. We give this procedure the name *ChangePoint-Detector*. A bootstrap procedure is proposed to assess statistical significance. This procedure is investigated on simulated data in Section 7. In Section 8 it is applied to several protein examples and compared to *Kink-Detector* (Mardia et al., 2018).

2 The Statistical Helix Model

If an arbitrary rotation and location are incorporated in Eq. 1.1, then the mathematical helix at a regularly spaced set of "times" takes the form

$$\boldsymbol{f}(t_i) = r\cos(t_i)\boldsymbol{u} + r\sin(t_i)\boldsymbol{v} + ct_i\boldsymbol{w} + \boldsymbol{b}, \quad i = n_1, n_1+1, \ldots, n_2-1, n_2. \tag{2.1}$$

Thus the number of points on the helix is $n = n_2 - n_1 + 1$. General indices $n_1 < n_2$ are allowed for the start and finish points in order to facilitate the parameterization when a change point is present; see Section 6. Here

- $\Gamma = \begin{bmatrix} \boldsymbol{u} & \boldsymbol{v} & \boldsymbol{w} \end{bmatrix}$ is a 3×3 orthogonal matrix whose three columns define the orientation of the helix. In particular the vector $\boldsymbol{w}$ defines the *helix axis*, and the vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ define the plane normal to the helix axis.
- $r > 0$ defines the *helix radius*,
- $2\pi c > 0$ defines the *helix pitch*, which is the vertical height of one turn of the helix,
- $\boldsymbol{b} \in \mathbb{R}^3$ is an *intercept*,
- $t_i = i\beta$ is a sequence of regularly-spaced *times* at which the helix is observed, where $\beta > 0$ defines the *turn angle* of the helix in radians (i.e. the angle between two consecutive points on the helix).

Helices can be regarded as right-handed or left-handed depending on whether $\det(\Gamma) = +1$ or -1 respectively (e.g. Campbell and Farrell, 2014). For the purpose of this paper we largely restrict a helix to be right-handed.

A *regular (statistical) helix* is a collection of points or landmarks

$$\{\boldsymbol{y}_i = [y_{i1}, y_{i2}, y_{i3}]^T,\ i = n_1, n_1 + 1, \ldots, n_2 - 1, n_2\}$$

in three dimensions obtained from a mathematical helix by adding noise,

$$\boldsymbol{y}_i = \boldsymbol{f}(t_i) + \boldsymbol{\epsilon}_i, \quad i = n_1, n_1 + 1, \ldots, n_2 - 1, n_2, \tag{2.2}$$

where the error terms

$$\boldsymbol{\epsilon}_i \sim N_3(\boldsymbol{0}, \sigma^2 I)$$

are assumed here to follow independent isotropic normal distributions. The adjective "regular" is used to distinguish model (2.2) from the change point helix model to be introduced in Section 4.

It is also convenient to let

$$\boldsymbol{g}(t) = ct\boldsymbol{w} + \boldsymbol{b}, \tag{2.3}$$

so that $\boldsymbol{g}(t) - \boldsymbol{b}$ denotes the projection of the centered true helix function $\boldsymbol{f}(t) - \boldsymbol{b}$ onto the helix axis $\boldsymbol{w}$. In particular, let

$$\boldsymbol{g}_i = ct_i\boldsymbol{w} + \boldsymbol{b}, \tag{2.4}$$

denote the axis values at the data times t_i.

For the purposes of this paper, we shall treat the turn angle β as known. It is well-known in the protein structure literature that the turn angle β along the helix can be treated as having a constant value very close to $\beta = 100°$; e.g. Dickerson and Geis (1969) give a value of 1.75 radians $= 100.3°$. Recent confirmation can be obtained from the detailed analysis of the maximum likelihood estimation of β carried out for 129 straight helices from crowdsourcing data in the web supplement (Web Figure 5(d)) to Mardia et al. (2018); for this data it was found that the mean estimate of β is $99.1°$ with a standard error of $1.2°$.

All of the other parameters will be regarded as unknown and needing estimation. However, for the purposes of developing an estimation algorithm, we shall first treat the case where the axis $\boldsymbol{w}$ is known.

The parameters of a helix can be divided into two types. The *registration* parameters are the orthonormal vectors $\boldsymbol{u}, \boldsymbol{v}$ and $\boldsymbol{w}$, and the intercept vector $\boldsymbol{b} = (b_1, b_2, b_3)^T$. The *shape* parameters are the radius r and the pitch c.

Consider a right-handed helix in Eq. 2.2 with orientation matrix Γ, so $\Gamma^T\Gamma = I$ and $\det(\Gamma) = +1$. The following definitions impose further restrictions on Γ. The helix is said to be in

H.1 *canonical coordinates* if $\Gamma = \Gamma_0$, where

$$\Gamma_0 = \begin{bmatrix} \boldsymbol{u}_0 & \boldsymbol{v}_0 & \boldsymbol{w}_0 \end{bmatrix} = I_3$$

is the identity matrix, so the three orientation vectors are given by the three standard coordinate directions in $\mathbb{R}^3$; in particular $\boldsymbol{w}_0 = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^T$ lies in the vertical direction.

H.2 *semi-canonical coordinates* if $\boldsymbol{w} = \boldsymbol{w}_0 = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^T$, without further restrictions on $\boldsymbol{u}$ and $\boldsymbol{v}$. They are used in Section 3.1.

H.3 *general coordinates* if there are no further restrictions on Γ. They are used in Section 3.2.

For a left-handed helix in canonical coordinates it is convenient to define

$$\Gamma_0 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Thus, looking at the horizontal plane from above in canonical or semi-canonical coordinates, a right-handed helix winds around the plane in a counter-clockwise direction as t increases; a left-handed helix winds in a clockwise direction.

3 Parameter Estimation for a Regular Helix

3.1. Known vertical helix axis Start with the assumption that a right-handed data helix is in semi-canonical coordinates, so that $\boldsymbol{w} = \boldsymbol{w}_0 = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^T$ is known to be vertical. Then $\boldsymbol{u}$ and $\boldsymbol{v}$ take the form

$$\begin{bmatrix} \boldsymbol{u} & \boldsymbol{v} \end{bmatrix} = \begin{bmatrix} \cos\tau & \sin\tau \\ -\sin\tau & \cos\tau \\ 0 & 0 \end{bmatrix}, \tag{3.1}$$

for some angle τ. In this case model (2.2) can be re-written as

$$\begin{aligned} \boldsymbol{y}_i &= r\cos(t_i - \tau)\boldsymbol{u}^{(0)} + r\sin(t_i - \tau)\boldsymbol{v}^{(0)} + ct_i\boldsymbol{w}^{(0)} + \boldsymbol{b} + \boldsymbol{\epsilon}_i, \\ &= \alpha_1(\cos t_i\, \boldsymbol{u}^{(0)} + \sin t_i\, \boldsymbol{v}^{(0)}) + \alpha_2(\sin t_i\, \boldsymbol{u}^{(0)} - \cos t_i\, \boldsymbol{v}^{(0)}) + ct_i\, \boldsymbol{w}^{(0)} + \boldsymbol{b} + \boldsymbol{\epsilon}_i, \\ &\quad i = n_1, n_1 + 1, \ldots, n_2 - 1, n_2 \end{aligned} \tag{3.2}$$

where $\alpha_1 = r\cos\tau,\ \alpha_2 = r\sin\tau$.

The model in Eq. 3.2 can be viewed as a multivariate linear regression model with a three-dimensional response on n observations. The regression parameters are $\alpha_1, \alpha_2, c, \boldsymbol{b}$. Since the error term is isotropic, the model can also be represented as a multiple regression model with $3n$ scalar responses, after stacking the 3 columns for the response. Further, maximum likelihood estimation reduces to least squares regression. The $3n \times 6$ design matrix is

$$X = \begin{bmatrix} \boldsymbol{c} & \boldsymbol{s} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \boldsymbol{s} & -\boldsymbol{c} & \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{t} & \mathbf{0} & \mathbf{0} & \mathbf{1} \end{bmatrix}, \tag{3.3}$$

where

$$\boldsymbol{c} = \begin{bmatrix} \cos t_{n_1} \\ \vdots \\ \cos t_{n_2} \end{bmatrix}, \quad \boldsymbol{s} = \begin{bmatrix} \sin t_{n_1} \\ \vdots \\ \sin t_{n_2} \end{bmatrix}, \quad \boldsymbol{t} = \begin{bmatrix} t_{n_1} \\ \vdots \\ t_{n_2} \end{bmatrix}, \quad \mathbf{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \quad \mathbf{0} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Write $\boldsymbol{y}_i = (y_{i1}, y_{i2}, y_{i3})^T$. Then the least squares estimators take the form

$$\begin{aligned} \hat{\alpha}_1 &= \textstyle\sum (c_i' y_{i1}' + s_i' y_{i2}')/\{n(1-\overline{R}^2)\}, \\ \hat{\alpha}_2 &= \textstyle\sum (s_i' y_{i1}' + c_i' y_{i2}')/\{n(1-\overline{R}^2)\}, \\ \hat{c} &= \textstyle\sum t_i' y_{i3}'/\{\sum (t_i - \overline{T})^2\}, \end{aligned} \tag{3.4}$$

in terms of the centered variables $\boldsymbol{y}_i' = \boldsymbol{y}_i - \overline{\boldsymbol{y}},\ c_i' = \cos t_i - \overline{C},\ s_i' = \sin t_i - \overline{S},\ t_i' = t_i - \overline{T}$, where

$$\overline{\boldsymbol{y}} = \frac{1}{n}\sum \boldsymbol{y}_i, \quad \overline{C} = \frac{1}{n}\sum \cos t_i, \quad \overline{S} = \frac{1}{n}\sum \sin t_i, \quad \overline{T} = \frac{1}{n}\sum t_i, \quad \overline{R} = \sqrt{\overline{C}^2 + \overline{S}^2}.$$

We can then derive $\hat{\tau}$ and $\hat{r}$ by

$$\hat{\tau} = \operatorname{atan2}(\hat{\alpha}_2, \hat{\alpha}_1), \quad \hat{r} = \sqrt{\hat{\alpha}_1^2 + \hat{\alpha}_2^2},$$

where atan2 is the two-argument arctan function, so that we have $(\hat{\alpha}_1, \hat{\alpha}_2) = r(\cos\hat{\tau}, \sin\hat{\tau})$. Further, the estimated shift vector $\hat{\boldsymbol{b}} = [\hat{b}_1, \hat{b}_2, \hat{b}_3]^T$ is given by

$$\begin{aligned} \hat{b}_1 &= \overline{y}_1 - \hat{\alpha}_1\overline{C} - \hat{\alpha}_2\overline{S}, \\ \hat{b}_2 &= \overline{y}_2 - \hat{\alpha}_1\overline{S} + \hat{\alpha}_2\overline{C}, \\ \hat{b}_3 &= \overline{y}_3 - \hat{c}\,\overline{T}, \end{aligned}$$

where $\overline{y}_1, \overline{y}_2$ and $\overline{y}_3$ are the means of each coordinate. Finally, the residual sum of squares (RSS) is given by:

$$RSS(\boldsymbol{w}_0) = \sum_{i=n_1}^{n_2} ||\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i||^2,$$

where $\hat{\boldsymbol{y}}_i$ is the i^{th} fitted value of $\boldsymbol{y}_i$, a vector of dimension 3. The residual sum of squares depends on the choice of helix axis, denoted here by $\boldsymbol{w}_0$.

In the case of a left-handed helix, just change the sign of one of the columns in Eq. 3.1; e.g. let

$$\begin{bmatrix} \boldsymbol{u} & \boldsymbol{v} \end{bmatrix} = \begin{bmatrix} \cos\tau & -\sin\tau \\ -\sin\tau & -\cos\tau \\ 0 & 0 \end{bmatrix}, \tag{3.5}$$

for some angle τ, with corresponding changes to the subsequent algebra. If it is unknown whether or not the helix is right- or left-handed, fit both models and choose the model with the smaller residual sum of squares. Unless σ^2 is extremely large, the correct choice will be obvious.

In addition, if the estimate of the pitch parameter c in Eq. 3.4 is negative, then it is necessary to change the sign of the helix axis $\boldsymbol{w}$ (plus the sign of either $\boldsymbol{u}$ or $\boldsymbol{v}$ to preserve the sign of $\det(\Gamma)$).

3.2. Known helix axis in general position Next let the axis $\boldsymbol{w}$ of a right-handed helix be a known unit vector, but not necessarily vertical. Let $G = G(\boldsymbol{w})$ be a 3×3 rotation matrix whose third column equals $\boldsymbol{w}$. Let $\boldsymbol{z}_i = G^T \boldsymbol{y}_i$, denote the rotated data, so that the known helix axis for the $\{\boldsymbol{z}_i\}$ is vertical. Then the estimation procedure of the previous section can be applied to the $\{\boldsymbol{z}_i\}$.

The plane spanned by the first two columns of G is determined by $\boldsymbol{w}$, but not the two columns themselves. A rotation of this plane about the $\boldsymbol{w}$ axis corresponds to changing the meaning of the angle τ in Eq. 3.2.

After fitting the helix by least squares, the quality of the fit can be summarized by the residual sum of squares, denoted RSS($\boldsymbol{w}$), say. The value of RSS($\boldsymbol{w}$) does not depend on the indeterminacy in the meaning of τ.

3.3. Unknown helix axis If $\boldsymbol{w}$ is unknown, an iterative method based on profile likelihood can be used to find the maximum likelihood estimators. The procedure works as follows:

(a) Start with an initial estimate $\boldsymbol{w}_{\text{init}}$, say. Two possibilities are suggested in the next section.

(b) Given $\boldsymbol{w}$, the maximized likelihood with respect to the remaining parameters (known as the "profile likelihood"),

$$-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\frac{\text{RSS}(\boldsymbol{w})}{n}) - \frac{n}{2},$$

is a monotone decreasing function of $\text{RSS}(\boldsymbol{w})$. A nonlinear optimization algorithm, e.g. the routine `nlm` in R (R Core Team, 2014), can be used to numerically minimize $\text{RSS}(\boldsymbol{w})$ over $\boldsymbol{w}$ lying on the unit sphere in $\mathbb{R}^3$. There is no mathematical guarantee that $RSS(\boldsymbol{w})$ possesses a unique local minimum. Hence it is important to choose a good starting point. For all the examples in the paper, convergence has never been as issue.

For the optimization in (b) it is helpful to rotate the initial estimate $\boldsymbol{w}_{\text{init}}$ to the north pole, $\boldsymbol{w}_0 = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^T$, and to represent $\boldsymbol{w}$ using an unconstrained coordinate system in $\mathbb{R}^2$, e.g. stereographic projection (p_1, p_2) where

$$(p_1, p_2)^T = (w_1, w_2)^T/(1 - w_3),$$

with inverse

$$\begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} = \begin{bmatrix} 2p_1 \\ 2p_2 \\ -1 + p_1^2 + p_2^2 \end{bmatrix} /(1 + p_1^2 + p_2^2).$$

As (p_1, p_2) ranges through $\mathbb{R}^2$, $\boldsymbol{w}$ covers the unit sphere minus the south pole, $\begin{bmatrix} 0 & 0 & -1 \end{bmatrix}^T$. In practice the minimizing value of $\boldsymbol{w}$ will usually be very close to the initial estimate $\boldsymbol{w}_0$.

The resulting MLEs for all the parameters can be termed the "Optimized least squares" (OptLS) estimates. An estimate of the error variance is

$$\hat{\sigma}^2 = \text{RSS}(\hat{\boldsymbol{w}})/(3n - 8),$$

where in the denominator we subtract 8 degrees of freedom from $3n$ since a regular helix model contains 8 regression parameters (6 for the linear regression model, plus 2 for the helix axis).

4 Initial Estimate of the Helix Axis

Several methods have been established in the literature to estimate a helix axis; see for example, Christopher et al. (1996) and Wilman (2014a). Here we limit discussion to two methods for getting initial estimates: Rotfit

(described in Christopher et al., 1996) and a new method based on modified least squares.

The principle behind Rotfit is easy to describe. Starting from an $n \times 3$ mathematical helix Y, let Y_{-1} denote Y without its first row, and Y_{-n} denote Y without its last row. Then Y_{-1} can be mapped onto Y_{-n} by a shift and rotation. Further the fixed axis of the rotation matrix is the desired axis $\boldsymbol{w}$.

When working with a statistical helix, the rotation matrix can be fitted using Procrustes analysis (e.g., Mardia et al., 1979 p. 416). Let $H_{n-1} = I_{n-1} - \frac{1}{n-1}\mathbf{1}\mathbf{1}^T$ denote the $(n-1) \times (n-1)$ centering matrix. Decompose the matrix $B = Y_{-1}^T H_{n-1} Y_{-n}$ using the singular value decomposition, $B = MLN^T$, where M and N are 3×3 orthogonal matrices and L is a diagonal matrix with positive entries. Then the Procrustes rotation matrix can be estimated by $R = MN^T$. Further, the fixed axis of R is the eigenvector with eigenvalue 1, and can be found by a spectral decomposition (the other two eigenvalues are complex).

The modified least squares method can be described as follows. Starting from Y, construct the *increments*

$$\boldsymbol{d}_i = \boldsymbol{y}_i - \frac{1}{2}\{\boldsymbol{y}_{i+1} + \boldsymbol{y}_{i-1}\}, \quad i = n_1 + 1, \ldots, n_2 - 1, \tag{4.1}$$

and combine them into an $(n-2) \times 3$ matrix D.

Set $E = D^T D$. In a mathematical helix, E will have a zero eigenvalue with eigenvector given by $\boldsymbol{w}$. In the statistical case, there will be one small eigenvalue, and two larger approximately equal eigenvalues.

Recall that if $\boldsymbol{w}$ is an eigenvector, so is $-\boldsymbol{w}$. Hence we need to specify its sign. That is, we need to choose the sign of $\boldsymbol{w}$ so that the fitted value of pitch parameter c in the helix model (2.1)–(2.2) will be positive. This task is straightforward when the level of noise σ^2 is not excessive. Just ensure the sign of $\boldsymbol{w}$ is chosen so that the difference between the endpoints, after projection onto the helix axis, is positive, $\boldsymbol{w}^T \boldsymbol{y}_n - \boldsymbol{w}^T \boldsymbol{y}_1 > 0$.

In practice, it does not matter which of these two initial estimates is used. However, simulations from the model (2.2) indicate that Rotfit is generally more accurate.

5 Assumptions About the Turn Angle β

The turn angle β is a key parameter in the statistical helix model, and there are several ways in which it can be treated.

Model A: the helix turn angle $\beta = 100°$ is known exactly. This is the choice made in this paper; see Section 2.

However, as emphasized by a referee, it is also of interest to consider what happens when this assumption is violated. There are two natural possibilities:

Model B: the turn angle β is constant within a helix, but is not known exactly. One course of action is to include it as one of the parameters to be estimated in the likelihood for a single helix, as in Mardia et al. (2018). Another is to carry out a sensitivity analysis to assess the effect on the conclusions of varying β. We carried out a small sensitivity analysis on simulated data and found that varying β in the range $98^o - 102^\circ$ had a negligible effect on the statistical analysis. This range has been chosen to be roughly $100^o \pm 2s$ where $s = 1.2^\circ$ is the standard error reported in Section 2 that was found by Mardia et al. (2018).

Model C: A more severe violation is to allow the turn angle to vary along the helix; that is, the incremental turn angle between x_j and x_{j+1}, denoted now by β_j, varies with j, $j = 1, \ldots, n-1$, where n is the number of landmarks in the helix. The simplest explicit model is to assume that the β_j are i.i.d. samples from $N(\beta_0, \tau_\beta^2)$, say, with $\beta_0 = 100^\circ$.

For illustration suppose $\tau_\beta = 4^\circ$. This choice is motivated from Web Figure 5(d) in the web supplement to Mardia et al. (2018), where the 129 straight helices have lengths ranging from 16 atoms to about 40 atoms with most helices having 20-25 atoms. Since the standard deviation of the 129 values for β is about 1°, and since each of these values stems from a data set with $n \geq 16$, one can conclude that the standard deviation of the distribution of the β_j is at least $\tau_\beta = 4^\circ$.

Next assume we have a helix consisting of $n = 2m + 1$ atoms, where the central atom is perfectly aligned with the model. The outermost atoms have cumulative turn angles

$$\alpha_1 = -\sum_{j=1}^{m} \beta_j \sim N(m\beta_0, m\tau_\beta^2), \quad \alpha_n = \sum_{j=m}^{n-1} \beta_j \sim N(m\beta_0, m\tau_\beta^2).$$

For example, this means that the outermost atoms in a helix consisting of $19 = 2 \times 9 + 1$ atoms will have turn angles which are distributed with a standard deviation of $\sqrt{9} \times 4^o = 12^\circ$ around the position assumed by a model with a constant turn angle. Hence errors of 10 - 15% in the cumulative turn angle can easily occur under Model C.

Model C was explored in the web supplement to Mardia et al. (2018), and was found to be unsupported by the data in the protein application. The view in biochemistry is that a constant incremental turn angle is more appropriate than Model C. In addition the external knowledge about β has a small enough standard error to justify using Model A instead of Model B.

6 The Change Point Model

We have already used the term *regular helix* model to describe the statistical helix in Eq. 2.2 in which all the data are modelled by a single helix. In this section we introduce and investigate a *change point helix* model in which there is a change point along the helix. That is, for some value of k, $n_1 < k < n_2$, we suppose that the points $\boldsymbol{y}_i$, $n_1 \le i \le k$ lie on one statistical helix, and the remaining points $\boldsymbol{y}_i, k+1 \le i \le n_2$ lie on another helix. All the regression parameters of the two helices are allowed to be different from one another, but they are assumed to have a common value of the error variance σ^2. Note that the change point model does not require the continuity of the curve consisting of the two helix pieces at the change point $t = k + \frac{1}{2}$.

Let m denote the smallest number of points along a helix we are willing to countenance for estimation purposes. Then the admissible values of k are

$$n_1 + m - 1 \le k \le n_2 - m.$$

Since a regular helix model contains 8 regression parameters, the smallest feasible choice of m is $m = 3$. However, it is reasonable to limit attention to larger values of m in order to avoid very short helices. Following Mardia et al. (2018), we take $m = 6$ in this paper.

6.1. Known change point index k Suppose the index k of a possible change point is known, and consider testing for the presence of a change point. The null hypothesis is H_0 : the data follow a regular helix model, and the alternative hypothesis is H_1 : the data follow a change point helix model.

Just as in the analysis of variance, the likelihood ratio test can be recast in terms of an F statistic. Let $RSS^{(0)}$ denote the optimized residual sum of squares under the null model after fitting a single helix to the whole data set by OptLS with $d^{(0)} = 3n - 8$ degrees of freedom. Similarly let $RSS^{(\ell)}(k)$ denote the optimized residual sum of squares after fitting a helix to the data points $i = n_1, \ldots, k$ for $\ell = 1$ and $i = k+1, \ldots, n_2$ for $\ell = 2$, respectively. Then the likelihood ratio test statistic can be recast as the F-statistic

$$F_k = \frac{SS_k^{(B)}/d^{(B)}}{SS_k^{(W)}/d^{(W)}} \sim F(d^{(B)}, d^{(W)}). \tag{6.1}$$

Here

$$SS_k^{(B)} = RSS^{(0)} - RSS^{(1)}(k) - RSS^{(2)}(k)$$

is the reduction in the residual sum of squares after fitting the alternative model with degrees of freedom $d^{(B)} = 8$, the number of extra estimated parameters under H_1. Note that $SS_k^{(B)}$ is analogous to the between-groups sum of squares in ANOVA. Similarly,

$$SS_k^{(W)} = RSS^{(1)}(k) + RSS^{(2)}(k)$$

represents the overall residual sum of squares under the alternative model with $d^{(W)} = 3n-16$ degrees of freedom, as we have estimated 8 helix parameters twice. Two estimates of σ^2 are given by the residual variance $\hat{\sigma}^2$ under the null hypothesis and pooled residual variance $\hat{\sigma}_p^2$ under the alternative hypothesis,

$$\hat{\sigma}^2 = \frac{SS_k^{(B)} + SS_k^{(W)}}{d^{(B)} + d^{(W)}}, \quad \hat{\sigma}_p^2 = \frac{SS_k^{(W)}}{d^{(W)}}. \tag{6.2}$$

If the error variance σ^2 is small, we expect that the dependence on the parameters in the helix model (2.2), and its change point counterpart, can be linearized through a Taylor series expansion, so that the model can be approximated by a standard linear model with normal errors. Hence from the standard ANOVA decomposition, we expect that when the change point position is known, the F_k-statistic will approximately follow the $F(d^{(B)}, d^{(W)})$ distribution, as suggested by the notation in Eq. 6.1. To test this expectation, a simulation study was carried out with $n = 30$, $r = 2.3$, c=$\frac{5.4}{2\pi}$ and $\sigma^2 = 0.05$. Then 10,000 helices were simulated from the null distribution and the statistic F_k was computed to look for a change point at $k = 15$. A Q-Q plot comparing the simulated F-statistics to the $F(8, 74)$ distribution is given in Fig. 1. Note that there is close agreement between the two distributions except in the upper tail, where the simulated test statistic has a shorter tail than the F-distribution.

6.2. Unknown change point index k In practice the location of the change point is generally unknown. The likelihood ratio statistic is now given by maximizing F_k over k,

$$F_{\max} = \max\{F_k : n_1 + m - 1 \le k \le n_2 - m\}, \tag{6.3}$$

where $m = 6$. Since k is no longer fixed, the distribution of $F_{\max}$ will be larger than the $F(d^{(B)}, d^{(W)})$ distribution. Hence standard statistical tables

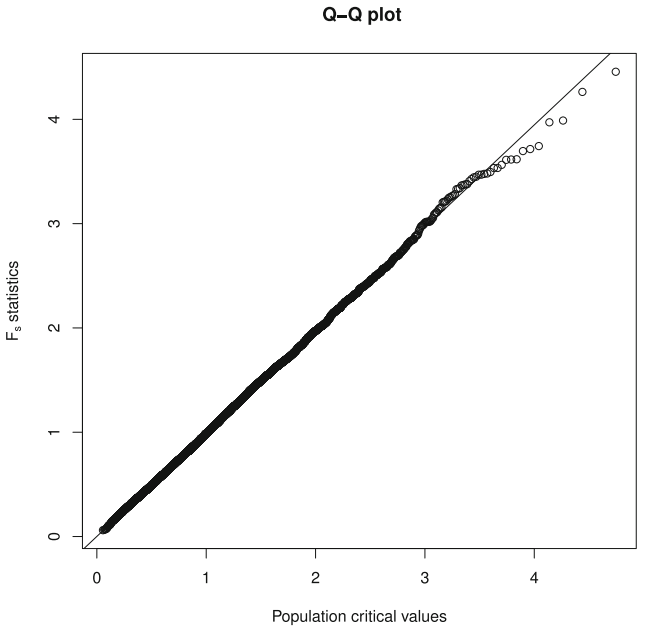

Figure 1: The Q-Q plot of 10,000 simulated F_k-statistics from Eq. 6.1 versus quantiles from the F-distribution for the example in Section 6.1

cannot be used for testing purposes. Instead the parametric bootstrap is used to assess significance.

The parametric bootstrap works as follows. Under the null hypothesis, estimate the regression parameters and σ^2. Using these values, simulate a large number n_{boot} of new helices with normally distributed error terms. For each simulation evaluate the value of $F^*_{\max}$, where * indicates a simulated bootstrap sample. Then the upper α tail of the simulated distribution, $F^{*(\alpha)}_{\max}$, say, gives the threshold of a statistical test of size α, e.g. with $\alpha = 0.05$.

If the null hypothesis is rejected, then we can try to pinpoint the ways in which the two sub-helices differ from one another. Let $\hat{k}$ denote the value of k maximizing (6.3). The following notation is useful. Let $\hat{\Gamma} = \left[\begin{array}{ccc} \hat{\boldsymbol{u}} & \hat{\boldsymbol{v}} & \hat{\boldsymbol{w}} \end{array} \right]$ denote the fitted orientation matrix under the null hypothesis, and let $\hat{\boldsymbol{g}}^{(\ell)}(t)$ and $\hat{\boldsymbol{f}}^{(\ell)}(t)$ denote the fitted axis and helix functions for each sub-helix. Then $\hat{\Gamma}^T \hat{\boldsymbol{g}}^{(\ell)}(t)$ and $\hat{\Gamma}^T \hat{\boldsymbol{f}}^{(\ell)}(t)$ denote the fitted axis and helix functions after rotation to canonical coordinates under the null hypothesis. Let

$$\boldsymbol{p}^{(\ell)} = (p_1^{(\ell)}, p_2^{(\ell)}, p_3^{(\ell)})^T = \hat{\Gamma}^T \hat{\boldsymbol{g}}^{(\ell)}(t_{\hat{k}+1/2})$$

$$\boldsymbol{q}^{(\ell)} = (q_1^{(\ell)}, q_2^{(\ell)}, q_3^{(\ell)})^T = \hat{\Gamma}^T \hat{\boldsymbol{f}}^{(\ell)}(t_{\hat{k}+1/2}), \quad \ell = 1, 2$$

denote the fitted axis position and fitted helix position, respectively, of a notional landmark at index $\hat{k} + \frac{1}{2}$ under each of the sub-helix models, after rotation to canonical coordinates under the null hypothesis.

The following six features can be constructed to compare fitted sub-helices. Note that they are not orthogonal to one another.

1. (Difference in helix axes.) Let

$$A_1 = 1 - \cos\hat{\theta}$$

where $\hat{\boldsymbol{w}}^{(1)T}\hat{\boldsymbol{w}}^{(2)} = \cos\hat{\theta}$ and where $0 \leq \hat{\theta} \leq \pi$ is the angle between the two sub-helix axes. The statistic A_1 contains the same information as $\hat{\theta}$ and measures the difference in helix axis for the two sub-helices, $\ell = 1, 2$. If the two helices point in the same direction then $\hat{\theta} = 0$ and $\cos\hat{\theta} = 1$ so that $A_1 = 0$. Further, A_1 increases as the directions get further apart. The angle $\hat{\theta}$ is the most important feature in practice when there is a change point.

2. (Shift parameter.) Let

$$A_2 = (p_3^{(1)} - p_3^{(2)})^2$$

denote the squared difference in the fitted axis positions at index $\hat{k}+\frac{1}{2}$, as measured along the fitted helix axis $\hat{\boldsymbol{w}}$ under the null model in canonical coordinates.

3. (Offset parameter.) Let

$$A_3 = (p_1^{(1)} - p_1^{(2)})^2 + (p_2^{(1)} - p_2^{(2)})^2$$

denote the squared Euclidean distance between the fitted axis positions at index $\hat{k} + \frac{1}{2}$, after projection onto the $(\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}})$-plane under the null model, which is a horizontal plane in canonical coordinates.

4. (Spin parameter). Let

$$A_4 = |\hat{\phi}|$$

where $\hat{\phi} = \text{atan2}(q_2^{(1)} - p_2^{(1)}, q_1^{(1)} - p_1^{(1)}) - \text{atan2}(q_2^{(2)} - p_2^{(2)}, q_1^{(2)} - p_1^{(2)})$, denotes the angle, treated as a number in $[-\pi, \pi)$, between the two fitted helix angles, after projection onto the $(\hat{\boldsymbol{u}} - \hat{\boldsymbol{v}})$-plane under the null model. This feature measures spin at the change point between the two sub-helices.

5. (Change in radius.) Let

$$A_5 = |\hat{r}^{(1)} - \hat{r}^{(2)}|.$$

6. (Change in pitch.) Let

$$A_6 = |\hat{c}^{(1)} - \hat{c}^{(2)}|.$$

Thus we have six features: difference in helix axes, shift parameter, offset parameter, spin parameter, change in radius, and change in pitch. In each case critical values for these features can be simulated using the parametric bootstrap. For all quantities $(A_1, A_2, A_3, A_4, A_5, A_6)$, an upper critical value is used. In each bootstrap sample, the values of $A_1, \ldots, A_6$ are computed using the bootstrap value of $k = k^*$ maximizing the bootstrap version of the F-statistic given by Eq. 6.3.

A simple simulation study was carried out to see how the distribution of $F_{\max}$ depends on σ^2 and n using parameters typical for protein α- helices. On fixing $n = 30$ and varying σ^2, with $n_{\text{boot}} = 1000$, we found that the bootstrap threshold values for $F^*_{\max}$, in Table 1, are very similar for all σ^2, which implies the distribution of the F statistic does not depend heavily on σ^2. On the other hand, if we fix $\sigma^2 = 0.05$ and alter n, then the threshold value does change somewhat; that is, the distribution of the $F_{\max}$ statistic does depend on the number of landmarks, just as the usual F distribution does.

For each data helix we can compare the computed F statistic with the threshold, $F^{*(\alpha)}_{\max}$. If $F < F^{*(\alpha)}_{\max}$, we conclude that there is no evidence our helix has a change point; otherwise, if $F \geq F^{*(\alpha)}_{\max}$, the point with the largest F statistic corresponds to the estimated index of the change point. We give our procedure for estimation and testing the name *ChangePoint-Detector*.

Table 1: Threshold $F^{*(0.05)}_{\max}$ with $n_{\text{boot}} = 1000$ simulations for various choices for n, the number of landmarks, and σ^2, the error variance

n	σ^2	$F^{*(0.05)}_{\max}$
15	0.05	2.877
25	0.05	3.073
30	0.05	3.008
30	0.01	3.008
30	0.1	3.015

7 Simulation Study for ChangePoint-Detector

In this section we illustrate the behavior of ChangePoint-Detector on two simulated helices, a regular helix and a change point helix.

- *Example 1.* A regular helix was simulated with $n = 27$ landmarks and parameters that mimic a protein α-helix (i.e. $r = 2.3,\ 2\pi c = 5.4,\ \beta = \frac{2\pi}{3.6}$ and $\sigma^2 = 0.05$).
- *Example 2.* A change point helix with change point $k = 12$, was constructed by simulating a regular helix in the same way as Example 1, but then introducing a rotation of the helix axis by an angle $\theta = 0.3$ radians $= 17.1°$, centered at time index $k + \frac{1}{2} = 12.5$.

Table 2 presents the results for the key parameters, the overall test statistic and the six features of interest. The p-values in parentheses are constructed using bootstrap sampling with $n_{\text{boot}} = 1000$.

First consider Example 1, the regular helix. As expected, there is no reason to reject the null hypothesis. The test statistic $F_{\max}$ is not significant, the estimates of σ^2 under the null hypothesis $\hat{\sigma}^2 = 0.039$ and under the alternative hypothesis $\sigma_p^2 = 0.039$, from Eq. 6.2, are approximately equal, and none of the features $A_1, \ldots, A_6$ are significant.

Next consider Example 2, the change point helix, for which the test statistic $F_{\max}$ is highly significant. For this example, ChangePoint-Detector

Table 2: The ChangePoint-Detector estimates of $\hat{\sigma}^2$, $\hat{\sigma}^2_{\text{p}}$, $\hat{k}$, $\hat{\theta}$, the statistics $F_{\max}, A_1, \ldots, A_6$, and the bootstrap p-values for Examples 1 and 2, the regular and the change point simulated helices

Helix	Regular Example 1	Change point Example 2
$\hat{\sigma}^2$	0.039	0.223
$\hat{\sigma}^2_{\text{p}}$	0.039	0.045
$F_{\max}$	0.924 (0.973)	26.3 **
$\hat{k}$	13	12
$\hat{\theta}$	2.5°	18.4°
A_1	1e-4 (0.765)	0.051 **
A_2	0.134 (0.455)	0.146 (0.447)
A_3	0.004 (0.981)	0.012 (0.954)
A_4	0.002 (0.965)	0.048 (0.392)
A_5	0.046 (0.666)	0.183 (0.125)
A_6	0.002 (0.919)	0.010 (0.724)

** indicates p-value < 0.001

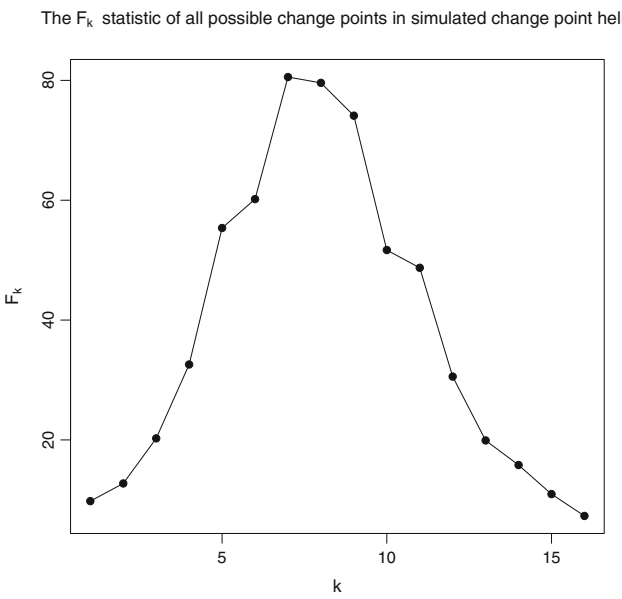

Figure 2: The F_k statistic against the possible choice of k for the simulated change point helix, where the maximum of F_k is at $k = 12$

estimates the change point correctly at $\hat{k} = 12$ as shown in Fig. 2. Table 3 shows the bootstrap distribution of k^* for $n_{\text{boot}} = 1000$ simulations to help judge the accuracy of $\hat{k}$. In particular, k^* equals the estimate $\hat{k} = 12$ most of the time, but occasionally overshoots by 1 or 2.

The estimated angle between the two sub-helices $\hat{\theta} = 0.32$ radians is close to the true value. Further the p-value for A_1 confirms that the change in angle is significantly different from 0, but none of the other features $(A_2, A_3, A_4, A_5, A_6)$ is significant.

Figure 3 shows the fitted helices. Panel (a) shows a fitted single helix for Example 1. Panel (b) shows the two fitted sub-helices for Example 2 after estimating the change point. The change in helix axis of size $\hat{\theta} = 18.4^\circ = 0.32$ radians at $k = 12$ is visible.

Table 3: The frequency table of k^* from 1000 bootstrap samples for Example 2

k	12	13	14
frequency	728	220	52

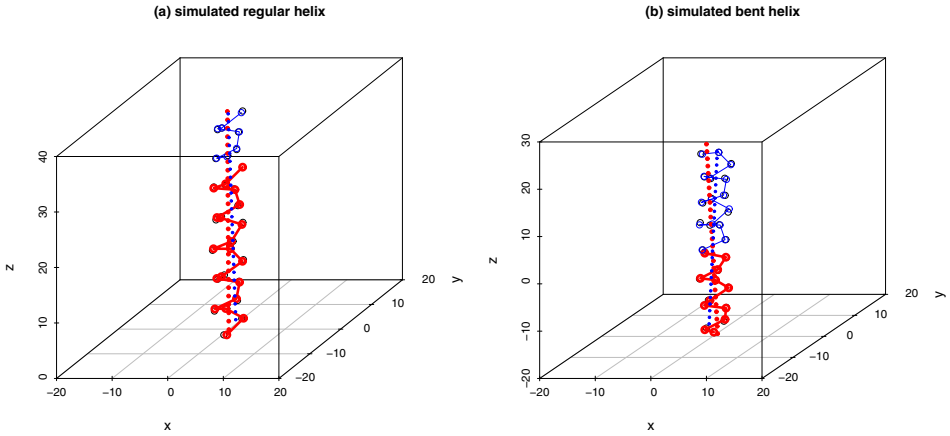

Figure 3: The two sub-helices found by ChangePoint-Detector for the simulated regular helix (Example 1) and the simulated change point helix (Example 2). In each case the two sub-helices are plotted using a thin line and a thicker line, respectively. Also plotted are the two fitted axes, as a dotted line with small dots and a dotted line with larger dots, respectively

8 Data Examples

In this section we look at nine α-helix protein structures from Mardia et al. (2018). Each helix comprises a sequence of C_α atoms; there are 3.6 equally-spaced C_α atoms per turn of the helix, i.e. one C_α atom every $\beta = \frac{2\pi}{3.6}$ radians (100 degrees). For the i^{th} landmark (C_α atom) we associate the time index $t_i = i\beta,\ i = n_1, \ldots, n_2$, where all the helices start at $n_1 = 1$ and and for each helix $n_2 = n$ denotes the total number of the landmarks.

Three other parameters in the helix model in the protein setting have ideal values: $2\pi c = 5.4$Å (the vertical distance of one complete turn), radius $r = 2.3$Å, and the variance $\sigma^2 = 0.056$Å^2 (e.g., Mardia et al., 2018). One Ångström (Å) equals 10^{-10} m. For the analysis in this paper, we treat β as known, but include r, c and σ^2 as unknown parameters in our estimation and testing procedures, so that the data can "speak for themselves" as far as possible.

In the study of protein helices, it is of particular interest to identify and locate *kinks*. Typically the number of kinks in protein helices is either 0 or 1 (Mardia et al., 2018). Mardia et al. (2018) developed an algorithm called *Kink-Detector* to identify the presence of a kink and to estimate its location. The Kink-Detector method is based on a local moving window (6 landmarks each side of the possible kink).

Following Mardia et al. (2018), we used their nine test helices here, labelled Helix 1 – Helix 9. Mardia et al. (2018) chose these helices partly because identification of any kinks was particularly challenging. Note that the conclusions from the Kink-Detector methodology were confirmed in a crowdsourcing assessment (Wilman et al., 2014b) by experts in the protein community.

Here we want to see if the kinks found by Mardia et al. (2018) coincide with any change points found by ChangePoint-Detector, or if the presence of a kink is unassociated with larger-scale structure in the helix.

The ChangePoint-Detector methodology was applied to each of the nine helices. The regular helix model H_0 consisting of a single helix and the change point helix model H_1 with one change point were fitted to the data. The $F_{\max}$ statistic of Section 6.1 was computed to test H_0 vs. H_1. Further the statistics A_1 to A_6 were computed to compare the two sub-helices under H_1. The results are summarized in Tables 4–5. These tables show for each helix: (a) the error variance estimates $\hat{\sigma}^2$ and $\hat{\sigma}_p^2$ from Eq. 6.2 (b) the overall $F_{\max}$ statistic; (c) the optimizing index $\hat{k}$ for the change point; (d) the angle $\hat{\theta}$ in degrees between the two sub-helix axes; and (e) the features $A_1, \ldots, A_6$; and (f) the p-values for $F_{\max}$ and $A_1, \ldots, A_6$ based on bootstrap sampling with $n_{\text{boot}} = 1000$.

Table 4: Test statistics and estimates from ChangePoint-Detector for Helices $1, \ldots, 4$: the estimates of variance $\hat{\sigma}^2$, pooled variance estimate $\hat{\sigma}^2_{\text{p}}$, the position $\hat{k}$, the angle $\hat{\theta}$ between the two sub-helices and the test statistics (and p-values) of $F_{\max}, A_1, \ldots, A_6$ for each helix

Helix	1	2	3	4
n	31	24	24	17
$\hat{\sigma}^2$	0.318	0.836	0.195	0.179
$\hat{\sigma}^2_{\text{p}}$	0.083	0.144	0.060	0.034
$F_{\max}$	30.9 **	39.5 **	18.8 **	24.0 **
$\hat{k}$	14	7	9	10
$\hat{\theta}$	10.7°	25.6°	9.2°	8.8°
A_1	0.017 **	0.098 **	0.013 **	0.012 **
A_2	0.005 (0.889)	0.454 (0.443)	0.076 (0.615)	0.012 (0.927)
A_3	1.983 **	1.518 (0.014)	0.343 (0.154)	0.880 **
A_4	0.352 **	5.039 (0.014)	0.436 (0.008)	0.352 (0.002)
A_5	0.014 (0.930)	0.001 (0.998)	0.012 (0.917)	0.038 (0.737)
A_6	0.007 (0.803)	0.008 (0.847)	0.010 (0.730)	0.039 (0.141)

** indicates p-value < 0.001

Table 5: Test statistics and estimates from ChangePoint-Detector for Helices $5, \ldots, 9$

Helix	5	6	7	8	9
n	24	23	19	15	27
$\hat{\sigma}^2$	0.200	0.108	0.122	0.061	1.684
$\hat{\sigma}^2_{\mathrm{p}}$	0.065	0.062	0.045	0.014	0.102
$F_{\max}$	17.6 **	6.7 (0.002)	11.5 **	16.7 **	142.9 **
$\hat{k}$	10	12	11	8	11
$\hat{\theta}°$	6.6°	5°	12.6°	9.6°	31.9°
A_1	0.007 (0.008)	0.004 (0.049)	0.024 **	0.014 **	0.151 **
A_2	0.191 (0.710)	1e-4 (0.988)	0.012 (0.975)	0.003 (0.978)	0.005 (0.946)
A_3	4.935 **	1.206 (0.002)	0.116 (0.572)	0.026 (0.597)	3.617 **
A_4	0.034 (0.599)	0.041 (0.519)	0.054 (0.342)	0.130 **	1.597 (0.013)
A_5	0.140 (0.322)	0.010 (0.964)	0.099 (0.410)	0.043 (0.566)	0.041 (0.813)
A_6	0.034 (0.267)	0.020 (0.514)	0.054 (0.052)	0.042 (0.025)	0.028 (0.409)

** indicates p-value < 0.001

From Tables 4–5 we see that all nine helices have a highly significant value of $F_{\max}$ (with p-value in each case less than 0.001), meaning that all helices are deemed to have a change point. Further, in general the change point seems to be due to a change in axis direction. For the most of the helices, the feature A_1 is significantly different from 0 with a p-value less than 0.001. Two mild exceptions are Helices 5 and 6 with p-values 0.008 and 0.049.

Some further evidence in support of the change point helix model is given by the residual variances. In most cases $\hat{\sigma}_p^2$ is close to the theoretical value of $\sigma^2 = 0.056$. The main exceptions are Helix 2 with $\hat{\sigma}_p^2 = 0.144$ too large and Helix 8 with $\hat{\sigma}_p^2 = 0.014$ too small. Figure 4 presents Helix 8 and the fitted helices using ChangePoint-Detector. The axis has a change point with $\hat{\theta} = 9.6°$ as the first sub-helix axis direction (bottom) is different than the second sub-helix axis direction (top). Figure 5 shows that the maximum F_k occurs at $k = 8$.

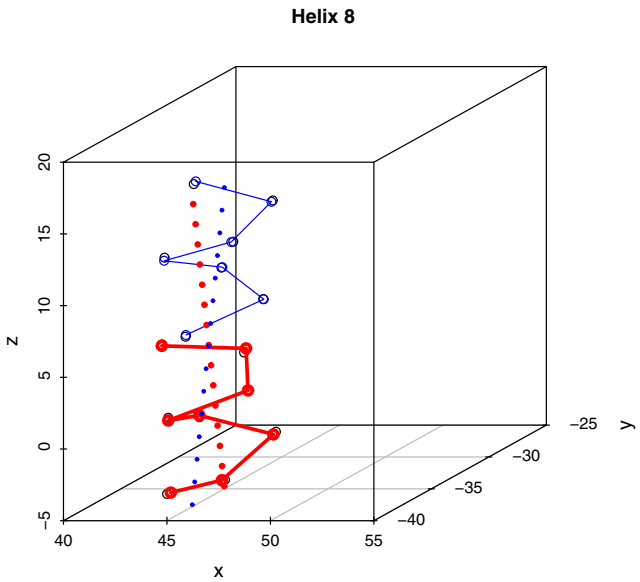

Figure 4: The two sub-helices found by ChangePoint-Detector for Helix 8. The two sub-helices are plotted using a thin line and a thicker line, respectively. Also plotted are the two fitted axes, as a dotted line with small dots and a dotted line with larger dots

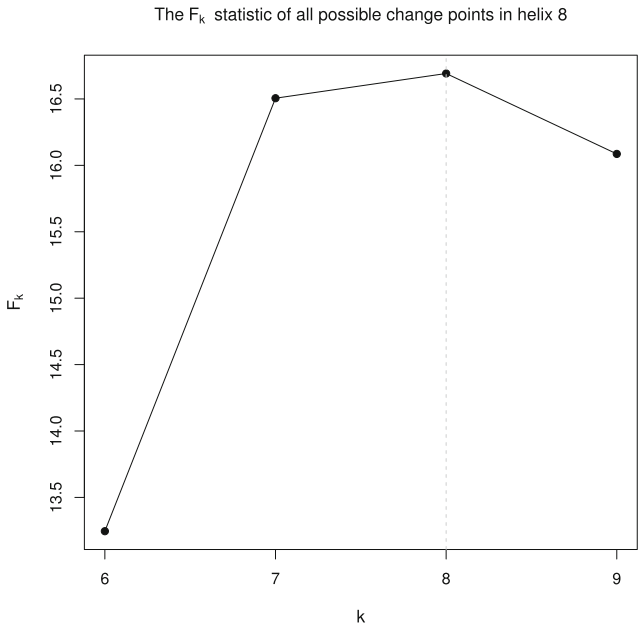

Figure 5: The plot of F_k statistic against the possible choice of k for helix 8; the maximum of F_k is at $k = 8$

It is also of interest to look at the other features. Features A_2 (shift), and A_5 (radius) never show any signs of significance. However, features A_3 (offset) and A_4 (spin) are occasionally very significant. Feature A_6 (pitch) is significant only in Helix 8. The reasons are unclear.

Next we compare the results of ChangePoint-Detector with the results of Kink-Detector (Table 6). It can be seen that there is little, if any, correspondence between the two methods. ChangePoint-Detector finds that all the helices have a change point; Kink-Detector finds only 6 out of 9 helices to be kinked. Further, when both methods do find a change point, the index k and the angle θ do not match.

This analysis leads to two general conclusions. First, Kink-Detector and ChangePoint-Detector find very different features in the data; the reason seems to be that Kink-Detector looks for local change whereas ChangePoint-Detector looks for global change. Further, although the change point model is certainly an over-simplification of the structure in the data, it has still managed to capture succinctly much of the variability for most of the helices in this study.

Table 6: The kink position $\hat{k}$, the angle between the two sub-helices $\hat{\theta}$ in degrees, and the classification by Kink-Detector (k=kinked, s= straight) and the classification by ChangePoint-Detector (c=change point, r= regular)

Helix	Kink-Detector			Change Point-Detector			
	$\hat{k}$	$\hat{\theta}°$	classification	$\hat{k}$	$\hat{\theta}°$	classification	$F_{\max}$
1	–	–	s	14	10.7°	c	30.9
2	–	–	s	7	25.6°	c	39.5
3	13	18.7°	k	9	9.2°	c	18.8
4	7	15.9°	k	10	8.8°	c	24.0
5	7	22.8°	k	10	6.6°	c	17.6
6	10	20.4°	k	12	5.0°	c	6.7
7	13	20.0°	k	11	12.6°	c	11.5
8	–	–	s	8	9.6°	c	16.7
9	9	30.5°	k	11	31.9°	c	142.9

9 Discussion

We have developed a ChangePoint-Detector procedure which tests if the helix has a change point; if the helix is deemed to have a change point, it estimates the location of the change point; and then fits the two sub-helices. In addition the method looks at six features at the change point, to investigate the reason for the change point.

In this paper, the errors are assumed to be independent and identically normally distributed with mean 0 and variance σ^2. In the simulation study we assumed $\sigma^2 = 0.05$, although ChangePoint-Detector performs well even for larger σ^2 (at least up to $\sigma^2 = 1$). In future work it would be interesting to incorporate dependence between the landmarks to model the effect of hydrogen bonds.

Another issue worthy of further study is the turn angle β. Although (Mardia et al., 2018) found no evidence of a varying turn angle in their work, it would be interesting to investigate this possibility in more detail, with a separate turn angle parameter for each successive pair of landmarks. Such a model could be investigated through likelihood methods (though fitting such a model faces challenges similar to the Neyman-Scott problem since the number of parameters increases with the number of landmarks) or through Bayesian hierarchical models.

Acknowledgments. We are grateful to Professor Charlotte Deane, University of Oxford, for providing the data and helpful discussions. We also wish to thank the referees for their helpful comments.

References

BANSAL, M., KUMAR, S. and VELAVAN, R. (2000). Helanal: a program to characterize helix geometry in proteins. *Journal of Biomolecular Structure and Dynamics* **17**, 5, 811–819.

BARLOW, D.J. and THORNTON, J.M. (1988). Helix geometry in proteins. *Journal of Molecular Biology* **201**, 601–619.

BLUNDELL, T., BARLOW, D., BORKAKOTI, N. and THORNTON, J. (1983). Solvent-induced distortions and the curvature of $\alpha -$ helices. *Nature* **306**, 281–283.

CAMPBELL, M.K. and FARRELL, S.O. (2014). *Biochemistry*, 6th edn. Lippincott Williams and Wilkins, Philadelphia.

CHRISTOPHER, J.A., SWANSON, R. and BALDWIN, T.O. (1996). Algorithms for finding the axis of a helix: fast rotational and parametric least-squares methods. *Computers & Chemistry* **20**, 3, 339–345.

DICKERSON, R.E. and GEIS, I. (1969). *The Structure and Action of Proteins*. W.A. Benjamin, California.

DRYDEN, I.L. and MARDIA, K.V. (2016). *Statistical Shape Analysis: With Applications in R.*, 2nd edn. Wiley, New York.

ENKHBAYAR, P., DAMDINSUREN, S., OSAKI, M. and MATSUSHIMA, N. (2008). Helfit: Helix fitting by a total least squares method. *Computational Biology and Chemistry* **32**, 4, 307–310.

MARDIA, K.V., KENT, J.T. and BIBBY, J. (1979). *Multivariate Analysis*. Academic Press, London.

MARDIA, K.V., MORRIS, R.J., WALDER, A.N. and KOENDERINK, J.J. (1999). Estimation of torsion. *Journal of Applied Statistics* **26**, 373–381.

MARDIA, K.V. (2013). Statistical approaches to three key challenges in protein structural bioinformatics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **62**, 3, 487–514.

MARDIA, K.V. (2014). *In-depth modelling of some angular shapes in proteins with applications: modelling conics and helices*. Presentation at ADISTA, Brussels.

MARDIA, K.V., SRIRAM, K. and DEANE, C.M. (2018). A statistical model for helices with applications. *Biometrics* **74**, 3, 845–854.

R CORE TEAM (2014). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna.

WILMAN, H.R. (2014a). *Computational Studies of Protein Helix Kinks*. PhD thesis, University of Oxford.

WILMAN, H.R., EBEJER, J.P., SHI, J.Y., DEANE, C.M. and KNAPP, B. (2014b). Crowdsourcing yields a new standard for kinks in protein helices. *Journal of Chemical Information and Modeling* **54**, 9, 2585–2593.

Mai F. Alfahad
John T. Kent
Kanti V. Mardia
Department of Statistics,
University of Leeds,
Leeds, LS2 9JT, UK
E-mail: mmmfa@leeds.ac.uk
J.T.Kent@leeds.ac.uk
K.V.Mardia@leeds.ac.uk

Kanti V. Mardia
Department of Statistics,
University of Oxford,
Oxford, OX1 3LB, UK

Paper received: 31 December 2017.