Original Paper — *Forma*, **23**, 9–18, 2008

Protrusion Fourier Descriptor: Skeleton-based Representation of Open Curves

Eiki Tanaka[1]*, Yoshiyasu Tamura[2], Masaki Hosoya[3] and Toshihiko Shiroishi[4]

[1]*Department of Statistical Science, The Graduate University for Advanced Studies, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569, Japan*
[2]*The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569, Japan*
[3]*Department of Genetics, The Graduate University for Advanced Studies, Yata 1111, Mishima, Shizuoka 411-8540, Japan*
[4]*Mammalian Genetics Laboratory, National Institute of Genetics, Yata 1111, Mishima, Shizuoka 411-8540, Japan*
**E-mail address: etanaka@ism.ac.jp*

(Received July 14, 2008; Accepted October 17, 2008)

If only some part of an object outline is of the matter of interest in statistical shape analysis, an appropriate open-curve descriptor is needed and *tangent Fourier descriptor* (TFD, also called *P-type Fourier descriptor*) is one such example. However, the TFD amplifies high frequency noise. In this paper we propose *protrusion Fourier descriptor* (PFD), an open-curve descriptor utilizing the skeletal information for an open curve, which is invariant under translation and rotation. Using regularized logistic regression model and generalized information criterion, we compare the PFD to the TFD in terms of capability to capture subtle variability of irregular shapes. The experiments with open curves extracted from nine inbred strains of mouse mandibular outlines have shown that different strains of data separate more clearly using the PFD than when using the TFD, and the PFD reflects inter-strain variability better.

Key words: Open Curve, Fourier Descriptor, Skeleton, Logistic Regression, Generalized Information Criterion

1. Introduction

The numerical description of shapes is an important task in the fields of statistical analysis of shapes and also shape recognition. During recent decades, a number of approaches have been presented to characterize shapes which are mostly represented as closed curves or regions. However, sometimes we need to describe open curves not closed curves. Here we think of the case that we need to analyze open curves, that is, the case that we are interested in only some part of an object outline. When we analyze shapes of plant organs, if the shapes contain artificial cutoff lines and worm-eaten defects, we need to analyze homologous contour segments in such a way that the influences of such artifacts and defects are taken off. For statistical analysis using methods of multivariate analysis, we need to numerically characterize open curves, and in this paper, we focus on numerical description of open curves.

Because there is no interior and exterior for open curves, we can not use region-based shape descriptions. Landmark-based descriptions have been commonly used in biological morphometrics for statistical analysis, however these methods have some problems which stem from ignoring information from the parts that are not selected as landmarks. When the sample shapes are very complicated and irregular and we can not easily see where the underlying essential feature is, we may lose useful information using the landmark-based method. Moreover, the method cannot be utilized when we can not set reliable homologous landmarks which are commonly found in all samples.

However, most contour-based methods capture whole contour information of irregular shapes such as those ordinarily seen in the forms of living organisms, and contour-based Fourier descriptor has been commonly used in biological morphometrics. In most Fourier descriptor methods, given a curve, after representing some geometric information associated with each point on the curve as a numerical value, the shape descriptor is defined as the Fourier transform of the sequence of the values ordered from the starting point to the ending point. The Fourier descriptor represents the information of the whole shape in frequency space.

In statistical analysis of closed curves, *complex Fourier descriptor* (Granlund, 1972) and *elliptic Fourier descriptor* (Kuhl and Giardina, 1982) are commonly used to describe contour shapes. Since Rohlf and Archie (1984) showed that, combining these descriptors with principal component analysis, one can evaluate the feature of a shape as principal component scores, these descriptors have been used in statistical shape analysis. To apply these closed-curve descriptors to open curves, there are several examples of making a closed curve from an open curve, connecting the endpoints of the original open curve with those of its half-turned curve (Kawamura and Yokota, 2005), or with those of the curve symmetric with respect to the line that goes through the two endpoints of the original open curve (Lestrel *et al.*, 2004). These techniques do not always work well. In the case that the shape of neighborhood of connecting points greatly affects the description of a curve, the descriptor may fail to capture the subtle feature of the shape.

An example of Fourier descriptor which can be used to characterize open curve as it is, is *tangent Fourier descriptor* (TFD, aka *P-type Fourier descriptor*) (Uesaka, 1984). The TFD has been applied to author attribution of ukiyoe

*The experiments in this research are approved by the Animal Experiment Committee of the National Institute of Genetics (NIG), and the care of the animals used in the experiments comply with the guidelines of the NIG.

works (Yamada and Hayakawa, 1997), recognition of human face profiles (Aibara *et al.*, 1991) and statistical shape analysis of lotus petal tip (Zheng and Tamura, 2005) and rice leaf (Zheng *et al.*, 2008), which showed that the descriptor can be used to accomplish those tasks. However, the TFD is defined as a Fourier transform of a sequence of differences between adjacent points, which means that high frequency variability and also noise are amplified, so fine and subtle features represented by relatively high frequency components may be lost in the noise. So the TFD method may lose useful information when the subtle and fine structure of a shape has an essential meaning.

The rest of this paper is organized as follows. Sec. 2 describes the TFD method. In Sec. 3, we propose a new Fourier-based open-curve descriptor, called *protrusion Fourier descriptor* (PFD), which utilizes an approximation of skeletal information at infinite resolution. Sec. 4 is the experimental part of this paper. In the experiments, with open curves extracted from mouse mandibular outlines of nine inbred strains, we compare the PFD to the TFD with respect to the ability to represent subtle feature of irregular shapes. Finally Sec. 5 concludes this study.

2. Tangent Fourier Descriptor

In this section we describe how to calculate the TFD. Suppose a given open curve is represented as the ordered sequence of pixel coordinates which are tracing the curve in an image. First of all, resampling from the sequence, we have to represent the curve as the set of $N+1$ equally-spaced points which are ordered from the starting point to the ending point, i.e.

$$\Big\{(x_0, y_0), \ldots, (x_N, y_N) | \sqrt{(x_{n+1}-x_n)^2 + (y_{n+1}-y_n)^2} = \delta \ \ \forall n \in \{0, \ldots, N-1\}\Big\},$$

where δ is some constant. Then we define complex functions w_n as

$$w_n = \frac{x_{n+1}-x_n}{\delta} + \sqrt{-1}\frac{y_{n+1}-y_n}{\delta} \ \ \forall n \in \{0, \ldots, N-1\}.$$

Lastly, we apply discrete Fourier transform to the sequence of $\{w_0, \ldots, w_{N-1}\}$

$$c_k = \frac{1}{N}\sum_{n=0}^{N-1} w_n \exp\left(-\sqrt{-1}\frac{2\pi kn}{N}\right) \ \ \forall k \in \{0, \ldots, N-1\}.$$

The set of these complex numbers $\{c_0, \ldots, c_{N-1}\}$ is the TFD and the set of $2N$ real numbers $\{\Re(c_0), \Im(c_0), \ldots, \Re(c_{N-1}), \Im(c_{N-1})\}$ is used as the feature vector of each shape in statistical analysis.

3. Protrusion Fourier Descriptor

We propose a new Fourier-based open-curve descriptor, *protrusion Fourier descriptor* (PFD). The detailed algorithm for computing the PFD and the method of inverse transform from the PFD to the corresponding curve shape is described in the appendix. Here in this section we only illustrate the geometric meaning and the general computing procedure of the PFD. Generally speaking, in the Fourier descriptor methods—which have an advantage of reflecting whole information of a shape—given a curve, first we compute numerical values so that each value represents *some kind of geometric information* associated with each point on the curve, and then define the shape descriptor as the Fourier transform of the sequence of the numerical values ordered from the starting point to the ending point. In our method, we use a degree of *protrusion* or *sticking out* at each point on the given open curve as the geometric information associated with each point on the curve. In other words, we compute numerical values in such a way that each value represent the degree of rightward or leftward protrusion (with respect to the direction from the starting point to the ending point) at each point on the curve. We then define the PFD as the Fourier transform of the ordered sequence of the numbers.

Now, we illustrate how to calculate the degree of leftward and rightward protrusions. See Fig. 1 for more detailed illustration. Figure 1(a) shows an open curve (the left endpoint is the starting point), The collections of maximal disks and skeletons for the open curve are seen in Figs. 1(b) and (c) respectively, and Fig. 1(d) shows the collection consists of the skeletons and the perpendiculars drawn from points on the skeletons to all points on the original open curve. We choose to call such graphs *skeleton-perpendicular graphs*.

Now, we should make clear the definition of maximal disks in the case of open curves. In the case of closed curves or regions, intuitive definition of maximal disks (MDs) is as follows: Any circle which is entirely within the object boundary and touches the boundary from the inside at more than one point. However, in the case of open curves, there is no interior or exterior for the curve, and in order to define the new open-curve descriptor, we should be able to distinguish whether a circle touches the curve from the left or the right (with respect to the direction from the starting point to the ending point). Here, we slightly modify the definition of MD to be applicable to the case of open curves as follows: *a left/right MD associated with a point on the curve is the maximal tangent circle that touches the curve at the point from the left/right but never intersects with the curve*. That is, there is a pair of left and right MDs associated with each point on the curve. In Fig. 1(b), the left MDs are blue-colored, and the right MDs are red. The word skeleton usually refers to a curve so that an arbitrary point on the curve is the center of a MD for the original open curve, that is, the skeleton consists of the centers of the MDs for the original open curve (Blum, 1973). Here in the case of open curves, we call the skeleton made of the centers of left MDs *left skeleton*, the other one is called *right skeleton*. In Fig. 1(c), the left skeletons are colored blue, and the right ones red.

Now, in order to define the PFD we use the skeleton-perpendicular graphs (Fig. 1(d)). The perpendicular drawn from a point on a skeleton to a point on the original open curve can be approximately represented by the line segment between the point on the original curve and its associated MD's center. So we can easily compute the skeleton-perpendicular graphs, and we call the skeleton-perpendicular graphs just *skeletons* from now on.

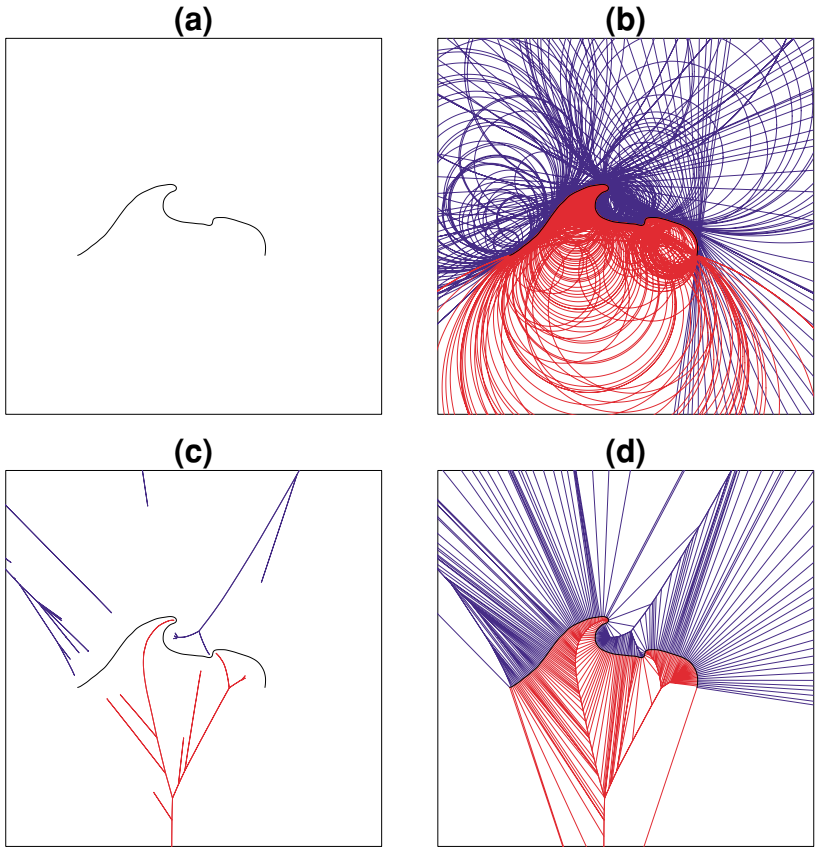

Fig. 1. Skeletonization of an open curve. (a) An open curve, (b) maximal disks, (c) skeletons, and (d) the skeleton-parpendicular graphs.

All left or right skeletons are not necessarily singly connected and each left or right skeleton has only one branch that goes to infinity. The branch is cut off at the point which is the center of a MD with a sufficiently large radius ρ, and the remaining skeleton becomes a tree graph of which the root is the center of a MD with a radius ρ and all its leafs are on the original open curve. For each point on the original open curve, we first compute the distance between the left root and the point on the curve through the left skeleton, and next, subtract ρ from the distance, then we call the obtained value *left protrusion*, which means rightward protrusion from the left at the point on the curve. *Right protrusion* is similarly defined. All protrusions converges to certain values respectively at the limit of $\rho \to \infty$, and can be assumed to be independent of ρ for sufficiently large ρ. Then a left/right protrusion is supposed to be 0 at a point on the original curve directly touched by a left/right MD with a radius ρ.

Suppose that given open curve is represented as the set of N equally-spaced points which are ordered from the starting point to the ending point in advance of computing the PFD, and the total length of the curve is normalized to 1. Figure 2(a) shows the sequences of left and right protrusions at points on the same open curve shown in Fig. 1(a) (the curve is represented as the ordered sequence of 200 equally-spaced points in advance, whereas it was represented as 201-point sequence in the case of TFD). Compared with the sequences of differences between adjacent points on the same open curve used in the TFD method (Fig. 2(b)), the sequences of protrusions at points on the curve which are to used in the PFD appear much smoother. Finally we define the PFD as the Fourier transform of the sequence of the complex numbers which real and imaginary components are left and right protrusions respectively.

4. Experiment

We implemented the experiment in order to compare the PFD to the TFD with respect to the abilities to represent subtle variability of irregular open curves. The data used in this experiment are open curves extracted from nine inbred strains of mouse mandible outlines. In the cases that the curve shapes are represented by TFD and PFD, we did the strain classification experiments using the same classification model. Comparing the classification accuracies in the two cases, we can see which descriptor represents inter-strain variability better. The logistic regression model returns the estimate of probability that a datum belongs to each class, the generalization ability (classification ability on unseen data) of the model can be easily evaluated by using the generalized information criterion (GIC) (Konishi and Kitagawa, 1996). That is, we do not have to do cross-validation. So we used the logistic regression model and the GIC. In this section, we illustrate the general procedure of the experiment and the detailed mathematical descriptions of logistic regression model and the GIC are in the appendix. We devised the PFD for the use in statistical analysis of irregular shapes such as usually seen in forms of living organisms, and also envisage its use in biological morpho-

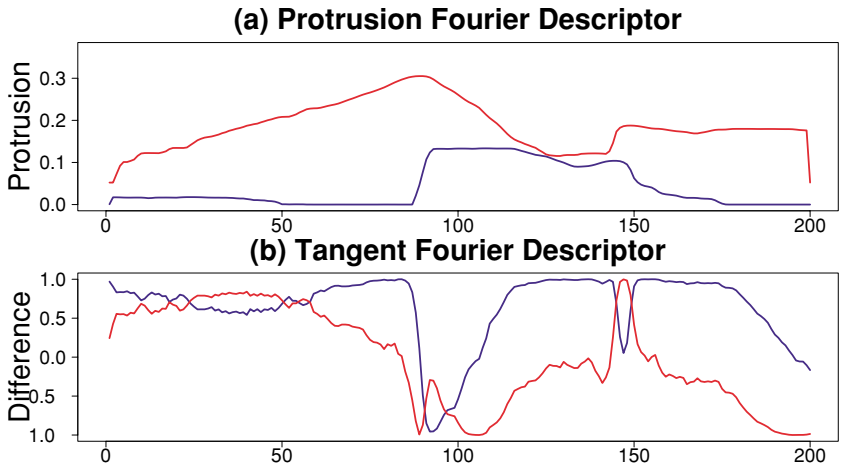

Fig. 2. Sequences to be transformed in the PFD and TFD methods (above: protrusions used in PFD, below: differences between adjacent points used in TFD).

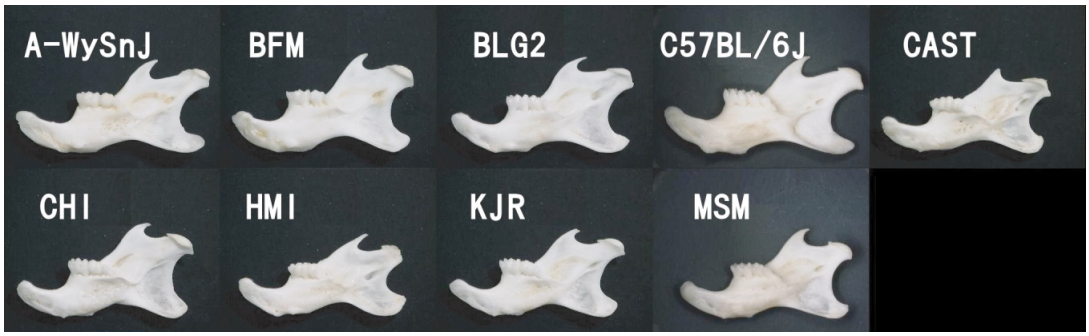

Fig. 3. Nine inbred strains of mouse mandibles (right halves).

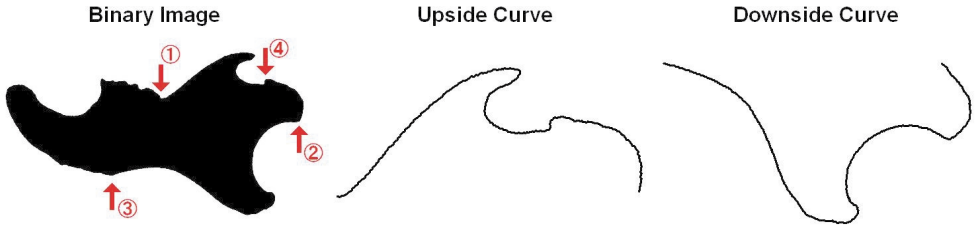

Fig. 4. Extraction of open curves from the binary image of a mouse mandibular shape (left). Tracing the boundary from 1 to 2 clockwise yields *upside curve* (center), and tracing the boundary from 3 to 4 counterclockwise yields *downside curve* (right).

metrics. We hope that the PFD do good for elucidation of morphological evolution of living things combined with genetic information which has become largely analyzed these days.

4.1 Materials and methods

It is well known that different strains of mouse mandibular shapes considerably differ (Festing, 1972), so we used mouse mandibles in this experiment (Fig. 3). Left and right halves of the mouse mandible are separated at the mandibular symphysis, and we used only right one. We placed the flat and white mouse mandibles on a dark-colored floor and photographed them from above. Then, enhancing the contrast of the images, we converted them to binary images (Fig. 4). The curve data used in this comparison are two kinds of open curves, *upside* and *downside curves* (Fig. 4), each of which are extracted from nine inbred strains of mouse mandibular outlines (Fig. 5). Each open curve was represented as the set of equally-spaced points, where $N = 201$ in the case of TFD, $N = 200$ in the case of PFD. Then we converted all open curves to 200-length feature vectors using the TFD and the PFD.

When we make strain classification, if the descriptor successfully reflects the subtle diversity of shapes among different strains, the classifier is expected to clearly separate different strains of feature vectors. In order to see which descriptor among the TFD and the PFD represents inter-strain variation of shapes more clearly, we implemented the following procedure: (1) given open curves (upside

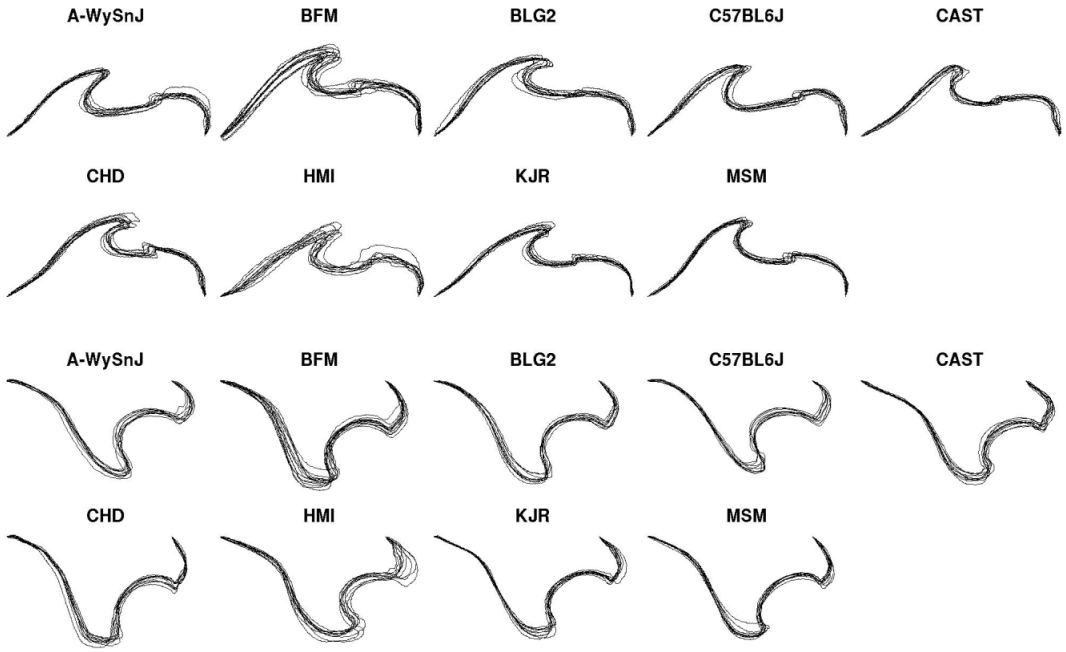

Fig. 5. 93 open curves extracted from mouse mandibular outlines of nine inbred strains with each strain consisting of 10–13 shapes. Curves of same strain are superposed (above: upside curves, below: downside curves).

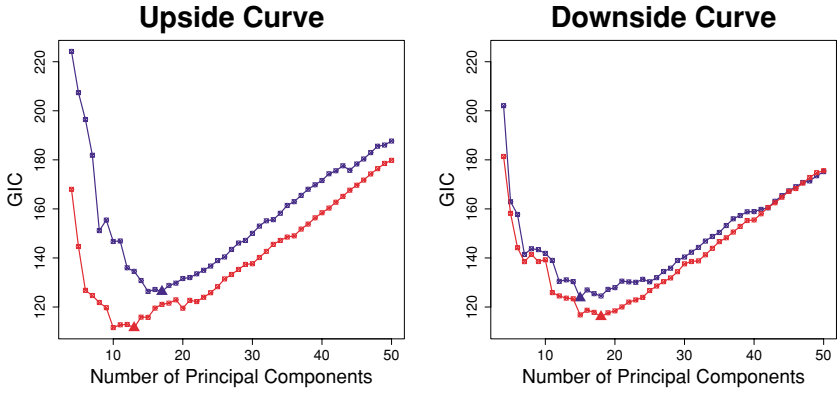

Fig. 6. GIC of the logistic regression model vs. the number of dimensions of the model, with the data represented by TFDs (blue graph) and PFDs (red graph). Triangles indicate the smallest peak values.

and downside curves) extracted from nine inbred strains of mouse mandibular shapes, represent all open curves as 200-dimensional feature vectors using the TFD and the PFD, (2) using principal component analysis, divide the whole variation in data into independent components, (3) fit the multinomial logistic regression model of arbitrary dimension K to the data using the first K princilal component scores (assuming that the essential difference among distinct inbred strains is contained in the first K principal component scores), (4) information-theoretically evaluate the goodness of the model using the GIC, and compare the GIC values in the cases of the TFD and the PFD. Using the descriptor which captures the inter-strain variability of shapes more clearly, the GIC value is supposed to be smaller compared with the one obtained with the other descriptor.

4.2 Experimental results

In the multidimensional shape space, we do not know which direction the essential difference among distinct inbred strains is represented most clearly. So, assuming that the essential feature is contained in the first K principal component scores, we fitted the model of arbitrary dimension K to the K dimensional data, and calculated GIC value of the model. And the dimension where the GIC value has the smallest peak value is the optimal dimension of the model. The graphs of GIC value vs. K (the dimension of the model) are shown in Fig. 6. Left and right figures show the results in the cases of upside and downside curves respectively. Red and blue graphs represent GIC values of the models with the data represented by PFDs and TFDs respectively. In either result for upside or downside curve, and for each number of dimensions of the data, GIC values of the models are smaller using PFDs than when using the TFDs, and the difference between the smallest peak values of GIC with the data represented by PFDs and TFDs appears more clearly in the case of upside curves than it is in the case of downside curves.

5. Conclusions and Discussion

We proposed the protrusion Fourier descriptor (PFD) method, the Fourier-based open-curve representation which utilizes the skeletal information for an open curve. And we compared the PFD to the tangent Fourier descriptor (TFD) with respect to the ability to represent the subtle variability of irregular shapes. Using the PFD and the TFD, We numerically represented the open curves extracted from different strains of mouse mandibular shapes, and fitted the strain classification model, the multinomial logistic regression model, with the data. Then we compared the goodnesses of the models in terms of the generalized information criterion (GIC) between the cases in which the data are represented by the PFD and the TFD. As a result, different strains of data were stably classified in higher belief using the PFD than when using the TFD, in other words, inter-strain variability is captured better using the PFD than when using the TFD. The difference between performances of the two Fourier descriptors is more pronounced in the case of upside curves than it is in the case of downside curves. Compared with downside curves, upside curves contain more corner-like shapes where the direction changes drastically, and such shapes seem to be characterized by relatively high frequency components. In such case, the TFD method tends to fail to represent subtle variability of shapes because of the tendency to amplify high frequency noise, whereas the PFD method is the Fourier transform of the smooth sequence of lengths and is robust against noise, so successfully represents the fine structures. However, the TFD has the advantage of easy reconstruction of the original shapes from the descriptor, so both methods are needed for statistical analysis of shapes. If genotype information from many markers on chromosomes for respective individuals is available, by setting the genotype as the class labels for every marker genotype, we can probably use the procedure implemented in this paper to detect locations of genes that have a large influence on morphology. We believe that the PFD is a useful method in biological morphometrics for elucidation of the processes of morphological evolution combined with genotype information.

Acknowledgments. The authors would like to thank the anonymous referee for helpful comments and suggestions.

Appendix A. Algorithm for computing protrusion Fourier descriptor

In this Appendix we present the algorithm of representing an open curve as the fixed-length feature vector. Given an open curve, suppose that which of the two endpoints is the starting point is determined, and the given open curve is represented as a set of N equally-spaced points which are ordered from the starting point to the ending point, that is

$$\{p_1, \ldots, p_N | \|p_{i+1} - p_i\| = \delta \ \ \forall i \in \{1, \ldots, N-1\}\},$$

where δ is some constant. For each point on the curve, there is a pair of left and right skeletons connected to the point. Each skeleton has a root and we call the left/right skeleton's root *left/right root*. The left/right root associated with a point on the curve is the center of a left/right maximal disk (MD) with a sufficiently large radius ρ. We first compute the distance between the point on the curve and its associated left/right root through the left/right skeleton, next subtract ρ form the distance, then we obtain the left/right protrusion associated with the point on the curve. Algorithm to obtain left and right protrusion at each point on the curve is divided into three main parts.

1. For all point on the original open curve, compute their associated left/right MDs.
2. Construct left/right skeletons and, for all MDs, measure the distance between the center of each MD and its root through its skeleton.
3. For each point on the curve, connect the information about its associated left and right MDs to the point on the curve.

First of all, in order to store information of the left and right MDs for every point on the curve, we make a list $\Omega = \{\omega_i^\alpha | a \in \{1, -1\}, i \in \{1, \ldots, N\}\}$, where each entry of the list is 9-tuple $\omega_i^\alpha = \{\Phi_i^\alpha, \Psi_i^\alpha, Q_i^\alpha, I_i^\alpha, J_i^\alpha, K_i^\alpha, c_i^\alpha, r_i^\alpha, a_i^\alpha\}$. The i is the index on the curve ordered from the starting point to the ending point, and whether $\alpha = 1$ or -1 means left or right with respect to the direction from the starting point to the ending point. And the datatypes and the meanings of those nine variables are as follows:

Φ_i^α	(logical value)"true" if p_i touches a MD, else"false".
Ψ_i^α	(logical value)"true" if the MD is directly connected to its root, else"false".
Q_i^α	(integer) Proper number of the MD.
I_i^α J_i^α K_i^α	(integer) Indexes of 3 points on the curve that the MD touches.
c_i^α	(2D vector) xy coodinates of the center of the MD.
r_i^α	(real number) Radius of the MD.
a_i^α	(real number) Protrusion associated with p_i.

Now, we compute the information of the left/right MDs associated with every point on the curve. In the next procedure, for arbitrary triple of points on the curve $\{p_i, p_j, p_k\}$, we check whether the circumcircle of $\Delta p_i p_j p_k$ is a tangent circle or not (whether the circle intersects the curve or not) and if so, whether it is a left tangent circle or a right one. And next, if the circle is a tangent circle, we compare the radius of the circle to the stored radii $r_i^\alpha, r_j^\alpha, r_k^\alpha$ which are associated with p_i, p_j, p_k respectively. If the present radius is larger than the stored radius, the present value is stored in there. When the computation is done for arbitrary triple of points on the curve, information of MD (radius, center coordinates, etc.) associated with every point on the curve is stored.

Procedure1: Computing information of MD associated with each point on the curve.

```
01  for α = 1, −1 do
02    for i = 1 to N do
03      ω_i^α :={"false","false",0, 0, 0, 0, 0, 0, 0};
```

```
    end
    m := 1
    for i := 1 to N − 2 do
      for j := i + 1 to N − 1 do
        for k := j + 1 to N do
          if ∠p_k p_j p_i > π [rad] then α := 1;
          if ∠p_k p_j p_i < π [rad] then α := −1;
          if ∠p_k p_j p_i = π [rad] then go to the next k;
          Set c′ the center of the circumcircle of
          Δp_i p_j p_k, and set r′ the radius of the circle;
          for l := 1 to N do
            if ‖p_l − c′‖ < r′ then go to the next k;
          end
          if r′ > r_i^α then
            ω_i^α :={"true","false",m, i, j, k, c′, r′, 0};
          end
          if r′ > r_j^α then
            ω_j^α :={"true","false",m, i, j, k, c′, r′, 0};
          end
          if r′ > r_k^α then
            ω_k^α :={"true","false",m, i, j, k, c′, r′, 0};
          end
          m := m + 1;
        end
      end
    end
  end
```

Then, only for MDs that directly connected to their roots, calculate the distances between the centers of the MDs and their roots. In the next procedure, first we set ρ sufficiently large value, and next, for each point on the curve, if any MD is associated with the point, we check whether a circle with a radius ρ can touch the point on the curve without intersecting the curve. If a circle with a radius ρ can touch, we store the linear distance between the center of the circle with a radius ρ and the center of the MD associated with the point on the curve.

Procedure2: Computing the distance from the root only for the points directly connected to their roots.

```
ρ := max{100 Σ_{i=2}^{N} ‖p_{i+1} − p_i‖,
      max{r_i^α|α ∈ {1, −1}, i ∈ {1, ..., N}}};
for α = 1, −1 do
  for l := 1 to N do
  if Φ_l^α = "false" then go to the next l;
    i := I_l^α;
    k := K_l^α;
    v := p_k − p_i;
    w := (p_i + p_k)/2+
          α√(ρ² − ‖v‖²/4) ( 0 −1 ) v/‖v‖;
                          ( 1  0 )
    /* w is the coordinate of the root. */
    for j := 1 to N do
      if ‖p_j − w‖ < ρ then go to the next l;
    end
    Ψ_l^α :="true";
    a_l^α := ‖c_l^α − w‖;
  end
end
```

Next, we construct the skeletons and measure depths of the center of each MD. Here, *depth of a point on a tree* means the distance between the point and the root through the tree. Next, to construct left and right skeletons, the other list $\tilde{\Omega} = \{\tilde{\omega}_i^\alpha | \alpha \in \{1, -1\} i \in \{1, \ldots, 3N\}\}$ is made from Ω. Each entry of $\tilde{\Omega}$ is, similar to that of the first list Ω, 9-tuple $\tilde{\omega}_i^\alpha = \{\tilde{\Phi}_i^\alpha, \tilde{\Psi}_i^\alpha, \tilde{Q}_i^\alpha, \tilde{I}_i^\alpha, \tilde{J}_i^\alpha, \tilde{K}_i^\alpha, \tilde{c}_i^\alpha, \tilde{r}_i^\alpha, \tilde{a}_i^\alpha\}$, so that these nine variables have the same datatype as the corresponding variables of the first list Ω. The second list $\tilde{\Omega}$ reflects the network structure of the skeletons.

Procedure3: Constructing the skeleton graphs from the information of MDs.

```
for α = 1, −1 do
  i := 1;
  for j = 1 to N do
    Ω̃_i^α := Ω_j^α; Ω̃_{i+1}^α := Ω_j^α; Ω̃_{i+2}^α := Ω_j^α;
    Ĩ_{i+1}^α := J_j^α; J̃_{i+1}^α := K_j^α;
    Ĩ_{i+2}^α := K_j^α; J̃_{i+2}^α := I_j^α + N;
    i := i + 3;
  end
  sort the array {ω̃_i^α|i ∈ {1, ..., 3N}} on
  the 4th variable Ĩ_i^α in an ascending order,
  for the ones which share the same value of
  Ĩ_i^α, sorting is done using the 5th variable J̃_i^α
  in descending order;
end
```

Next, for each point on the skeletons, we measure the shortest distance from the root through the skeleton, that is, the length of the shortest path that never goes through same node multiple times.

Procedure4: Measuring the distance from the root for each point on the skeleton.

```
for α := 1, −1 do
  j := 1;
  for i := 1 to 3N do
    if Ψ̃_i^α ="true" then
      j := i;
      go to the next i;
    end
    if i ≥ j + 2 then
      for k := j + 1 to i − 1 do
        if Q̃_k^α = Q̃_i^α then
          ã_i^α := ã_k^α;
          go to the next i;
        end
      end
    end
    ã_i^α := ã_{i−1}^α + ‖c̃_i^α − c̃_{i−1}^α‖;
  end
end
```

Next, for each point on the original open curve, we associate the distance from its left/right root through the left/right skeleton with the point on the curve.

Procedure5: Associating the left/right protrusion with each point on the curve.

```
for α := 1, −1 do
```

```
02    for i := 1 to N do
03      if Ψ̃_i^α ="false" then go to the next i;
04      for j := 1 to 3N do
05        if Q̃_j^α = Q_i^α then
06          a_i^α := ã_j^α + r̃_j^α − ρ;
07          go to the next i;
08        end
09      end
10    end
11  end
```

We define a complex function w_i using left and right protrusions $a_n^{(1)}, a_n^{(-1)}$ so that

$$w_n = a_n^{(1)} + \sqrt{-1}a_n^{(-1)} \ \ \forall n \in \{1, \ldots, N\}.$$

Then, we apply the discrete Fourier transform to the sequence of the complex functions $\{w_1, \ldots, w_N\}$

$$c_k = \frac{1}{N}\sum_{n=0}^{N-1} w_{n+1} \exp\left(-\sqrt{-1}\frac{2\pi kn}{N}\right)$$
$$\forall k \in \{0, \ldots, N-1\}.$$

The set of these N complex numbers $\{c_0, \ldots, c_{N-1}\}$ is protrusion Fourier descriptor and we use the set of $2N$ real numbers $\{\Re(c_0), \Im(c_0), \ldots, \Re(c_{N-1}), \Im(c_{N-1})\}$ as the feature vector of each shape in statistical analysis.

Appendix B. Inverse Transform from PFD to Curve Shape

When we think about curves of constant curvature, it can be easily imagined that there exist trillions of curves that have the same protrusions but different total lengths. So if the total length is unknown, the curve shape associated with a given PFD can not be uniquely identified. Assume that we know the total length. Suppose the curve to be obtained from the PFD is represented as $C = \{p_1, \ldots, p_N\}$, a set of N equally-spaced points which are ordered from the starting point to the ending point, and $\theta_i \angle p_{i+1} p_i p_{i-1}$ for $i \in \{2, \ldots, N-1\}$. The curve shape is specified by $N-2$ parameters $\{\theta_2, \ldots, \theta_{N-1}\}$. Using the inverse Fourier transform from the PFD, we can obtain a_i^L and a_i^R, the left and right protrusions associated with each point $p_i \in C$ as shown in Fig. 2(a). We devised the physical model so that, starting from the arbitrary curve shape, the shape changes in such a way that the protrusions gradually come close to the corresponding target protrusions. Suppose the changing curve shape is represented as $\tilde{C} = \{\tilde{p}_1, \ldots, \tilde{p}_n\}$, and $\tilde{a}_i^L$ and $\tilde{a}_i^R$ are respectively the left and right protrusions at $\tilde{p}_i \in \tilde{C}$. In this model, the point $\tilde{p}_i$ is acted on by two forces F_i^L and F_i^R, which magnitudes are respectively $a_i^L - \tilde{a}_i^L$ and $-(a_i^R - \tilde{a}_i^R)$, and their directions are the normal direction of the curve at the point (the left of the curve is the positive direction). That is, each point of the curve is acted on by normal forces from the left and the right, in such a way that the point is pushed or pulled whether the protrusion associated with the point is smaller or larger than the corresponding target protrusion. And in this model, $\tilde{p}_i \in \{\tilde{p}_2, \ldots, \tilde{p}_{N-1}\}$ is acted on by the the force moment propotional to the following value.

$$N_i^\alpha = \sum_{j=1}^{N} (-1)^{I(i<j)} \phi^{|i-j|} (\tilde{p}_j - \tilde{p}_i) \times (F_j^\alpha - F_i^\epsilon)$$
$$\text{for } \alpha = L, R$$

where $\times$ indicates vector product, $I(\cdot)$ is indicator function, and ϕ is set as a uniform random number $0 < U < 1$ at each time step. $\phi^{|i-j|}$ represents the effect that influence of the force attenuates as an exponential function of the number of joints which exist between $\tilde{p}_i$ and $\tilde{p}_j$, and ϕ represents the *hardness* of the curve. For $\tilde{p}_i$, we randomly choose either N_i^L or N_i^R, and add the small value propotional to the chosen moment to $\tilde{\theta}_i = \angle \tilde{p}_{i+1} \tilde{p}_i \tilde{p}_{i-1}$. By the way, the points belong to the convex hull need special handling. Left and right convexes are defined in the following way: $\tilde{p}_i$ is left convex if $\angle \tilde{p}_k \tilde{p}_i \tilde{p}_j < 180^\circ$ for any (j, k) where $1 \le j < i < k \le N$, and similarly $\tilde{p}_i$ is right convex if $\angle \tilde{p}_k \tilde{p}_i \tilde{p}_j > 180^\circ$ for any (j, k). When $\tilde{p}_i$ is left convex, if $\tilde{\theta}_i$ is slightly changed subject to N_i^L, N_i^L does not come close to zero. And so, when $\tilde{p}_i$ is left convex, we set $N_i^L = 0$. The right side is also handled similarly.

Iterating this process gradually decreases the whole stress produced by forces acting on the curve, and the curve shape continues to move unless the whole stress becomes to zero. By the way, the reason why we do not used $N_i^L + N_i^R$ and randomly choose either N_i^L or N_i^R when updating the $\tilde{\theta}_i$, is to avoid the case that the moment equilibrium is established throughout the curve, and the curve shape stop chaning although the shape is not optimal. Figure B.1 shows the process of finding the curve shape corresponding to a given PFD. The initial state of the shape is set as the straight line, the curve shape changes little by little using the above procedure, and gradually come close to the original curve shape. In this way, if the total length is known in advance, we can find the optimal curve shape for a given PFD.

Appendix C. Fitting Regularized Logistic Regression Model

Suppose $x = (1, x_1, \ldots, x_k, \ldots, x_K)$ is a feature vector of each object, and whole variation in data is divided into independent components using principal component analysis in advance. The fact that a object belongs to a certain class j is represented by M dimensional 0/1 valued vector $y = (y_1, \ldots, y_M)$ in which $y_j = 1$ and all other entries are 0. Logistic regression model returns a estimatie of probability that a datum x belongs to a class j' in the following way:

$$p(y_{j'} = 1 | x, B) = \frac{\exp(\beta_{j'}^{\mathrm{T}} x)}{\sum_{j=1}^{M} \exp(\beta_j^{\mathrm{T}} x)}. \qquad \text{(C.1)}$$

This model is parameterized by a vector $B = (\beta_1^{\mathrm{T}}, \ldots, \beta_M^{\mathrm{T}})^{\mathrm{T}}$ with each parameter vector β_j corresponding to the class j : $\beta_j = (\beta_{j1}, \ldots, \beta_{j,(K+1)})^{\mathrm{T}}$. Since probabilities must sum to 1: $\sum_j p(y_j = 1 | x, B) = 1$, one of β_j can be set as $\beta_j = 0$ without affecting the generality. The goodness of the model can be evaluated via GIC, and does not have to perform cross-validation. When the number of training examples is small

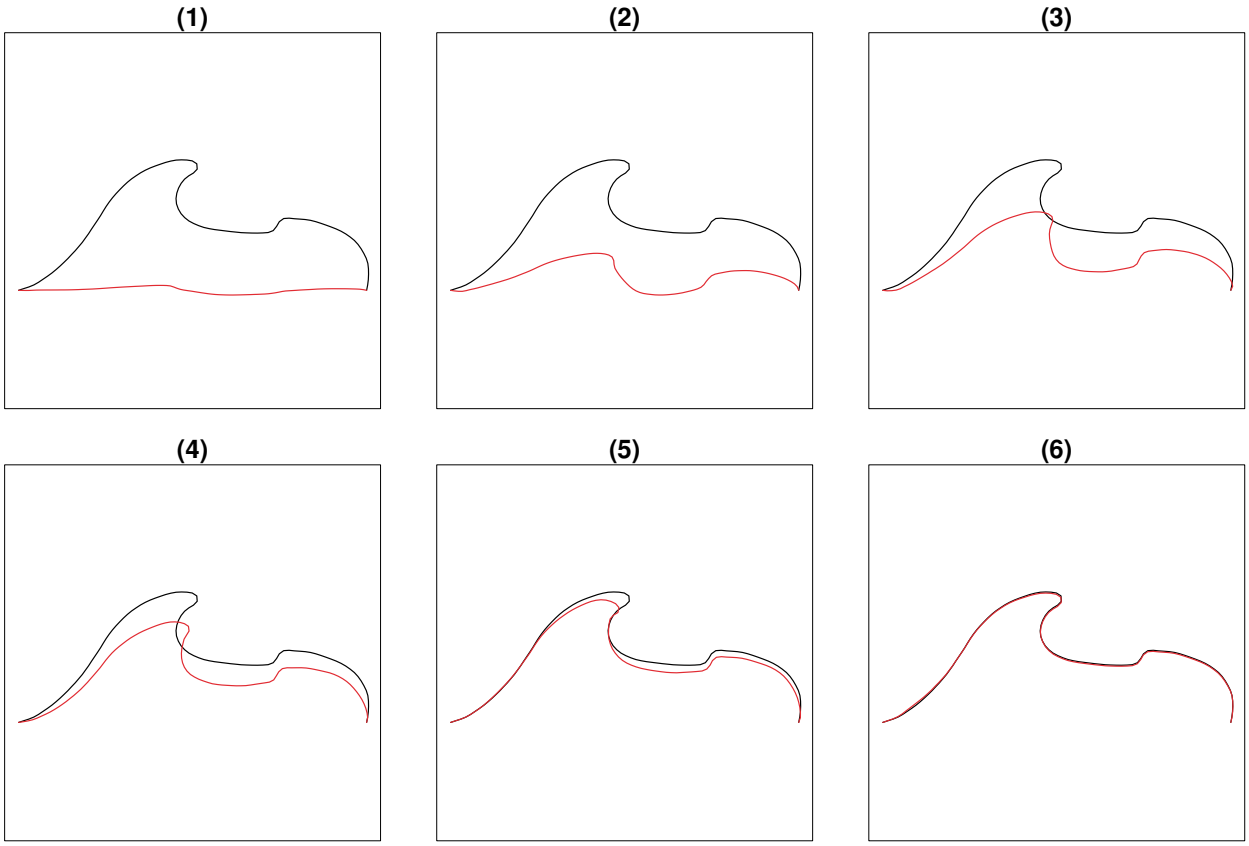

Fig. B.1. Process of the inverse transform from a PFD to its corresponding curve shape. Black line is the original curve and red line is the curve changing its shape.

and the number of dimensions of feature vectors is large, regularization is required to avoid overfitting. We estimated parameters of the model by maximizing the following L_2-regularized log-likelihood with dataset $D = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(i)}, y^{(i)}), \ldots, (x^{(N)}, y^{(N)})\}$,

$$\begin{aligned} & l(B|D) \\ &= \sum_{i=1}^{N} \left\{ \sum_{j=1}^{M-1} y_j^{(i)} \beta_j^{\mathrm{T}} x^{(i)} - \log \left[1 + \sum_{j=1}^{M-1} \exp(\beta_j^{\mathrm{T}} x^{(i)}) \right] \right\} \\ & \quad -\lambda N \sum_{j=1}^{M-1} \beta_j^{\mathrm{T}} \beta_j, \end{aligned} \tag{C.2}$$

where λ is a regularization parameter, a positive constant we must specify in advance. As mentioned in (Perkins and Theiler, 2003), All the features used as input to the model should have a similar scale, because the regularizer in Eq. (A.2) penalizes all weights in the model uniformly. We standardized all explanatory variables (the principal component scores) to have mean 0 and variance 1 before training the model, and solved the above optimization problem with Newton-Raphson iterations.

Appendix D. Hyper-Parameter Setting

The purpose of this experiments is to roughly compare the degrees of inter-strains separateness in feature space between when using TFD and when using PFD, not to identify the true model generating the data. So we fitted the model to the data with the hyperparameter λ fixed in a default value and did not tune λ. Now we explain the default value setting of the hyperparameter directly from the training data. Regularized log-likelihood of this model is

$$\log \left[\frac{\exp(\beta_{j'}^{\mathrm{T}} x)}{1 + \sum_{j=1}^{M-1} \exp(\beta_j^{\mathrm{T}} x)} \right] - \lambda \sum_{j=1}^{M-1} \beta_j^{\mathrm{T}} \beta_j.$$

Since all explanatory variables are standardized in advance, assuming that the same class of data exist in almost the same direction, for estimated β_j, if j is the class which x belongs to, $|\beta_j^{\mathrm{T}} x| \approx \|\beta_j\| \cdot \|x\|$, or else, $|\beta_j^{\mathrm{T}} x|$ is a small value, then $\exp(\beta_j^{\mathrm{T}} x) \approx 1$. Suppose b is the typical l^2-norm of β_j and m is the mean l^2-norm of the data, roughly we have the following relationship.

$$\frac{\partial}{\partial b} \left[\log \left(\frac{\exp(bm)}{M - 1 + \exp(bm)} \right) - \lambda (M-1) b^2 \right] = 0.$$

Solving for λ yields

$$\begin{aligned} \lambda &= \frac{M-1}{M-1+\exp(bm)} \cdot \frac{m}{2(M-1)b} \\ &= \frac{m^2}{2(M-1)} (1-p) \left\{ \log \left[\frac{p}{1-p} (M-1) \right] \right\}^{-1}, \end{aligned} \tag{D.1}$$

where

$$p = \frac{\exp(bm)}{M - 1 + \exp(bm)},$$

p is an arbitrary value of probability we wish to have as an output for typical data. We set $p = .9$ in this experiment. Since all explanatory variables are standardized, $m^2 = K + 1$ (number of explanatory variables plus 1 for the constant term). Regularization parameter can be set from the mean l^2-norm of the data using Eq. (C.3). This hyperparameter setting is loosely inspired by the method proposed to choose the regularization parameter of support vector machines in (Cherkassky and Ma, 2004) .

Appendix E. Generalized Information Criterion

GIC is a model evaluation criterion which is introduced by Konishi and Kitagawa (1996) as an estimate of Kullback-Leibler divergence between a supposed statistical model and an underlying true model. The GIC can be used even if the parametric family of a supposed statistical model does not contain the true model which generates the data, or if the model was estimated via regularization (maximum penalized likelihood estimation). Suppose $\hat{\theta}_\lambda$ is the estimate of a parameter vector θ, which is obtained by maximizing the penalized log-likelihood $\sum_{i=1}^{N} \log f(x^{(i)}|\theta) - \lambda N k(\theta)$, where $f(x|\theta)$ is the density function of a supposed statistical model, λ is a regularization parameter, and $k(\theta)$ is a penalty function. Taking $\psi(z, \theta) = \partial\{\log f(z|\theta) - \lambda k(\theta)\}/\partial\theta$, the GIC for the model $f(x|\hat{\theta}_\lambda)$ is

$$GIC = -2\sum_{i=1}^{N} \log f(x^{(i)}|\hat{\theta}_\lambda) + 2\mathrm{tr}(J^{-1}I),$$

where

$$I = \frac{1}{N}\sum_{i=1}^{N} \psi(x^{(i)}, \hat{\theta}_\lambda) \frac{\partial \log f(x^{(i)}|\theta)}{\partial\theta^{\mathrm{T}}}\Big|_{\hat{\theta}_\lambda},$$

$$J = -\frac{1}{N}\sum_{i=1}^{N} \frac{\partial\psi(x^{(i)}, \theta)^{\mathrm{T}}}{\partial\theta}\Big|_{\hat{\theta}_\lambda}.$$

The GIC which evaluates the goodness of the logistic regression model (Eq. (C.2)) is

$$\begin{aligned} &GIC \\ &= -2\sum_{i=1}^{N}\left\{\sum_{j=1}^{M-1} y_j^{(i)}\hat{\beta}_j^{\mathrm{T}}x^{(i)} - \log\left[1 + \sum_{j=1}^{M-1}\exp(\hat{\beta}_j^{\mathrm{T}}x^{(i)})\right]\right\} \\ &\quad +2\mathrm{tr}(J^{-1}I), \end{aligned} \tag{E.1}$$

where I and J are $(K + 1)(M - 1)$-by-$(K + 1)(M - 1)$ matrices, the (n, m) entries of I and J are as follows, (where $n = (j' - 1)(K + 1) + k'$, $m = (j'' - 1)(K + 1) + k''$, and j', $j'' \in \{1, \ldots, M - 1\}$, and k', $k'' \in \{1, \ldots, K + 1\}$,)

$$I_{nm} = \frac{1}{N}\sum_{i=1}^{N}[(y_{j'}^{(i)} - p_{j'}^{(i)})x_{k'}^{(i)} - 2\lambda\hat{\beta}_{j'k'}][(y_{j''}^{(i)} - p_{j''}^{(i)})x_{k''}^{(i)}],$$

$$J_{nm} = -\frac{1}{N}(p_{j''}^{(i)} - \delta_{j'j''})p_{j'}^{(i)}x_{k'}^{(i)}x_{j''}^{(i)} + 2\lambda\delta_{j'j''}\delta_{k'k''},$$

where

$$p_{j'}^{(i)} = \frac{\exp(\hat{\beta}_{j'}^{\mathrm{T}}x^{(i)})}{\sum_{j=1}^{M}\exp(\hat{\beta}_{j'}^{\mathrm{T}}x^{(i)})}, \quad \delta_{ab} = \begin{cases} 1 & \text{if } a = b \\ 0 & \text{otherwise.} \end{cases}$$

References

Aibara, T. Ohue, K. and Matsuoka, Y. (1991) Human face recognition by P-type Fourier descriptor, in *Proc, SPIE* Vol. 1606 (ed. K.-H. Tzou and T. Koga), SPIE, pp. 198–203.

Blum, H. (1973) Biological shape and visual science (part I), *J. Theoretical Biology*, **38**, 205–287.

Cherkassky, V. and Ma, Y. (2004) Practical selection of SVM parameters and noise estimation for SVM regression, *Neural Netw.*, **17(1)**, 113–126.

Festing, M. F. W. (1972) Mouse strain indentification, *Nature*, **238**, 351–352.

Granlund, G. H. (1972) Fourier preprocessing for hand print character recognition, *IEEE Trans. on Computers*, **C-21(2)**, 195–201.

Kawamura, Y. and Yokota, Y. (2005) I-type Fourier descriptor: a new Fourier descriptor applicable to open curves, *IEICE Transactions on Information and Systems*, Pt. 2 (Japanese Edition), **88(10)**, 2021–2028.

Konishi, S. and Kitagawa, G. (1996) Generalized information criteria in model selection, *Biometrika*, **83**, 875–890.

Kuhl, F. P. and Giardina, C. R. (1982) Elliptic Fourier features of a closed contour, *Comp. Graph. Image Process*, **18**, 236–258.

Lestrel, P., Takahashi, O. and Kanazawa, E. (2004) A quantitative approach for measuring crowding in the dental arch: Fourier descriptors, *American Journal of Orthodontics and Dentofacial Orthopedics*, **125(6)**, 716–725.

Perkins, S. and Theiler, J. (2003) Online feature selection using grafting, in *ICML*, pp. 592–599.

Rohlf, F. J. and Archie, J. W. (1984) A comparison of Fourier methods for the description of wing shape in mosquitoes (diptera: culicidae), *Systematic Zoology*, **33(3)**, 302–317.

Uesaka, Y. (1984) A new Fourier descriptor applicable to open curves, *Transaction of IEICE A*, **J67-A(3)**, 166–173 (in Japanese).

Yamada, S. and Hayakawa, M. (1997) Some experiments on identifying ukiyo-e painters from profiles: a preliminary study, *IPSJ SIG Notes*, **97(108)**, 37–42 (in Japanese).

Zheng, Z. and Tamura, Y (2005) Cultivar identification of lotus (nelumbo nucifera gaertn) by p-type Fourier descriptors with petal tip contour(breeding & germplasm resources), *Horticultural research*, **4(4)**, 385–390 (in Japanese).

Zheng, Z., Iwata, H., Hirata, Y. and Tamura, Y. (2008) Quantitative evaluation of the degree of sprout leaf bending of rice cultivars using p-type Fourier descriptors and principal component analysis, *Euphytica*, DOI:10.1007/s10681-007-9642-9.