1 Introduction

KAJAL's (and grady) Lit review

2 Notation

Consider a closed curve:

$$C : \mathbb{S} \to \mathbb{R}^2$$

Shapes should be invariant across several transformations including translation, scaling, rotation, and reparameterizaion. Specifically, the shape C is defined to be the equivalence class

$$C = \{\sigma OC(\gamma(t)) + a, \gamma \in \Gamma, O \in SO(2), a \in \mathcal{R}^2, \sigma > 0\}$$

3 Notation

Consider a closed curve, C, in two dimensions:

$$C : \mathbb{S} \to \mathbb{R}^2$$

.

Since, shapes should be invariant across several transformations including translation, scaling, rotation, and reparameterizaion, the shape C is defined to be the equivalence class

$$C = \{\sigma OC(\gamma(t)) + a, \gamma \in \Gamma, O \in SO(2), a \in \mathcal{R}^2, \sigma > 0\}$$

.

Consider a set of shapes $\mathcal{C}$, containing some fully observed shapes and some partially observed shapes. We define the set of fully observed shapes to be called

$$\mathcal{C}^o = (C_1^o \cdots C_{n_o}^o)$$

whereas the set of partially observed shapes is the referred to as

$$\mathcal{C}^m = (C_1^m \cdots C_{n_m}^m)$$

.

Each C_j^m consists of a the observed part $C_{j,obs}^m$ and the missing part $C_{j,mis}^m$ such that $C_j^m = (C_{j,obs}^m, C_{j,mis}^m)$. That is

$$C_{j,mis}^m = C_j : [a,b] \to \mathbb{R}^2$$

where

$$[a,b] = x \in \mathcal{S} | R_j'(x) = 1$$

.

So

$$\mathcal{C}_{obs}^m = C_{1,obs}^m \cdots C_{n_m,obs}^m$$

and

$$\mathcal{C}_{mis}^m = C_{1,mis}^m \cdots C_{n_m,mis}^m$$

We then have

$$\mathcal{C} = (\mathcal{C}^o, \mathcal{C}^m) = (\mathcal{C}^o, \mathcal{C}_{mis}^m, \mathcal{C}_{obs}^m)$$

and $n = n_o + n_m$.

In traditional missing data settings, a missingness indicator is defined such that it is 1 when a data point is missing and 0 otherwise. In this setting we define a function indicating which part of the function is observed and which is missing. This function is called R and is defined as follows:

$$R : \mathbb{S} \to \{0,1\}^2$$

While this function allows for a curve to be missing the x or y value individually at a given point in $\mathbb{S}$, in our setting both the x and y values are either both observed or both missing, and the indicator function in our setting is defined as follows:

$$R' : \mathbb{S} \to \{\{0,0\},\{1,1\}\}$$

We will work with R' as a missingness indicator function here, and it is 1 when the values of $(x(t), y(t))$ are both unobserved and 0 if both $(x(t), y(t))$ are observed where $t \in \mathbb{S}$. The set of all missingness functions is defined to be $\mathcal{R} = (R_1', \cdots, R_n')$.

Next, we want to consider the joint distribution of the set of curves $\mathcal{C}$ and the set of missingness indicator functions $\mathcal{R}$.

Next we want to consider the joint distribution of the curves with the missingnes indicator.

$$P(\mathcal{C}, \mathcal{R}) = P(\mathcal{C}^o, \mathcal{C}^m_{mis}, \mathcal{C}^m_{obs}, \mathcal{R})$$

In order to perform imputations we first need to define a model for the likelihood:

$$P(\mathcal{C}^o, \mathcal{C}^m_{mis}, \mathcal{C}^m_{obs}, \mathcal{R}|\theta, \phi)$$

where θ is a vector of parameters associated with modeling the data (i.e. $\mathcal{C}$) and ϕ are parameters associated with the model of the missing data mechanism (i.e. $\mathcal{R}'$).

We can then place a prior distribution on the parameters $p(\theta, \phi)$ which will lead to a posterior probability:

$$P(\theta, \phi|\mathcal{C}^o, \mathcal{C}^m_{mis}, \mathcal{C}^m_{obs}, \mathcal{R}) \propto P(\mathcal{C}^o, \mathcal{C}^m_{mis}, \mathcal{C}^m_{obs}, \mathcal{R}|\theta, \phi) \times p(\theta, \phi)$$

We then seek this distribution:

$$\int\int P(C_{mis}|\theta, \phi)P(\theta, \phi|C_{obs}, C_{mis}, R)d\theta d\phi = P(C_{mis}|C_{obs}, R)$$

Random draws from the distribution $P(C_{mis}|C_{obs}, R)$ can then be used to to fill in the missing part of the shape.

Each partially observed shape can be completed M times. This yields M sets of completed shapes.

4 Combining

4.1 Shape Analysis

Let's say we wanted to ask a statistical question where the answer was a shape (e.g What is the average shape?). I propose to answer this question, you first find the karcher mean within each of the M completed data sets and then you find the karcher mean of the M means across the completed data sets. This is analagous to Rubin's conbining rules.

It would be nice to try to come up with something like a "95 percent shape shadow". Like a band around the tooth showing the uncertainty around the mean shape.

4.2 Traditional Analysis

There are other types of analysis, like classification, where the results might be probabilities or other traditional statistical quantitities. I believe that these can be combined across imputatations using Rubin's classic combining rules.

5 Theory

6 Approximating $P(C_{mis}|C_{obs}, R)$

The distribution $P(C_{mis}|C_{obs}, R)$ can be approximated non-parametrically by employing a hot deck type procedure similar to the idea of predictive mean matching in traditional missing data imputation.

We first perform a matching step. Consider a shape $C_i^m \in \mathcal{C}^m$ and a

7 Simulations Study

8 Results

9 Conclusions

What if we took all the complete shapes and aligned them

We can represent shapes using finite representations.

An issue here though is that these are really just one realization from an equivalence class.

MAR here means that the proability of missingness is a function of only the observes part of the shape. This seems inrealistic in our setting.

deRuiter, Brophy, Lewis, Churchill, and Berger (2008)

References

deRuiter, D., J. Brophy, P. Lewis, S. Churchill, and L. Berger (2008). Faunal assemblage composition and paleoenvironment of plovers lake, a middle stone age locality in gauteng province, south africa. *Journal of Human Evolution 55*(5), 1102–1117.