Understanding Variability Between Groups of Sequences Using a Bayesian Object-Oriented Data Model

Maria Tackett* Dan Spitzner†

Abstract

Optimal Matching is an algorithm used to analyze object-oriented data, specifically sequences. In this article, we propose a framework inspired by classical Analysis of Variance to estimate the variability between groups of sequences using Optimal Matching. To estimate between group variability, we explore approaches based on established methods- bootstrapping and Hidden Markov Models. We also propose a Bayesian object-oriented model. These three methods are examined and demonstrated on a well-known dataset of historical English dance sequences, with a mind towards application to life course data.

Key Words: Bayesian analysis, life course data, Optimal Matching, sequence analysis, Analysis of Variance, object-oriented data

1. Introduction

Taking a holistic approach to understanding life course data has become a popular point of research in the social sciences. In this field of study, a person's entire life course trajectory becomes a single unit (Barban and Billari, 2012). Because the observations are sequences, it is challenging to compare observations and derive basic measures of center and spread. In current research, many social scientists use Optimal Matching to compare sequences and create clusters of similar observations (Salmela et.al, 2011, Barban and Billari, 2012).

Optimal Matching is the measure of the "distance" between two sequences (though in some formulations it is asymmetric and therefore not a true distance). It was originally proposed by Needleman and Wunsch (1970) in their research comparing amino acid sequences of proteins. In the beginning, Optimal Matching was used primarily in computational biology, but Abbott and Forrest pioneered using this algorithm in the social sciences in their 1986 paper. Since then, it has become the most widely known (Barban and Billari, 2012) and most commonly used (Salmela et. al, 2011) algorithm in the social sciences to analyze life course data.

It is well known that Optimal Matching may be used to quantify variability between and within groups of sequences. In this article, we extend the range of techniques that may be employed with Optimal Matching for this purpose. However, a secondary interest in this study is to investigate the potential for more interpretable fine tuning of the basic scheme. Specifically, formulation of Optimal Matching distance requires specification of a set of "cost" parameters, whose values are often determined arbitrarily or using more ad-hoc methods (Barban and Billari, 2012). Our investigation posits that by couching Optimal Matching methods within a formal statistical model, the costs become fully interpretable as parameters of a probability distribution that are to be estimated. Though actual estimation of the costs proves to be challenging and is not achieved here, we are able to make a first step by verifying the efficacy of a formal model in quantifying variability.

*Maria Tackett is Ph.D. Student, Department of Statistics, University of Virginia, P.O. Box 400135, Charlottesville, VA 22904-4135 USA

†Dan Spitzner is Associate Professor, Department of Statistics, University of Virginia, P.O. Box 400135, Charlottesville, VA 22904-4135 USA

To calculate the between group variation, we will need to derive a centroid (center) sequence overall and within each group. We propose three methods for obtaining these centroid sequences: bootstrapping, a Bayesian object-oriented model, and a Hidden Markov Model. The primary focus of this paper is on sequence analysis; however, these ideas apply more generally to any type of object-oriented data with an algorithm to measure the discrepancy between observations.

In Section 2, we briefly describe the Optimal Matching algorithm used throughout the remainder of the paper. In Section 3.1, we provide details for our proposed ANOVA formulation and the quantity we're interested in estimating. In Section 3.2, we provide details for the three methods we propose for estimating the centroid sequences. In Section 4, we show the results of our methods using a dataset of English folk dance sequences from Abbott and Forrest (1986). In Section 5, we conclude with discussion about the proposed methods and next steps in the research.

2. Optimal Matching

To measure the "distance" between two sequences, we use Optimal Matching. The goal of Optimal Matching is to find the alignment between two sequences that minimizes the "cost" . The total cost of this alignment is called the *Optimal Matching distance* (Abbott and Forrest 1986, Salmela et. al, 2011).

In each step of Optimal Matching, we perform an action using a character from an *alphabet*. The *alphabet* is the set of all possible characters in a sequence, i.e. the state space (Barban and Billari, 2012). The three possible actions in each step of the algorithm are to substitute one character for another, insert a character, or delete a character. There is a predefined cost associated with each one of these actions. For example, with an alphabet consisting of only the letters *A* and *B*, the sequence *ABA* becomes *ABBA* with the insertion of the letter *B* in the second position, and then becomes *ABBB* with the substitution of the last letter *A* with the letter *B*. The distance between *ABA* and *ABBB* is determined by the total cost of making an insertion followed by a substitution. Currently, the costs are defined in a variety of ways. Some of the most common are to set all of the costs equal, to use transition frequencies (Barban and Billari, 2012), or to use theory from previous social science research (Salmela et. al, 2011).

Insertion and deletion costs are each stored in vectors of length k, where k is the number of elements in the alphabet. Substitution costs are stored in a $k \times k$ matrix. In the substitution matrix, the entry (i, j) represents the cost for substituting the j^{th} character from the alphabet in place of the i^{th} character. In most analyses using Optimal Matching, the substitution matrix is symmetric. In this paper, however, we will allow for asymmetry in the substitution matrix. This allows for a situation in which it is more costly to substitute the j^{th} character for the i^{th} character than it is to do the reverse, or vice versa. Because of this asymmetry, for the remainder of this paper, we will no longer use the term Optimal Matching *distance*, but rather we will call the alignment that minimizes cost the Optimal Matching *deviance*.

We denote the deviance between two sequences y_1 and y_2 by $d_\phi(y_1, y_2)$, where ϕ represents the substitution, insertion and deletion costs. Note that due to the asymmetry allowed in the costs, it is not necessarily true that $d_\phi(y_1, y_2) = d_\phi(y_2, y_1)$. We will use this Optimal Matching deviance throughout the remainder of this paper. The methods proposed in this paper could be generalized to include other measures of discrepancy between observations in an object-oriented data context.

Table 1: Analysis of Variance of Sequences

Source	DF	SS	MS
Between	$g-1$	$\sum_{i=1}^{g} n_i d_\phi(\hat{\mu}_i, \hat{\mu}_0)$	$\frac{SSB}{g-1}$
Within	$n-g$	$\sum_{i=1}^{g}\sum_{j=1}^{n_i} d_\phi(y_{ij}, \hat{\mu}_i)$	$\frac{SSW}{n-g}$
Total	$n-1$	$\sum_{i=1}^{g}\sum_{j=1}^{n_i} d_\phi(y_{ij}, \hat{\mu}_i) + d_\phi(\hat{\mu}_i, \hat{\mu}_0)$	

3. Understanding Variability

3.1 Set Up

To estimate variability between groups of sequences, we begin with a framework that is motivated by the traditional Analysis of Variance (ANOVA). In our framework, $\hat{\mu}_0$ is the overall centroid, $\hat{\mu}_i$ is the centroid for group i, and $d_\phi(y, \mu)$ is the Optimal Matching deviance from y to μ given the substitution, insertion and deletion costs represented by ϕ.

For a given sequence y_{ij}, we can formulate a model, breaking down its sources of variability. Reflecting the model $y_{ij} = \mu_i + \epsilon_{ij}$, where $\mu_i = \mu_0 + \alpha_i$ of traditional one-way ANOVA, we imagine there is an overall centroid sequence, represented by μ_0, and group-specific centroid sequences, represented for group i by μ_i. With the traditional one-way ANOVA as our guide, we develop Table 1, an analog of an ANOVA table, to find between and within group variability.

The key difference between Table 1 and a traditional ANOVA table is that our ANOVA framework does not include an F-test. Our goal is not to use testing to find significant sources of variation, but to get an estimate of the magnitude of the variability. This idea of estimating variability is discussed further in Gelman (2006). Instead of testing, he calculates confidence intervals for an estimate of variation. He proposes a *finite population standard deviation* s_m to estimate the relevant source of variation. We use this quantity as the motivation for our parameter of interest λ^*, shown in (1).

$$\lambda^* = \sqrt{\frac{1}{g}\sum_{i=1}^{g} n_i d_\phi(\hat{\mu}_i, \hat{\mu}_0)} \tag{1}$$

The parameter in (1) is derived as an analog to the classic result $\frac{SSB}{\phi} \sim \chi^2_{g-1}(\frac{g\lambda^2}{\phi})$. The proof of this result can be easily derived using properties of classical ANOVA. The parameter λ^* is an estimate of λ in the noncentrality parameter of this χ^2 distribution. The derivation of λ^* may be found in the appendix.

3.2 Proposed Methods to Estimate Variability

In order to estimate (1), we must estimate the group centroids ($\hat{\mu}_i$) and the overall centroid ($\hat{\mu}_0$). We propose three methods for finding these centroids.

3.2.1 Bootstrap

The first proposed method for estimating λ^* is to derive confidence intervals using bootstrapping. The bootstrap approach we use follows the classical bootstrapping method in

Efron (1982).

The steps for the bootstrap sample are as follows. Given fixed costs, for each run:

1. Draw sample with replacement from each group.
2. Choose the overall centroid to be the sequence that minimizes the deviance from all other sequences in the bootstrap sample.
3. In each group, choose the group centroid to be the sequence that minimizes the total deviance from all other sequences.
4. Given the centroids, calculate λ^* using (1).

Repeating this process n times, we can find the 50% and 95% confidence intervals for λ^*.

The greatest advantage to this method is its computational efficiency. Because it does not rely on a formal model, it does not need to run MCMC algorithms to derive the distributions for the centroids like the other proposed methods. This is ideal for estimating variation, especially for large datasets.

One limitation to this approach is that it uses the observations in the dataset to determine the group and overall centroids in each run. Therefore, there is the potential that the true centroid is missed. To remedy this, a search algorithm could be used to explore all possible sequences to find each centroid. Taking this approach would cost a lot of time, so it is not recommended for estimating the variation. If the costs for the characters in the alphabet are similar, then the results will not change significantly. Therefore, the advantage of using a search algorithm to find the centroids does not outweigh the extra computational time.

3.2.2 Bayesian Object-Oriented Model

The second method we propose is to use a Bayesian object-oriented model to obtain posterior distributions of the group and overall centroids. We start with the probability mass function for a sequence y given a centroid μ and costs ϕ.

$$\pi(y|\mu, \phi) = N(\mu, \phi)e^{-\frac{1}{2}d_\phi(y,\mu)} \tag{2}$$

This model takes the form used to study random graphs in Banks and Constantine (1998). Using this model, we can calculate the posterior distributions of the group and overall centroids. This gives us additional information about the structure of our data.

We can now write the posterior distribution for a given centroid. We start with the hyperprior distribution:

$$\pi(\mu_0) \propto e^{-\frac{1}{2}d_{\phi_0}(\mu_0,\varnothing)} \tag{3}$$

where ϕ_0 is the set of cost parameters associated with the prior for the overall centroid. Typically these costs are set low to allow for a lot of variation in the distribution. The centroid sequence for the hyperprior distribution is $\varnothing$, which represents the NULL sequence. Similar to univariate objects, the NULL sequence is empty, i.e. the sequence that has no characters. Because we are working with a distribution that is centered at the NULL sequence, the distribution in (3) is an analog to a uniform hyperprior distribution.

Given an overall centroid μ_0, the prior distribution for the group centroid μ_i is as follows:

$$\pi(\mu_i|\mu_0) \propto e^{-\frac{1}{2}d_\phi(\mu_i,\mu_0)} \tag{4}$$

Based on the model in (2), the likelihood of the observations $\mathbf{y}$ given the group centroids $\mu_1, \ldots, \mu_g$ is

$$L(\mathbf{y}|\mu_1, \ldots, \mu_g) \propto e^{-\frac{1}{2}\sum_{i=1}^{g}\sum_{j=1}^{n_i} d_\phi(y_{ij},\mu_i)} \tag{5}$$

where g represents the number of groups. Thus, using (3), (4), and (5), we have the posterior distribution of a centroid μ_i

$$L(\mu_i|\mu_0, \boldsymbol{\mu}_{[-i]}, \mathbf{y}) \propto e^{-\frac{1}{2}[\sum_{i=1}^{g}\sum_{j=1}^{n_i} d_\phi(y_{ij},\mu_i)+d_\phi(\mu_i,\mu_0)+d_{\phi_0}(\mu_0,\varnothing)]} \tag{6}$$

where, $\boldsymbol{\mu}_{[-i]}$ is the set of group centroids, excluding the centroid for group i.

Using this posterior distribution, we can draw a sample of group and overall centroids to calculate λ^*.

3.2.3 Hidden Markov Model

The final method we consider is a Hidden Markov Model (HMM) proposed by Churchill and Lazareva (1999). This is a well established method used in computational biology. More specifically, Churchill and Lazareva (1999) look at how this methodology is used in molecular biology to study the evolution of DNA sequences. We use the *Mutation-deletion-insertion model* they propose to compute the posterior distribution of the group and overall centroids. This model is based on a series of actions - substitution, insertion, and deletion - so it is comparable in some ways to the Bayesian object-oriented model using Optimal Matching.

The one key difference between the HMM and the Bayesian object-oriented model, is that the goal of HMM is not to find the centroid that minimizes the distance from all other sequences in the dataset. A Hidden Markov Mode is designed to capture the evolution from a centroid sequence to all other sequences, so its goal is to find the centroid that is the "average" distance from all other sequences in the data. Therefore, we expect the posterior distribution of the centroid to be slightly different using this model than with our proposed Bayesian object-oriented model.

The quantity in (1) for estimating variability uses the Optimal Matching deviance. Therefore, we will use HMM to obtain the posterior distribution of the centroid, then we will use the Optimal Matching deviance in each run to calculate λ^*. This "ad-hoc" approach for calculating λ^* using centroids determined by HMM did not have a major impact on the results in this analysis. One should proceed with caution if using this method to estimate variability between groups of sequences that have a wide variety of sequence lengths.

4. Results

4.1 Data

To illustrate the performance of the methods described above, we will use a dataset of English folk dances from Abbott and Forrest (1986). The data contains 27 dance sequences from the village Ilmington in Warwickshire, England. Though there were up to 75 different possible steps in the dances from this village, we will consider 22 different moves which consist of various dance patterns such as partner moves, turns, footwork, etc. Each dance step is a character in the alphabet. The dance patterns are from four popular dances from this village: "Shepard's Hey", "Black Joke", "Maid of the Mill", and "Bumpus o'Stretton." There are sequences from the years 1887 and 1906 that were observed by historians and sequences from the years 1867 and 1945 that were constructed by historians to mimic the dances of those time periods. The full dataset and list of dance steps in the alphabet may

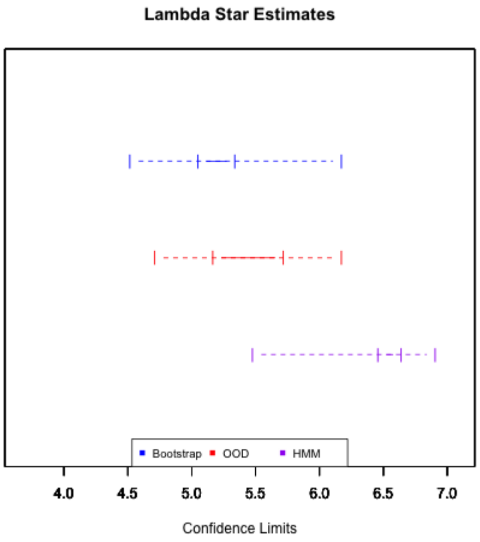

Figure 1: 50% and 95% confidence intervals of λ^* for $n = 1000$

be found in Abbott and Forrest (1986). We also use the costs used by Abbott and Forrest. Insertion and deletion costs equal one for each character in the alphabet. The substitution costs follow a hierarchical structure that is described in detail in their paper.

The goal of this analysis is to use the proposed methods in Section 3 to estimate the variability between the group of observed dances and the group of dances constructed by historians. We will use λ^* from (1) as the estimate of between group variation. Understanding this between group variation can help us assess how closely historians recreated these historical dance sequences.

4.2 Results

The confidence intervals of λ^* for each of the proposed methods are shown in Figure 1. Our proposed Bayesian object-oriented model performed similarly to the bootstrap approach. This indicates that both methods chose similar group and overall centroids in the 1000 runs. This gives confidence that the model we proposed is behaving stably in obtaining the posterior distribution for the centroids.

The Hidden Markov Model estimated between group variation slightly higher than the other methods, on average. The difference is unsurprising and has an easy explanation. Given that the Hidden Markov Model obtains the posterior distribution based on the average path between sequences, instead of the minimal path, it is expected that algorithm would choose centroid sequences that are further apart. Even with the fundamental difference between the Hidden Markov Model and the other methods, all three methods have the same similar widths for the 95% confidence intervals.

5. Discussion

Overall, each method performed comparably in estimating variability between groups of sequences. The bootstrap approach is extremely computationally efficient, so it is ideal for large datasets. Since it is not based on a formal model, one can gain only limited information from its results. Therefore, it can not be used to address some of the criticisms of Optimal Matching, such as estimating costs.

The Bayesian object-oriented model we propose performed comparably to the bootstrap method in estimating variability. With our formal model, we were able to calculate a posterior distribution of the group and overall centroids. This provides some additional information about the data, and it could potentially be used to estimate the cost parameters. Using Metropolis-Hastings within Gibbs, we could conceivably simulate the posterior distribution of the cost parameters. However, because the cost parameters ϕ are included in the normalizing constant of the probability mass function, the normalizing constant must be derived in order to obtain the posterior distribution of the costs. This is the primary challenge of working with the object-oriented model for the purpose of estimating costs: an exact expression for the normalizing constant is extremely challenging to obtain.We've explored simple techniques of approximation, but our results are inconclusive at this time.

Another approach to estimate the cost parameters is to use a MCMC method that is designed to handle intractable normalizing constants. There is a variety of literature on this topic; however, we have explored the MLE-MH algorithm proposed by Liang and Jin (2013). Further exploration is required to use this algorithm to estimate costs.

Finally, using the Hidden Markov Model proposed by Churchill and Lazareva (1999), we were able to calculate posterior distributions for the group and overall centroids. We used the centroids from these distributions to estimate variability. The primary advantages of this method are that is computationally efficient, and it is an established approach currently used in computational biology. Because it relies on averaging the deviance between sequences instead of minimizing it, it is structured differently from the way social scientist currently approach research in life course data. In order to incorporate this method in their research, social scientists would have to make a shift in their research objectives.

In this paper, we showed how Optimal Matching can be used to estimate the variability between groups of sequences. We proposed a formal model for this sequence data, and showed that it performs comparably to more established methods. There are still questions open about the intractable normalizing constant in the model and how to estimate costs. However, we have shown preliminary results that suggest this Bayesian Object-Oriented model is worth further exploration.

6. Appendix: Derivation of λ^*

Below is the derivation of (1).

$$E\Big(\frac{SSB}{\phi}\Big) = g - 1 + \frac{g\lambda^2}{\phi} \Rightarrow E(SSB) = \phi(g-1) + g\lambda^2$$

$$\Rightarrow \hat{\lambda^2} = \frac{SSB - \phi(g-1)}{g}$$

$$\Rightarrow \hat{\lambda} = \sqrt{\frac{SSB - \phi(g-1)}{g}} = \sqrt{\frac{g-1}{g}[MSB - \phi]}$$

ϕ is fixed, so we estimate

$$\lambda^* = \sqrt{\frac{1}{g}SSB} = \sqrt{\frac{1}{g}\sum_{i=1}^{g} n_i d_\phi(\mu_i, \mu_0)}$$

REFERENCES

Abbott, A. and Forrest J. (1986), “Optimal Matching Methods for Historical Sequences,” *Journal of Interdisciplinary History*, 16, 471-494.

Banks, D. and Constantine, G. M. (1998), “Metric Models for Random Graphs,” *Journal of Classification*, 15, 199-223.

Barban, N. and Billari, F. C. (2012), “Classifying Life Course Trajectories: a A Comparison of Latent Class and Sequence Analysis,” *Journal of the Royal Statistical Society*, 61, 765-784.

Churchill, G. and Lazareva, B. (1999), “Bayesian Restoration of a Hidden Markov Chain with Applications to DNA Sequencing,” *Journal of Computational Biology*, 6, 261-277.

Efron, B. (1982), *The Jackknife, the Bootstrap, and Other Resampling Plans*, Philadelphia: Society for Industrial and Applied Mathematics.

Gelman, Andrew (2006), “Analysis of Variance - Why It Is More Important Than Ever,” *The Annals of Statistics*, 33,1-53.

Liang, F. and Jin, I. (2013), “A Monte Carlo Metropolis-Hastings Algorithm for Sampling from Distributions with Intractable Normalizing Constants,” *Neural Computation*, 25, 2199-2234.

Mosteller, F. and Tukey, J. (1977), *Data Analysis and Regression: a Second Course in Statistics*, Reading: Addison-Wesley Publishing Company.

Needleman, S. and Wunsch, C. (1970), “A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins,” *Journal of Molecular Biology*, 48, 443-453.

Salmela-Aro, K., Kiuru, N., Nurmi, J., and Eerola, M. (2011), “Mapping Pathways to Adulthood Among Finnish University Students: Sequences, Patterns, Variations in Family - and Work- Related Roles,” *Advances in Life Course Research*, 16, 25-41.

Burnham, K. P., and Anderson, D. R. (1998), *Model Selection and Inference*, New York: Springer.