Combinatorial Structure of the Deterministic Seriation Method with Multiple Subset Solutions

Mark E. Madsen

Department of Anthropology, Box 353100, University of Washington, Seattle WA, 98195 USA

Carl P. Lipo

Department of Anthropology and IIRMES, 1250 Bellflower Blvd, California State University at Long Beach, Long Beach CA, 90840 USA

Abstract

Seriation methods order a set of descriptions given some criterion (e.g., unimodality or minimum distance between similarity scores). Seriation is thus inherently a problem of finding the optimal solution among a set of permutations of objects. In this short technical note, we review the combinatorial structure of the classical seriation problem, which seeks a single solution out of a set of objects. We then extend those results to the iterative frequency seriation approach introduced by Lipo et al. (1997), which finds optimal subsets of objects which each satisfy the unimodality criterion within each subset. The number of possible solutions across multiple solution subsets is larger than **n**!, which underscores the need to find new algorithms and heuristics to assist in the deterministic frequency seriation problem.

Keywords: seriation, combinatorics

1. Single Seriation Combinatorics

Seriation, whether employing class frequencies or simple occurrence to order assemblages, yields solutions which are permutations of the set of assemblages. Because we cannot determine the "polarity" of a seriation solution—which ends represent early and late—from the class data alone, a unique seriation solution is thus

Email addresses: mark@madsenlab.org (Mark E. Madsen), Carl.Lipo@csulb.edu (Carl P. Lipo)

URL: http://notebook.madsenlab.org (Mark E. Madsen), http://lipolab.org (Carl P. Lipo)

formally a pair of mirror-image permutations:

$$\{\mathbf{a}, \mathbf{d}, \mathbf{b}, \mathbf{c}, \mathbf{e}\} \equiv \{\mathbf{e}, \mathbf{c}, \mathbf{b}, \mathbf{d}, \mathbf{a}\} \tag{1}$$

This means that a set of $\mathbf{n}$ assemblages can yield $\mathbf{n!/2}$ distinct solutions, regardless of whether solutions are composed of ordered similarity matrices or"Fordian" frequency curves. With small numbers of assemblages, enumeration and testing of all possible solutions is easy, even without parallel testing across many processors. The ability to test solutions by enumeration quickly breaks down with only a modest number of assemblages. Table 1 gives the number of unique solutions for selected problem sizes between 4 and 100 assemblages, and estimates of processing time to enumerate and test all solutions, assuming a cluster of 64 cores, and 0.005 seconds per solution test.[1] With 10 assemblages, we can test all solutions quickly enough that even a serial algorithm on a single core will be adequate to find the global best solution in a matter of hours, with parallelism improving this to real time responses.

A typical characteristic of many combinatorial algorithms is that small changes in problem size can have massive changes in processing time. 13 assemblages will turn out to be the practical limit for direct enumeration, even given parallel processing with circa-2012 technology, with total processing time of nearly 3 days running 64 cores at full capacity.[2] Problems involving 14 and 15 assemblages reach the point where large clusters require more than a month and 19 months respectively, to solve. Beyond 15 assemblages, a "combinatorial explosion" sets in, with 20 assemblages requiring more than 3 million years, before solution times quickly exceed the lifetime of the universe.

In short, top-down enumerative methods are feasible for small sets of assemblages, and given widespread availability of multiple core computers, seriation packages should employ enumeration for small problems, or to build and test smaller parts of larger seriation solutions.

[1] These assumptions concerning per-trial processing time and parallelism are arbitrary but within reach of social scientists given Amazon's EC2 cloud computing infrastructure, without requiring formal "supercomputer" access. Modification by a factor of 10 has little effect on the results, perhaps shifting feasibility upward slightly before combinatorial explosion occurs.

[2] Realistically, almost nobody would contemplate doing this, given the expense of the computing time relative to the value of guaranteeing the optimal solution, but the hypothetical example demonstrates that such solutions are *feasible*.

N	Seriation Solutions	Seconds	Years
4	12	0.00094	3e-11
6	3.6e+02	0.028	8.9e-10
8	2e+04	1.6	5e-08
10	1.8e+06	1.4e+02	4.5e-06
12	2.4e+08	1.9e+04	0.00059
13	3.1e+09	2.4e+05	0.0077
14	4.4e+10	3.4e+06	0.11
15	6.5e+11	5.1e+07	1.6
16	1e+13	8.2e+08	26
20	1.2e+18	9.5e+13	3e+06
40	4.1e+47	3.2e+43	1e+36
60	4.2e+81	3.3e+77	1e+70
80	3.6e+118	2.8e+114	8.9e+106
100	4.7e+157	3.6e+153	1.2e+146

Table 1: Number of unique seriation solutions and parallel processing time for sets of assemblages $\mathbf{4 < n < 100}$, testing solutions across 64 cores, assuming 5ms per trial

2. Deterministic Seriation with Multiple Solution Groups

In an earlier paper (Lipo et al., 1997), we introduced an iterative method for finding deterministic solutions to the frequency seriation problem by partitioning assemblages into subsets, each of which meets the unimodal ordering principle, within tolerance limits governed by sample size. Lipo (2001) extended and refined the method in his dissertation research. Our initial work on the method employed a combination of automated calculations (e.g., bootstrap significance tests for pairwise orderings), and manual sorting of assemblages into groups and specific positions (using an Excel macro package available at `http://lipolab.org/seriation.html`). Figure 1 is an example of seriation with multiple solution groups, from Lipo's dissertation research in the Lower Mississippi Valley.

Our initial work suggests assemblages seriate together into groups reflecting variation in the intensity of cultural transmission among assemblages, over their duration of accumulation. In most cases, solution groups tend to be spatiotemporally compact, and form clusters when mapped on the landscape, although long-distance connections between past communities can also yield patterns which are more complex and less cohensive when mapped. Madsen's dissertation research is aimed at tying the properties seriation solution groups to their causes in regional patterns of

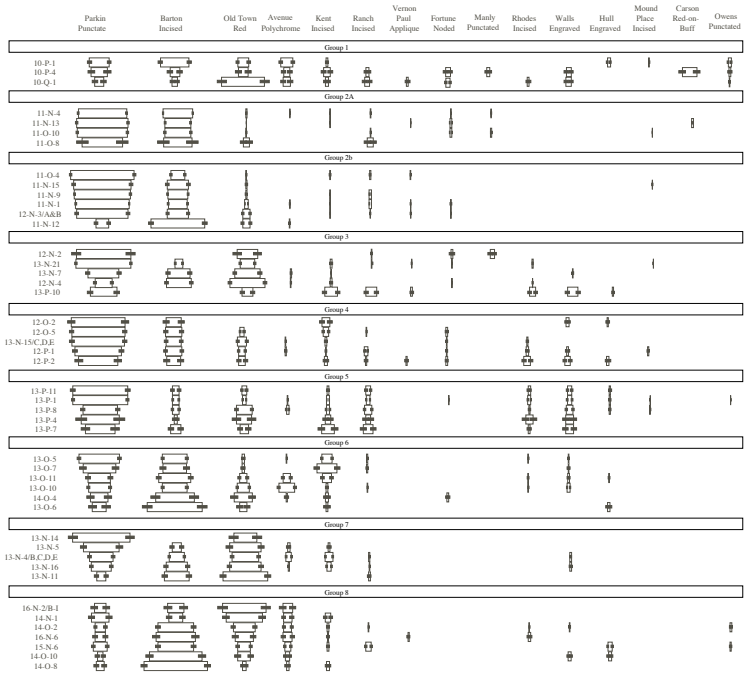

Figure 1: Example of a deterministic frequency seriation with assemblages partitioned into multiple subsets or solution groups. From Lipo (2001), Figure 4.4.

interaction and the dynamics of specific cultural transmission models.

# of Solution Groups (m)	20	40	60
3	5.8e+08	2e+18	7.1e+27
4	4.5e+10	5e+22	5.5e+34
6	4.3e+12	1.8e+28	6.8e+43
8	1.5e+13	3.2e+31	3.8e+49
10	5.9e+12	2.4e+33	2.7e+53
15		2.9e+34	2.2e+58
20		1.6e+32	1.7e+59
25			3.7e+57
30			9.6e+53

Table 2: Number of ways to form m subsets (seriation solutions) from 20, 40, and 60 assemblages

In this section, the goal is to understand the complexity of the multiple seriation groups problem, constructing reasonable upper bounds for a given problem size, even if some problems encountered in real analyses do not approach the worst case. From a combinatorial standpoint, seriation with multiple solution groups has the following structure. We begin with $\mathbf{n}$ assemblages in total, and seek a solution or solutions whereby we end up with $\mathbf{m}$ solution groups, where $\mathbf{m} < \mathbf{n}$. Each solution must have at least one assemblage, and in practice will often have 3 or more (singletons may indicate assemblages which simply do not "fit" with anything else in the data set). The number of ways that $\mathbf{n}$ objects can be partitioned into $\mathbf{m}$ non-empty subsets (or solution groups) is given by the Stirling numbers of the second kind, which are given by the recursion equation:

$$\left\{ \begin{matrix} \mathbf{n} \\ \mathbf{m} \end{matrix} \right\} = \mathbf{m} \left\{ \begin{matrix} \mathbf{n-1} \\ \mathbf{m} \end{matrix} \right\} + \left\{ \begin{matrix} \mathbf{n-1} \\ \mathbf{m-1} \end{matrix} \right\} \tag{2}$$

Table 2 gives the number of ways to form a specific number of subsets (or seriation solution groups) from sets of assemblages ranging from 20 to 60. Each column runs from 3 solution groups to half of the number of assemblages, since the number of possible subsets is maximized just before $\mathbf{n/2}$ and declines thereafter (Figure 2).

We can immediately see that there are an enormous number of possible subsets for any assemblage size. There are fewer subsets, of course, than complete permutations of the set of assemblages since subsets are unordered (i.e., $\{ \begin{smallmatrix} \mathbf{n} \\ \mathbf{m} \end{smallmatrix} \} < \mathbf{n}!$ for all $\mathbf{m}$). However, in the multiple seriation group problem, the problem size is larger than the corresponding Stirling number because we do not know in advance how many

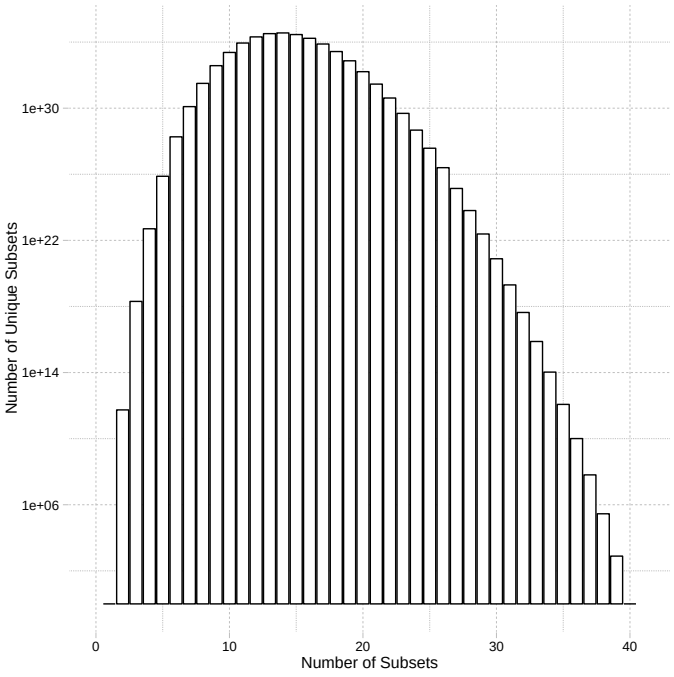

Figure 2: Number of Unique Solution Sets for 40 Assemblages When Partitioned Into **m** Solution Groups

groups (subsets) a set of assemblages will seriate into. Thus, the total number of unique subsets which might contain the optimal solution is the total of the number of subsets, across all subset sizes:

$$\sum_{\mathbf{i=1}}^{\mathbf{n}} \left\{ \begin{matrix} \mathbf{n} \\ \mathbf{i} \end{matrix} \right\} \tag{3}$$

This result is still smaller than the total permutations for a set of **n** assemblages. For example, given 40 assemblages, $\mathbf{n!} = \mathbf{8.159 \times 10^{47}}$, whereas the total from Equation 3 for 40 assemblages is $\mathbf{1.575 \times 10^{35}}$.

Another factor to consider is that each of these unique subsets resulting from a partition of **n** assemblages into seriation groups is still unordered. For example, if we partition 10 assemblages into 3 solution groups, there are 9330 unique ways of assigning the 10 assemblages to the 3 solution groups. Each group within a partition will have $\mathbf{n_i}$ members, where $\sum \mathbf{n_i} = \mathbf{n}$. The number of unique seriations for each of the 3 solution groups is $\mathbf{n_i!/2}$, but we cannot assume that solution groups will have

a balanced or equal number of assemblages (as Figure 1 does). Partitions such as:

$$\{1, 2, 3, 4, 5, 6\}\{7, 8\}\{9, 10\}$$

are common in seriating real assemblages (Lipo, 2001).

Since the factorial function grows so quickly, the computational cost of determining the correct permutation within a given seriation solution group is controlled by the size of the largest subset, especially if the other subsets are relatively small, as in the previous example. At worst, for a solution set with $\mathbf{m}$ solution groups, $\mathbf{m-1}$ solution groups will contain 1 assemblage each, and the last solution group will consist of the remaining $\mathbf{n-m-1}$ assemblages. This means, of course, that the worst case would involve consideration of on the order of $(\mathbf{n-m-1})!$ permutations within each solution group, for each of the subsets given by Equation 3. This yields:

$$\sum_{m=1}^{n} \left\{ {n \atop m} \right\} (n - m - 1)! \tag{4}$$

Table 3 gives the total number of possible solutions for assemblages ranging from 4 to 100, where solutions may fall into multiple seriation groups of any size.

N	Total Solutions	Seconds	Years
4	15	0.0012	3.7e-11
6	4.7e+02	0.037	1.2e-09
8	5.2e+04	4	1.3e-07
10	1.5e+07	1.1e+03	3.6e-05
12	8.5e+09	6.6e+05	0.021
13	2.6e+11	2e+07	0.64
14	8.9e+12	7e+08	22
15	3.5e+14	2.8e+10	8.7e+02
16	1.6e+16	1.2e+12	3.9e+04
20	1.7e+23	1.3e+19	4.2e+11
40	9e+65	7e+61	2.2e+54
60	5.1e+116	4e+112	1.3e+105
80	5.1e+172	4e+168	1.3e+161
100	4.4e+232	3.4e+228	1.1e+221

Table 3: Number of total solutions with multiple seriation groups and processing time for sets of assemblages $4 < n < 100$, testing solutions across 64 cores

3. Discussion

Clearly, in the worst case, the combinatorial complexity of the multiple seriation groups problem is much worse than even the straight factorial case involved in single solution permutations. The feasibility of parallelized enumerative methods still explodes after 13 assemblages, but much more steeply. The goal of a new algorithm for deterministic multiple group seriations is, therefore, to employ heuristics to drastically reduce the size of the solution space. Vast amounts of the solution space involve partial orders which violate unimodality, but of course we cannot easily identify those regions of solution space *a priori* without testing possibilities. But given small partial solutions which do meet the seriation model, we can easily test solutions which are "adjacent" to the partial solutions, suggesting that agglomerative heuristics may be the best approach to finding a computationally feasible method.

References Cited

Lipo, C., Madsen, M., Dunnell, R., Hunt, T., 1997. Population structure, cultural transmission, and frequency seriation. Journal of Anthropological Archaeology 16, 33.

Lipo, C.P., 2001. Science, Style and the Study of Community Structure: An Example from the Central Mississippi River Valley. British Archaeological Reports, International Series, no. 918, Oxford.