Navarro, D. J. (2003). Regarding the complexity of additive clustering models:
Comment on Lee (2001). Journal of Mathematical Psychology, 47, 241-243
http://dx.doi.org/10.1016/S0022-2496(02)00026-3

Regarding the complexity of additive clustering models: Comment on Lee (2001)

Danielle J. Navarro
Department of Psychology
University of Adelaide

Abstract

The additive clustering approach to modeling pairwise similarity of entities is a powerful tool for deriving featural stimulus representations. In a recent paper, Lee (2001) proposes a statistically principled measure for choosing between clustering models that accounts for model complexity as well as data fit. Importantly, complexity is understood to be a property, not merely of the number of clusters, but also their size and pattern of overlap. However, some caution is required when interpreting the measure, with regard to the applicability of the Hadamard inequality to the complexity matrix.

Additive clustering (Shepard & Arabie, 1979) represents a simple and effective method for modeling the similarity between a set of n stimuli. Clustering algorithms that fit the additive clustering model input a matrix of pairwise similarities $\mathbf{S} = [s_{ij}]$ and derive a stimulus representation in the form of a set of m saliency-weighted clusters (sometimes interpreted as features). Formally, the model consists of an $n \times m$ feature matrix $\mathbf{F} = [f_{ik}]$ where f_{ik} is 1 if the i-th stimulus possesses the k-th feature, and 0 if it does not, as well as a vector of nonnegative saliency weights $\mathbf{w} = [w_k]$. An additive clustering representation estimates the similarity between two stimuli by the sum of the weights of shared features: that is to say, $\hat{s}_{ij} = \sum_k w_k f_{ik} f_{jk}$. It is common practice to include a nonnegative "additive constant", added to all similarity estimates, which can be regarded as a mandatory extra cluster encompassing all stimuli.

An important theoretical issue in additive clustering regards how to choose between featural representations. In a recent paper, Lee (2001) proposes a measure that approximates the Bayesian posterior probability by employing an established variant on Laplace's method (see Kass & Raftery, 1995). This measure provides a trade off between goodness-of-fit and model complexity, and importantly, the measure of complexity is sensitive to the interaction between clusters, as well as to their number. The key component of this measure is the determinant of the *complexity matrix* $\mathbf{G} = [g_{xy}]$, the $m \times m$ matrix such that $g_{xy} = g_{yx} = \sum_{i<j} f_{ix} f_{jx} f_{iy} f_{jy}$. In other words, the xy-th element of $\mathbf{G}$ counts the number of pairs of stimuli that share the x-th feature and the y-th feature. Main diagonal elements of $\mathbf{G}$ are given by the number of pairs of stimuli that share a single feature: that is, g_{xx} reduces to $\sum_{i<j} f_{ix} f_{jx}$. Correspondingly, a row (or column) in $\mathbf{G}$ reflects both the size of a

cluster and the extent of its overlap with other clusters. This complexity matrix has broad applicability, also appearing in expressions for the Stochastic Complexity (Rissanen, 1996; see also Lee, 2002) and the related Geometric Complexity (Myung, Balasubramanian, & Pitt, 2000) for additive clustering representations.

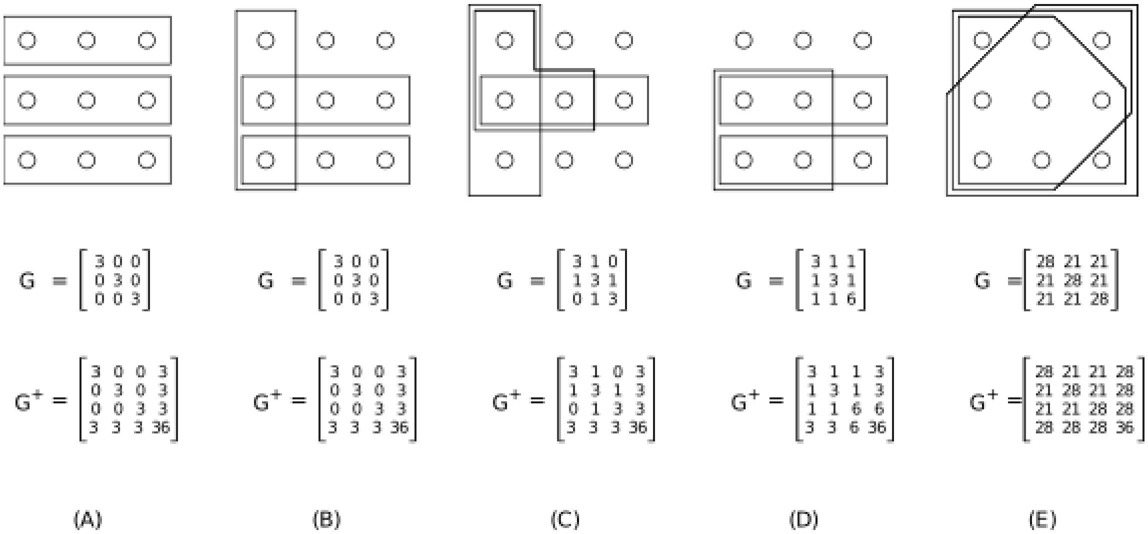
Lee demonstrates that for non-degenerate feature structures, **G** is positive definite, and applies Hadamard's inequality (Bellman, 1970, pp. 129-130), which states that the determinant of **G** is less than or equal to the product of its main diagonal,

$$|\mathbf{G}| \leq \prod_x \sum_{i<j} f_{ix} f_{jx}$$

with equality occurring when all off-diagonal elements are zero. From this observation Lee argues that, for a fixed number of clusters, the most complex representation is a partition, in which every stimulus belongs to precisely one cluster, since these models have diagonal complexity matrices. However, it should be observed that although all partitions have diagonal complexity matrices, not all diagonal complexity matrices correspond to partitions. A diagonal complexity matrix results whenever no two stimuli ever share two or more features. Therefore, so long as each pair of clusters has no more than a single stimulus in common, **G** remains diagonal.

A concrete example of this is illustrated by Figure 1, in which feature structures A and B yield precisely the same (diagonal) complexity matrix. All features have 3 stimuli and hence $\binom{3}{2} = 3$ stimulus pairs, and no two features are shared by any two stimuli, even though only feature set A is a partition. This concern notwithstanding, when the clusters share pairs of stimuli without changing size, as in feature structure C, the determinant of the complexity matrix decreases in accordance with Hadamard's inequality: $|\mathbf{G}|$ for structures A and B is 27, whereas $|\mathbf{G}|$ for structure C equals 21. If the additive constant is included, the determinant of the expanded matrix $\mathbf{G}^+$ is 729 for A and B, and 621 for C. In general, the more that a pair of clusters overlap (in terms of stimulus pairs) the less complexity is introduced, since each cluster makes a smaller unique contribution to $|\mathbf{G}|$.

The second caveat that attaches to Lee's discussion is that Hadamard's inequality applies only if the product of the main diagonal elements remains constant: that is, when the number of stimuli (and hence pairs of stimuli) in each cluster remains constant. Hadamard's inequality does not indicate what happens to the model's complexity as the number of stimuli in a cluster changes. Therefore, although Lee identifies encompassment and overlap as sources of model complexity, arguments based on Hadamard's inequality only take overlap into account. In some situations, these two factors can be varied independently: for example, a stimulus that does not belong to any cluster can be added to one of them without causing any change in the off-diagonal elements of **G**. Similarly, the comparison between feature structures A and C in Figure 1 involves manipulating the overlap between clusters without changing their size. Nevertheless, such independence is hardly the norm, and it is not immediately obvious what happens to complexity when a feature is enlarged at the expense of introducing more overlap. Consider feature structures A, D and E in Figure 1. Two of the features in A and D are identical, but the third feature in D contains four stimuli rather than three, and shares one stimulus pair with each of the other two features. As it turns out, D is the more complex representation, with $|\mathbf{G}| = 44$ and $|\mathbf{G}^+| = 1224$ (compared to 27 and 729 for A). Feature structure E involves larger clusters and more overlap, as there



$$\mathbf{G} = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 3 \end{bmatrix} \quad \mathbf{G} = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 3 \end{bmatrix} \quad \mathbf{G} = \begin{bmatrix} 3 & 1 & 0 \\ 1 & 3 & 1 \\ 0 & 1 & 3 \end{bmatrix} \quad \mathbf{G} = \begin{bmatrix} 3 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 6 \end{bmatrix} \quad \mathbf{G} = \begin{bmatrix} 28 & 21 & 21 \\ 21 & 28 & 21 \\ 21 & 21 & 28 \end{bmatrix}$$

$$\mathbf{G}^+ = \begin{bmatrix} 3 & 0 & 0 & 3 \\ 0 & 3 & 0 & 3 \\ 0 & 0 & 3 & 3 \\ 3 & 3 & 3 & 36 \end{bmatrix} \quad \mathbf{G}^+ = \begin{bmatrix} 3 & 0 & 0 & 3 \\ 0 & 3 & 0 & 3 \\ 0 & 0 & 3 & 3 \\ 3 & 3 & 3 & 36 \end{bmatrix} \quad \mathbf{G}^+ = \begin{bmatrix} 3 & 1 & 0 & 3 \\ 1 & 3 & 1 & 3 \\ 0 & 1 & 3 & 3 \\ 3 & 3 & 3 & 36 \end{bmatrix} \quad \mathbf{G}^+ = \begin{bmatrix} 3 & 1 & 1 & 3 \\ 1 & 3 & 1 & 3 \\ 1 & 1 & 6 & 6 \\ 3 & 3 & 6 & 36 \end{bmatrix} \quad \mathbf{G}^+ = \begin{bmatrix} 28 & 21 & 21 & 28 \\ 21 & 28 & 21 & 28 \\ 21 & 21 & 28 & 28 \\ 28 & 28 & 28 & 36 \end{bmatrix}$$

Figure 1. Five feature structures for a nine-stimulus domain. Feature set A has a partitioning structure, whereas B is an example of a non-partitioning structure that also has a diagonal complexity matrix. In set C, each cluster still encompasses three stimuli, but some overlap emerges. Feature structure D introduces a small amount of overlap at the expense of increasing the size of one cluster, whereas the features in E are large and overlap extensively. Two complexity matrices are given for each: $\mathbf{G}$ is the complexity matrix for the features shown, whereas $\mathbf{G}^+$ incorporates the additive constant. The xy-th element of a complexity matrix is obtained by counting the number of stimulus pairs common to the x-th and y-th clusters.

are 8 stimuli in each cluster and 7 stimuli shared between all pairs of clusters, yielding $|\mathbf{G}| = 3430$. Once the additive constant is introduced, it is no longer possible to have larger features or more overlap without including the same feature twice (which is degenerate), and $|\mathbf{G}^+|$ for this representation is 8232. In this example at least, representations with smaller clusters are simpler than those with larger clusters, even though it comes at the expense of reduced overlap.

It is also worthwhile to note that, for a fixed number of clusters the simplest representation is one consisting only of clusters containing two stimuli. The complexity matrix for this representation is the identity, and therefore has determinant 1. Since $\mathbf{G}$ is positive definite, its determinant must be positive, and since the elements of $\mathbf{G}$ are integers, no complexity matrix can ever have a determinant smaller than 1. This argument does not incorporate the additive constant, but it is heartening to note that a representation of nine stimuli using three two-stimulus clusters has $|\mathbf{G}^+| = 33$, making it simpler than any of those displayed in Figure 1.

To summarize, the solid statistical foundation of Lee's (2001) approximation to the Bayesian posterior lends it considerable status as a selection criterion for additive clustering models. The determinant of the complexity matrix $\mathbf{G}$ is a function of the size and overlap of features in the representation, but caution is required when applying Hadamard's inequality, which only takes overlap into account: in the examples presented here, representations with smaller clusters and less overlap were simpler than those with larger clusters but more overlap.

Acknowledgements

The research in this paper was financially supported by the Communications Division of the Australian Defence Science and Technology Organisation, and by an Australian Postgraduate Award. The author wishes to thank Michael Lee, Barbara Mellers, and an anonymous reviewer for helpful comments on an earlier version of this article.

References

Bellman, R. (1970). *Introduction to Matrix Analysis* (2nd ed.). New York: McGraw-Hill.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773-795.

Lee, M. D. (2001). On the complexity of additive clustering models. *Journal of Mathematical Psychology*, *45*, 131-148.

Lee, M. D. (2002). Generating additive clustering models with limited stochastic complexity. *Journal of Classification*, *19*, 69-85.

Myung, I. J., Balasubramanian, V., & Pitt, M. A. (2000). Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences USA*, *97*, 11170-11175.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, *42*(1), 40-47.

Shepard, R. N., & Arabie, P. (1979). Additive clustering representations of similarities as combinations of discrete overlapping properties. *Psychological Review*, *86*(2), 87-123.