An Application of Minimum Description Length Clustering to Partitioning Learning Curves

Daniel J. Navarro
Department of Psychology
University of Adelaide, SA 5005, Australia
Email: daniel.navarro@adelaide.edu.au

Michael D. Lee
Department of Psychology
University of Adelaide, SA 5005, Australia
Email: michael.lee@adelaide.edu.au

***Abstract*—We apply a Minimum Description Length–based clustering technique to the problem of partitioning a set of learning curves. The goal is to partition experimental data collected from different sources into groups of sources that are statistically the same. We solve this problem by defining statistical models for the data generating processes, then partitioning them using the Normalized Maximum Likelihood criterion. Unlike many alternative model selection methods, this approach which is optimal (in a minimax coding sense) for data of any sample size. We present an application of the method to the cognitive modeling problem of partitioning of human learning curves for different categorization tasks.**

I. Introduction

Clustering is one of the most basic and useful methods of data analysis. It involves treating groups of objects as if they were the same, and describing how the groups relate to one another. Clustering summarizes and organizes data, provides a framework for understanding and interpreting the relationships between objects, and proposes a simple description of these relationships that has the potential to generalize to new or different situations. For these reasons, many different clustering models have been developed and used in fields ranging from computer science and statistics to marketing and psychology. In the current paper, we consider a specific applied clustering problem from an information theoretic perspective.

Imagine an experiment in which we collect data from T different sources, but we suspect that these sources fall naturally into $K \leq T$ groups. To verify this suspicion, we need a tool to partition the data in a principled way. Since each of the T sources is a sample of data, we refer to this as a "sample partitioning" problem. As with any partitioning, the goal is to extract a set of equivalence relations (which form a *class* or *cluster*), while remaining agnostic about the relationships between the classes. The sample-partitioning problem arises in a number of applied situations. Throughout this paper, we use a common problem in cognitive modeling as a concrete example. The cognitive modeling problem arises when comparing the way people's performance improves over the course of many learning experiments. The nature of the different experimental tasks can lead to similar performance or very different learning performance. The inference problem is determining which tasks give rise to curves that are essentially the same and which give rise to curves that are inherently different.

The partitioning problem can be decomposed into two elements: searching for partitions and choosing between them. In this paper we concentrate on the latter, while applying standard methods [6] for the former. Broadly speaking, there are two approaches to choosing between clustering solutions. In "distance-based" clustering, one defines a proximity measure between items, and then seeks to minimize within-group distances and/or maximize between-group distances [4]. In contrast, the "model-based" approach treats a partition as a statistical model that assigns some likelihood to the observed data. ¿From this vantage point, we can choose a partition using information theoretic techniques such as Minimum Description Length (MDL; [5], [19], [21]). An innovative approach for doing this is developed in [10]. In this paper we apply this technique to the problem of partitioning learning curves, and discuss the statistical question of model complexity in data clustering.

II. Minimum Description Length Clustering

The MDL principle states that the goal in statistical modelling is to use the regularities present in a data set to compress it the greatest possible extent, describing the data in the most economical manner possible [5]. From a clustering perspective, we choose the clustering solution that allows the greatest possible compression of the data. In what follows, we adopt a model-based clustering procedure closely related to the one used in [10], which seeks to maximize the Normalized Maximum Likelihood criterion (NML; [21]),

$$\mathrm{NML}(\boldsymbol{X}) = \frac{p(\boldsymbol{X} \mid \hat{\boldsymbol{\theta}}_{\boldsymbol{X}})}{\sum_{\boldsymbol{X}} p(\boldsymbol{X} \mid \hat{\boldsymbol{\theta}}_{\boldsymbol{X}})}.$$

where $p(\cdot \mid \boldsymbol{\theta})$ is the model class associated with a partition. In this expression, $\hat{\boldsymbol{\theta}}_{\boldsymbol{X}}$ is the maximum likelihood estimate (MLE) for data set $\boldsymbol{X}$, and the sum is taken over all possible data sets. The denominator term is the *regret* for the model, denoted $\mathcal{R} = \sum_{\boldsymbol{X}} p(\boldsymbol{X} \mid \hat{\boldsymbol{\theta}}_{\boldsymbol{X}})$. The NML criterion is optimal in the sense that the stochastic complexity of the data SC $= -\ln(\mathrm{NML})$ gives the length of an idealized prefix code that minimizes the expected codelength for data generated by the "worst possible" source [21],

$$q^* = \min_q \max_g E_g \left[\ln \frac{p(\boldsymbol{X} \mid \hat{\boldsymbol{\theta}}_{\boldsymbol{X}})}{q(\boldsymbol{X})} \right].$$

Thus, from a coding perspective the Shannon-Fano code corresponding to the NML distribution q^* represents the minimax optimal method of encoding the data with the help of the model class $p(\cdot \mid \boldsymbol{\theta})$. By choosing the partition that maximizes NML, we find the most economical expression of the structure in the data, which is precisely our goal in clustering.

A. Fixed Partitions as Model Classes

Under a model-based clustering procedure, the data are assumed to be the outcome of some random process. A clustering solution is thus treated as a *model* for the data, and the adequacy of that solution can be assessed using statistical model selection tools. In this section we outline a clustering model for discrete data that is appropriate to the applied problem of partitioning learning curves.

Suppose that we have a discrete data set made up of T samples, each of which is an M-variate discrete probability over H response options. For instance, we might have T participants who solve M different kinds of problems, and each problem has H possible answers. Note that since each class of problem may have a different number of potential responses, H should technically be denoted H_m. However, this subscript will be dropped, since it will be clear from context. A particular partitioning of these T samples might be expressed in the following way. If we assume that there are K clusters, we might let D_k indicate how many of the original samples fall into the kth cluster. So D_k represents the size of the cluster, and thus $\sum_k D_k = T$. As before, we will generally drop the subscript k when discussing D.

We represent the data $\boldsymbol{X}$ in terms of the statistics $x_{11}^{11} \dots x_{DH}^{KM}$, where x_{dh}^{km} counts the number of observations that fall into the hth response category on the mth dimension for the dth sample that belongs to the kth cluster. In the example given earlier, x_{dh}^{km} would denote the number of times that participant d of cluster k gave the response h to a problem of type m. It will be convenient to define y_h^{km} and w^{km} as,

$$y_h^{km} = \sum_{d=1}^{D} x_{dh}^{km}, \; w^{km} = \sum_{h=1}^{H} y_h^{km}.$$

In the example discussed, y_h^{km} is the number of times that someone in the kth cluster gave the answer h to a problem in m, while w_{km} is the total number of times that a problem of type m was presented to group k.

A partitioning model for $\boldsymbol{X}$ consists of the set of K clusters $\boldsymbol{C} = (\boldsymbol{c}_1 \dots \boldsymbol{c}_K)$. In this expression, $\boldsymbol{c}_k$ denotes the set of (indices of) samples that belong to the kth cluster. The model parameters $\boldsymbol{\theta} = (\theta_1^{11}, \dots \theta_H^{MK})$ correspond to the probabilities with which each of the responses are chosen. Accordingly, θ_h^{mk} gives the probability with which response h is predicted to occur in trials belonging to cluster k and dimension m. Thus the likelihood $p(\boldsymbol{X} \mid \boldsymbol{\theta})$ is,

$$p(\boldsymbol{X} \mid \boldsymbol{\theta}) = \prod_{m=1}^{M} \prod_{k=1}^{K} \prod_{d=1}^{D} \prod_{h=1}^{H} (\theta_h^{km})^{x_{dh}^{km}} = \prod_{m=1}^{M} \prod_{k=1}^{K} \prod_{h=1}^{H} (\theta_h^{km})^{y_h^{km}}.$$

Note the y_h^{km} values are sufficient statistics for the data, assuming that the model $\boldsymbol{C}$.

Besides the stipulation that observations come partially pre-clustered in samples, the main difference between this model class and that in [10] is they employ a finite mixture model, in which the assignment of items to clusters is assumed to be the result of a latent probabilistic process. Motivated by the learning curves problem, we assume that a cluster is a *fixed* grouping of samples. Since the category structures that elicit the samples are derived from the fixed representational structure of the stimuli [23], it makes little sense in this context to propose a model class in which object assignments are assumed to result from a probabilistic process.

B. Calculating NML

We briefly discuss how the NML computations are performed, and show that the results in [10] apply to the current model. For this clustering model, the MLE is given by,

$$\hat{\theta}_h^{km} = \frac{y_h^{km}}{w^{km}}.$$

Substituting the MLE values into the likelihood function gives the maximized likelihood,

$$p(\boldsymbol{X} \mid \hat{\boldsymbol{\theta}}) \;\; = \;\; \prod_{m=1}^{M} \prod_{k=1}^{K} \left(\frac{\prod_{h=1}^{H} (y_h^{km})^{y_h^{km}}}{(w^{km})^{w^{km}}} \right).$$

The regret for a clustering model $\boldsymbol{C}$ is given by,

$$\mathcal{R}_{\boldsymbol{C}} \;\; = \;\; \sum_{y_1^{11}+\ldots+y_H^{11}=w^{11}} \cdots \sum_{y_H^{KM}+\ldots+y_H^{KM}=w^{KM}} \left[\prod_{m=1}^{M} \prod_{k=1}^{K} \frac{w^{km}!}{\prod_{h=1}^{H} y_h^{km}!} \right] \left[\prod_{m=1}^{M} \prod_{k=1}^{K} \frac{\prod_{h=1}^{H} (y_h^{km})^{y_h^{km}}}{(w^{km})^{w^{km}}} \right],$$

where the first square-bracketed term counts the number of data sets that have the sufficient statistics $y_1^{11} \dots y_H^{KM}$, and the second square-bracketed term gives the maximized likelihood to any such data set. After rearranging:

$$\mathcal{R}_{\mathbf{C}} \;\; = \;\; \sum_{y_1^{11}+\ldots+y_H^{11}=w^{11}} \cdots \sum_{y_1^{KM}+\ldots+y_H^{KM}=w^{KM}} \left[\prod_{m=1}^{M} \prod_{k=1}^{K} \frac{w^{km}!}{(w^{km})^{w^{km}}} \prod_{h=1}^{H} \frac{(y_h^{km})^{y_h^{km}}}{y_h^{km}!} \right].$$

Notice that any particular inner term depends on only a single value of m and k. Thus terms where $m = 1$ and $k = 1$ may be moved forward. Now, notice that all of the nested terms do not depend on the values of $y_1^{11} \dots y_H^{11}$, so they can be removed as a factor. Repeating this for all m and k allows the regret to be factorized, yielding

$$\mathcal{R}_{\mathbf{C}} \;\; = \;\; \prod_{m=1}^{M} \prod_{k=1}^{K} \left[\sum_{y_1^{mk}+\ldots+y_H^{mk}=w^{mk}} \frac{w^{km}!}{(w^{km})^{w^{km}}} \prod_{h=1}^{H} \frac{(y_h^{km})^{y_h^{km}}}{y_h^{km}!} \right].$$

Since individual clusters and dimensions are assumed to be independent, it is not surprising to see the regret factorize. The inner term corresponds to the regret $\mathcal{R}(H, w)$ for a one-dimensional multinomial with H options and a sample size of w. That is, $\mathcal{R}_{\mathbf{C}} = \prod_m \prod_k \mathcal{R}(H_m, w^{mk})$. The problem of calculating multinomial regret is addressed in [10], so it suffices simply to restate their result:

$$\begin{aligned}\mathcal{R}(H, w) &= \sum_{r_1+r_2=w} \left(\frac{w!}{r_1!r_2!}\right)\left(\frac{r_1^{\,r_1} r_2^{\,r_2}}{w^w}\right) \\ &\qquad \mathcal{R}(J_1, r_1)\,\mathcal{R}(J_2, r_2),\end{aligned}$$

where J_1 and J_2 are any two integers between 1 and $H-1$ such that $J_1 + J_2 = H$. They use this result to calculate $\mathcal{R}(H, w)$ efficiently using a recursive algorithm. In essence, we start by calculating all the binomial regrets $\mathcal{R}(2, 1), \ldots \mathcal{R}(2, w)$. This is reasonably fast since there are comparatively few ways of dividing a sample across two responses. Once these are known, they can be used to construct the regrets for larger multinomials. For example, if we needed $H = 14$, we would set $J_1 = 2$ and $J_2 = 2$ to arrive at the regrets for $H = 4$. We could then set $J_1 = 4$, and $J_2 = 4$ to get $H = 8$. Then $J_1 = 8$ and $J_2 = 4$ gives $H = 12$, and finally $J_1 = 12$ and $J_2 = 2$ would give the regret for $H = 14$. Obviously, at each step we need to calculate the sum over r_1 and r_2, but this can be done quickly by constructing tables of regret values. Once we have the regrets for the various multinomials, we merely need to take the appropriate product to get the regret for the clustering model.

III. Partitioning Learning Curves

The applied problem comes from Shepard et al. [23], in which human performance was examined on a category learning task involving eight stimuli divided evenly between two categories. The stimuli were generated by varying exhaustively three binary dimensions such as (black, white), (small, large) and (square, triangle). If the dimensions are interchangeable, there are only six fundamental category structures, illustrated in Figure 1. Empirically, there are robust differences in the way in which each of the six fundamental category structures is learned. In particular, Type 1 is learned more easily than Type 2, which in turn was learned more easily than Types 3, 4 and 5, and that Type 6 is the most difficult to learn. More recently, Nosofsky et al. [16] replicated the experiment and reported the detailed learning curves shown in Figure 2. Nevertheless the ordinal constraint $1 < 2 < (3, 4, 5) < 6$ is the most theoretically important result from the experiment.

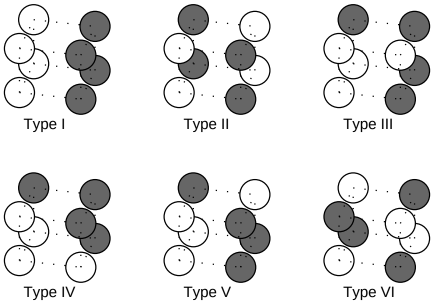

Fig. 1. The six possible category learning tasks for eight stimuli (the spheres) defined in terms of three binary-valued features (arranged as a cube). Each task divides the stimuli into four dark colored and four light colored stimuli.

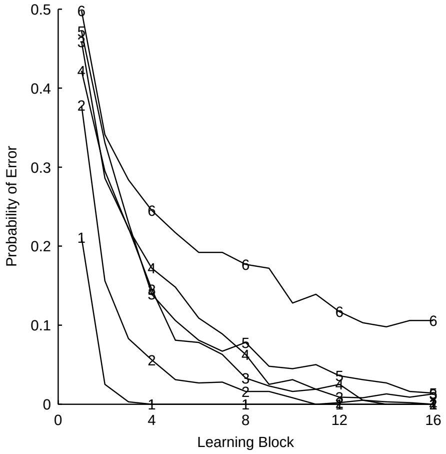

Fig. 2. Empirical learning curves for the Shepard, Hovland and Jenkins task.

A. Clustering Models for the Shepard Curves

The data have the following properties: each “data point” is a pooled set of $n = 40 \times 16 = 640$ binary observations (implying $H = 2$), assumed to be the outcome of independent Bernoulli trials. Each “curve” consists of $M = 16$ data points, corresponding to 16 different measurement intervals. There are $T = 6$ such curves, and the task is to find a set of equivalence relations among the T curves. In previous analyses [3], [16], [23], the extraction of the partition from data has only been done subjectively, by visual inspection of the curves in Figure 2. The empirical partition is then used as a set of strong ordinal constraints on potential models. That is, any cognitive model of people’s learning must show the same clustering of learning task performance [3], [11], [16], [23]. Given this important role in understanding human category learning , it would clearly be preferable to extract the partition using principled statistical methods. This becomes especially important for data sets that do not lend themselves to simple visual displays.

For clusters containing 1, 2, …, 6 curves, the log-regrets are approximately 57.3, 62.3, 65.3, 67.4, 69.1 and 70.5 respectively. We then applied an average-link clustering procedure to find six candidate partitions, with $K = 1, 2, \ldots, 6$. The results, shown in Table I, agree with the intuition that the correct clustering should be (1)(2)(3,4,5)(6). However, it suggests that (1)(2)(3,5)(4)(6), with an MDL value differing by 10, is the closest competitor, though since this is on a logarithmic scale it implies that the normalized likelihood assigned to the data by this model is actually $e^{10} \approx 22,000$ times lower. Inspection of Figure 2 agrees with this, since the curve for Type 4 is a little different from those for Types 3 and 5, but the discrepancy is not of the same order as those corresponding to Types 1, 2

TABLE I

SIX CANDIDATE CLUSTERING SOLUTIONS FOR THE SHEPARD TASK.

Partition	(Mis)Fit	Complexity	MDL
(1, 2, 3, 4, 5, 6)	16,337	70	16,408
(1, 2, 3, 4, 5)(6)	15,399	126	15,525
(1, 2)(3, 4, 5)(6)	14,772	185	14,957
(1)(2)(3, 4, 5)(6)	*14,597*	*237*	*14,834*
(1)(2)(3, 5)(4)(6)	14,553	291	14,844
(1)(2)(3)(4)(5)(6)	14,518	343	14,861

and 6. In short, the clustering procedure behaves appropriately for this data set.

B. Human Category Learning and Stimulus Coding

Theoretical work on human classification stresses the importance of data compression. It is assumed that humans make classification decisions not only to make predictions about the world, but to efficiently code the information in the environment. In the last section we demonstrated that there is strong empirical evidence that the Shepard curves should indeed be partitioned as (1)(2)(3,4,5)(6). In [3] the corresponding partial order 1<2<(3,4,5)<6 was shown to reflect the amount of information carried by each category structure, so it appears that the rate at which humans acquire a category is well-predicted by the informational content of the category. Given the obvious theoretical importance of this regularity, an interesting test of the validity of category learning models is the extent to which they preserve this regularity across their parameter spaces. If different parameterizations of a model are intended to correspond to different kinds of plausible human performance, then they should not violate this ordering too severely.

We tested this proposition with regard to the classic ALCOVE model [11], which learns by backpropagating the error made by an adaptive kernel density estimator. In order to search ALCOVE's parameter space, we used the Markov chain Monte Carlo algorithm proposed in [9] to find the different partial orders predicted by the model (see [18] for details). In total, there are only eleven stable orderings that occupy a substantial proportion of the parameter space, one of which is the empirically observed order. From this, it is clear that Types 3 and 4 are always (11 of 11) predicted to be learned at about the same rate, and Type 5 is usually (9 of 11) also about the same. Type 6, on the other hand, is mostly learned slower than 3, 4 and 5 (8 of 11). Types 1 and 2 are usually (8 of 11) slower than 3–5. So, not only is the empirically-observed ordering among the most common predictions, but the other high-frequency predictions generally preserve most of the pairwise relations implied by the empirical data. The exception to this claim regards the relationship between Types 1 and 2. In this regard, the model predictions are ambiguous. It might be that $1 < 2$ (4 of 11), or $1 = 2$ (4 of 11), or even $2 < 1$ (3 of 11). In this case, ALCOVE does not make a strong prediction about the relationship between informational content and category learning.

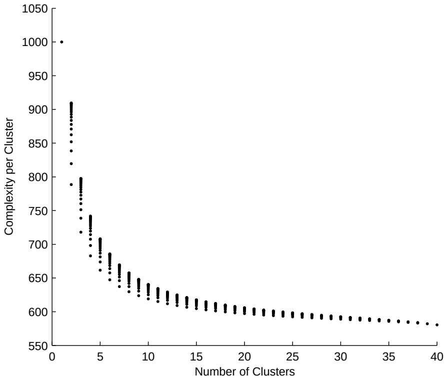

Fig. 3. Model complexity per cluster $(1/K)\ln\mathcal{R}$ is not constant, either as the number of clusters changes or within a fixed model order.

IV. EXACT MEASURES OF MODEL COMPLEXITY

Accounting for model complexity is an important topic in statistics [7] with clustering models receiving particular attention in applied work [13], [14]. Unfortunately, most approaches to model selection rely on asymptotic criteria (e.g., AIC [1], BIC [22]), or else do not provide a measure of model complexity (e.g., Bayes factors [8]). As a result, a great deal of the discussion of complexity and model selection has relied on asymptotic measures (e.g., [17]) that can be misleading in finite samples or when regularity conditions are violated [15], [12]. In contrast the NML criterion is exact, and optimal (in the minimax coding sense discussed earlier) for data of any sample size. Moreover, it supplies a natural complexity measure (i.e., $\ln\mathcal{R}$). Taken together, these two properties allow us to measure complexity properly and discuss it accurately.

It has often been argued [17], [13], [14] that model complexity is not the same as model order. However, these assertion have usually relied on asymptotic criteria: in a clustering context, Lee [13] used a Laplace approximation to the Bayes factor [8], while Lee and Navarro [14] used the Fisher information approximation to MDL [20]. Using the recursive algorithm to calculate exact NML complexities for clustering models, it is worth briefly revisiting the question. Figure 3 plots NML complexity per cluster $(1/K)\ln R_{\boldsymbol{C}}$ against the number of clusters K for every possible partition of $T = 40$ samples, with $H = 20$ response options, $N = 100$ observations per cell, and $M = 16$ dimensions. If complexity is well-captured by the number of parameters, $(1/K)\ln R_{\boldsymbol{C}}$ should be constant. Figure 3 shows that complexity per cluster is not constant as K increases, nor is it constant across models with the same number of clusters. As suggested in [13], some partitions are indeed more complex than others even when the total number of clusters remains constant.

The reason for this pattern becomes clearer when we consider the relationship between the size of a cluster (i.e., the number of samples assigned to it) and its complexity. Figure 4 plots this relationship for clusters of the same data sets referred to in Figure 3 (i.e., $T = 40$, $H = 20$, $N = 100$ and $M = 15$). The dotted line is the predicted curve if complexity

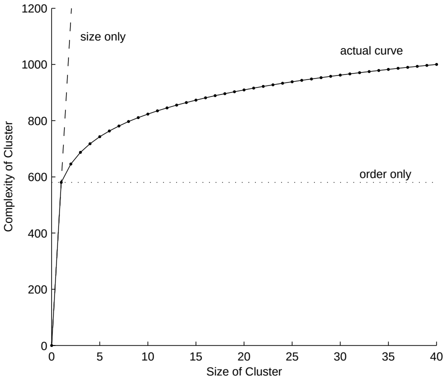

Fig. 4. Complexity associated with a particular cluster increases with size (solid line). The dotted line (“order only”) shows the predicted curve if only the number of clusters contributed to complexity. The dashed line (“size only”) shows the predicted curve if complexity related only to the size of the clusters.

were a constant function of model order, and the dashed line shows the prediction if complexity were a constant function of cluster size (in fact, if the dashed line were accurate, then each observation would contribute equally to complexity irrespective of how they were partitioned, and all clustering solutions would be of equal complexity). However, the figure shows that complexity is a concave increasing function of cluster size. If model complexity were equivalent to model order, this function would be constant, ensuring that all clusters contribute the same amount of complexity irrespective of size. Since the function is increasing, two clusters of size 1 are simpler than two clusters of size 2. Moreover, since the function is concave, complexity is subadditive. As a result, complexity is always decreased by transferring an observation from a small cluster to a large one, implying that the least complex solution is one in which all clusters except one are of size 1, while the remaining cluster is of size $T - K + 1$. This agrees with results based on Laplacian approximations [13].

V. Conclusion

The MDL-based clustering procedure introduced in [10] and applied here allows optimal clustering solutions to be found without relying on asymptotic expressions. In psychology it enables safe inferences when comparing learning curves, and when assessing theoretical accounts of cognitive processes. More speculatively, it is interesting to note that the MDL analysis of the Shepard data using the NML coding produces the same partial order as a simpler coding scheme applied to the representations of the categories themselves [3]. This suggests that human inference in many tasks is sensitive to the information inherent in the task, and that this sensitivity is strongly reflected in empirical data. If so, it is likely that information theory will remain a useful tool in not just analyzing psychological data, but also in building psychological theories and models (e.g., [2], [3], [14], [17]).

Acknowledgements

This work was partially completed while the first author was employed at Ohio State University. The research was supported by Australian Research Council grant DP-0451793, NIH grant R01-MH57472, and a grant from the Office of Research at OSU. We thank Jay Myung and Mark Pitt for helpful comments and discussions.

References

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds), *Second International Symposium on Information Theory*. Budapest: Akademiai Kiado. 1973.

[2] N. Chater & P. M. B. Vitányi. The generalized universal law of generalization. *Journal of Mathematical Psychology*, vol. 47, pp 346–369, 2003.

[3] J. Feldman. Minimization of Boolean complexity in human concept learning. *Nature*, vol. 407, pp. 630–633, 2000.

[4] A. D. Gordon. *Classification* (2nd ed.). Boca Raton, FL: Chapman and Hall. 1999.

[5] P. Grünwald. *The Minimum Description Length Principle and Reasoning under Uncertainty*. Ph.D. Thesis, ILLC Dissertation Series DS 1998-03, CWI, the Netherlands. 1998.

[6] J. A. Hartigan. *Clustering Algorithms*. New York: Wiley. 1975.

[7] T. Hastie, R. Tibshirani & Friedman. *The Elements of Statistical Learning*. New York: Springer. 2001.

[8] R. E. Kass & A. E. Raftery. Bayes factors. *Journal of the American Statistical Society* vol. 90, pp. 773–795, 1995.

[9] W. Kim, D. J. Navarro, M. A. Pitt & I. J. Myung. An MCMC-based method of comparing connectionist models in cognitive science. In S. Thrun, L. Saul & B. Schölkopf (Eds.), *Advances in Neural Information Processing Systems*, vol. 16, pp. 937–944. Cambridge, MA: MIT Press. 2004.

[10] P. Kontkanen, P. Myllymäki, W. Buntine, J. Rissanen & H. Tirri. An MDL framework for data clustering. In P. Grünwald, I. J. Myung & M. A. Pitt (Eds.) *Advances in Minimum Description Length: Theory and Applications*, pp. 323–354. Cambridge, MA: MIT Press. 2005.

[11] J. K. Kruschke. ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, vol. 99, pp. 22–44, 1992.

[12] A. D. Lanterman. Schwarz, Wallace, and Rissanen: Intertwining themes in theories of model selection. *International Statistical Review*, vol. 69, pp. 185-212. 2001.

[13] M. D. Lee. On the complexity of additive clustering models. *Journal of Mathematical Psychology*, vol. 45, pp. 131–148, 2001.

[14] M. D. Lee & D. J. Navarro. Minimum description length and psychological clustering models. In P. Grünwald, I. J. Myung, & M. A. Pitt (Eds.), *Advances in Minimum Description Length: Theory and Applications*. pp. 355–384. Cambridge, MA: MIT Press. 2005.

[15] D. J. Navarro. A note on the applied use of MDL appproximations. *Neural Computation*, vol. 16, pp. 1763–1768, 2004.

[16] R. M. Nosofsky, M. A. Gluck, T. J. Palmeri, S. C. McKinley & P. Glauthier. Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition*, vol. 22, pp. 352–369, 1994.

[17] M. A. Pitt, I. J. Myung & S. Zhang. Toward a method of selecting among computational models of cognition. *Psychological Review*, vol. 109, pp. 472–491, 2002.

[18] M. A. Pitt, D. J. Navarro, W. Kim & J. I. Myung. Global model analysis by parameter space partitioning. Submitted manuscript.

[19] J. Rissanen. Modeling by the shortest data description. *Automatica*, vol. 14, pp. 465–471, 1978.

[20] J. Rissanen. Fisher information and stochastic complexity. *IEEE Transactions on Information Theory* vol. 42, pp. 40–47, 1996.

[21] J. Rissanen. Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory*, vol. 47, pp. 1712–1717, 2001.

[22] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, vol. 6, pp. 461–464. 1978.

[23] R. N. Shepard, C. I. Hovland & H. M. Jenkins. Learning and memorization of classifications. *Psychological Monographs*, vol. 75(13), whole no. 517, 1961.