

# Calculating Geometric Complexity for Three Categorization Models: PRT, GCM, and GCM- $\gamma$

Daniel J. Navarro  
Department of Psychology  
Ohio State University

July 6, 2004

## Three Categorization Models

Consider one prototype model and two exemplar models. All models assume the probability of deciding that the  $i$ th stimulus  $S_i$  belongs to category  $C_K$ ,  $p(C_K|S_i, \theta)$  is multinomial, with probabilities proportional to the similarity of stimulus  $S_i$  to category  $C_K$ . Stimuli are represented as points in an  $m$ -dimensional Minkowski space, and similarity decays exponentially with distance. So,

$$s_{ij} = \exp \left[ -\lambda \left( \sum_{k=1}^m w_k |x_{ik} - x_{jk}|^r \right)^{\frac{1}{r}} \right], \quad (1)$$

where  $x_{ik}$  is the co-ordinate value of  $S_i$  in dimension  $k$ . In this equation,  $w_k$  denotes the proportion of attention applied to the  $k$ th dimension,  $r$  determines the metric, and  $\lambda$  denotes the steepness of the exponential decay, called the generalization gradient.

The prototype model (PRT; Reed, 1972) assumes that each category is represented by a single prototype  $S_K$ . Under this model,

$$p(C_K|S_i, \theta) = s_{iK} / \sum_J s_{iJ}, \quad (2)$$

where  $\theta = (w_1, \dots, w_{m-1}, r, \lambda)$ . It is generally assumed that the prototype behaves as if it were a stimulus whose co-ordinate value on the  $k$ th dimension is equal to  $(1/n) \sum_{j \in C_K} x_{jk}$ , where  $n$  is the number of stimuli that belong to  $C_K$ .

In the Generalized Context Model (GCM, Nosofsky, 1986), each category is represented by a set of stored exemplars. The total similarity between a stimulus and a category is simply the summed similarity between the presented stimulus and the exemplars. Hence,

$$p(C_K|S_i, \theta) = \sum_{x \in C_K} s_{ix} \bigg/ \sum_J \sum_{y \in C_J} s_{iy}. \quad (3)$$

The original GCM has the same parameters as PRT. GCM- $\gamma$  has an extra parameter, since it assumes that the summed similarities are raised to some power  $\gamma$ , yielding,

$$p(C_K|S_i, \theta) = \left( \sum_{x \in C_K} s_{ix} \right)^\gamma \bigg/ \sum_J \left( \sum_{y \in C_J} s_{iy} \right)^\gamma. \quad (4)$$

### Geometric Complexity

The geometric complexity of a model (Balasubramanian 1997, Myung, Balasubramanian & Pitt 2000) is given by,

$$G = \frac{k}{2} \ln \left( \frac{N}{2\pi} \right) + \ln \int_{\Theta} \sqrt{\det I(\theta)} d\theta, \quad (5)$$

where  $k$  denotes the number of parameters in the model,  $N$  is the sample size of the data set, and  $\Theta$  denotes some parameter bounds. In this expression  $I(\theta)$  is the Fisher information matrix of sample size one, whose  $uv$ th element is given by the expected value of the second partial derivatives of the negative log-likelihood,

$$I_{uv}(\theta) = -E \left[ \frac{\partial^2 \ln f(X|\theta)}{\partial \theta_u \partial \theta_v} \right], \quad (6)$$

where  $f(X|\theta)$  is the likelihood function for the data set  $X$ . In this expression  $\theta_u$  and  $\theta_v$  correspond to the  $u$ th and  $v$ th model parameters ( $u$  may equal  $v$ ). Since PRT, GCM and GCM- $\gamma$  are all multinomial, a standard result (e.g., Schervish 1995, p110-115; Su, Myung & Pitt, in press) allows the  $uv$ th element of the Fisher information matrix to be written

$$I_{uv}(\theta) = \sum_{S_i} \sum_{C_K} \frac{1}{p(C_K|S_i, \theta)} \frac{\partial p(C_K|S_i, \theta)}{\partial \theta_u} \frac{\partial p(C_K|S_i, \theta)}{\partial \theta_v}.$$

As a result, all we need are the partial derivatives of  $p(C_K|S_i, \theta)$  with respect to each of the model parameters (see below). Once  $I(\theta)$  can be calculated, all that is needed is the integration  $\int_{\Theta} \sqrt{\det I(\theta)} d\theta$  over the parameter space  $\Theta$ . The integrals are not very high-dimensional, so simple Monte Carlo methods suffice. We use the numerical approximation,

$$\int_{\Theta} \sqrt{\det I(\theta)} d\theta \approx \frac{1}{q} \left( \sum_{i=1}^q \sqrt{\det I(\theta^{(i)})} \right) \times V_{\Theta},$$

where  $V_{\Theta}$  denotes the volume of the parameter space, and the  $\theta^{(i)}$  values are  $q$  independent samples from a uniform distribution over  $\Theta$ .

### Partial Derivatives for PRT

For PRT, the chain rule yields the following form for the partial derivative with respect to some parameter  $\theta_u$ ,

$$\frac{\partial p(C_K|S_i, \theta)}{\partial \theta_u} = \frac{\left(\frac{\partial s_{iK}}{\partial \theta_u}\right) \left(\sum_J s_{iJ}\right) - s_{iK} \left(\sum_J \frac{\partial s_{iJ}}{\partial \theta_u}\right)}{\left(\sum_J s_{iJ}\right)^2}.$$

However, the partial derivatives of similarity differ for each parameter, and are given by,

$$\frac{\partial s_{iJ}}{\partial \theta_u} = \begin{cases} s_{iJ} \cdot (-\lambda/r) \cdot T_{iJ}^{(1-r)/r} \cdot (|x_{ik} - x_{jk}|^r - |x_{im} - x_{jm}|^r) & \text{if } \theta_u \equiv w_k \\ s_{iJ} \cdot \ln s_{iJ} \cdot \left(-\frac{1}{r^2} \cdot \ln T_{iJ} + \frac{\sum_{t=1}^{m-1} w_t |x_{it} - x_{jt}|^r \ln |x_{it} - x_{jt}|}{r T_{iJ}}\right) & \text{if } \theta_u \equiv r \\ s_{iJ} \cdot (-T_{iJ}^{(1-r)/r}) & \text{if } \theta_u \equiv \lambda \end{cases},$$

where

$$T_{iJ} = \sum_{t=1}^{m-1} w_t |x_{it} - x_{jt}|^r.$$

### Partial Derivatives for GCM

For the GCM, application of the chain rule yields,

$$\frac{\partial p(C_K|S_i, \theta)}{\partial \theta_u} = \frac{\left(\sum_{z \in C_K} \frac{\partial s_{iz}}{\partial \theta_u}\right) \left(\sum_J \sum_{y \in C_J} s_{iy}\right) - \left(\sum_{z \in C_K} s_{iz}\right) \left(\sum_J \sum_{y \in C_J} \frac{\partial s_{iy}}{\partial \theta_u}\right)}{\left(\sum_J \sum_{y \in C_J} s_{iy}\right)^2}.$$

In this case the partial derivatives of the similarity are,

$$\frac{\partial s_{ij}}{\partial \theta_u} = \begin{cases} s_{ij} \cdot (-\lambda/r) \cdot T_{ij}^{(1-r)/r} \cdot (|x_{ik} - x_{jk}|^r - |x_{im} - x_{jm}|^r) & \text{if } \theta_u \equiv w_k \\ s_{ij} \cdot \ln s_{ij} \cdot \left(-\frac{1}{r^2} \cdot \ln T_{ij} + \frac{\sum_{t=1}^{m-1} w_t |x_{it} - x_{jt}|^r \ln |x_{it} - x_{jt}|}{r T_{ij}}\right) & \text{if } \theta_u \equiv r \\ s_{ij} \cdot (-T_{ij}^{(1-r)/r}) & \text{if } \theta_u \equiv \lambda \end{cases},$$

where

$$T_{ij} = \sum_{t=1}^{m-1} w_t |x_{it} - x_{jt}|^r.$$

(The derivation of GCM partial derivatives was originally provided by Yong Su).

## Partial Derivatives for GCM- $\gamma$

GCM- $\gamma$  has the extra scaling parameter. In this case, the partial derivatives are:

$$\frac{\partial p(C_K|S_i, \theta)}{\partial \theta_u} = \begin{cases} \frac{A-B}{E} & \text{if } \theta_u \equiv \gamma \\ \frac{C-D}{E} & \text{otherwise} \end{cases}$$

where

$$\begin{aligned} A &= \left( \sum_{z \in C_K} s_{iz} \right)^\gamma \left( \ln \left( \sum_{z \in C_K} s_{iz} \right) \right) \sum_J \left( \sum_{y \in C_J} s_{iy} \right)^\gamma \\ B &= \left( \sum_{z \in C_K} s_{iz} \right)^\gamma \sum_J \left( \sum_{y \in C_J} s_{iy} \right)^\gamma \ln \left( \sum_{y \in C_J} s_{iy} \right) \\ C &= \gamma \left( \sum_{z \in C_K} s_{iz} \right)^{\gamma-1} \left( \sum_{z \in C_K} \frac{\partial s_{iz}}{\partial \theta_u} \right) \sum_J \left( \sum_{y \in C_J} s_{iy} \right)^\gamma \\ D &= \gamma \left( \sum_{z \in C_K} s_{iz} \right)^\gamma \sum_J \left( \sum_{y \in C_J} s_{iy} \right)^{\gamma-1} \left( \sum_{y \in C_J} \frac{\partial s_{iy}}{\partial \theta_u} \right) \\ E &= \left( \sum_J \left( \sum_{y \in C_J} s_{iy} \right)^\gamma \right)^2. \end{aligned}$$

Note that  $s_{ij}$  does not depend on  $\gamma$ , and that its partial derivative with respect to  $\gamma$  never appears in the above expression. Correspondingly, the formulas for  $\frac{\partial s_{ij}}{\partial \theta_u}$  are identical to those used in the regular GCM.

## References

- Balasubramanian, V. (1997). Statistical inference, Occam's razor and statistical mechanics on the space of probability distributions. *Neural Computation*, *9*, 349-368.
- 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.
- Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship, *Journal of Experimental Psychology: General*, *115*, 39-57.
- Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*, 382-407.
- Schervish, M. J. (1995). *Theory of Statistics*. New York: Springer.
- Su, Y., Myung, I. J. & Pitt, M. A. (in press). Minimum description length and cognitive modeling. To appear in P. Grünwald, I. J. Myung and M. A. Pitt (eds), *Advances in Minimum Description Length: Theory & Applications*. MIT Press.