Minimum Message Length Clustering of Spatially-Correlated Data with Varying Inter-Class Penalties

Gerhard Visser
Clayton School of I.T.,
Monash University, Clayton,
Vic. 3168, Australia,
(gvis1@student.monash.edu.au)

David L. Dowe
Clayton School of I.T.,
Monash University, Clayton,
Vic. 3168, Australia

Abstract

We present here some applications of the Minimum Message Length (MML) principle to spatially correlated data. Discrete valued Markov Random Fields are used to model spatial correlation. The models for spatial correlation used here are a generalisation of the model used in (Wallace 1998) [14] for unsupervised classification of spatially correlated data (such as image segmentation). We discuss how our work can be applied to that type of unsupervised classification. We now make the following three new contributions. First, the rectangular grid used in (Wallace 1998) [14] is generalised to an arbitrary graph of arbitrary edge distances. Secondly, we refine (Wallace 1998) [14] slightly by including a discarded message length term important to small data sets and to a simpler problem presented here. Finally, we show how the Minimum Message Length (MML) principle can be used to test for the presence of spatial correlation and how it can be used to choose between models of varying complexity to infer details of the nature of the spatial correlation.

1. Introduction and Minimum Message Length

1.1. Spatially Correlated Data

Often the elements of a data set have coordinates associated with them or perhaps some form of distance function is defined over them. For such data important questions to ask are, does the data exhibit spatial correlation, what is the degree of spatial correlation and what is the nature of the spatial correlation. To answer the first two questions one needs to compare hypotheses which assume spatial correlation with ones that do not. To answer the last, one needs to compare hypotheses which model spatial correlation in different ways. This task requires comparing models with different degrees of complexity and different numbers of parameters. The inference method we use is Minimum Message Length (MML).

1.2. Minimum Message Length

Minimum Message Length (MML) [15, 16] is a Bayesian inference method with an information-theoretic interpretation. By minimising the length of a two-part message of Hypothesis (H) followed by Data given Hypothesis ($D|H$), we seek a quantitative trade-off between the desiderata of model simplicity and goodness of fit to the data. This can be thought of as a quantitative version of Ockham's razor [11] and is compared to Kolmogorov complexity and algorithmic complexity [13, 10, 3] in [18].

For further discussions of these issues and for contrast with the much later Minimum Description Length (MDL) principle [12], see [18], other articles in that 1999 special issue of the *Computer Journal* and [4, sec. 11.4].

1.3. MML compared to other methods

Maximum Likelihood. (ML) chooses the hypotheses H which give the largest likelihood $Pr(D|H)$ for the observed data. A problem with ML is that it can not reliably choose between models of different complexity. A greater likelihood is almost always attained by models with a greater number of parameters. Unlike maximum likelihood, MML works well when comparing models with different numbers of parameters [20, sec. 6].

Even when Maximum Likelihood's tendency to overfit is reduced by introducing the penalty term in Akaike's Information Criterion (AIC), we still find both theoretical and empirical reasons for preferring MML over AIC [5].

Alternative Bayesian approaches. As well as being in general statistically consistent [6, 15, 5], another advantage of MML over alternative Bayesian approaches is that it is statistically invariant [17] - meaning that the inference is preserved under 1-to-1 transformations of the parameter space. The posterior median is only defined in one dimension, the posterior mean is not invariant and - when, as usual, it is used to maximise a density rather than a probability - the posterior mode (or Maximum A Posteriori, or MAP) is also not invariant. See [19, secs. 5,6] and [4].

2. Inference of spatial correlation

2.1. A model for spatial correlation

Let $G = (N, E)$ be an undirected graph with N and E the sets of nodes and edges respectively. We will used i and j to denote nodes in N while $e_{i,j} \in E$ is the edge between i and j. With each edge is associated a distance $w_{i,j}$ satisfying the usual triangle inequality $w_{ik} \leq w_{ij} + w_{jk}$ and $w_{ii} = 0$. Denote the set of neighbours of node i by δ_i. G is undirected, so if $i \in \delta_j$ then $j \in \delta_i$.

Let $x \in X$ be a vector of discrete valued variables $x_i \in \{1, 2, ..., K\}$ indexed by the set of nodes $i \in N$. We assume that the probability distribution over x forms a *Markov Random Field* (MRF). That is, for each node $\Pr(x_i|\phi, x_j, j \neq i) = \Pr(x_i|\phi, x_j, j \in \delta_i)$ and $\Pr(x|\phi) > 0$ for all possible assignments of x. ϕ denotes the model parameters to be inferred. So each variable x_i is conditionally independent of all others given its neighbours.

A *Gibbs Random Field* (GRF) for the graph G and configuration space X is a probability distribution over X which can be written in terms of an energy function $q(x)$. Equations 1 and 2 show the form of this distribution for first order GRFs. The Hammersley-Clifford theorem, proved in [9, 1, 8], states that MRFs and GRFs are equivalent hence our distribution over x can be written as:

$$\Pr(x|\phi) = Z(\phi)e^{-q(x,\phi)} \tag{1}$$

$$q(x, \phi) = \sum_{i \in N} q_i(x_i, \phi) + \sum_{e_{i,j} \in E} q_{i,j}(x_i, x_j, \phi) \tag{2}$$

where $Z(\phi)$ is a normalisation term known as the partition function and $q(x, \phi)$ is the energy of state x. Thus the probability of a state is determined by its energy with more probable states having lower energies. The terms q_i are known as first order clique potentials (corresponding to individual nodes) and the terms $q_{i,j}$ are known as second order clique potentials (corresponding to edges). A clique of G is any subset of nodes in N (of size equal to the clique order) which are all neighbours of each other. While clique potentials of order three and higher are sometimes used in MRF image models, we will not consider higher order cliques.

2.2. Detecting spatial correlation

We now describe how the Minimum Message Length principle can be used to determine if an observed vector x for a given graph G displays spatial correlation. There are two hypotheses to be compared.

1. H1: there is no spatial correlation, $\Pr(x_i|\phi_1, x_j, j \neq i) = \Pr(x_i|\phi_1)$
2. H2: there is spatial correlation, $\Pr(x_i|\phi_2, x_j, j \neq i) = \Pr(x_i|\phi_2, x_j, j \in \delta_i)$

To choose between these two models we can calculate the minimum message lengths for H1 and H2 and choose the one which leads to the most compression. The difference in message lengths gives us a degree of preference for one model over the other.

For H1, we can simply set all second order clique potentials to zero and let $Z(\phi) = 1$. This leads to a distribution where $\Pr(x_i|\phi_1) = e^{-q_i(x_i,\phi_1)}$. Note this is simply a discrete multi-state distribution. For H1 the model parameters are $\phi_1 = (a_1, a_2, ..., a_{K-1})$ where $e^{-q_i(k,\phi_1)} = a_k$ and $\sum_{k=1}^{K} a_i = 1$. This leads to the following message length [15, chap. 5]:

$$\begin{array}{c} I_{H1} = -\log(h(H1)) + \frac{K-1}{2}(\log(\frac{|N|}{12}) + 1) \\ -\log(h(\phi_1)) - \sum_{k=1}^{K}(n_k + \frac{1}{2})\log(a_k) \end{array} \tag{3}$$

where n_k is the number of x_i equal to k, $h(H1)$ is the a priori probability of $H1$ being true and $h(\phi_1)$ is the a priori density over the model parameters a_k.

For H2, the model parameters are $\phi_2 = (\beta, a_1, a_2, ..., a_{K-1})$ where the parameters a_k are defined as before while β is a parameter (not used in ϕ_1) that determines the degree of spatial correlation. The second order clique potentials are,

$$q_{i,j}(x_i, x_j, \phi) = \frac{\beta}{1 + w_{i,j}^{2.5}} c(x_i, x_j) \tag{4}$$

where $c(x_i, x_j)$ is a suitably chosen measure of the difference between x_i and x_j. If the values of x_i are categorical (or nominal or multi-state) we can define $c(x_i, x_j)$ as 0 if they are the same and 1 otherwise. If they are ordinal (ordered but discrete, for example integers) something like $|x_i - x_j|$ for example can be used. The term $\frac{1}{1+w_{i,j}^{2.5}}$ may be replaced by other reasonable functions of $w_{i,j}$ depending on the application. The message length for H2 is:

$$\begin{array}{c} I_{H2} = -\log(h(H2)) + \frac{K}{2}(\log(\frac{1}{12}) + 1) \\ -\log(h(\phi_2)) + \frac{1}{2}\log(F(\phi_2)) - \log(\Pr(x|\phi_2)) \end{array} \tag{5}$$

Let $L = -\log(\Pr(x|\phi_2))$ be the (negative) log likelihood. Let $F(\phi_2)$ be the determinant of the matrix of expected second derivatives of L with respect to the parameters ϕ_2 [15, chap. 5] [20]. For this problem the term $\Pr(x|\phi_2)$ can not be calculated easily [14, sec. 5.6], and $F(\phi_2)$ is presumably harder to evaluate. The following two subsections describe how they can be approximated.

For H1, given an a priori density over ϕ_1 the optimal estimates for those parameters can often be calculated directly [15, 19]. For H2 the message length can only be calculated using numerical approximations and for optimization of the parameters ϕ_2 we use simple but slow search algorithms.

2.3. Calculating the likelihood term

We describe here two numerical approximations for the likelihood term $\Pr(x|\phi)$ needed for our message length calculations. Note that ϕ is notationally used here as this applies to both ϕ_2 and ϕ_3 (introduced in a later section) for H2 and H3 respectively.

The first numerical approximation is the numerical approximation used in [14, secs. 5.5,5.6]. The second is a simpler method which is computationally less expensive when the number of nodes is small (about 300 or less). The focus of this paper is on smaller data sets as inferring the presence and properties of spatial correlation becomes more difficult when little data is available. Note that both these methods rely on Gibbs sampling from the distribution over x defined by the parameters ϕ. The time needed to generate such samples reliably depends on the graph used. The graphs used in our tests were all laid out on two dimensional planes (using the distance between points as the edge weights) and presented no such problems. Both approximations have relatively large variances, however they seem to settle on the same value when run slowly enough.

For large data sets. This numerical approximation is the one used in [14, secs. 5.5, 5.6]. Let $Z(T,\phi)e^{-q(x,\phi)/T}$ be the distribution over x at temperature T. As T increases this distribution reaches its maximum possible entropy. It can be shown that $\frac{dH}{dT} = \frac{dQ}{dT}/T$ (hence $dH = dQ/T$) where $H(T,\phi)$ is the entropy of the distribution over x at temperature T and $Q(T,\phi)$ is the expected energy at this temperature. Gibbs sampling can be used to sample random states of x given T and ϕ, and hence $Q(T,\phi)$ can be approximated at any temperature. [8] describes how Gibbs sampling can be used to sample from Markov Random Fields.

We know that at $T = \infty$ the entropy $H(T,\phi)$ attains its maximum value, which can be easily calculated. The entropy of the distribution at temperature $T = 1$ can be calculated as follows. Starting at $T = 1$ and slowly incrementing it up to some value high enough to give a distribution similar to that attained at $T = \infty$, calculate dQ at each temperature increment. By subtracting the term dQ/T at each increment from the maximum entropy $|N|\log(K)$ we end with a good estimate of $H(1,\phi)$. It can be shown that $\Pr(x|\phi) = H(1,\phi) - Q(1,\phi) + q(x,\phi)$, this gives us an approximation of the likelihood. Note that using Gibbs sampling to sample from the distribution at each temperature is computationally expensive and to get a good estimate requires that small increments be used [14, Sec. 5.6].

For small data sets. Let $x = \{x_1, x_2, ...x_n\}$ using some arbitrary ordering over the elements of x. Let $x_D = (x_{i+1}, x_{i+2}, ..., x_n)$ be the vector of descendants of i and let $x_A = (x_1, x_2, ..., x_{i-1})$ be the ancestors. The likelihood term can be written as:

$$\Pr(x|\phi) = \prod_{i=1}^{n} \Pr(x_i|\phi, x_A) \tag{6}$$

The terms in the left hand product can be approximated by:

$$\begin{array}{c} \Pr(x_i|\phi, x_A) = \\ \sum_{x_D \in K^{n-i}} \Pr(x_i|\phi, x_D, x_A) \Pr(x_D|\phi, x_A) \\ \approx \frac{1}{|S_i|} \sum_{x_D \in S_i} \Pr(x_i|\phi, x_D, x_A) \end{array} \tag{7}$$

where S_i is a set of assignments for $x_D \in K^{n-i}$ randomly sampled from the distribution $\Pr(x_D|\phi, x_A)$. These samples can be obtained by fixing the values of x_A and performing Gibbs sampling on the remaining elements of x. As $|S_i|$ increases the accuracy of the approximation improves. This sampling process is performed for each node and the resulting approximations of $\Pr(x_i|\phi, x_A)$ can be put together (equation 6) to give a value for $\Pr(x|\phi)$.

2.4. Precision of the model parameters

The term $\frac{1}{2}\log(F(\phi))$ in equation 5 is from the MML approximation introduced in [20] and arises because in an optimal code the parameters ϕ need only be stated to finite precision. In [14, Sec. 5.4] the $F(\phi)$ term was discarded because of its relatively low contribution to the message length. For our problem and in general for small data sets this is no longer true.

Calculating the determinant of the matrix $F(\phi)$ directly is impractical. For this term we approximate the likelihood using a pseudo-likelihood function [2, sec. 3.3]:

$$\Pr(x|\phi) \approx \prod_{i \in N} \Pr(x_i|\phi, x_j, j \in \delta_i) \tag{8}$$

The second derivatives of this function with respect to the parameters ϕ can be calculated for given values of x and ϕ. The expectation of these second derivatives can then be approximated by sampling values of x from $\Pr(x|\phi)$.

For this approximation to be useful it is necessary that the second derivatives of the pseudo-likelihood function with respect to the parameters of ϕ are close to the second derivatives of the true likelihood. Note also under some conditions the standard MML approximation from [20] used here does break down. MML approximations other than the one presented in [20] (e.g., Dowe's MMLD/I_{1D} [15, sec. 4.10]) may provide a better solution however they have as yet proven to be too computationally expensive.

3. Comparing models of varying complexity

One significant strength of MML over other methods is its ability to choose between models with different numbers and types of parameters. We will now present a model which is more complex than H2, whose hypotheses we will denote by H3. Note that such models are not uncommon in image analysis and texture analysis. The model parameters are $\phi_3 = (\beta, a_1, a_2, ..., a_{K-1})$ where β is now a $K \times K$ symmetric matrix, where we recall from section 2.1 that K denotes the number of states of the variables x_i. All diagonal entries are forced to be zero for the sake of simplicity. This leaves $K(K-1)/2$ parameters in β to be estimated. The difference between this model and the one presented in section 2.2 and equation (4) is simply that the energies associated with the edges of the graph are now defined as:

$$q_{i,j}(x_i, x_j, \phi_3) = \frac{\beta_{x_i,x_j}}{1 + w_{i,j}^{2.5}} c(x_i, x_j) \qquad (9)$$

That is the energy penalty due to conflicting neighbours now depends on the values of those neighbours according to parameters to be inferred. Similarly to H2 presented in section 2.2 equation 5 the message length for H3 is:

$$\begin{aligned} I_{H3} = &-\log(h(H3)) + \tfrac{(K+2)(K-1)}{4}(\log(\tfrac{1}{12}) + 1) \\ &-\log(h(\phi_3)) + \tfrac{1}{2}\log(F(\phi_3)) - \log(\Pr(x|\phi_3)) \end{aligned} \qquad (10)$$

Near optimal estimates for ϕ_3 can be found as described for H2 and the resulting message length I_{H3} can then be compared with I_{H2} and I_{H1} to decide which model to choose.

4. Preliminary tests on artificial data

To begin we generate artificial data from the models H1, H2 and H3 to see if our criterion can select the correct model given enough data. We generate a random graph of order (number of nodes) S by assigning to each node a two dimensional coordinate randomly selected from a uniform density over a square with side lengths $\sqrt{S}$. Only the $4S$ shortest possible edges (using edge weights $w_{i,j}$ equal to the Euclidean distance between i and j) are included in the graph. The values $x_i \in \{1, 2, ..., K\}$ for each node are then generated by sampling the state configuration x from some chosen model using this graph.

4.1. Comparing H1 and H2

Using data from H1. A random graph generated as described above is used. The data x is sampled from the non-spatial model H1 with uniform distribution $\phi_{true} = (0.25, 0.25, 0.25, 0.25)$ over $K = 4$ states. Parameters for H1 and H2 are then inferred and message lengths calculated for both. For these tests we use ϕ_{true} to denote the parameters of the model used to generate the data while ϕ_1 and ϕ_2 denote the inferred parameters obtained by assuming H1 and H2 respectively. H1 and H2 are assumed to be equally likely a priori. The prior over the parameters of H1 ϕ_1 is uniform $h(\phi_1) = (K-1)!$. The prior over the parameters of H2 $\phi_2 = (\beta, a_1, a_2, ..., a_{K-1})$ is $h(\phi_2) = 0.5(K-1)!$ if $0 < \beta < 2$ else zero. The function $c(x_i, x_j)$ from section 2.2 equation 4 is defined here as 0 if $x_i = x_j$ else 1.

This test was repeated 20 times, with a new graph and assignment for x in each case, using graphs of order (number of nodes) 20, 30, 50 and 80. The averages for each set of twenty runs are recorded in the table below in nits (where 1 nit = $\log_2 e$ bits). The first column shows the data set size, the next three show the average inferred message lengths assuming H1, the three after that show the average inferred message lengths assuming H2. The rightmost two columns show the average difference between the message lengths for H1 and H2 and the number of times that the correct hypothesis was chosen out of 20. We denote the length of encoding the hypothesis by (H), the length of the data given the hypothesis by $(D|H)$ and the message length by $I_H = (H) + (D|H)$.

size	$(H1)$	$(D\|H1)$	I_{H1}	$(H2)$	$(D\|H2)$	I_{H2}	$diff$	$correct$
20	7.0	25.9	32.9	9.0	25.9	34.9	2.0	20
30	7.5	40.1	47.6	8.8	40.0	48.8	1.2	20
50	8.2	67.6	75.8	10.8	67.5	78.3	2.5	20
80	8.9	109.5	118.4	12.0	109.3	121.3	2.9	20

For each test performed here the correct model was selected. In most cases the inferred value of β was close to zero (less than 0.1). In the most extreme case (using 30 nodes) the inferred value was quite high ($\beta = 0.68$).

Using data from H2. This test was performed as in the immediately preceding subsection however, this time the data was generated from H2 with parameters $\phi_{true} = (\beta, a_1, a_2, a_3, a_4) = (0.9, 0.25, 0.25, 0.25, 0.25)$. The priors over the inferred parameters for H1 and H2 are the same as before.

size	$(H1)$	$(D\|H1)$	I_{H1}
20	7.5	20.6	28.1
30	7.9	34.7	43.6
50	8.5	60.9	69.4
80	9.1	100.3	109.4

$(H2)$	$(D\|H2)$	I_{H2}	$diff$	$correct$
8.0	21.8	28.8	0.7	8
9.3	34.5	43.8	0.2	11
9.9	52.2	61.1	-8.3	19
11.1	87.5	98.6	-10.8	20

In most cases for the sets of size 50 and 80 the inferred values of β fell between 0.6 and 1.1. We can see here that for the cases where the set size was 30 or less the simpler explanation was often preferred.

4.2. Comparing H1, H2 and H3

In this test, data was generated from H3 with node states taking three possible values $K = 3$. The parameters ϕ_{true} used to generate the data are $a_1 = a_2 = a_3 = \frac{1}{3}$ and:

$$\beta = \begin{bmatrix} 0 & .2 & 1 \\ .2 & 0 & .2 \\ 1 & .2 & 0 \end{bmatrix}$$

Parameters for H1, H2 and H3 were inferred for the generated data and their message lengths calculated. The prior for H3 over ϕ_3 is uniform over the region where the three non-diagonal entries of β each fall between 0 and 2. The priors for H1 and H2 were as with the previous tests. This test was repeated 10 times using data sets of size 100 and 150. The message length averages are in the table below in the same format as before.

size	$(H1)$	$(D\|H1)$	I_{H1}
100	5.9	101.3	107.2
150	6.3	156.2	162.5

$(H2)$	$(D\|H2)$	I_{H2}
8.6	97.7	106.3
9.3	151.3	160.6

$(H3)$	$(D\|H3)$	I_{H3}	$correct$
11.4	94.6	106.0	4
12.6	144.7	157.3	10

For the size 100 case there seems to be no preference for H2 or H3, the correct model (H3) was selected 4 out of 10 times. The non-spatial model was not selected in any of these runs. As expected with more data (size 150) preference for the true model (H3) increases.

Note also that we could not find any data sets for which H3 was selected for data generated from H1 or H2.

5. Applications to classification

This section describes how our work can be applied to the technique given in [14] for the classification of spatially correlated data. In that paper the positions of the data elements was assumed to lie on a rectangular grid. The second order clique potential parameter β was a single scalar (as with H2) as opposed to the $K \times K$ matrix used in H3. Research on applying our extensions to this sort of MML classification of spatially correlated data is still in progress. We present here an outline of the theory behind it.

Assume now that for our given graph, each node i has associated with it a class $x_i \in \{1, 2, ..., K\}$ and a vector y_i. The form of the vectors y_i depends on the specific problem. As before the discrete-valued elements of x are spatially correlated according to one of the models H1, H2 or H3. Let $\gamma \in \{H_1, H_2, H_3\}$ denote the inferred spatial correlation model. Again the parameters ϕ defining this distribution need to be inferred. However in this problem we can not observe the values x_i directly, instead we wish to infer them based on the observed values y_i. On top of this we are not given the number of classes K and need to infer that as well.

To make things practical it is assumed that $\Pr(y_i|\theta, x, y_j, j \neq i) = \Pr(y_i|\theta, x_i)$. The form of the likelihood function used to model the classes $\Pr(y_i|\theta, x_i)$ will depend on the particular application. Here $\theta = \{\theta_1, \theta_2, ..., \theta_K\}$ denotes the parameters defining the K different classes. It is necessary that MML estimates for θ given y and some assignment for x exist.

An MML code explaining the observed data y in terms of the parameters K, γ, ϕ, x and θ consists of:

1. A statement of the inferred number of classes K.
2. A statement specifying the type of spatial correlation model chosen γ (for H1, H2 and H3).
3. The parameters for the spatial correlation ϕ.
4. The parameters θ defining the K different classes.
5. A statement specifying the class assignments x given K and ϕ.
6. A statement (known as the detail) of the observed data y given the above listed parameters.

The length of this code can be calculated as:

$$\begin{aligned} I = I_1(K) + I_2(\gamma) + I_3(K, \gamma, \phi) + I_4(K, \theta) \\ + I_5(K, \phi, x) + I_6(K, \theta, x, y) - \log K! \end{aligned} \tag{11}$$

where the terms I_1 through to I_6 are the lengths of the message fragments listed above in that order. The term $-\log K!$ appears because the numbers assigned to the K different classes is arbitrary.

The functions I_1 and I_2 depend on our priors over the possible values of K and γ respectively. Note that when $K \leq 2$, H3 does not apply. The sum $I_2 + I_3 + I_5$ is equal to the message lengths form equations 3, 5 and 10 that we have given for H1, H2 and H3 respectively. The functions for I_4 and I_6 depend on how the classes are modelled and we refer the reader to [15] and [19] for examples.

Finding optimal assignments for the parameters K, γ, ϕ, x and θ is a difficult problem. The search algorithm below describes how ϕ, x and θ can be inferred given assumed values K and γ. To infer K and γ this algorithm needs to be repeated for different assignments for them and the resulting message lengths need to be compared.

The justification for this (EM-like) algorithm as a search algorithm for a minimal message length I is complex and we refer the reader to [14, sec. 5.1] for a more thorough explanation.

1. Create some initial set of assignments x.
2. Re-estimate the spatial correlation parameters ϕ.
3. Re-estimate the class parameters θ.
4. Update x by sampling randomly from the distribution $Pr(y|x)Pr(x)$ using Gibbs sampling.
5. If a stable solution is reached then terminate, else return to step 2.

The difference that our extensions make to this algorithm is in step 2. The parameters ϕ here need to be optimized according to equations 3, 5 and 10. Estimating the parameters θ in step 3 depends on the class models used. Standard MML techniques for that problem exist for many interesting class model types [15, 19, 7].

6. Conclusion and further work

We have introduced a criterion for analysing spatial correlation for discrete-valued variables associated with the nodes of a weighted, undirected graph (based on the work of (Wallace 1998) [14] which used rectangular a grid). The general class of graphs for which this can be performed efficiently has not been thoroughly investigated. We have shown how the Minimum Message Length (MML) principle can be used to infer not only the presence of spatial correlation but also how it can select between spatial models of varying complexity. We have refined [14] slightly by including a discarded message length term important to small data sets. While the tests shown here are preliminary they indicate that this criterion is resistant to over-fitting.

We have argued that MML inference is well suited to this problem as it is statistically consistent [6, 15, 5], statistically invariant [17] and capable of comparing models with different numbers of parameters [20, sec. 6][18].

We have outlined here how our work can be applied to the method of unsupervised classification of spatially correlated data introduced in [14]. Finally, further work needs to be done to make this more scalable. This might be achieved by finding classes of spatial models which allow for faster sampling. Another possibility is to improve or replace the numerical approximations to the likelihood term given in section 2.3.

References

[1] J. E. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society*, B36(2):192–236, 1974.

[2] J. E. Besag. Statistical analysis of non-lattice data. *The Statistician*, 24(3):179–195, 1975.

[3] G. J. Chaitin. On the length of programs for computing finite binary sequences. *Journal of the Association of Computing Machinery*, 13:547–569, 1966.

[4] J. W. Comley and D. L. Dowe. Minimum message length and generalized Bayesian nets with asymmetric languages. In P. Grünwald, M. A. Pitt, and I. J. Myung, editors, *Advances in Minimum Description Length: Theory and Applications*, pages 265–294. M.I.T. Press, April 2005.

[5] D. L. Dowe, S. Gardner, and G. R. Oppy. Bayes not bust! Why simplicity is no problem for Bayesians. *British J. Philosophy of Science*, 2007. to appear, forthcoming.

[6] D. L. Dowe and C. S. Wallace. Resolving the Neyman-Scott problem by Minimum Message Length. In *Proc. Computing Science and Statistics - 28th Symposium on the interface*, volume 28, pages 614–618, 1997.

[7] T. Edgoose and L. Allison. MML Markov classification of sequential data. *Statistics and Computing*, 9:269–278(10), November 1999.

[8] S. Geman and D. Geman. Stochastic relaxations, Gibbs distributions and the Bayesian restoration of images. *IEEE Tran. On PAMI*, PAMI-6:721–741, 1984.

[9] G. R. Grimmett. A theorem about random fields. *Bull. London Math. Soc.*, 5:81–84, 1973.

[10] A. N. Kolmogorov. Three approaches to the quantitive definition of information. *Problems of Information Transmission*, 1:1–17, 1965.

[11] S. L. Needham and D. L. Dowe. Message length as an effective Ockham's razor in decision tree induction. In *Proc. 8th Int. Workshop of Artificial Intelligence and Statistics (AISTATS 2001)*, pages 253–260, Key West, FL, 2001.

[12] J. Rissanen. Modelling by the shortest data description. *Automatica*, 14:465–471, 1978.

[13] R. J. Solomonoff. A formal theory of inductive inference. *Information and Control*, 7:1–22,224–54, 1964.

[14] C. S. Wallace. Intrinsic classification of spatially correlated data. *Computer Journal*, 41(8):602–611, 1998.

[15] C. S. Wallace. *Statistical and Inductive Inference by Minimum Message Length.* Springer, 2005.

[16] C. S. Wallace and D. M. Boulton. An information measure for classification. *Computer Journal*, 11:185–195, 1968.

[17] C. S. Wallace and D. M. Boulton. An invariant Bayes method for point estimation. *Classification Society Bulletin*, 3(3):11–34, 1975.

[18] C. S. Wallace and D. L. Dowe. Minimum message length and Kolmogorov complexity. *Computer Journal*, 42(4):270–283, 1999.

[19] C. S. Wallace and D. L. Dowe. MML clustering of multistate, Poisson, von Mises circular and Gaussian distributions. *Statistics and Computing*, 10:73–83, January 2000.

[20] C. S. Wallace and P. R. Freeman. Estimation and inference by compact coding. *Journal of the Royal Statistical Society. Series B*, 49(3):240–265, 1987.