Efficient recursions for general factorisable models

By R. REEVES and A. N. PETTITT

School of Mathematical Sciences, Queensland University of Technology, GPO Box 2434, Brisbane, QLD, 4001, Australia.

r.reeves@qut.edu.au a.pettitt@qut.edu.au

Summary

Let n S-valued categorical variables be jointly distributed according to a distribution known only up to an unknown normalising constant. For an unnormalised joint likelihood expressible as a product of factors, we give an algebraic recursion which can be used for computing the normalising constant and other summations. A saving in computation is achieved when each factor contains a lagged subset of the components combining in the joint distribution, with maximum computational efficiency as the subsets attain their minimum size. If each subset contains at most $r+1$ of the n components in the joint distribution, we term this a lag-r model, whose normalising constant can be computed using a forward recursion in $O(S^{r+1})$ computations, as opposed to $O(S^n)$ for the direct computation. We show how a lag-r model represents a Markov random field and allows a neighbourhood structure to be related to the unnormalised joint likelihood. We illustrate the method by showing how the normalising constant of the Ising or autologistic model can be computed.

Some key words: Autologistic distribution; Gibbs distribution; Ising model; Normalising constant; Partition function; Markov chain Monte Carlo.

1. Introduction

High-dimensional summations occur in finding normalising constants for, and marginalising over, discrete probability distributions. The direct evaluation of these sums becomes intractable for nontrivial problems. We propose an application of forward recursion to these summations for what we call general factorisable models, which make problems of useful size tractable.

A number of techniques have been proposed for estimating the normalising constant (Ogata & Tanemura, 1981; Penttinen, 1984; Younes, 1989; Moyeed & Baddeley, 1991; Geyer & Thompson, 1992; Jerrum & Sinclair, 1993; Gelman & Meng, 1998; Huang & Ogata, 1999; Gu & Zhu, 2001). A direct method for evaluating the autologistic normalising constant for a lattice with cylinder boundary conditions and a small number of rows has also been presented in previous work (Pettitt et al., 2003).

Here we present a direct summation generalised from the principal of forward recursion, commonly applied to the computation of posterior distributions in product form arising from hidden Markov models (Scott, 2002). The technique is applicable to lattices with a small number of rows, up to about 20, and increases in computational time in direct proportion to the number of columns.

We consider a joint probability distribution for discrete random variables $y = (y_1, \ldots, y_n)$ such that the unnormalised probability distribution can be written as a product of terms $q_i(y_i, y_{i+1}, \ldots, y_{i+j})$. We call this a general factorisable model. We find computationally efficient recursions for summing over the state space for such a model, and apply them to the computation of the normalising constant. This factorised definition of the joint probability is fundamentally related to the conditional independence of neighbourhood structures and, with suitable indexing, applies to any discrete Markov random field. The initial motivation

for the development is the study of the autologistic distribution on the lattice.

2. A general factorisable model

2·1. *Recursion for a general factorisable model*

Let a general unnormalised probability distribution for a discrete-valued vector y be given by $q(y)$. Let the components of y be ordered in such a way that

$$q(y) = q_1(y_1, y_2, \ldots, y_{r+1}) q_2(y_2, y_3, \ldots, y_{r+2}) \ldots q_k(y_k, y_{k+1}, \ldots, y_n) \tag{1}$$

is a valid factorisation of $q(y)$, where $r < n$ and $k = n - r$. We call this a lag-r model, so that a lag-0 model would correspond to independent $y_1, \ldots, y_n$. With the notation that the vector $(y_i, y_{i+1}, \ldots, y_j)$ is denoted by y_i^j, the normalising constant Z is given by

$$\begin{aligned} Z &= \sum_y q(y) \\ &= \sum_{y_{k+1}^n} \sum_{y_k} q_k(y_k^n) \sum_{y_{k-1}} q_{k-1}(y_{k-1}^{n-1}) \ldots \sum_{y_1} q_1(y_1^{r+1}) \end{aligned} \tag{2}$$

and this can be evaluated recursively as follows. Let

$$Q_1(y_2^{r+1}) = \sum_{y_1} q_1(y_1^{r+1}), \tag{3}$$

$$Q_t(y_{t+1}^{r+t}) = \sum_{y_t} q_t(y_t^{r+t}) Q_{t-1}(y_t^{r+t-1}), \text{ for } t = 2, \ldots, k. \tag{4}$$

Then the normalising constant is given by

$$Z = \sum_{y_{k+1}^n} Q_k(y_{k+1}^n). \tag{5}$$

Since there are $k = n - r + 1$ recursions in (3) and (4), with $q(\cdot)$ evaluated S^{r+1} times at each, the recursion is performed in $O\{(n - r + 1)S^{r+1}\}$ computations, with an additional $O(S^r)$ computations in the final summation (5), where each y_i can take one of S possible values.

Equation (3) provides the unnormalised marginal likelihood distribution for y_2^n by finding

$$Q_1(y_2^{r+1})q_2(y_2^{r+2})\dots q_k(y_k^n),$$

while $Q_t(y_{t+1}^{r+t})$ from (4) when multiplied by $q_{t+1}(y_{t+1}^{t+r+1})\dots q_k(y_k^n)$ gives the unnormalised marginal distribution for y_{t+1}^n. Note that no probabilistic interpretation need be given to the functions $q_i(\cdot)$ in (1). The method is essentially an algebraic method, and the exact factorisation is arbitrary, though ordering y to minimise the lag r reduces the computational complexity. In exponential models, the factorisation can be easily made by assigning terms to the appropriate functions $q_i(\cdot)$, as we illustrate in § 3 with the autologistic model. The model is motivated if we note that, given y_i^{i+r-1}, then y_1^{i-1} and y_{i+r}^n are conditionally independent, lending the model to specialisation for time series and spatial data. When $r = 1$, these recursions correspond to the so-called forward recursions defined for hidden Markov models, see for example Zucchini & Guttorp (1991), with y_i being the state of a Markov chain with S states, at time step i.

2·2. *A general factorisable model as a Markov random field*

A Markov random field on a set of nodes $\{1,\dots,n\}$ is defined by conditional probabilities for each node that depend only on a subset of the remaining nodes. This subset of the remaining nodes constitutes the neighbourhood of the node. The lag-r general factorisable model defines a Markov random field with the neighbourhood determined by $N_j = \{j - r,\dots,j-1,j+1,\dots,j+r\}$, where N_j is the neighbourhood of node j. For a first-order Markov random field on a rectangular lattice, we require the lag r to be equal to the minimum of either the number of rows or columns. The neighbourhood of a first-order Markov random field is a subset of N_j. Thus a first-order Markov random field on an $m \times n$ lattice is a special case of the more general lag-r model with $r = \min(m,n)$. Similarly, a second-order Markov random field on a rectangular $m \times n$ lattice requires $r = \min(m,n) + 1$, and once again the

neighbourhood is a subset of the full general factorisable model neighbourhood.

The full conditionals are easily picked out from the factorised form of the joint probability,

$$p(y_j|y_{-j}) \propto q(y_{j-r}, \dots y_j) \dots q(y_j, \dots y_{j+r}),$$

and the conditional normalising constant is trivially obtained by the sum over y_j.

2·3. *Backward recursion*

The stochastic backward recursion, see, for example Scott (2002), can be generalised to the general factorisable model, to produce the joint likelihood in terms of a product of conditional probabilities,

$$p(y_1^n) = p(y_{k+1}^n)p(y_k|y_{k+1}^n)p(y_{k-1}|y_k^{n-1}) \dots p(y_1|y_2^{r+1}). \tag{6}$$

Once the normalising constant has been found through forward recursion, the probabilities in the product of conditionals are given by

$$p(y_{k+1}^n) = \frac{1}{Z} Q_k(y_{k+1}^n),$$

$$p(y_i|y_{i+1}^n) = \frac{q_i(y_i^{i+r}) Q_{i-1}(y_i^{i+r-1})}{Q_i(y_{i+1}^{i+r})},$$

for $i = 2, \dots, k$, and

$$p(y_1|y_2^n) = \frac{q_1(y_1^{1+r})}{Q_1(y_2^{r+1})}.$$

This result is obtained by recognising that $p(y_i|y_{i+1}^n)$ is proportional to $p(y_i^n)$, which can be obtained by marginalising $p(y_1^n)$ over y_1^{i-1}. An additional summation over y_i provides the conditional normalising constant, and the product $\frac{1}{Z} q_k(y_k^{k+r}) \dots q_{i+1}(y_{i+1}^{i+1+r})$ cancels from numerator and denominator.

Bartolucci & Besag (2002) also present a recursive algorithm for directly computing the likelihood of a Markov random field, in the form of a product of conditional probabilities. While the Bartolucci & Besag algorithm is similar in spirit to the recursions presented here,

the details are quite different. They employ a conditional probability lemma which allows them to build up the terms of the recursion from the full conditionals. When applied to an autologistic lattice with r rows, the two methods have much the same order of computational complexity, with our method increasing with $O(S^{r+1})$, while Bartolucci & Besag give an upper bound of $O(S^{r+2})$ for their method. This is consistent with Bartolucci & Besag reporting working with autologistic lattices with up to 12 rows or columns, which our algorithm extends by several rows or columns. We have found that normalising constants for autologistic lattices of 20 rows can be computed feasibly with a desktop PC, though, in Markov chain Monte Carlo algorithms where the normalising constant is computed repeatedly, this makes for rather slow iterations. Markov chain Monte Carlo for the parameters of autologistic lattices with 15 rows is however relatively painless.

An advantage of our method is that the full cycle of forward and then backward recursion is not required for normalising constant computations, which require only the forward recursion. Thus the conditional probabilities of (6) do not need to be computed, as they would be if the Bartolucci & Besag algorithm were used for finding the normalising constant. This would be an advantage, for example, in Markov chain Monte Carlo algorithms for autologistic parameter estimation, where the ratio of normalising constants must be evaluated at each proposal for the parameter values. While not significantly altering the computational complexity and hence computation time, it does simplify the programming task, which may be an advantage in these cases.

2·4. *Permuting the index for minimum lag*

The problem of finding the minimum lag index is equivalent to permuting the rows of a sparse matrix in order to concentrate non zero entries around the diagonal. Methods for doing this are well established, including the reverse Cuthill-McKee algorithm (George & Liu, 1981).

As an illustration, suppose we have the product

$$q(y_1, y_7, y_9)q(y_2, y_4, y_8)q(y_3, y_5, y_7)q(y_4, y_6, y_8).$$

This can be represented as a 4×9 matrix, with each row constructed from a function $q(\cdot)$, by placing a 1 in the columns corresponding to the indices of each argument, and 0 elsewhere. For example, the first row corresponding to $q(y_1, y_7, y_9)$ has a 1 in columns 1,7 and 9. Rows of zeros are then added to produce a square matrix. The reverse Cuthill-McKee algorithm, as, for example, implemented in the Matlab function *symrcm*, gives the index permutation $(9 \to 1, 1 \to 2, 7 \to 3, 3 \to 4, 5 \to 5, 2 \to 6, 8 \to 7, 4 \to 8, 6 \to 9)$, resulting in the product with permuted indices

$$q(y_2, y_3, y_1)q(y_6, y_8, y_7)q(y_4, y_5, y_3)q(y_8, y_9, y_7).$$

With the permutation, a lag-8 model has been reduced to a lag-2 model.

By definition of a suitable indexing scheme, optimised in this way, the lag-r general factorisable model can be applied to irregular arrays and neighbourhood structures arising, for example, from polygon regions of a geographical map.

3. Application to the autologistic model on the lattice

Let y be binary with $y_i \in \{-1, 1\}$, and defined on a rectangular lattice, with m rows and n columns. Let the index $i \in \{1, 2, ..., mn\}$ be ordered from top to bottom in each column, from left to right. Then the unnormalised likelihood for the autologistic model is given by

$$q(y|\theta) = \exp\{\theta_0 V_0(y) + \theta_1 V_1(y)\}. \tag{7}$$

For a first-order neighbourhood model defined with free boundaries we define the abundance statistic, V_0, and the association statistic, V_1, as

$$V_0 = \sum_{i=1}^{mn} y_i,$$

$$V_1 = \sum_{j=0}^{n-1} \sum_{i=jm+1}^{(j+1)m-1} y_i y_{i+1} + \sum_{j=0}^{n-2} \sum_{i=jm+1}^{(j+1)m} y_i y_{i+m}. \tag{8}$$

The two terms in the association statistic V_1 are then simply the within-column interactions between neighbours, and the between-column interactions. As a result of the interaction between neighbours in adjacent columns, we note terms of the form $y_i y_{i+m}$ in (8), indicating that a lag of $r = m$ is the minimum possible. The exponentiated terms of (7) are then distributed amongst the functions $q_i(\cdot)$. There is no unique way of doing this, but the exact method is immaterial. We adopt the method illustrated in Fig. 1. The first between-column interaction term from V_1, all the V_1 interaction terms within the first column and all the V_0 terms up to and including y_{m+1} are allocated to $q_1(\cdot)$. Subsequent functions $q_i(\cdot)$ add the additional within-column and between-column terms involving y_{m+i} to the product, and the additional term y_{m+i} from V_0. In the case where subscript i corresponds to the top row of the lattice, there is no within-column term for that particular $q_i(\cdot)$. Then

$$q_1(y_1, y_2, \ldots, y_{m+1}) = \exp\{\theta_0 \sum_{i=1}^{m+1} y_i + \theta_1 \sum_{i=1}^{m-1} y_i y_{i+1} + \theta_1 y_1 y_{m+1}\},$$

$$q_i(y_i, y_{i+1}, \ldots, y_{i+m}) = \exp\{\theta_0 y_{i+m} + \theta_1 (y_{i+m-1} y_{i+m} + y_i y_{i+m})\},$$

for $i = 2, \ldots, mn - m$, except that, when i corresponds to the top row of the lattice, i.e. $i = km + 1$, where $k \in \{1, 2, ...n - 2\}$,

$$q_i(y_i, y_{i+1}, \ldots, y_{i+m}) = q_{\text{top}}(\cdot) = \exp\{\theta_0 y_{i+m} + \theta_1 y_i y_{i+m}\}.$$

The normalising constant is then found by application of (3),(4) and (5), with $r = m$.

4. Discussion

Many datasets used in spatial statistics are small enough to have their normalising constants computed directly by the algorithm we propose, thus eliminating the need to use inefficient approximations such as the pseudolikelihood or importance sampling in maximum likelihood

estimation or Bayesian inference for the autologistic parameters, as in, for example, Huffer & Wu (1998). We expect that, in particular, spatial analysis of binary and categorical data that exhibit spatial clustering of categories will be advanced by our approach. For example, Green & Richardson (2002) apply a hierarchical model based on a hidden Markov random field to epidemiological data. To overcome the problem of computing the normalising constant for a Potts model, they precompute it on a discrete set of parameter values, using the path sampling approach of Gelman & Meng (1998), an approach followed by Low Choy in her 2001 Ph.D. thesis from the Queensland University of Technology. A prior probability distribution is constructed to limit the association parameter to the same discrete set of values. The technique we have proposed could be used to reduce or eliminate the stochastic variability in their normalising constant estimation, either directly or in combination with path sampling along the lines of Friel & Pettitt (2004).

The method we propose can be viewed as a complementary approach for Markov random fields to that of Bartolucci & Besag (2002), defining them in terms of joint probabilities, instead of conditional probabilities. Whereas the approach of Bartolucci & Besag applies, in theory, to any probability model, it presupposes that the full conditionals are compatible with a valid joint distribution. For conditionals that are not derived from a known, though possibly unnormalised, joint distribution, compatibility must be checked; see for example Casella (1996) and Arnold et al. (2001). Indeed, one method for checking compatibility would be to execute the Bartolucci & Besag algorithm for all possible recursive sequences, checking that the same valid joint distribution results in each case. Just such an approach based on similar recursive use of Bartolucci & Besag's Lemma 1, was suggested by Meng (1996). Kaiser & Cressie (2000) consider the question of defining Markov random fields with arbitrary conditionals, and give necessary and sufficient conditions which such conditionals must fulfil. Their method, which relaxes the requirement of positivity, is based on checking

for permutation invariance in the indices of clique associated terms of the Gibbs potential. In either case, checking for compatibility is a nontrivial computational task, avoided at risk of invalid statistical inference from a conditionally specified model.

Our method, starting with the unnormalised joint density, avoids questions of compatibility. However, it provides no advantage without a valid lag-r factorisation. Such a factorisation arises, for example, from the local neighbourhood structure of a Markov random field, and it is the associated reduction in dependence we exploit to produce efficient recursions. While the general factorisable model still applies if cylindrical or toroidal boundary conditions are imposed, the lag of the resulting model after index permutation can be much increased over the free-boundary case. In these cases, the Bartolucci & Besag algorithm may prove more useful.

Finally, we note that the generalised recursions we have proposed are also applicable to marginalisation of hidden partially ordered Markov models (Cressie & Davidson, 1998), which generalise the Markovian dependence structure to a directed acyclic graph.

5. Acknowledgement

Improvements in the work resulted from discussions with Jesper Møller, and anonymous referees to whom we are grateful. We also thank the editor for his constructive remarks.

References

Arnold, B. C., Castillo, E. & Sarabia, J. M. (2001). Conditionally specified distributions: An introduction (with Discussion). *Statistical Science* **16**, 249–74.

Bartolucci, F. & Besag, J. (2002). A recursive algorithm for Markov random fields. *Biometrika* **89**, 724–30.

Casella, G. (1996). Statistical inference and Monte Carlo algorithms (with Discussion). *Test* **5**, 249–344.

Cressie, N. & Davidson, J. (1998). Image analysis with partially ordered Markov models. *Comp. Statist. Data Anal.* **29**, 1–26.

Friel, N. & Pettitt, A. N. (2004). Likelihood estimation and inference for the autologistic model. *J. Comp. Graph. Statist.* To appear.

Gelman, A. & Meng, X.-L. (1998). Simulating normalizing contants: from importance sampling to bridge sampling to path sampling. *Statist. Sci.* **13**, 163–85.

George, A. & Liu, J. (1981). *Computer Solution of Large Sparse Positive Definite Systems.* Englewood Cliffs, N.J: Prentice-Hall.

Geyer, C. J. & Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with Discussion). *J. R. Statist. Soc.* B **54**, 657–99.

Green, P. J. & Richardson, S. (2002). Hidden Markov models and disease mapping. *J. Am. Statist. Assoc.* **97**, 1055–70.

Gu, M. G. & Zhu, H.-T. (2001). Maximum likelihood estimation for spatial models by Markov chain Monte Carlo stochastic approximation. *J. R. Statist. Soc.* B **63**, 339–55.

Huang, F. & Ogata, Y. (1999). Improvements of the maximum pseudo-likelihood estimators in various spatial statistical models. *J. Comp. Graph. Statist.* **8**, 510–30.

Huffer, F. W. & Wu, H. (1998). Markov chain Monte Carlo for autologistic regression models with application to the distribution of plant species. *Biometrics* **54**, 509–25.

Jerrum, M. & Sinclair, A. (1993). Polynomial-time approximation algorithms for the Ising model. *SIAM J. Comp.* **22**, 1087–116.

Kaiser, M. S. & Cressie, N. (2000). The construction of multivariate distributions from Markov random fields. *J. Mult. Anal.* **73**, 199–220.

Meng, X. (1996). Discussion of 'Statistical inference and Monte Carlo algorithms', by G.

Casella. *Test* **5**, 310–8.

Moyeed, R. & Baddeley, A. (1991). Stochastic approximation of the MLE for a spatial point pattern. *Scand. J. Statist.* **18**, 39–50.

Ogata, Y. & Tanemura, M. (1981). Estimation of interaction potentials of spatial point patterns through the maximum likelihood procedure. *Ann. Inst. Statist. Math.* **33**, 315–38.

Penttinen, A. (1984). *Modelling interactions in spatial point processes: parameter estimation by the maximum likelihood method.* Jyvaskyla studies in computer science, economics, and statistics, vol. 7. Jyvaskyla, Finland: Jyvaskylan yliopisto.

Pettitt, A. N., Friel, N. & Reeves, R. (2003). Efficient calculation of the normalising constant of the autologistic and related models on the cylinder and lattice. *J. R. Statist. Soc.* B **65**, 235–46.

Scott, S. L. (2002). Bayesian methods for hidden Markov models: Recursive computing in the 21st century. *J. Am. Statist. Assoc.* **97**, 337–51.

Younes, L. (1989). Parametric inference for imperfectly observed Gibbsian fields. *Prob. Theory Rel. Fields* **82**, 625–45.

Zucchini, W. & Guttorp, P. (1991). A hidden Markov model for space time precipitation. *Water Resour. Res.* **27**, 1917–23.

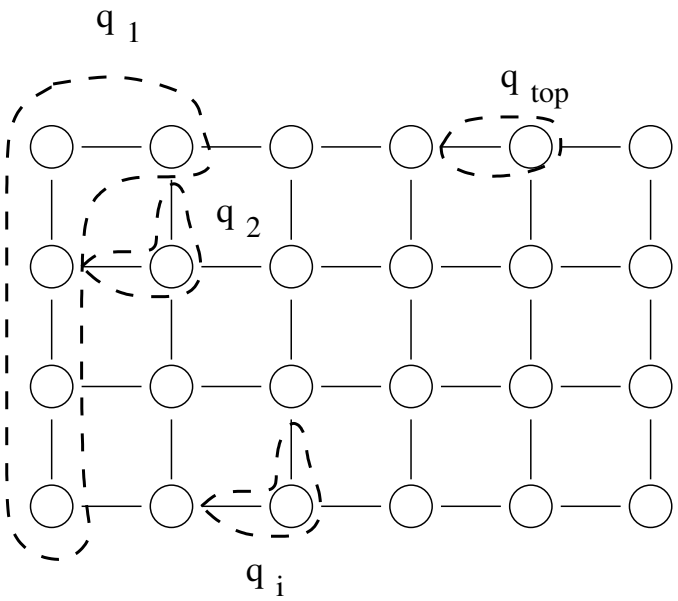

Figure 1: The allocation of the autologistic terms to the factors $q_i(\cdot)$ of the joint distribution. Terms from V_0 corresponding to y_i are shown as a circle. Terms from V_1 corresponding to within-column association, $y_i y_{i+1}$, are shown as vertical lines. Terms from V_1 corresponding to between-column association, $y_i y_{i+m}$, are shown as horizontal lines. Dashed lines show how these terms are grouped into factors.