Spatial Statistics 52 (2022) 100708

Contents lists available at ScienceDirect

Spatial Statistics

journal homepage: www.elsevier.com/locate/spasta

Ordered conditional approximation of Potts models

Anirban Chakraborty [a], Matthias Katzfuss [a,*], Joseph Guinness [b]

[a] Department of Statistics, Texas A&M University, United States of America
[b] Department of Statistics and Data Science, Cornell University, United States of America

ARTICLE INFO

Article history:
Received 19 October 2021
Received in revised form 26 September 2022
Accepted 27 September 2022
Available online 9 October 2022

Keywords:
Categorical data
Distributed computation
Image analysis
Ising model
Spatial grid
Vecchia approximation

ABSTRACT

Potts models, which can be used to analyze dependent observations on a lattice, have seen widespread application in a variety of areas, including statistical mechanics, neuroscience, and quantum computing. To address the intractability of Potts likelihoods for large spatial fields, we propose fast ordered conditional approximations that enable rapid inference for observed and hidden Potts models. Our methods can be used to directly obtain samples from the approximate joint distribution of an entire Potts field. The computational complexity of our approximation methods is linear in the number of spatial locations; in addition, some of the necessary computations are naturally parallel. We illustrate the advantages of our approach using simulated data and a satellite image.

1. Introduction

Potts models can be used to describe spatial grids for which each location falls into one of a fixed number of classes. The Ising model (Ising, 1925) can be viewed as a special case of the Potts model with two classes. Since both of these models were originally developed to describe the magnetization of atoms in crystalline solids, they involve an inverse temperature parameter β that controls the strength of magnetization. The idea of Potts model also coincides with the autologistic models introduced by Besag (1974) for analyzing categorical data on spatial grids. Potts models have been used in a variety of areas beyond statistical physics, including neuroscience (Roudi et al.,

* Corresponding author.
E-mail address: katzfuss@gmail.com (M. Katzfuss).

https://doi.org/10.1016/j.spasta.2022.100708
2211-6753/

2009) and quantum computing (King et al., 2018). The hidden Potts model, in which the model for the data depends on an unobserved Potts model configuration, has also seen multiple applications, including in medical image processing (Li et al., 2017) and satellite image analysis (Moores et al., 2020).

Potts models are computationally challenging, as model probabilities involve a normalizing constant that becomes intractable for large spatial grids. To address this intractability, Besag (1975) introduced a pseudo-likelihood consisting of the product of full conditional distributions, which, due to a cancellation, does not require calculation of the normalizing constant. However, the asymptotic variance of the maximum-pseudo-likelihood estimator is high when the size of the spatial grid tends to infinity (Stoehr, 2017). Pettitt et al. (2003) performed an algebraic simplification of the normalizing constant in the observed Potts model using its Markov property. Although this can work efficiently for small spatial grids, the computational expense increases exponentially as the grid size increases. Several authors have proposed MCMC algorithms and variational methods (as reviewed by Stoehr, 2017) for inference on the parameter β of the observed Potts model. Some of these methods can also be used to sample the spatial grid for given values of the parameter β.

Inference for hidden Potts models is even more involved due to a doubly intractable likelihood. Many works have been dedicated to developing approximate Bayesian computation (ABC) algorithms for speeding up the computation by assuming an approximate structure of the likelihood of a hidden Potts model that follows certain properties. Recently, Moores et al. (2020) have introduced a parametric functional approximate Bayesian (PFAB) algorithm for inference on the temperature parameter in the hidden Potts model, which surpasses the computational efficiency of previous methods by a long margin while preserving accurate inference.

Here, we propose an ordered conditional approximation (OCA) that essentially approximates the joint density of a Potts models as a product of conditional distributions, for which we order the spatial locations by their coordinates and then condition each point on its nearest previously ordered neighbors. Our OCA is motivated by the simpler Vecchia approximation that has been highly successful for speeding up Gaussian-process inference (e.g., Vecchia, 1988; Stein et al., 2004; Datta et al., 2016; Guinness, 2018; Katzfuss and Guinness, 2021; Katzfuss et al., 2020). However, unlike the Gaussian-process inference, where the conditional distributions can be calculated as closed-form multivariate Gaussian distributions, conditional distributions under the Potts model do not yield such closed-form expressions and require computing sums whose cost increases exponentially with the grid size. To overcome this computational issue, we restrict the calculation of the sums in the conditional distributions to a sub-region of the whole grid to construct the OCA. Because of its construction, our OCA can take advantage of distributed computation structures for fast calculations. Additionally, we propose a simple mechanism using our method to draw random samples from the joint distribution of a Potts field for specific values of the parameter. We extend our approach to hidden Potts models, to allow fast evaluation of the marginal likelihood and to enable scalable inference on the hidden configuration of the spatial grids. We also propose a Gibbs sampler as an optional tool for inference in this setting. All OCA inference scales linearly in the total number of spatial locations for fixed tuning parameters.

The remainder of this article is organized as follows. In Section 2, we describe the Potts model and introduce its OCA. In Section 3, we extend our OCA to the hidden Potts model. In Sections 4 and 5, we present numerical results using simulations and a satellite image, respectively. We conclude in Section 6.

2. The Potts model

2.1. Definition

Let $\mathbf{s}_1, \ldots, \mathbf{s}_n$ be a set of locations on a two-dimensional rectangular spatial grid of size $n = n_1 \times n_2$. At each location $\mathbf{s}_i$, we observe one of K classes, $z(\mathbf{s}_i) = z_i \in \{1, \ldots, K\}$. The Potts model assumes that the likelihood of a class label, say k, at a particular location increases with the number of neighboring locations on the grid falling into the same class. This likelihood is governed by an

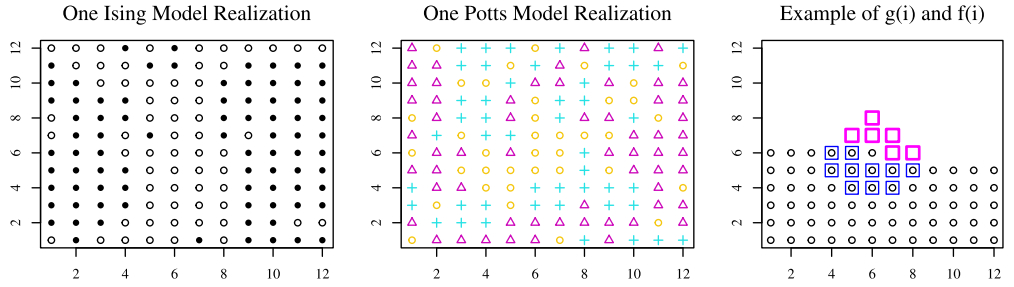


Fig. 1. For a spatial field of size $n = 12 \times 12 = 144$, realizations of a Potts model with inverse temperature $\beta = 0.35$ and $K = 2$ (i.e., Ising model) at the left and $K = 3$ classes in the middle panel. Right panel: For the $i = 66$th location, the points that are previous in lexicographic ordering (circles), and the $m_g = 10$ nearest previously ordered points (blue boxes) and the $m_f = 6$ nearest subsequently ordered points (pink boxes). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

inverse temperature parameter β, which hence controls the strength of the spatial dependence in the grid.

Specifically, the joint density (i.e., the joint probability mass function) for a particular grid configuration $\mathbf{z} = (z_1, \ldots, z_n)^\top$ is given by,

$$p(\mathbf{z}|\beta) = \frac{\exp(-H(\mathbf{z}, \beta))}{N_\beta}, \tag{1}$$

where

$$H(\mathbf{z}, \beta) = \beta S(\mathbf{z})$$

is the Hamiltonian function,

$$S(\mathbf{z}) = \sum_{i \sim j} 1_{\{z_i = z_j\}} \tag{2}$$

is the summary statistics, and

$$N_\beta = \sum_{\mathbf{a} \in \{1,\ldots,K\}^n} \exp(-H(\mathbf{a}, \beta)) \tag{3}$$

is the normalizing constant.

In (2), $i \sim j$ means that s_i and s_j are neighbors. Throughout, we consider a first-order neighborhood structure consisting, for each (interior) pixel, of the four pixels immediately above, below, left, and right. More on the neighbor structure can be found at Stoehr (2017). The Ising model is a special case of the Potts model with $K = 2$ states. See Fig. 1 for examples of realizations of the Ising and Potts models.

For a particular grid configuration, calculation of the joint density in (1) involves evaluating the normalizing constant in (3), which requires summing over the K^n possible states, and is thus computationally infeasible for even moderately large n. Besag (1975) introduced a pseudo-likelihood consisting of the product of full conditional distributions, $\prod_{i=1}^n p(z_i|\mathbf{z}_{-i}, \beta)$, which, due to a cancellation, does not require calculation of the normalizing constant.

2.2. Ordered conditional approximation of the Potts model

To introduce our ordered conditional approximation (OCA), we first assume that the locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$ (and hence the corresponding observations $z_1, \ldots, z_n$ in $\mathbf{z}$) follow a lexicographic ordering according to their coordinates, as illustrated in the right panel of Fig. 1.

Next, we obtain an ordered conditional expression of the joint density (1) as a product of conditional densities:

$$p(\mathbf{z}|\beta) = \prod_{i=1}^{n} p(z_i|\mathbf{z}_{1:i-1}, \beta), \tag{4}$$

where the conditional distributions for the Potts model can be shown to be

$$p(z_i|\mathbf{z}_{1:i-1}, \beta) = \frac{\sum_{\mathbf{z}^*_{i+1:n}\in\{1,\dots,K\}^{n-i}} \exp\big(-H((\mathbf{z}_{1:i-1}, z_i, \mathbf{z}^*_{i+1:n}), \beta)\big)}{\sum_{(z_i^*,\mathbf{z}^*_{i+1:n})\in\{1,\dots,K\}^{n-i+1}} \exp\big(-H((\mathbf{z}_{1:i-1}, z_i^*, \mathbf{z}^*_{i+1:n}), \beta)\big)}. \tag{5}$$

We propose to approximate the expression in (5) by considering subvectors of $\mathbf{z}_{1:i-1}$ and $\mathbf{z}^*_{i+1:n}$. Define $g(i) \subset \{1, \dots, i-1\}$ and $f(i) \subset \{i+1, \dots, n\}$ (i.e., the blue and magenta boxes in Fig. 1) of size $m_g = |g(i)|$ and $m_f = |f(i)|$, respectively. Then define the modified Hamiltonian for $V_i := \{g(i), i, f(i)\}$ as

$$H_i(\mathbf{z}, \beta) = \beta \sum_{\substack{j\sim k \\ j,k\in V_i}} 1_{\{z_j=z_k\}}. \tag{6}$$

Since H_i depends only on a subset of the full set of states $\mathbf{z}$, if we replace H with H_i in (5), the sums over $\{1, \dots, K\}^{n-i}$ and $\{1, \dots, K\}^{n-i+1}$ can be replaced by sums over $\{1, \dots, K\}^{m_f}$ and $\{1, \dots, K\}^{m_f+1}$, yielding

$$\hat{p}(z_i|\mathbf{z}_{1:i-1}, \beta) = \frac{\sum_{\mathbf{z}^*_{f(i)}\in\{1,\dots,K\}^{m_f}} \exp\big(-H_i((\mathbf{z}_{1:i-1}, z_i, \mathbf{z}^*_{i+1,n}), \beta)\big)}{\sum_{(z_i^*,\mathbf{z}^*_{f(i)})\in\{1,\dots,K\}^{m_f+1}} \exp\big(-H_i((\mathbf{z}_{1:i-1}, z_i^*, \mathbf{z}^*_{i+1:n}), \beta)\big)}, \tag{7}$$

which is much less costly to evaluate, as long as m_f is small. Note that while the expressions include $\mathbf{z}^*_{i+1:n}$, the states that are not part of $\mathbf{z}_{f(i)}$ do not enter into H_i, so their values do not matter. In addition, since $\mathbf{z}_{g(i)}$ is a smaller vector than $\mathbf{z}_{1:i-1}$, there are fewer terms in each individual evaluation of H_i, providing further computational speedups. Furthermore, since the states $\mathbf{z}_{1:i-1\setminus g(i)}$ appear in both numerator and denominator in (5), we can ignore their contribution while defining the modified Hamiltonian in (6). A detailed derivation of (7) from (5) can be found in Appendix A. The state vector $\mathbf{z}_{g(i)}$ captures the dependence of individual z_i on its previously ordered nearest neighbors. Although the calculation time is linear in $m_g = |g(i)|$, choosing a large value of m_g will not yield better results, since calculations involving the previously ordered neighbors cancel out from both the numerators and denominators in (7). Thus, a reasonable choice of m_g can be $m_g = 2 \times m_f$.

Thus, the OCA of the Potts density in (4) can be expressed as,

$$\hat{p}(\mathbf{z}|\beta) = \prod_{i=1}^{n} \hat{p}(z_i|\mathbf{z}_{1:i-1}, \beta). \tag{8}$$

Evaluating this expression requires $\mathcal{O}(n(m_f + m_g + 1)K^{m_f})$ time, which is linear in the grid size n. Furthermore, each of the n terms in (8) can be computed independently of the remaining $n-1$ terms. Hence, we can also take advantage of parallel computation for calculating the joint log-likelihood.

Inference on β can proceed by replacing the exact likelihood or density $p(\mathbf{z})$ by the OCA $\hat{p}(\mathbf{z})$, both of which implicitly depend on β. Both maximizing the (log) likelihood or Bayesian inference on β are possible. In contrast to the pseudo-likelihood of Besag (1975), the OCA has the virtue that it converges to the exact likelihood as m_g and m_f increase to n; we show in Section 4.1.1 that the OCA can be more accurate than the pseudo-likelihood even when using very small conditioning sets.

As a brief aside, it has been shown that other orderings (e.g., the maximum-minimum-distance, or maximin ordering) can improve over lexicographic ordering in terms of the accuracy of the Vecchia approximation of Gaussian processes (e.g., Guinness, 2018; Katzfuss and Guinness, 2021). However, we carried out exploratory numerical experiments that indicated that both maximin and reverse maximin ordering was less accurate than lexicographic ordering for the OCA of the Potts model.

2.3. Sampling from the Potts model using OCA

We also provide an algorithm for joint sampling from a Potts field using OCA. Unlike many existing algorithms based on iterative techniques such as Markov chain Monte Carlo (MCMC), our approach directly draws a joint sample from the (approximate) Potts model for the whole grid (i.e., from (8)) in a single iteration. Thus, OCA does not require a large number of iterations to ensure mixing and convergence. For each location i, (7) can be rewritten as,

$$\begin{aligned}\hat{p}(z_i = k|\mathbf{z}_{1:i-1}, \beta) &= \frac{\sum_{\mathbf{z}^*_{f(i)}\in\{1,\dots,K\}^{m_f}} \exp\big(-H_i((\mathbf{z}_{1:i-1}, z_i = k, \mathbf{z}^*_{i+1,n}), \beta)\big)}{\sum_{(z_i^*,\mathbf{z}^*_{f(i)})\in\{1,\dots,K\}^{m_f+1}} \exp\big(-H_i((\mathbf{z}_{1:i-1}, z_i^*, \mathbf{z}^*_{i+1:n}), \beta)\big)} \\ &= \frac{\sum_{\mathbf{z}^*_{f(i)}\in\{1,\dots,K\}^{m_f}} \exp\big(-H_i((\mathbf{z}_{1:i-1}, z_i = k, \mathbf{z}^*_{i+1,n}), \beta)\big)}{\sum_{z_i^*=1}^{K}\sum_{\mathbf{z}^*_{f(i)}\in\{1,\dots,K\}^{m_f}} \exp\big(-H_i((\mathbf{z}_{1:i-1}, z_i^* = k, \mathbf{z}^*_{i+1:n}), \beta)\big)} \\ &= \frac{c_{i,k}(\mathbf{z}_{1:i-1})}{\sum_{k=1}^{K} c_{i,k}(\mathbf{z}_{1:i-1})},\end{aligned}$$

for a K-state Potts model. Thus, sequentially for each $i = 1, \dots, n$, we evaluate $c_{i,k}(\mathbf{z}_{1:i-1})$ for $k = 1, \dots, K$, and then sample z_i from this discrete distribution. Simulation studies to examine this method are performed in Section 4.1.2.

3. The hidden Potts model

3.1. Definition

Let $\mathbf{z}$ denote a spatial grid generated from a Potts model as in Section 2.1, but now assume that $\mathbf{z}$ is hidden (i.e., latent), and instead we observe a vector $\mathbf{y} = (y_1, \dots, y_n)^\top$. We assume that $y_1, \dots, y_n$ are conditionally independent given $z_1, \dots, z_n$:

$$p(\mathbf{y}|\mathbf{z}) = \prod_{i=1}^{n} p(y_i|z_i).$$

Various distributional assumptions for the likelihood $p(y_i|z_i)$ are possible, including a normal distribution (Geman and Geman, 1984), a Poisson distribution (Green and Richardson, 2002), or a multivariate normal distribution for vector-valued y_i. For example, for the normal case, we could assume that

$$y_i \mid \{z_i = k\} \overset{ind}{\sim} \mathcal{N}(\mu_k, \sigma_k^2), \qquad k = 1, \dots, K, \quad i = 1, \dots, n. \tag{9}$$

Some realizations from such a hidden Potts model are shown in Fig. 2.

3.2. OCA of the marginal likelihood

Similar to Section 2.2, we begin by writing the joint marginal density of the data $\mathbf{y}$ (i.e., the integrated likelihood) in an ordered conditional form,

$$p(\mathbf{y}|\beta) = \prod_{i=1}^{n} p(y_i|\mathbf{y}_{1:i-1}, \beta),$$

which can be written as the ratio of joint distributions

$$\begin{aligned}p(y_i|\mathbf{y}_{1:i-1}, \beta) = \frac{p(\mathbf{y}_{1:i}|\beta)}{p(\mathbf{y}_{1:i-1}|\beta)} &= \frac{\sum_{\mathbf{z}^*\in\{1,\dots,K\}^n} p(\mathbf{z}^*|\beta)\prod_{j=1}^{i} f(y_j|z_j^*)}{\sum_{\mathbf{z}^*\in\{1,\dots,K\}^n} p(\mathbf{z}^*|\beta)\prod_{j=1}^{i-1} f(y_j|z_j^*)} \\ &= \frac{\sum_{\mathbf{z}^*\in\{1,\dots,K\}^n} \exp(-H(\mathbf{z}^*, \beta))\prod_{j=1}^{i} f(y_j|z_j^*)}{\sum_{\mathbf{z}^*\in\{1,\dots,K\}^n} \exp(-H(\mathbf{z}^*, \beta))\prod_{j=1}^{i-1} f(y_j|z_j^*)}.\end{aligned}$$

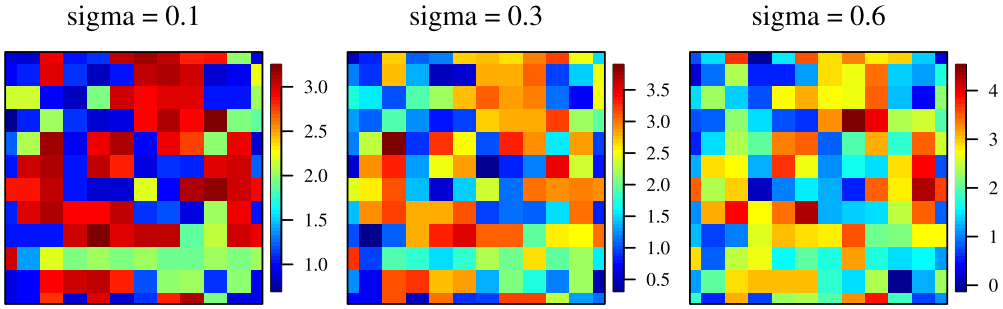


Fig. 2. For a single simulated hidden Potts model configuration with $K = 3$ classes on a grid of size $n = 12 \times 12 = 144$, noisy realizations of a hidden Potts model with a normal likelihood as in (9) with class means $\mu_1 = 1$, $\mu_2 = 2$, $\mu_3 = 3$ and standard deviations $\sigma_1 = \sigma_2 = \sigma_3 = \sigma$. The classes are clearly separated into red, green, and blue for low noise level $\sigma = 0.1$ (left panel), but they are more difficult to distinguish for larger $\sigma = 0.6$ (right). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

As before, we replace the H with H_i, which allows us to sum over a much smaller set,

$$\hat{p}(y_i|\mathbf{y}_{1:i-1}, \beta) = \frac{\sum_{\mathbf{z}^*_{g(i),i,f(i)} \in \{1,\dots,K\}^{m_g+m_f+1}} \exp(-H_i(\mathbf{z}^*, \beta)) f(y_i|z_i^*) \prod_{j \in g(i)} f(y_j|z_j^*)}{\sum_{\mathbf{z}^*_{g(i),i,f(i)} \in \{1,\dots,K\}^{m_g+m_f+1}} \exp(-H_i(\mathbf{z}^*, \beta)) \prod_{j \in g(i)} f(y_j|z_j^*)}.$$

Hence, by virtue of OCA, the final approximated integrated likelihood can be written as,

$$\hat{p}(\mathbf{y}|\beta) = \prod_{i=1}^{n} \hat{p}(y_i|\mathbf{y}_{g(i)}, \beta).$$

The OCA integrated likelihood can be evaluated in $\mathcal{O}(n(m_g + 1)(m_f + m_g + 1)K^{m_f+m_g+1})$ time, which is also linear in n.

3.3. OCA inference on the hidden Potts model

We now consider the joint posterior distribution of a hidden K-state Potts model, which can be written as

$$p(\mathbf{z}|\mathbf{y}, \beta) = \prod_{i=1}^{n} p(z_i|\mathbf{z}_{1:i-1}, \mathbf{y}, \beta) = \prod_{i=1}^{n} p(z_i|\mathbf{z}_{1:i-1}, y_i, \mathbf{y}_{i+1:n}, \beta), \tag{10}$$

since, for each location i, the latent state z_i depends on $\mathbf{y}_{1:i-1}$ only through $\mathbf{z}_{1:i-1}$, and hence z_i is independent of $\mathbf{y}_{1:i-1}$ given $\mathbf{z}_{1:i-1}$.

The conditional distributions in (10) can be written as:

$$\begin{aligned} p(z_i = m \mid \mathbf{z}_{1:i-1}, y_i, \mathbf{y}_{i+1:n}) &= \frac{p(\mathbf{z}_{1:i-1}, z_i = m, y_i, \mathbf{y}_{i+1:n})}{\sum_{z_i^* \in \{1,\dots,K\}} p(\mathbf{z}_{1:i-1}, z_i^*, y_i, \mathbf{y}_{i+1:n})} \\ &= \frac{\sum_{\mathbf{z}^*_{i+1:n} \in \{1,\dots,K\}^{n-i}} \prod_{j=i}^{n} f(y_j \mid z_j^*) p(\mathbf{z}_{1:i-1}, m, \mathbf{z}^*_{i+1:n})}{\sum_{m \in \{1,\dots,K\}} \sum_{\mathbf{z}^*_{i+1:n} \in \{1,\dots,K\}^{n-i}} \prod_{j=i}^{n} f(y_j \mid z_j^*) p(\mathbf{z}_{1:i-1}, z_i^*, \mathbf{z}^*_{i+1:n})}. \end{aligned}$$

However, computation of exact posterior probability is computationally infeasible because of calculating the full Hamiltonian H for all $\mathcal{O}(K^n)$ state configurations.

To overcome this, we again replace H by H_i, and write the approximate likelihood as:

$$\hat{p}(\mathbf{z}|\mathbf{y}, \beta) = \prod_{i=1}^{n} \hat{p}(z_i|\mathbf{z}_{1:i-1}, y_i, \mathbf{y}_{f(i)}, \beta),$$

where

$$\hat{p}(z_i = k|\mathbf{z}_{1:i-1}, y_i, \mathbf{y}_{f(i)}, \beta) = \frac{\sum_{\mathbf{z}^*_{f(i)} \in \{1,\dots,K\}^{m_f}} \prod f(y_j|z^*_j) \exp(-H_i((\mathbf{z}_{1:i-1}, z_i = k, \mathbf{z}^*_{i+1:n}), \beta))}{\sum_{(z_i, \mathbf{z}^*_{f(i)}) \in \{1,\dots,K\}^{m_f+1}} \prod f(y_j|z^*_j) \exp(-H_i((\mathbf{z}_{1:i-1}, z_i, \mathbf{z}^*_{1+1:n}), \beta))}.$$

The modified Hamiltonian H_i will only take the observations corresponding to the index set $V_i = \{g(i), i, f(i)\}$ and hence the remaining observations will not enter the calculation. This approximate posterior probability calculation has a complexity of $\mathcal{O}(n(m_g + 1)K^{m_f+1})$.

3.4. Gibbs sampler for hidden Potts model

We also provide tool for Bayesian inference on hidden Potts models and parameters using the OCA that can proceed using a Gibbs sampler, for which we can sample the hidden field $\mathbf{z}$ from its joint (OCA) full-conditional distribution. Let us consider the following prior distributions for the parameters of a K-state hidden Potts model described in Section 3.1.

$$\mu_j \overset{ind}{\sim} \mathcal{N}(c_j, \sigma^2), \quad \sigma_j^2 \sim \mathcal{IG}(\alpha, \eta), \quad j = 1, \dots, K. \tag{11}$$

In addition, we assume a uniform prior on β on the positive real line. The Gibbs sampler consists of the following steps.

1. Sample $\mathbf{z}$, the hidden field, given $\mathbf{y}$, β, and the μ_j, σ_j, as described in Section 3.3.
2. Update the μ_j, σ_j using the current hidden configuration $\mathbf{z}$, by sampling from the following normal-inverse gamma posterior:

$$\sigma_j^2|\mathbf{z}, \mathbf{y}, \beta \sim \mathcal{IG}(\hat{\alpha}_j, \hat{\eta}_j),$$
$$\mu_j|\mathbf{z}, \mathbf{y}, \beta, \sigma_j \sim \mathcal{N}(\hat{c}_j, \hat{\sigma}_j^2),$$

   where $\hat{\alpha}_j = \alpha + (n_j - 1)/2$, $\hat{\eta}_j = \eta + \frac{1}{2}\sum_{z_i=j}(y_i - \bar{y}_j)^2$, $\hat{c}_j = \frac{n_j\bar{y}_j/\sigma_j^2 + c_j/\sigma^2}{n_j/\sigma_j^2 + 1/\sigma^2}$, $\hat{\sigma}_j^2 = \frac{1}{n_j/\sigma_j^2 + 1/\sigma^2}$.
3. Update β using a Metropolis update based on the following un-normalized pdf:

$$p(\beta|\mathbf{z}) \propto p(\mathbf{z}|\beta)p(\beta),$$

   where $p(\mathbf{z}|\beta)$ is approximated as in (8).
4. Repeat from step 1.

This algorithm has been illustrated with simulation studies in Section 4.2.

4. Numerical experiments

4.1. Simulation for observed Potts model

4.1.1. Parameter estimation using the OCA likelihood

We consider the (directly observed) Potts model and its OCA described in Section 2. We demonstrate the use of the OCA likelihood in (8) for inference on the inverse temperature parameter β.

Fig. 3 shows the log-likelihood for an Ising model (i.e., Potts with $K = 2$) on a spatial field of size $n = 12 \times 12 = 144$ for different conditioning-set sizes m_g and m_f. As the set size increased, the log-likelihood seemed to converge, with differences between the curves decreasing. For $m_f > 4$, the OCAs log-likelihoods (and their maxima) appeared quite accurate.

For a more systematic comparison, we fixed the true value of β at 0.35 and simulated 180 datasets from the Ising model. After that, we obtained estimates of β for each dataset by maximizing both the OCA likelihood and the pseudolikelihood described in Besag (1974). We then compared the quality of these estimates using the root mean squared error (RMSE). As shown in Fig. 4(a), the OCA estimates converged quickly as m_f increased, resulting in a smaller RMSE than for the

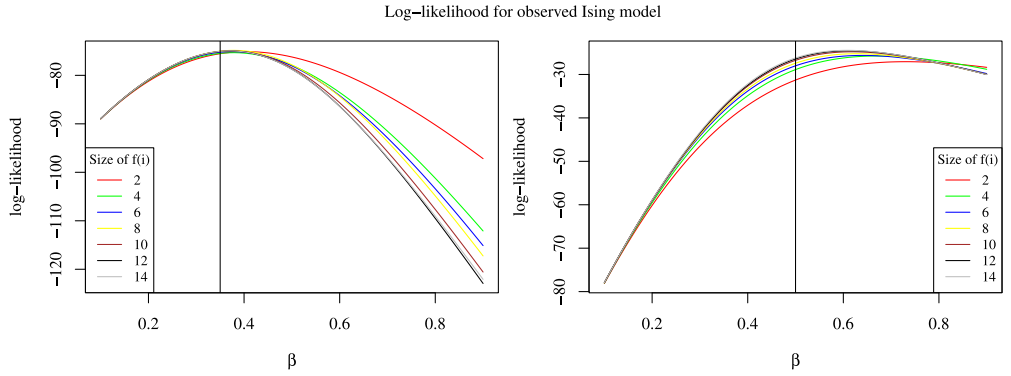


Fig. 3. OCAs of the likelihood $\hat{p}(\mathbf{z}|\beta)$ in (8) as a function of β, for different condition set sizes m_f (represented by different colors) and $m_g = 2m_f$. Data $\mathbf{z}$ was simulated from an Ising model (i.e., $K = 2$) on a 12×12 grid with true β values indicated by vertical lines ($\beta = 0.35$ in the left panel; $\beta = 0.5$ in the right panel). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

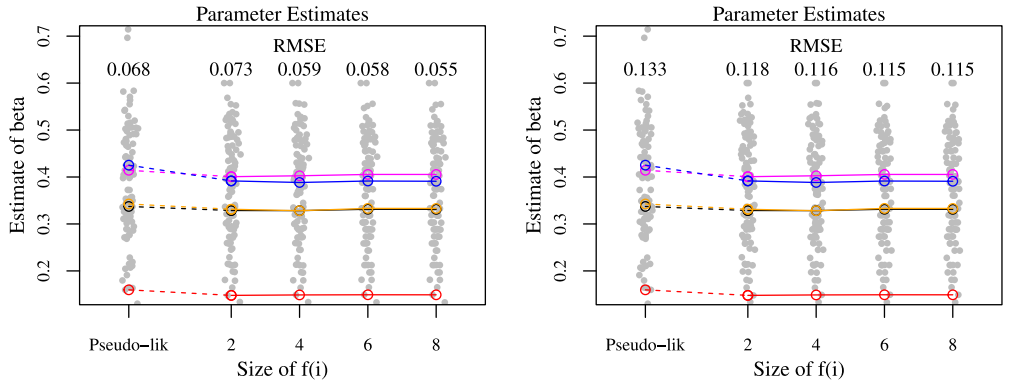


Fig. 4. Parameter estimates from pseudo-likelihood and OCA for each of the 180 simulated datasets on a 12×12 grid with $\beta = 0.35$ from (a) an Ising model (left panel), and (b) a 3-state Potts model (right panel) as gray dots, with estimates for five datasets highlighted in color. OCA converges to the exact likelihood as size of $f(i)$ increases, and so OCA parameter estimates converge to maximum likelihood estimates. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

pseudo-likelihood. The OCA likelihood evaluations were computationally feasible as well; even with $m_f = 10$, OCA Ising likelihood evaluations were obtained in less than 60 s on a laptop.

We repeated the same set of simulations in Fig. 4(b), but instead of using the Ising model, we simulated datasets from a 3-state Potts model using the `PottsUtils` (Feng and Tierney, 2018) package in R. OCA yielded better RMSEs than the pseudolikelihood method.

While we do not have theoretical guarantees for the consistency of our OCA method, we can see from our simulation experiments that the estimators are close to the observed value of β and the RMSE decreases with increasing m_f.

4.1.2. Sampling from observed Potts model

We simulated 60 samples from a 3-state observed Potts model for different values of β using the OCA likelihood on a 50×50 spatial grid. The required steps for sampling have been described in Section 2.3. We also considered several sizes m_f of the future condition sets, while setting $m_g = 2m_f$.

We considered the Swendsen–Wang algorithm (Swendsen and Wang, 1987) for comparison purposes. Since the model configurations are nominal variables that indicate only the class labels and

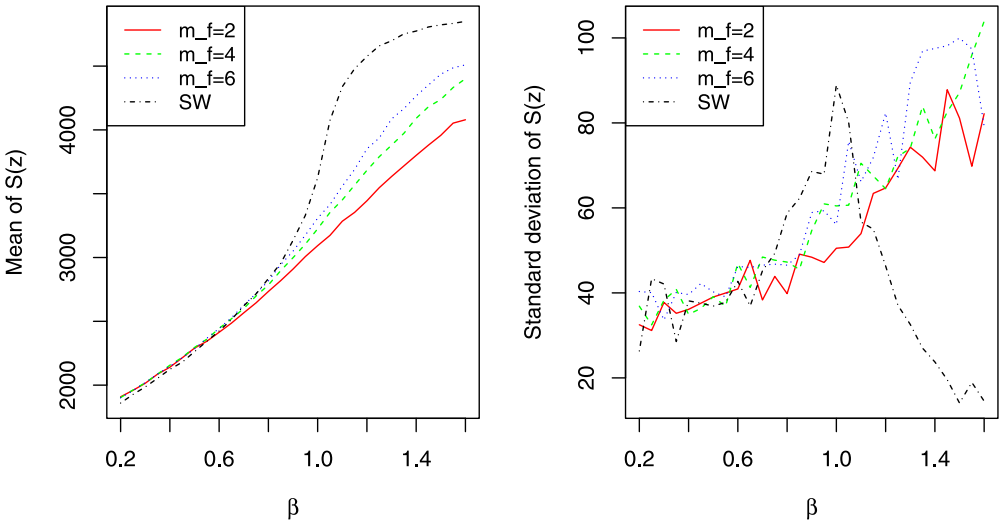


Fig. 5. Mean and standard deviation of summary statistics S(**z**) for 60 observations simulated from a 3-state Potts models on a 50 × 50 spatial grid. m_f denotes different future conditional set sizes, and SW denotes the Swendsen–Wang algorithm.

cannot be compared individually, the summary statistics S(**z**) has been computed. Our simulation results are presented in Fig. 5.

From the plot, it can be concluded that while our method performed well below the critical temperature ($\beta \leq \log(1+\sqrt{K}) \approx 1$), the estimates of both mean and standard deviation for higher values of β can differ from the Swendsen–Wang algorithm more substantially. Yet because of its simple construction, OCA likelihood can be a useful tool when sampling from small values of β. Fixing different values of the future condition set size m_f = 2, 4, and 6, a single draw of random sample from a 3-state Potts model on a 50 × 50 spatial grid on an average required 0.01, 0.1, and 2.4 s, respectively.

4.2. Simulation for hidden Potts model

We followed our assumptions in Section 3.1 to generate realizations from a 3-state hidden Potts model on a 12 × 12 spatial field. We assumed $\mu_j = j$ for all $j = 1, \ldots, K = 3$. Generally speaking, for simulating data from a hidden Potts model, we first simulated the latent Potts field using the `PottsUtils` (Feng and Tierney, 2018) package, and then we simulated the noisy observed data conditional on the Potts field class at each pixel using the `rnorm` function in `R`.

4.2.1. Integrated likelihood

Fig. 6 shows integrated log-likelihoods for hidden Potts models. For data generated from a 3-state and a 5-state Potts model for two different β values (indicated by vertical lines) with noise level $\sigma = 0.25$, we calculated the OCA likelihood of the data at different values of β. The computational expense for calculating the log-likelihood grows exponentially with the size of the conditioning sets. On average, a single evaluation of the integrated likelihood takes 2 s with $m_f = 2$, 4 min with $m_f = 4$, and 10 min with $m_f = 6$ for a 3-state hidden Potts model.

As expected, the error in estimating β increased with the noise level in Fig. 7. While the RMSE was approximately 0.014 for estimating β on a hidden 3-state Potts model with low noise, it increased multiple folds when more noise was added. For $m_f = m_g = 2$, we obtained an estimate of β for the hidden Potts model within 2 min on a laptop. Computation of the likelihood as well as optimization become expensive as m_g increases.

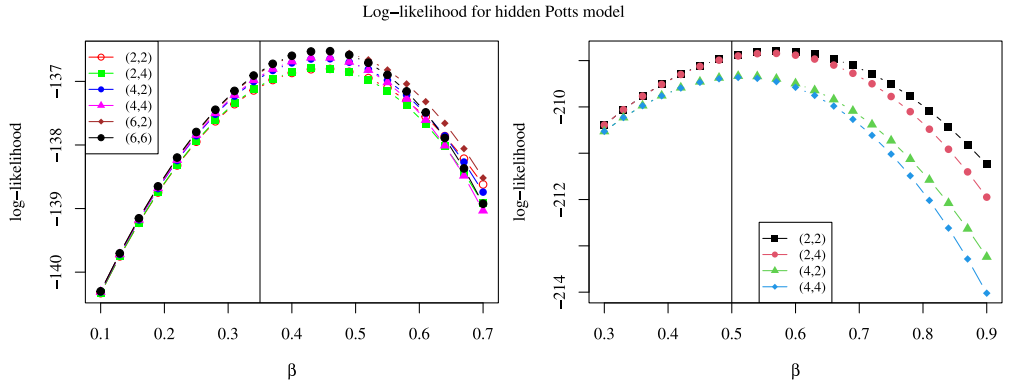


Fig. 6. Log-likelihood of a 3-state (at the left) and a 5-state (at the right) hidden Potts Model on a 12 × 12 grid with $\beta = 0.35, 0.5$, respectively. Log-likelihoods for two different condition set sizes have been plotted through two different colors as well as points. Different colors correspond to different combinations of (m_g, m_f). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

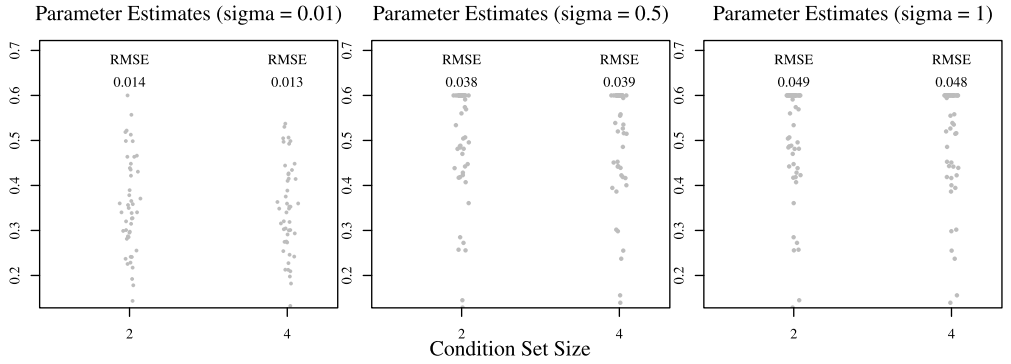


Fig. 7. OCA estimates for β for each of 100 simulated datasets from a hidden 3-state Potts model on a 12 × 12 grid with $\beta = 0.35$. The three panels represent simulation settings with different observation noise standard deviations. Here, $m_g = 2m_f$.

4.2.2. Gibbs sampler for the hidden field

We considered the Gibbs sampler described in Section 3.4 for a 3-state hidden Potts model (i.e., $K = 3$). The data were generated with different noise levels, which can be found in Fig. 2. For parameters of the prior distributions assumed in (11), we considered $\sigma = 0.1$, $\alpha = 1.5$, $\eta = 0.135$, and $c_j = j$ for $j \in \{1, 2, 3\}$.

We ran the Gibbs sampler for 8000 iterations, with the first half considered burn-in. Our numerical results are plotted in Fig. 8, including the highest-posterior-probability (HPP) map, which for each pixel shows the class that had the highest posterior probability. The results show that, even with moderately large noise in data, our algorithm can generally extract the hidden spatial structure. Also, using this Gibbs sampler, we have obtained examples of posterior samples of the Potts model parameter β for one simulated dataset from each noise setting. The posterior densities of β in Fig. 9 contain the true value of β in all of the settings considered.

4.2.3. Comparison with other models

We used the Brier score as a tool of comparison between our method and the finite mixture model of Gaussian distributions (a brief overview of this Gaussian mixture model can be found in

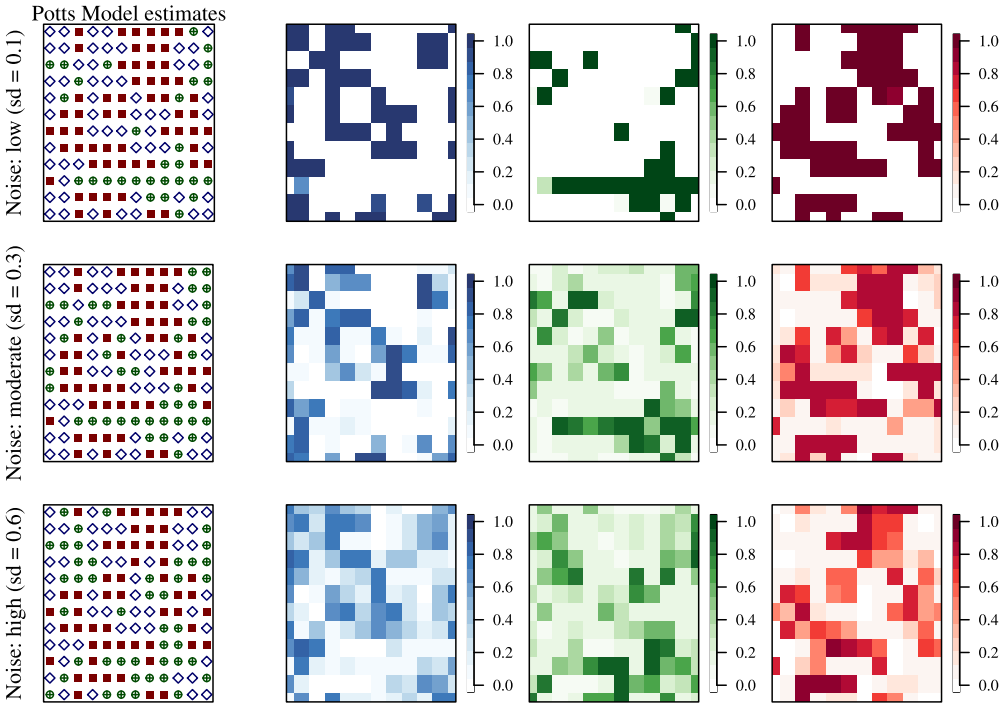
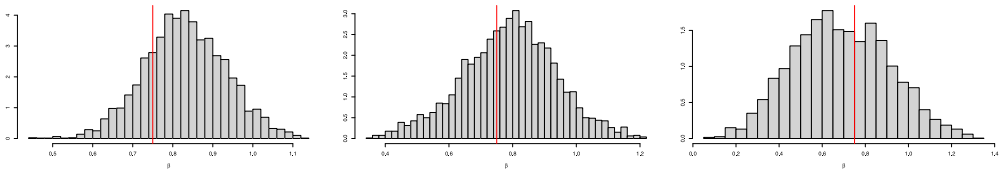


Fig. 8. Inference on the latent field for a 3-state hidden Potts model with three different noise levels (corresponding to the three rows). The observed data are shown in the left panel of Fig. 2. The first column shows the inferred highest-posterior-probability (HPP) map, while the remaining three columns show the posterior probabilities for classes 1, 2, 3, respectively.



(a) Posterior density of β (low noise) (b) Posterior density of β (medium noise) (c) Posterior density of β (high noise)

Fig. 9. For the 3-state hidden Potts model on a 12 $\times$ 12 grid, the posterior density of β for (a): low noise (sd = 0.1), (b): medium noise (sd = 0.3), and (c) high noise (sd = 0.3). The setup is as in Section 4.2.2 and Fig. 8. The true value of the Potts model parameter β is displayed as the solid red vertical line. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Appendix B), and the PFAB algorithm illustrated in Moores et al. (2020). The Brier score measures the accuracy of probabilistic prediction for a set of mutually exclusive discrete outcomes. It is defined by,

$$BS = \frac{1}{N}\sum_{t=1}^{n}\sum_{j=1}^{K}(f_{j,t} - o_{j,t})^2,$$

Table 1
Brier score for three algorithms — Gaussian mixture model (GMM), Ordered conditional approximation (OCA) and Parametric functional approximate Bayesian (PFAB) in three different signal to noise ratio (SNR)'s. σ_j denotes the noise for the jth pixel. In our simulation, we have considered same noise level for all the pixels.

	$\sigma_j = 0.1$	$\sigma_j = 0.3$	$\sigma_j = 0.6$
GMM	0.705	0.687	0.684
OCA	0	0.156	0.560
PFAB	0	0.075	0.328

Table 2
Computation time (in s) of one evaluation of the integrated hidden-Potts likelihood for different data sizes n and different numbers of classes K.

n	$K = 3$			$K = 6$	
	$m_f = 2$	$m_f = 4$	$m_f = 6$	$m_f = 2$	$m_f = 4$
100	0.12	13.62	741.94	5.05	6447.95
10,000	6.23	542.52	36596.41	280	329852.41

where $f_{j,t}$ is probability of forecast in jth class for tth observation, and $o_{j,t}$ is an indicator function which attains the values of 1 if the tth observation truly belongs to class j, and 0 otherwise. Hence, by its construction, Brier score takes on small positive values when there is little mismatch between the forecast and the test data (i.e., when the calibration is good).

Using the results found in Section 4.2.2, we calculated the Brier score for the three different methods. The result is presented in Table 1.

From our simulation results, we can interpret that in all of the cases, our method (OCA) worked well compared to the Gaussian mixture model (GMM). However, it performs poorly with respect to Moores et al. (2020) (PFAB) in terms of calibration. However, the PFAB algorithm requires β to be a scalar random variable. If β is vector-valued (e.g., as in the modified Potts model of Li et al., 2017), the current PFAB algorithm does not work, whereas our method could be extended to such a scenario.

5. Analysis of a satellite image

5.1. Image classification

We applied our method to the `Menteith` data in Robert et al. (2013). This is a 100×100 satellite image of the lake of Menteith, near Stirling, Scotland. Following the description by Robert et al. (2013), our objective is to use the noisy satellite data of the lake and the land bodies surrounding it, to classify each pixel into one of six different classes of surface types, using the Gibbs sampler described in Section 3.4. We have performed a K-means algorithm with six classes on the dataset and used the corresponding class means and standard deviations to find the optimal value of β based on the log-likelihood described in Section 3. The corresponding value was then used as the initial value of β for a Gibbs sampler as described in Section 3.3. While calculation of the Potts model and conditional distribution of the hidden Potts model is extremely fast, computation of the likelihood for the hidden Potts model can be slow due to its high computational complexity, and results in a total of 6 h to find the optimal value of β for $n = 10{,}000$ and $m_f = 2$. Hence, we have restricted m_f and m_g to 2 in Sections 2.2 and 3.3 while calculating the log-likelihood of β and sampling from the posterior distributions to mitigate the computational expense. Table 2 provides likelihood computation times in various scenarios, including the setup considered in this section.

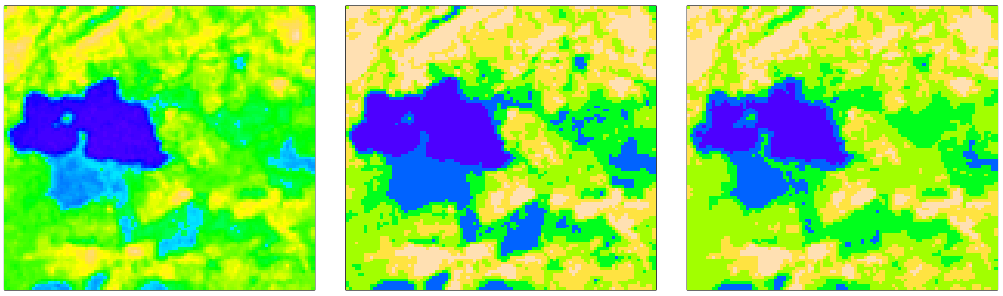


Fig. 10. The leftmost panel displays the original (noisy) Menteith data, the middle panel displays the highest posterior probability configuration of the latent Potts field (based on 50 Gibbs iterations), and the rightmost panel displays a single sample from the latent field.

Since our log-likelihood calculation is able to take advantage of distributed computation, we believe that this calculation can be done faster using a system with higher number of processing cores. A single draw of the random sample from the conditional distribution of hidden Potts model only takes 2 s. All timing results were obtained on a personal laptop with an Intel i5 7th gen CPU and 2 cores.

The original satellite picture and the HPP image are provided in Fig. 10, along with one image sampled from the posterior to give a sense of the posterior uncertainty for the latent field. It is interesting to note that, while our method is able to distinguish between different classes, all the pixels are classified into a single class when we use a Gaussian mixture model (see Appendix B) with six classes or components.

5.2. Prediction for held-out pixels

Using the same setup as in Section 5.1, we effectively held out 10% randomly chosen pixels of the 100×100 spatial grid (i.e., 1000 pixels) by assuming each of them to have a very large (known) standard deviation of 100, which means that the information in these 1000 data points was down-weighted and largely ignored. We then obtained the posterior distribution of the entire latent field as in step 1 in Section 3.4, in effect generating out-of-sample predictions for the "held-out" 1000 pixels.

To overcome the computational limitation on a personal laptop, we considered only 100 MCMC iterations. After performing steps 1 and 2 of Section 3.4 in each iteration, we consider the previously chosen pixels, and draw 100 normal samples using the predicted means and standard deviations of the corresponding classification labels for each of those pixels that we get from step 2. For example, if one of the 1000 pixels has a prediction as label 3, then for that particular pixel we draw 100 normal samples using μ_3 and σ_3 that we get from Step 2 of the Gibbs sampler in Section 3.4. Then, we consider continuous rank probability score (CRPS) to measure the difference between the originally observed value of the pixels and the 100 samples. Given a probability distribution $\mathbf{F}$ and a sample y, the CRPS is given by,

$$CRPS(y, \mathbf{F}) = \int_{-\infty}^{\infty} (\mathbf{F}(z) - \mathbf{1}\{y \leq z\})^2 dz.$$

We repeat this process of randomly selecting pixels and perform following steps 10 times, and take average CRPS of the whole process. We do the same experiment using the Gaussian mixture model (GMM).

From our experiment, the Gibbs sampler method using the OCA approach yields an average CRPS of 5.43, while the GMM yielded a much higher average CRPS of 20.36. From the definition of CRPS,

it is clear that low CRPS implies that the absolute difference between the original distribution and the empirical distribution of the samples are closely situated, which means better prediction. Using this fact, we can conclude in this section that, when less information about the pixels are present (which also means high standard deviation), OCA performs better than GMM in terms of prediction.

6. Conclusions

We introduced an ordered conditional approximation (OCA) of the likelihood of both the Potts model and hidden Potts model. OCA does not require calculation of the intractable normalizing constant. We have also devised a simple algorithm to sample from the observed Potts model using OCA and numerically shown that it can be used to sample from the Potts model for a range of values of the parameter β. Our method could straightforwardly be extended to higher spatial dimensions (i.e., more than 2 coordinates) because of its simple structure of calculating multiple conditional likelihoods (refer to Section 2.2). A Gibbs sampler for inference on the spatial configuration and the parameter β has also been suggested. We have tested these methods on both simulated data and real observations.

Inference using our methods requires linear time in n, total number of locations in the spatial grid. We have used root mean squared error (RMSE) to show superiority of our method over pseudo-likelihood. Additionally, we have used the Brier score to show better consistency of our method compared to a Gaussian mixture model. As calculation of the likelihood slows down with increasing conditioning-set size, inference becomes infeasible for large datasets in personal laptops. As a side-note, OCA should work well if applied to the algorithm developed by Moores et al. (2020); however, because of its redundancy, we have not performed this step.

One of the possible future directions of this work might be developing theoretical properties of OCA based on varying conditional set size. Schäfer et al. (2021) have provided some theoretical bounds of conditional likelihood approximations in the context of Gaussian processes, but, to the best of our knowledge, the same results cannot be applied directly in the context of Potts model. Another direction is to extend our approach to the modified Potts models, where the parameters vary based on the class labels (Li et al., 2017). As the latter involves more than one parameter, gradient-based optimization may be implemented to obtain the maximum likelihood estimators.

Acknowledgments

MK and JG were partially supported by National Science Foundation (NSF), USA Grant DMS–1953005. MK was also partially supported by NSF, USA Grants DMS–1654083 and CCF–1934904.

Appendix A. Calculation of Potts model likelihood

Here our objective is to approximate the conditional likelihood described in (5). It is to be noted that, for each fixed i, calculation of the conditional likelihood takes approximately $\mathcal{O}(K^{n-i})$ unit of time, which is infeasible for small values of i. For this reason, we truncate the calculation of our conditional likelihoods up to a subset of future indices for every i. So, we define, $f(i) \subset \{i+1, \ldots, n\}$ (i.e., the magenta boxes in Fig. 1) of size $m_f = |f(i)|$. Then, we define the first modified Hamiltonian as,

$$H_i^*(\mathbf{z}, \beta) = \beta \sum_{\substack{j \sim k \\ j,k \in \{1,\ldots,i,f(i)\}}} 1_{z_j = z_k}. \tag{12}$$

Here we consider the first-order nearest neighborhood structure; that is, $1_{z_j=z_k}$ will take the value 1 only when $j \in \{k - n_2, k - 1, k + 1, k + n_2\}$. A closer look in Fig. 1 suggests that the modified Hamiltonian in (12) can be written as,

$$H_i^*(\mathbf{z}, \beta) = H_i'(\mathbf{z}, \beta) + H_i(\mathbf{z}, \beta),$$

where $H_i'(\mathbf{z},\beta) = \sum_{\substack{j\sim k \\ j,k\in\{1,\ldots,i-1\}\setminus g(i)}} 1_{z_j=z_k} + \sum_{\substack{j\sim k \\ k\in g(i), j\in\{1,\ldots,i-1\}\setminus g(i)}} 1_{z_j=z_k}$, and $H_i(\mathbf{z},\beta) = \sum_{\substack{j\sim k \\ j,k\in V_i=\{g(i),i,f(i)\}}} 1_{z_j=z_k}$ for suitable choice of $g(i)\subset\{1,\ldots,i-1\}$ with $m_g = |g(i)|$, since the locations with indices in f(i) will only have nearest neighbors from the locations with indices in $g(i)$ (the blue boxes in Fig. 1). Finally we can approximate the conditional likelihood as,

$$\begin{aligned}
\hat{p}(z_i|\mathbf{z}_{1:i-1},\beta) &= \frac{\sum_{\mathbf{z}^*_{f(i)}\in\{1,\ldots,K\}^{m_f}} \exp\big(-H_i^*((\mathbf{z}_{1:i-1}, z_i, \mathbf{z}^*_{i+1,n}),\beta)\big)}{\sum_{(z_i^*,\mathbf{z}^*_{f(i)})\in\{1,\ldots,K\}^{m_f+1}} \exp\big(-H *_i ((\mathbf{z}_{1:i-1}, z_i^*, \mathbf{z}^*_{i+1:n}),\beta)\big)} \\
&= \frac{\sum_{\mathbf{z}^*_{f(i)}\in\{1,\ldots,K\}^{m_f}} \exp\big(-(H_i'((\mathbf{z}_{1:i-1}, z_i^*, \mathbf{z}^*_{i+1:n}),\beta) + H_i((\mathbf{z}_{1:i-1}, z_i^*, \mathbf{z}^*_{i+1:n}),\beta)),\beta\big)}{\sum_{(z_i^*,\mathbf{z}^*_{f(i)})\in\{1,\ldots,K\}^{m_f+1}} \exp\big(-(H_i'((\mathbf{z}_{1:i-1}, z_i^*, \mathbf{z}^*_{i+1:n}),\beta) + H_i((\mathbf{z}_{1:i-1}, z_i^*, \mathbf{z}^*_{i+1:n}),\beta))\big)} \\
&= \frac{\sum_{\mathbf{z}^*_{f(i)}\in\{1,\ldots,K\}^{m_f}} \exp\big(-H_i((\mathbf{z}_{1:i-1}, z_i^*, \mathbf{z}^*_{i+1:n}),\beta)\big)}{\sum_{(z_i^*,\mathbf{z}^*_{f(i)})\in\{1,\ldots,K\}^{m_f+1}} \exp\big(-H_i((\mathbf{z}_{1:i-1}, z_i^*, \mathbf{z}^*_{i+1:n}),\beta)\big)},
\end{aligned} \tag{13}$$

since the first part of the modified Hamiltonian H_i' is constant in both numerator and denominator. The final expression in (13) yields the approximation stated in (7).

Appendix B. Gaussian mixture model (GMM)

Suppose we have n observations $\mathbf{y} = (y_1, y_2, \ldots, y_n)$, and the corresponding state vector $\mathbf{z} = (z_1, z_2, \ldots, z_n)$ that denotes individual class labels. We assume that the observations follow the following distribution.

$$y_i \mid \{z_i = k\} \overset{ind}{\sim} \mathcal{N}(\mu_k, \sigma_k^2), \qquad k = 1,\ldots,K, \quad i = 1,\ldots,n,$$

where K is the maximum number of unique states.

The goal is to make inference about the class labels and the corresponding parameters. However, we do not have the opportunity to directly observe the class labels. Instead we only observe $\mathbf{y}$. For this reason, we consider some priors on the class labels and the parameters using a framework called the Gaussian mixture model (GMM).

First, we consider the following priors for the individual means and standard deviations:

$$\mu_j \overset{ind}{\sim} \mathcal{N}(c_j, \sigma^2), \quad \sigma_j^2 \sim \mathcal{IG}(\alpha,\eta), \quad j = 1,\ldots,K.$$

The above distributions have followed the same prior structure with the Bayesian framework illustrated in Section 3.4. However, here we do not assume that $\mathbf{z} \sim \text{Potts}(\beta)$. Instead, we assume that,

$$z_i|\pi_1,\ldots\pi_K \overset{iid}{\sim} \textit{multinominal}(\pi_1,\ldots,\pi_K), \quad i = 1,\ldots,n.$$

where $\Pi = \{\pi_1,\ldots\pi_K\}$ denote the individual class probabilities. We assume that the class probabilities are not known either (since we do not have any information about the class labels, as discussed earlier), and so we consider a Dirichlet prior on Π. We assume,

$$\Pi \sim \textit{Dirichlet}(\alpha_1,\ldots,\alpha_K),$$

where $\alpha_1,\ldots,\alpha_K$ are the tuning parameters that control the mean and variances of corresponding parameters. We have fixed $\alpha = \frac{1}{K}$ as the tuning parameter in the numerical simulations. For example, $\alpha = \frac{1}{3}$ in Section 4.2.3 and $\frac{1}{6}$ in Section 5. The posterior distributions for the GMM follows.

$$\begin{aligned}
\Pi|\mathbf{y},\mathbf{z} &\sim \textit{Dirichlet}(\alpha_1+n_1,\ldots\alpha_K+n_K), \\
\sigma_j^2|\mathbf{z},\mathbf{y} &\sim \mathcal{IG}(\hat{\alpha}_j,\hat{\eta}_j), \\
\mu_j|\mathbf{z},\mathbf{y},\sigma_j &\sim \mathcal{N}(\hat{c}_j,\hat{\sigma}_j^2), \\
z_i|\mathbf{y},\mu_j,\sigma_j &\overset{ind}{\sim} p(z_i = k|\mathbf{y},\mu_j,\sigma_j), i = 1,\ldots,n,
\end{aligned}$$

where $n_j = \sum_{i=1}^{n} 1_{z_i=j}$, $j = 1, \ldots, n$, $\hat{\alpha}_j = \alpha + (n_j - 1)/2$, $\hat{\eta}_j = \eta + \frac{1}{2}\sum_{z_i=j}(y_i - \bar{y}_j)^2$, $\hat{c}_j = \frac{n_j \bar{y}_j/\sigma_j^2 + c_j/\sigma^2}{n_j/\sigma_j^2 + 1/\sigma^2}$, $\hat{\sigma}_j^2 = \frac{1}{n_j/\sigma_j^2 + 1/\sigma^2}$. More details of this Gaussian mixture model can be found, for example, in Rasmussen et al. (1999, Sec. 2).

Appendix C. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.spasta.2022.100708.

References

Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems. J. R. Stat. Soc. Ser. B Stat. Methodol..

Besag, J., 1975. Statistical analysis of non-lattice data. J. R. Stat. Soc.. Ser. D (the Statistician) 24 (3), 179–195.

Datta, A., Banerjee, S., Finley, A.O., Gelfand, A.E., 2016. Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. J. Amer. Statist. Assoc. 111 (514), 800–812.

Feng, D., Tierney, L., 2018. PottsUtils: Utility functions of the Potts models. R package version 0.3-3.

Geman, S., Geman, D., 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. PAMI-6 (6), 721–741.

Green, P.J., Richardson, S., 2002. Hidden Markov models and disease mapping. J. Amer. Statist. Assoc. 97 (460), 1055–1070.

Guinness, J., 2018. Permutation and grouping methods for sharpening Gaussian process approximations. Technometrics 60 (4), 415–429.

Ising, E., 1925. Beitrag zur theorie des ferromagnetismus. Z. Für Phys. 31 (1), 253–258.

Katzfuss, M., Guinness, J., 2021. A general framework for Vecchia approximations of Gaussian processes. Statist. Sci. 36 (1), 124–141.

Katzfuss, M., Guinness, J., Gong, W., Zilber, D., 2020. Vecchia approximations of Gaussian-process predictions. J. Agric. Biol. Environ. Stat. 25 (3), 383–414.

King, A.D., Carrasquilla, J., Raymond, J., Ozfidan, I., Andriyash, E., Berkley, A., Reis, M., Lanting, T., Harris, R., Altomare, F., Boothby, K., Bunyk, P.I., Enderud, C., Fréchette, A., Hoskinson, E., Ladizinsky, N., Oh, T., Poulin-Lamarre, G., Rich, C., Sato, Y., Smirnov, A.Y., Swenson, L.J., Volkmann, M.H., Whittaker, J., Yao, J., Ladizinsky, E., Johnson, M.W., Hilton, J., Amin, M.H., 2018. Observation of topological phenomena in a programmable lattice of 1,800 qubits. Nature 560 (7719), 456–460.

Li, Q., Yi, F., Wang, T., Xiao, G., Liang, F., 2017. Lung cancer pathological image analysis using a hidden Potts model. Cancer Inform. 16, 1176935117711910, PMID: 28615918.

Moores, M., Nicholls, G., Pettitt, A., Mengersen, K., 2020. Scalable Bayesian inference for the inverse temperature of a hidden Potts model. Bayesian Anal. 15 (1), 1–27.

Pettitt, A.N., Friel, N., Reeves, R., 2003. Efficient calculation of the normalizing constant of the autologistic and related models on the cylinder and lattice. J. R. Stat. Soc. Ser. B Stat. Methodol. 65 (1), 235–246.

Rasmussen, C.E., et al., 1999. The infinite Gaussian mixture model. In: NIPS, Vol. 12. Citeseer, pp. 554–560.

Robert, C.P., Dauphine, U.P., Marin, J.-M., 2, U.M., 2013. Bayess: Bayesian essentials with R. R package version 1.4.

Roudi, Y., Tyrcha, J., Hertz, J., 2009. Ising model for neural data: Model quality and approximate methods for extracting functional connectivity. Phys. Rev. E - Stat., Nonlinear, Soft Matter Phys. 79 (5).

Schäfer, F., Katzfuss, M., Owhadi, H., 2021. Sparse cholesky factorization by Kullback-Leibler minimization. SIAM J. Sci. Comput. 43 (3), A2019–A2046.

Stein, M.L., Chi, Z., Welty, L., 2004. Approximating likelihoods for large spatial data sets. J. R. Stat. Soc. Ser. B Stat. Methodol. 66 (2), 275–296.

Stoehr, J., 2017. A review on statistical inference methods for discrete Markov random fields. arXiv:1704.03331.

Swendsen, R.H., Wang, J.-S., 1987. Nonuniversal critical dynamics in Monte Carlo simulations. Phys. Rev. Lett. 58, 86–88.

Vecchia, A., 1988. Estimation and model identification for continuous spatial processes. J. R. Stat. Soc. Ser. B Stat. Methodol. 50 (2), 297–312.