Bayesian Parameter Estimation in Ising and Potts Models: A Comparative Study with Applications to Protein Modeling

Xiang Zhou and Scott C. Schmidler*
Department of Statistical Science
Duke University

December 6, 2009

Abstract

Ising and Potts models are discrete Gibbs random field models originating in statistical physics, which are now widely used in statistics for applications in spatial modeling, image processing, computational biology, and computational neuroscience. However, parameter estimation in these models remains challenging due to the appearance of intractable normalizing constants in the likelihood. Here we compare several proposed approximation schemes for Bayesian parameter estimation, including multiple Monte Carlo methods for approximating ratios of normalizing constants based on importance sampling, bridge sampling, and recently proposed perfect simulation methods. On small lattices where exact recursions can be used for comparison, we evaluate the accuracy and rate of convergence for these methods, and compare to a pseudo-likelihood based method. We conclude that a pseudo-likelihood approximation to the posterior performs surprisingly well, and is the only method that scales to realistic-size problems. We demonstrate this approach for statistical protein modeling, and compare the results on a protein fold recognition experiment, where it significantly outperforms knowledge-based statistical potentials based on the 'quasi-chemical approximation' commonly used in structural bioinformatics.

* *Corresponding author:* Scott C. Schmidler, Department of Statistical Science, Duke University, Durham, NC 27708-0251. Tel: (919) 684-8064; Fax: (919) 684-8594; Email: schmidler@stat.duke.edu

1 Introduction

Ising and Potts models are discrete Gibbs random fields first developed in statistical physics as models for ferromagnetism (Brush, 1967; Potts and Domb, 1952), and now widely used in statistics for applications in spatial modeling, image processing, and computational neuroscience among others (Banerjee et al., 2004; Geman and Geman, 1984; Green and Richardson, 2002; Hopfield, 1982). Here we consider their use for problems in computational structural biology. Protein structure prediction remains a central question in structural biology, and a key challenge is finding energy functions whose minima correspond to proteins' native states (Anfinsen, 1973). Currently, the most successful approaches rely on 'knowledge-based' or statistical potentials derived from large datasets, either by a "quasi-chemical" approximation (Miyazawa and Jernigan, 1985; Sippl, 1993), or by maximizing the native structure stability among an ensemble of alternative structures (Maiorov and Crippen, 1992; Mirny and Shakhnovich, 1996). Here we consider model-based statistical inference for such potentials using a generative, Potts-type model.

Estimating the parameters of Ising and Potts models, including our protein model, is notoriously difficult by likelihood-based methods due to appearance of an intractable normalizing constant in the likelihood. A variety of approximate methods and computational schemes have been proposed, but it remains unclear which to prefer for practical use. For certain graphs and small lattices, the normalizing constant can be calculated exactly using standard graphical model recursions (Lauritzen and Spiegelhalter, 1988; Reeves and Pettitt, 2004), but these algorithms take time exponential in the size of the graph for lattices.

The most commonly used methods therefore involve pseudolikelihood estimation (Besag, 1974, 1975), or Monte Carlo approximation of the likelihood (Geyer, 1991, 1992; Geyer and Thompson, 1992, 1995). Neither is considered ideal, as the pseudo-likelihood approximation is known to introduce non-negligible bias, and Monte Carlo approximation of normalizing constants is notoriously difficult and computationally expensive. However pseudo-likelihood methods are asymptotically consistent (Mase, 2000; Comets, 1992; Guyon and Kunsch, 1992), and thus practical adequacy depends on empirical rate of convergence. Similarly, significant progress has been made on Monte Carlo methods for normalizing constants (see e.g. Meng and Wong (1996); Gelman and Meng (1998) for overviews). In addition, Møller recently proposed an intriguing approach for posterior sampling without direct calculation of the normalizing constant in the Ising model (Møller et al., 2006) which relies on the ability to generate exact samples from the model. Ising models are one of the few cases where this can be achieved in practice (Propp and Wilson, 1996).

Finally, in the computational biology and protein biophysics literature it is common to use a "quasi-chemical approximation" or pairing frequency estimate for determining parameters of a knowledge-based statistical potential based on Boltzmann's law (Miyazawa and Jernigan, 1985; Sippl, 1993). However, theoretical properties of this estimator do not appear to have been studied. In Section 4.4 we show that it appears to be an inconsistent estimator.

This paper is organized as follows. In Section 2 we review the Ising and Potts models, and introduce a generative model for protein structure using a Potts-like potential. In Section 3, we review several parameter estimation methods proposed in the literature, as well as some relevant Monte Carlo methods. In Section 4, we compare three Bayesian parameter estimation methods and two point estimators by a simulation study, where the true model is known. In Section 5, we apply the conclusions to our protein model, and test the performance of the resulting estimated energy function in a protein fold recognition experiment.

2 Preliminaries

2.1 Ising and Potts Models

Let $X = \{X_1, \ldots, X_m\}$ be a set of random variables taking values $X_j \in \{1, \ldots, k\}$ in a discrete set of 'colors', and let $G = (X, E)$ be an undirected graph with edge set $E \subset X \times X$. A Gibbs random field (GRF) is a joint distribution over $\mathcal{X} = \mathbb{Z}_k^m$ defined by a potential U:

$$P(x) = Z^{-1} \mathrm{e}^{-\frac{1}{k_B T} U(x)}$$

where $Z = \sum_{x \in X} \mathrm{e}^{-\frac{1}{k_B T} U(x)}$ is normalizing constant or partition function. A *Potts model* is a GRF having potential U with respect to the neighborhood system G of the form:

$$U(x) = -\sum_{i \sim j} J_{ij} \delta(x_i, x_j) - \sum_i H_i x_i$$

Here $i \sim j$ denotes neighbors $(i, j) \in E$ in the graph, δ is Kronecker's delta, and the H's and J's are marginal and interaction parameters of the model, called the external magnetic field and coupling constants respectively. When $J_{ij} \equiv J$ and $H_i \equiv H$, we have a *homogeneous* Potts model. The special case of $k = 2$ is called the *Ising model*; labeling the colors $\{1, -1\}$

and taking $J' = J/2$ then gives:

$$U(x) = -J' \sum_{i \sim j} x_i x_j - H \sum_i x_i.$$

The case $J \geq 0$ is called the *ferromagnetic* Ising (or Potts) model, and $J < 0$ *anti-ferromagnetic*.

2.2 Protein Model

For modeling protein sequence-structure relationships, we adopt a slight generalization of the $k = 20$ color Potts model. (Each color represents one of the 20 naturally-occurring amino acids.) Let $s = (s_1, \ldots, s_n)$ be the amino acid sequence of a protein and c the protein conformation. Denote by S and C the space of possible sequences and conformations, respectively. Let $E_c(s)$ be the energy of conformation c taking sequence s. We assume a generative model of sequence given structure, which specifies that the sequence follows a Boltzmann law:

$$p(s \mid c) = Z_c^{-1} \mathrm{e}^{-\frac{E_c(s)}{k_B T}} \tag{1}$$

where $Z_c = \sum_{s \in S} \mathrm{e}^{-\frac{E_c(s)}{k_B T}}$ is the partition function and k_B is Boltzmann's constant.

The generative model (1) can be viewed as a *protein design* model, with sequences chosen to generate a desired conformation according to their corresponding free energies (Shakhnovich, 1994; Meyerguz et al., 2004; Fromer and Yanover, 2008). It has been argued that this model also reflects the evolutionary process, which selects sequences to adopt a given functional conformation (Meyerguz et al., 2004; Fromer and Yanover, 2008; Kleinman et al., 2006). Indeed, studies have shown that the native protein sequences are close to optimal for their structures (Kuhlman and Baker, 2000), and that this model can be used to select sequences which fold into a desired structure (Kuhlman et al., 2003). Formally it is sometimes justified by use of the more widely accepted Boltzmann law in the reverse direction (Seno et al., 1996; Micheletti et al., 1998), which gives the probability of s taking conformation c at temperature T:

$$p(c \mid s) = \frac{1}{Z_s} \mathrm{e}^{-\frac{E_s(c)}{k_B T}}$$

where $E_s(c)$ is the energy of sequence s adopting conformation c and $Z_s = \sum_{c \in C} \mathrm{e}^{-\frac{E_s(c)}{k_B T}}$, and then applying Bayes rule assuming sequences $s \in S$ are uniformly likely *a priori*, and that

$Z_s \approx Z$ for all $s \in S$.

Our potential is a slight generalization of the $k = 20$ color Potts model, given by

$$E_c(s) = \sum_{j \sim i} \theta(s_i, s_j) + \sum_i \mu(s_i)$$

where s_i is the ith amino acid in s, and θ is a 20×20 matrix of contact energies, with $\mu(s_i)$ the marginal energetic preference for amino acid s_i. The neighborhood graph is defined by spatial proximity in the 3D conformation c. We define contacts via distance between C_β atoms as suggested by a comparative study (Melo et al., 2002), with two amino acids considered to be in contact if their C_β distance is less than 8Åand they are not neighbors in sequence. (A virtual C_β is created for Gly by Procrustes superimposition of (N,C,C_α,C_β) positions from a standard Alanine (Maiorov and Crippen, 1992; Thomas and Dill, 1996a).)

We assume θ symmetric, and constrain $\theta(Ala, Ala) = 0$ and $\mu(Ala) = 0$ for identifiability in parameter estimation. (For the fold recognition experiment we instead constrain $\sum_{i=1}^{210} \theta_i = 0$ and $\sum_{i=1}^{20} \mu_i = 0$, for reasons discussed in Section 5.) We set $k_B T = 1$.

3 Parameter Estimation

Both Ising and Potts models are examples of Gibbs random fields, exponential families where the normalizing constant is a Laplace transform than cannot be calculated exactly for large graphs. More generally, when the likelihood involves an unknown normalization:

$$f(y \mid \theta) = q_\theta(y)/Z_\theta \qquad \text{where} \qquad Z_\theta = \sum_y q_\theta(y)$$

then likelihood-based inference is challenging. Both likelihood maximization (Besag, 1974; Geyer and Thompson, 1992) and calculation of a Bayesian posterior:

$$\pi(\theta \mid y) \propto q_\theta(y)\pi_0(\theta)/Z_\theta$$

under prior $\pi_0(\theta)$ are challenging, due to the appearance of the unknown Z_θ in the likelihood. Although for Ising and Potts models on small graphs the normalizing constant can be calculated directly, for even moderate size graphs this quickly becomes infeasible.

3.1 Bayesian Estimation via Monte Carlo Methods

Monte Carlo sampling from the posterior distribution is a standard tool for Bayesian computation. However, Monte Carlo methods typically require large numbers of evaluations of the posterior density function, which is significantly complicated by the appearance of the unknown Z_θ. For example, Metropolis-Hastings methods for sampling $\pi(\theta \mid y)$ using a proposal distribution $p(\theta' \mid \theta)$ require computation of the acceptance ratio:

$$H(\theta' \mid \theta) = \frac{q_{\theta'}(y)\pi_0(\theta')Z_\theta p(\theta \mid \theta')}{q_\theta(y)\pi_0(\theta)Z_{\theta'}p(\theta' \mid \theta)}$$

which involves the ratio of normalizing constants $Z_\theta/Z_{\theta'}$.

In this section we consider three alternatives for addressing this problem: (i) Approximate evaluation of the ratio $Z_\theta/Z_{\theta'}$ via Monte Carlo methods; (ii) An auxiliary variable method proposed by Møller that avoids evaluation of $Z_\theta/Z_{\theta'}$; and (iii) Sampling from a "pseudoposterior" based on a pseudolikelihood approximation of the likelihood which does not involve Z_θ. The practicality and effectiveness of these different approaches is studied through simulation experiments in Section 4, followed by application to the protein problem in Section 5.

3.1.1 Estimating Ratios of Normalizing Constants by Importance Sampling

A direct approach to the problem is to approximate the normalizing constants Z_θ by Monte Carlo sampling whenever the likelihood involving $f(y \mid \theta)$ needs to be evaluated. That is, the normalizing constant is estimated by:

$$\hat{Z}_\theta = \frac{1}{B}\sum_{t=1}^{B}\frac{q_\theta(x^{(t)})}{g(x^{(t)})} \qquad \text{for } x^{(1)},\ldots,x^{(B)} \sim g(x) \tag{2}$$

for some distribution $g(x)$ satisfying $\text{supp}(g) \supseteq \text{supp}(q_\theta)$ from which samples $x^{(t)}$ can be drawn efficiently. Note that this relies on $g(x)$ being normalized itself; if $g(x)$ is also unnormalized then (2) estimates Z_θ/Z_g instead, and therefore (2) is often called the importance sampling estimator (see e.g. Meng and Wong (1996)). This approach was explored for likelihood maximization by Geyer (Geyer, 1991, 1992; Geyer and Thompson, 1992, 1995), and has also been used for Bayesian inference in state-space models. The variance of the estimator (2), and therefore the number of samples needed, is determined by variance of the ratio q_θ/g. Therefore g should be chosen to be as close to the target distribution as possible;

a convenient choice is another member of the parametric family, i.e. $g = f(x \mid \tilde{\theta})$ for some $\tilde{\theta}$.

Using (2), the ratio of two normalizing constants needed for a Metropolis acceptance can be estimated by:

$$\frac{Z_\theta}{Z_{\theta'}} \approx \frac{\hat{Z}_\theta}{\hat{Z}_{\theta'}} = \frac{\frac{1}{B_1}\sum_{t=1}^{B_1} \frac{q_\theta(x_1^{(t)})}{q_{\tilde{\theta}}(x_1^{(t)})}}{\frac{1}{B_2}\sum_{t=1}^{B_2} \frac{q_{\theta'}(x_2^{(t)})}{q_{\tilde{\theta}}(x_2^{(t)})}} \tag{IS1}$$

where $x_1^{(1)}, \ldots, x_1^{(B_1)}$ and $x_2^{(1)}, \ldots, x_2^{(B_2)}$ are samples from $f(x \mid \tilde{\theta})$; taking g the same for both θ and θ' avoids the requirement that g itself be normalized. Since $f(x \mid \tilde{\theta})$ needs to be close to both $f(x \mid \theta)$ and $f(x \mid \theta')$, we take $\tilde{\theta}$ to be the maximum pseudolikelihood estimate (see Section 3.1.3) as in IS3 below.

Alternatively, taking $g(x) = f(x \mid \theta')$ in (2) gives the estimator:

$$\frac{Z_\theta}{Z_{\theta'}} \approx \frac{1}{B}\sum_{t=1}^{B} \frac{q_\theta(x^{(t)})}{q_{\theta'}(x^{(t)})} \tag{IS2}$$

where the $x^{(t)}$ are draws from $p(x \mid \theta')$.

Finally, the inverse normalizing constant can be approximated instead:

$$\frac{1}{\hat{Z}_\theta} = \frac{1}{B}\sum_{t=1}^{B} \frac{g(x^{(t)})}{q_\theta(x^{(t)})}$$

using draws from $f(x \mid \theta)$ itself, again requiring g normalized. This yields yet another approximation of the ratio:

$$\frac{Z_\theta}{Z_{\theta'}} \approx \frac{\hat{Z}_\theta}{\hat{Z}_{\theta'}} = \frac{\frac{1}{B_1}\sum_{t=1}^{B_1} \frac{q_{\tilde{\theta}}(x_1^{(t)})}{q_{\theta'}(x_1^{(t)})}}{\frac{1}{B_2}\sum_{t=1}^{B_2} \frac{q_{\tilde{\theta}}(x_2^{(t)})}{q_\theta(x_2^{(t)})}} \tag{IS3}$$

where $x_1^{(t)}$ are draws from $f(x \mid \theta')$ and $x_2^{(t)}$ are draws from $f(x \mid \theta)$. Møller (Møller et al., 2006) suggests taking $\tilde{\theta}$ to be the maximum pseudolikelihood estimate (see Section 3.1.3). IS3 is a special case of a *bridge sampling* estimate (Meng and Wong, 1996), with $\alpha(x) = \frac{q_{\tilde{\theta}(x)}}{q_\theta(x) q_{\theta'}(x)}$ and is a multiple-sample version of Møller's method described in the next section.

3.1.2 Møller's Auxiliary Variable Method

The above methods for Monte Carlo approximation of the ratio of normalizing constants is computationally very demanding, requiring many samples and needing to be repeated for each iteration of the Metropolis-Hastings algorithm on $\pi(\theta \mid y)$. Recently, Møller proposed an alternative method for sampling such posterior distributions using an auxiliary variable method which avoids this approximation, which is applicable when the model itself can be sampled from exactly (Møller et al., 2006).

To sample from the posterior $\pi(\theta \mid y) \propto \pi_0(\theta) f(y \mid \theta)$ where $f(y \mid \theta) = \frac{1}{Z_\theta} q_\theta(y)$, Møller introduces an auxiliary variable x with conditional distribution $g(x \mid \theta, y)$, and constructs a Metropolis-Hastings chain on with target distribution $\pi(\theta, x \mid y) \propto \pi_0(\theta) g(x \mid \theta, y) \frac{1}{Z_\theta} q_\theta(y)$. This chain proposes a new state (θ', x') jointly by drawing θ' from $p(\theta' \mid \theta, x)$ and then x' from $f(x' \mid \theta') = \frac{1}{Z_{\theta'}} q_{\theta'}(x')$. Then the problematic normalizing constants cancel in the acceptance ratio:

$$H(\theta', x' \mid \theta, x) = \frac{g(x' \mid \theta', y)\pi_0(\theta')q_{\theta'}(y)q_\theta(x)p(\theta \mid \theta', x')}{g(x \mid \theta, y)\pi_0(\theta)q_\theta(y)q_{\theta'}(x')p(\theta' \mid \theta, x)} \tag{3}$$

This method then produces Markov chain with stationary distribution exactly $\pi(\theta, x \mid y)$, requiring no approximation of the acceptance ratio. An extension to this approach has been given by Murray et al. (2006)

As pointed out by Møller, (3) multiplies the acceptance on the original state by $\frac{q_\theta(x)/g(x|\theta,y)}{q_{\theta'}(x)/g(x'|\theta',y)}$ which can also be viewed as approximating the ratio $Z_\theta / Z_{\theta'}$ using two single-sample importance sampling estimates. Note that this provides a biased estimate of the ratio, a fact often considered negligible for large sample sizes but potentially important for single samples. This connection motivates consideration of (IS3) as a potential improvement.

Exact Sampling with Bounding Chains Møller's method requires the ability to generate exact samples from $f(x \mid \theta)$. For the ferromagnetic Ising model, this can be achieved by *coupling from the past* (CFTP) (Propp and Wilson, 1996), and this idea has been extended to the homogeneous Potts model using *bounding chains* (Huber, 2004). The latter constructs a bounding chain y with state space $(2^k)^m$ along with the original chain x on k^m, such that $x_i^{(-t)} \in y_i^{(-t)}$ ensures $x_i^{(-t+1)} \in y_i^{(-t+1)}$ for all i and time t. Then we need only consider the possible transitions of x from states bounded by $y^{(-t)}$, to obtain $y^{(-t+1)}$. For $-t$ such that $|y_i^{(0)}| = 1$ for all i, the chains have coalesced and $x^{(0)} = y^{(0)}$ is an exact sample. A straightforward extension of Huber's algorithm to the non-homogeneous Potts model is given in Algorithm 1. As with CFTP, the chain is constructed starting backwards in time

at $-T$, until T is large enough to ensure coalescence at time zero.

The above algorithms are constructed from an underlying heat-bath (or Gibbs sampler) Markov chain, which is known to exhibit critical slowing down at certain values of J/T (the *critical temperature*). Because Møller's method involves embedding this CFTP in the loop of the larger MCMC sampling many values of J (or in our case, the matrix θ), the efficiency is a concern. In particular, if the posterior distribution has mass near the critical values of J, the perfect sampling steps will take exponentially long and thus so will the overall Markov chain. We explore this issue via simulation studies in Section 4.1.

Algorithm 1 Exact sampling from non-homogeneous k-color Potts model on m node graph.

$T = 1$, $u_t = \text{runif}(T, 0, 1)$ and $t_m = \max_{i\in\{1,\cdots,m\};a,b\in\{1,\cdots,k\}}\{\sum_{i\sim j}\theta(a,b)\}$
while $|y_i| \neq 1$ for some i **do**
  for $t = -T$ to 0 **do**
    Choose $i \in \{1, \cdots, m\}$ randomly, let N_i be the neighbors of i, let $y_i = \emptyset$
    repeat
      Choose $c \in \{1, \cdots, k\}$
      Let $b_c = \sum_{j\in N_i, j:y_j=\{c_j\}}\theta(c, c_j)$
      Let $\min_c = \sum_{j\in N_i, j:|y_j|>1}\min_{c_w\in y_j}[\theta(c, c_w)]$ and $\max_c = \sum_{j\in N_i, j:|y_j|>1}\max_{c_w\in y_j}[\theta(c, c_w)]$
      if $u_t \leq \mathrm{e}^{b_c+\max_c - t_m}$ **then**
        $y_i = y_i \cup \{c\}$
      end if
    until $u_t \leq \mathrm{e}^{b_c+\min_c - t_m}$
  end for
  $T = 2T$ and $u_t = (\text{runif}(T, 0, 1), u_t)$
end while

3.1.3 Pseudolikelihood Approximation of Likelihood Function

When a maximum likelihood estimator is not easy to obtain on a graphical model, a general alternative is the use of a maximum pseudolikelihood estimator (MPLE) (Besag, 1974, 1975). The pseudolikelihood function is given by the product of local conditional densities, and for a k color Potts model on a lattice of m nodes takes the form:

$$L_p(\theta) = \prod_{i=1}^{m} p(x_i \mid x_j : j \sim i) = \prod_{i=1}^{m} \frac{\mathrm{e}^{-(\sum_{i\sim j} J_{ij}\delta(x_i,x_j)+H_i x_i)}}{\sum_{x=1}^{k}\mathrm{e}^{-(\sum_{i\sim j} J_{ij}\delta(x,x_j)+H_i x)}} \tag{4}$$

Pseudolikelihood estimation is convenient because it avoids the difficulties of evaluating normalizing constants, making maximization straightforward. Although pseudolikelihood

estimators are biased, they have been shown to be asymptotically unbiased and consistent (Comets, 1992; Guyon and Kunsch, 1992; Jensen and Kunsch, 1994; Mase, 2000; Baddeley and Turner, 2000).

Here we also consider using the pseudolikelihood function in place of the likelihood function for defining a "pseudoposterior", approximating the posterior distribution by

$$\pi(\theta \mid y) \propto L(\theta)\pi_0(\theta) \approx L_p(\theta)\pi_0(\theta) \tag{5}$$

This circumvents the intractable normalizing constant problem, replacing the computationally difficult likelihood function with the simpler pseudolikelihood function. We have not found much discussion of this approach in the literature, but it seems an obvious approach and we highly doubt we are the first to use it.

3.2 'Quasi-chemical' Approximation for Protein Contact Potentials

A point estimator commonly used for protein contact potentials is the pairing frequency estimator, or pair-wise statistical potential. This knowledge-based statistical potential is generated by a "quasi-chemical" approximation using Boltzmann's law (Miyazawa and Jernigan, 1985; Sippl, 1993). It is derived by assuming that all amino acids are in a gas phase, and by the Boltzmann relation, the free energy between amino acid i and j is given by:

$$w(i,j) = -k_B T \log(\frac{\rho_{ij}}{\rho_{ij}*})$$

where ρ_{ij} is the pairing frequency of amino acids i and j, and $\rho_{ij}*$ is the pairing frequency at the reference state, which depends on the concentration of each amino acid. Let m_{ij} be the number of pairs of amino acid i and j, $m = \sum_{i,j} m_{ij}$ be the total number of pairs, $n_i = \sum_j m_{ij}$ be the number of amino acid i, and $n = \sum_i n_i$ be the total number of amino acids. Again taking $k_B T = 1$, we have:

$$\hat{\rho}_{ij} = \frac{m_{ij}}{m} \qquad \hat{\rho}_{ij}* = \begin{cases} \frac{n_i n_j}{n^2} & i \neq j \\ \frac{n_i n_j}{2n^2} & i = j \end{cases}$$

4 Numerical Results

Before comparing the three Bayesian posterior parameter estimation methods described in Section 3.1, we first explore two intermediate questions: the runtime of the exact sampling step (Section 3.1.2), and the number of samples needed in the importance sampling approximations (Section 3.1.1).

4.1 Runtime of Exact Sampling with Bounding Chains

Application of Møller's method (Section 3.1.2) to our protein model requires the generation of exact samples using the bounding chains CFTP algorithm (Algorithm 1). Because a typical protein has a few hundred contacts in its neighborhood graph, we first performed a simulation experiment to investigate the efficiency of this algorithm for reasonable lattice sizes. Simulations were run on an Intel Pentium Dual-Core Processor E2140, with code written in C.

Fig. 1a shows the mean runtime for generating exact samples from a homogeneous Ising model on a 10×10 lattice, as a function of interaction parameter, estimated from 10 samples at each value. We see that the runtime is reasonable for parameters in the range $J \in (-0.60, 0.60)$, but gets exponentially large outside this range. As described above, this is to be expected as the exact sampling is based on the heat bath algorithm, and so also suffers from critical slowing down.

Fig. 1b shows the mean runtime curve for a $k = 20$ color homogeneous Potts model, which corresponds to our protein model with $\theta(i, i) \equiv \theta$ and $\theta(i, j) = 0$ for $i \neq j$. Again, 10 exact samples were generated for various θ's. We see that the runtime is fast in the range $-1.8 \leq \theta \leq 1.5$, and again grows exponentially outside this range. Generating an exact sample took as long as 7min for $\theta \approx -1.93$ and 5.5 min for $\theta \approx 1.55$.

4.2 Comparison of Importance Sampling Ratio Approximations

To determine the number of samples required to accurately estimate the ratio $Z_\theta / Z_{\theta'}$ by importance sampling (Section 3.1.1), we performed a simulation experiment to compare the three schemes given by equations (IS1), (IS2), and (IS3) (hereafter called IS1, IS2, and IS3, respectively). Using each method, we estimated the ratios $Z_{0.10}/Z_{0.15}$, $Z_{0.20}/Z_{0.25}$, and $Z_{0.30}/Z_{0.35}$ for a homogeneous Ising model on a 10×10 lattice. These ratios were chosen because: (i) perfect samples are easily generated for θ in this range (Fig. 1); and (ii) when using these methods in the context of parameter estimation MCMC algorithms, the newly

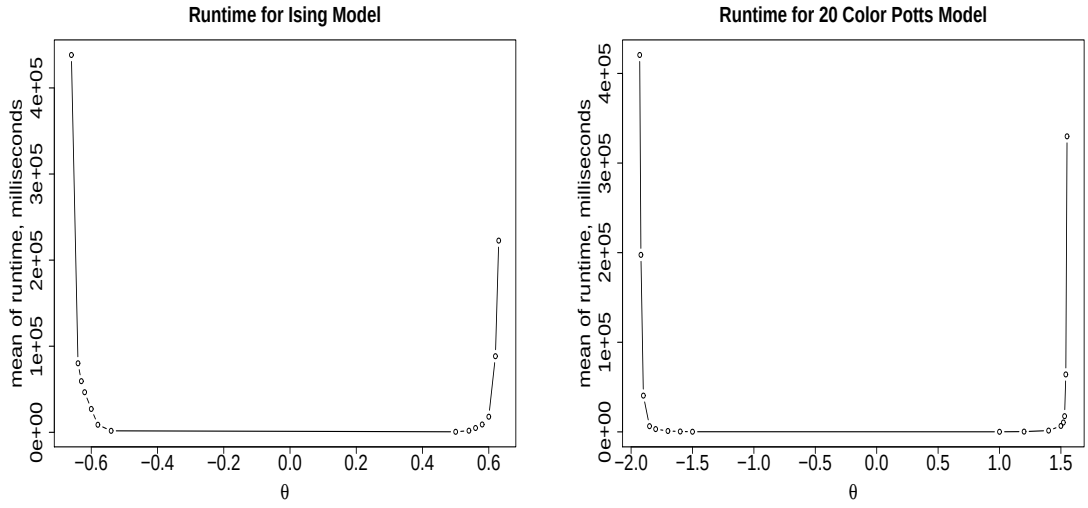


Figure 1: Expected runtime of exact sampling as a function of interaction strength for (a) 10×10 Ising model, and (b) 20 color Potts model.

proposed θ' will typically not be too far from the current θ. For methods IS1 and IS3, we use $\tilde{\theta} = \theta_1 - 0.05$ to estimate $Z_{\theta_1}/Z_{\theta_2}$, again because in the context of parameter estimation the use of the MPLE means $\tilde{\theta}$ will typically be near but not equal to the posterior mode.

For each method, we repeatedly estimated the ratio $Z_\theta/Z_{\theta'}$ using independent Monte Carlo samples of sizes 10, 100 and 1000. Fig. 2 shows boxplot of the resulting estimates, compared to the exact value which can be computed for a lattice of this size by recursive methods (Reeves and Pettitt, 2004). It can be seen that all three methods converge to the true value as the sample size increases, but IS2 appears to do so more quickly, performing well even with only 10 samples. In the MCMC context, method IS2 will also have the advantage that the samples used are drawn from a different, nearby distribution at each iteration. Thus in what follows we use method IS2 to evaluate the performance of IS methods for use in parameter estimation.

4.3 Comparison of Bayesian Estimation Methods

We are now in a position to empirically compare the performance of the three Bayesian parameter estimation schemes introduced in Section 3.1. In this section, we compare their ability to recover parameters for data generated from a model with known parameters, for the Ising and then Potts models. In the next section we look at performance for the protein

model with real data. We denote the methods by the shorthand:

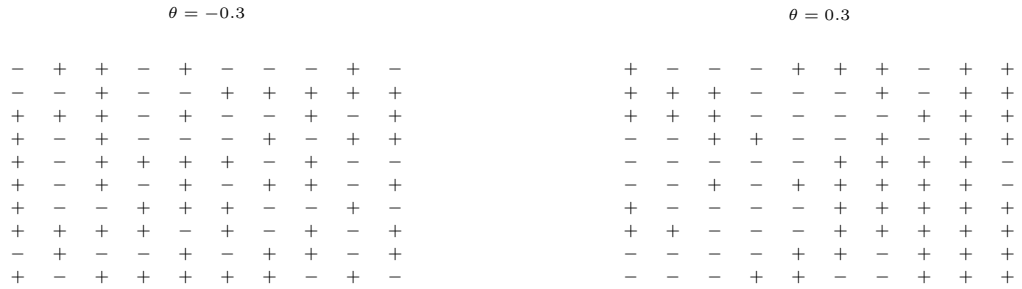
IS method: Metropolis-Hastings algorithm with $Z_\theta/Z_{\theta'}$ estimated by importance sampling as in Section 3.1.1. Based on the results of Section 4.2, we use IS2 with 100 samples.

M method: Møller's auxiliary variable algorithm described in Section 3.1.2.

PP method: Metropolis-Hastings algorithm targeting the "pseudoposterior" (5) described in Section 3.1.3.

4.3.1 Ising model comparison

We generated two perfect samples from 10×10 Ising models with parameters $J = -0.3$ and $J = 0.3$, respectively:



All methods use a random-walk Metropolis algorithm on θ, with proposal distribution $\theta' \sim N(\theta, \sigma^2)$ where σ^2 was chosen to achieve the acceptance rates ≈30-40%; ($\sigma^2 = 0.15$ for IS2, $\sigma^2 = 0.1$ for M, and $\sigma^2 = 0.2$ for PP). All chains were initialized at the MPLE, and $\tilde{\theta}$ for the M method is also the MPLE. Priors are uniform: $\pi(\theta) \sim U(0, 0.7)$ for $\theta = 0.3$, and $\pi(\theta) \sim U(-0.7, 0)$ for $\theta = -0.3$. MCMC chains are run for 10,000 iterations, discarding the first 2000 iterations as the burn-in period.

For comparison, we calculated the posterior distribution and quantiles numerically, by discretization of θ at a resolution of 10^{-6}. At each value, the posterior density can be directly calculated using the forward recursion method (Reeves and Pettitt, 2004), as can the quantiles. Figures 3 and 4 show traceplots, posterior histograms, autocorrelation-plots and quantile-quantile plots for the three different methods on the $\theta = -0.3$ and $\theta = 0.3$ Ising model, respectively. We can see that the IS and M methods both approximate the theoretical distribution well in each case. However, the M method suffers higher autocorrelation, while the IS method is computationally intensive, requiring 100 exact samples at each iteration.

Interestingly, the PP method approximates the theoretical posterior distribution reasonably well, but does display a small bias in the antiferromagnetic case.

4.3.2 Generalized Potts model comparison

The Ising model is a particularly simple special case. For protein modeling it is common to use hundreds of observations, each containing a few hundred contacts, to estimate the 210 contact potential parameters. Even for a few dozen structures, there will be on the order of dozens of observed contacts per interaction parameter. To design a simulation experiment comparable to this situation and still simple enough to generate perfect samples, we use a 3 color Potts model, generalized to have non-zero interactions on the off-diagonal of θ as in the protein model, on a 10×10 lattice. This model has 5 parameters to be estimated, and a single realization contains 180 total contacts.

We generated an exact sample from this model with parameters

$$\theta = \begin{pmatrix} 0.0 & 0.1 & 0.2 \\ 0.1 & 0.4 & 0.3 \\ 0.2 & 0.3 & 0.5 \end{pmatrix}$$

resulting in the sample observation:

2	3	1	2	3	3	3	2	1	1
2	1	2	3	2	3	1	2	3	2
2	2	2	1	3	3	1	3	1	1
3	2	2	3	3	2	1	1	3	2
3	3	3	1	3	3	3	3	2	3
2	2	1	2	1	3	3	2	2	3
3	2	2	3	3	3	1	3	3	3
3	3	2	3	1	3	3	2	3	3
3	2	2	1	2	3	3	3	3	2
3	1	3	1	3	3	2	3	3	1

Although this is a dramatic simplification of the $k = 20$ protein model, with elements of θ chosen relatively small to ensure fast exact sampling, nevertheless the IS method takes several days to run the MCMC chain. The major limitation is the generation of 100 exact samples during each MCMC iteration. Moreover, the number 100 was chosen based on the Ising model simulations of Section 4.2 and is probably insufficient to accurately estimate $Z_\theta/Z_{\theta'}$ for the current model (checking by numerical calculation using exact recursions is prohibitive in this case). We therefore conclude that the IS method is impractical for parameter estimation in our protein model.

Therefore, we compare only the M and PP methods here. As before, the variance σ^2 of the random-walk proposal $\theta' \sim N(\theta, \sigma^2)$ is chosen to achieve acceptance rate ≈30-40%. ($\sigma^2 = 0.4$

for PP method). However, Møller's method suffers from low acceptance rates for all σ^2, and we are only able to achieve acceptance rate around $\approx 20\%$ with small $\sigma^2 = 0.02$. As before, $\tilde{\theta}$ and initial states are chosen to be the MPLE. Prior distributions are $\pi_0(\theta) \sim U(-1, 2)$. We run the MCMC chains for 100,000 iterations, and discard the first 5000 iterations as the burn-in period.

Fig.5 shows traceplots, posterior histograms, and autocorrelation-plots of the two methods for two representative parameters $\theta(1, 2)$ and $\theta(3, 3)$. (Results for the other 3 parameters are similar.) We see that even for this simple Potts model, the M method suffers from high autocorrelation and slow mixing. Because the acceptance decreases with dimension of x, the chain is likely to be unacceptably slow in our protein model where x entails tens of thousands of sites (hundreds of observations each containing hundreds of sites).

More importantly, exact sampling is too slow even in this simple case: the Møller chain takes about 48hrs to run. Since it takes significantly longer to generate exact samples from a $k = 20$ color Potts model with a wide range of θ's, we conclude that this sampling speed limitation precludes use of Møller's method and its extensions to complicated models such as the protein design model.

In contrast, the PP method runs very fast (only a few minutes, regardless of the range of parameters) and shows good mixing. Although the exact posterior distribution cannot be calculated numerically by the forward recursion method in this case, we can compare the PP results to those from Møller's method which has the exact posterior as its limiting distribution. Fig. 6 shows quantile-quantile plots comparing the samples from these two methods for $\theta(1, 2)$ and $\theta(3, 3)$. We see that they are in close agreement, although the PP method appears to exhibit a small bias as expected (it is difficult to attribute differences purely to the PP method due to the small effective sample size obtainable by the M method). These results suggest that the PP method may approximate the posterior distribution reasonably well, especially for larger sample sizes which mitigate the bias (see Section 4.4 below). Therefore the PP method appears to be a reasonable choice for satisfactory parameter estimation in our protein model. Given the intractable computational requirements of the IS and M methods for larger, more complex models with larger sample sizes, the PP method also appears to be the only viable alternative.

4.4 Convergence of Point Estimators in Potts Models

Lastly, we evaluated the performance of the pairing frequency estimators (PFE) commonly used in the protein science literature (see Section 3.2) by a simulation experiment, and

compared the performance to the MPLE.

Writing the pseudolikelihood function for Potts models (4) for the generalized model gives:

$$l_p = \sum_{i=1}^{m\times n}\{-\sum_{j\sim i}\theta(x_i,x_j) - \log[\sum_{x=1}^{k} \mathrm{e}^{-\sum_{j\sim i}\theta(x,x_j)}]\}$$

and since each $p(x_i \mid x_j : j \sim i, \theta)$ is log-concave, the pseudolikelihood function is log-concave, and the global maximum found efficiently by numerical optimization. We use a Newton-Raphson algorithm, where the Hessian matrix is easily derived:

$$\frac{\partial l_p}{\partial \theta_{ab}} = \sum_{i=1}^{m\times n}\{\frac{\sum_{x=1}^{k} m_i^{ab}\mathrm{e}^{-\sum_{j\sim i}\theta(x,x_j)}}{\sum_{x=1}^{k}\mathrm{e}^{-\sum_{j\sim i}\theta(x,x_j)}} - m_i^{ab}\}$$

$$\frac{\partial 2 l_p}{\partial \theta_{ab}\partial\theta_{cd}} = \sum_{i=1}^{m\times n}\{\frac{[\sum_{x=1}^{k} m_i^{ab}\mathrm{e}^{-\sum_{j\sim i}\theta(x,x_j)}\sum_{x=1}^{k} m_i^{cd}\mathrm{e}^{-\sum_{j\sim i}\theta(x,x_j)}]}{[\sum_{x=1}^{k}\mathrm{e}^{-\sum_{j\sim i}\theta(x,x_j)}]^2} - \frac{[\sum_{x=1}^{k}\mathrm{e}^{-\sum_{j\sim i}\theta(x,x_j)}][\sum_{x=1}^{k} m_i^{ab}m_i^{cd}\mathrm{e}^{-\sum_{j\sim i}\theta(x,x_j)}]}{[\sum_{x=1}^{k}\mathrm{e}^{-\sum_{j\sim i}\theta(x,x_j)}]^2}\}$$

Here m_i^{ab} is the number of neighbors of x_i taking value b if $x_i = a$, and 0 otherwise.

We compare the ability of PFE and MPLE estimators to recover parameters for a $k = 5$ color 10×10 lattice Potts model. Elements of the parameter matrix θ were generated uniformly on $[-0.5, 0.5]$, subject to the constraints (symmetry, $\theta(1,1) = 0$ and $w(1,1) = 0$). The small range of parameters around zero ensured that observations could be generated by exact sampling with reasonable speed. n independent samples of the 10×10 lattice were then generated and the PFE and MPLE estimators applied to the sample of size n. This was repeated 20 times in order to examine the sampling distribution of these point estimators.

Figures 7a and 7b show the results for the two estimators for sample sizes $n = 10$ and $n = 100$, respectively. We see that the sampling distribution of the MPLEs are centered at the true values, and converge quickly as n increases. Although the MPLEs are known to be biased, they are asymptotically unbiased. In contrast to applications to single images or spatial datasets, our application involves independent samples of the entire graph. It appears that the asymptotics kick in quickly in this case. By comparison, the pairing frequency estimates appear to be inconsistent: they are biased estimates and do not converge to the true parameters.

We repeated this for the $k = 20$ model with parameters generated uniformly from $[-0.3, 0.3]$, this time performing 10 replications for sample sizes $n = 100$ and $n = 1000$.

Fig. 8 shows the distributions of the errors $(\hat{\theta} - \theta)$. Although the intervals for both estimates contain 0 for $n = 100$, when uncertainty is reduced ($n = 1000$) we see that once again the MPLE converges to the true values while PFE converges to incorrect values.

We conclude that for independent samples from Potts models the MPL estimates converge quickly to the true values as sample size increases, while the pairing frequency estimates appear to be inconsistent.

5 Application to Protein Model

We now consider parameter estimation in the protein model of Section 2.2. Recall that we concluded from the simulation studies of Section 4 that, of the Bayesian estimation schemes described in Section 3, only the pseudoposterior method was sufficiently accurate and efficient to scale to the protein problem. Thus in this section we evaluate the performance of the PP method, along with the point estimators MPLE and PFE, on a protein modeling problem using real data. The optimization procedure for finding the MPLE (also used to initialize the MCMC for PP) is given in the Appendix.

5.1 Protein Data Sets and Pseudolikelihood Estimates

The protein structure dataset used is listed in Table 1. All structures are obtained from the Protein Data Bank (`www.rcsb.org`), have $> 2.5\mathring{A}$ resolution, and share $< 25\%$ sequence identity. 23 large proteins (> 255 residues) serve as templates for building alternative structures for the threading experiment, and 37 proteins with < 255 residues serve as the training set; these sets are taken from (Maiorov and Crippen, 1992; Thomas and Dill, 1996a). The test set contains 174 proteins chosen from a set of 302 proteins with little sequence similarity (Hobohm and Sander, 1994). To distinguish between compact and non-compact structures, only proteins with a relatively small radius of gyration are selected, following (Maiorov and Crippen, 1992): only proteins with $e_g < 1.3$ are selected, where e_g is ratio of the radius of gyration of the putative reference structure to $r_{\min}$ the minimal radius of gyration over the set of all its alternatives. r_{min} is calculated by a linear regression $r_{min} = -1.26 + 2.79(N)^{1/3}$, where N is the number of residues in the protein. This process results in 174 proteins in the training set.

Training Set of 37 Small Proteins							
155C	1ACX	1BDS	1CC5	1CRN	1CSE.I	1ECD	1FDX
1HIP	1HMQA	1HOE	1HVP.A	1LH4	1PP2.R	1REI.A	1RN3
1SN3	2AZA.A	2B5C	2C2C	2CDV	2HHB.A	2HHB.B	2OVO
2PAB.A	2PKA.B	2RHE	2SNS	2SSI	2STV	351C	3ADK
3EBX	4PTI	5CYT	5RXN	9WGA.A			
Set of 23 Large Proteins for Building Alternative Structures							
1ABP	1CSE.E	1CTS	1HMG.A	1PFK.A	1PHH	1PYP	1RHD
2CAB	2CYP	2TAA.A	3GPD.G	3GRS	3PGK	4APE	4MDH.A
4RHV.1	4TLN	5CPA	6LDH	7API.A	8ADH	8CAT.A	
Testing Set of 174 Low Similarity Proteins							
1AAK	1AAP.A	1ABA	1ABK	1ABM.A	1APS	1ATX	1AYH
1BAA	1BAB.B	1BAR.B	1BBA	1BBH.A	1BBL	1BBO	1BBP.A
1BBT.2	1BGC	1BOV.A	1BRD	1BW4	1C2R.A	1C5A	1CAU.A
1CAU.B	1CBN	1CD8	1CDT.A	1CID	1CIS	1CMB.A	1COB.A
1CPC.A	1CPC.L	1CTA.A	1D66.A	1DFN.A	1DHR	1DNK.A	1EAF
1ECO	1EGF	1END	1ERP	1FAS	1FC2.C	1FCS	1FDD
1FHA	1FIA.B	1FXI.A	1GKY	1GLA.F	1GMF.A	1GMP.A	1GPS
1GRC.A	1GSS.A	1HCC	1HDD.C	1HIV.A	1HLE.B	1IFC	1ISU.A
1IXA	1L92	1LE4	1LTS.A	1LTS.C	1LTS.D	1MDA.A	1MDC
1MHU	1MS2.A	1MUP	1NXB	1OFV	1OIA.A	1OVB	1PAZ
1PDC	1POA	1POC	1PPF.E	1PPN	1R1A.2	1RBP	1RCB
1RND	1RPR.A	1RRO	1SGT	1SHA.A	1SHF.A	1SNC	1TAB.I
1TEN	1TFG	1TFI	1TGL	1TGS.I	1THO	1TIE	1TLK
1TNF.A	1TRE.A	1TRO.A	1TTB.A	1UTG	1VAA.B	1WSY.A	1YCC
256B.A	2ACH.B	2ATC.B	2AVI.A	2BDS	2BOP	2BPA.2	2BPA.3
2CBH	2CBP	2CCY.A	2CPL	2CRD	2CRO	2ECH	2GB1
2HIP.A	2HSD.A	2IHL	2LAL.A	2LAL.B	2MAD.L	2MHR	2MHU
2MSB.A	2PDE	2PF2	2PHY	2PRF	2RN2	2SAS	2SCP.A
2SGA	2SN3	2SNV	2ZTA.A	3B5C	3CHY	3CLA	3DFR
3GAP.A	3IL8	3MON.A	3PGM.A	3RUB.S	3SC2.B	3SGB.I	3SIC.I
4BLM.A	4CPA.I	4FXN	4GCR	4HTC.I	4SBV.A	4SGB.I	4TGF
5P21	7API.B	7ZNF	8I1B	8RXN.A	9RNT		

Table 1: List of Proteins Used in the Work

5.2 Threading Experiment and Results

We evaluate the parameter estimates for the protein model via a protein structure prediction experiment. Although the "true" parameters are unknown, it is assumed that correct parameters will identify the native protein structure of a given sequence as the minimum energy among a pool of alternative structures. (This is a test of the model as much as the estimation scheme; but allows for comparison between estimation schemes for the given model). To do so we perform gapless threading, a simple and common way to generate alternative structures (see e.g. Thomas and Dill (1996a)). Briefly, the sequence of any small protein is mounted on the backbone of the 23 largest proteins without gaps, and slides alone the backbone one amino acid at a time, generating $N_l - N_s + 1$ alternative structures for small and large proteins of length N_s and N_l respectively. The energy of each alternative structure, as well as the native structure, is then calculated for each sequence using the estimated energy function parameters.

Because the fold recognition experiment involves evaluation of the energy on native and alternative structures containing varying numbers of contacts, care is needed in imposing identifiability constraints. (For example, adding a large constant to all contact energies will not affect the MLE or MPLE, but then a protein with a small number of contacts generally have lower energy than a protein with a large number of contacts.) To exclude the effects of different numbers of contacts, here we use constraints $\sum_{i=1}^{210} \theta_i = 0$ and $\sum_{i=1}^{20} \mu_i = 0$ instead.

Table 5.1 shows the maximum pseudolikelihood estimates obtained from the training set. We compare the predictive performance of our parameter estimates with those obtained using the pairing frequency estimates described in Subsection 3.2 (a pairwise 'empirical" or 'statistical' potential), and a discriminative model given algorithmically in (Thomas and Dill, 1996a). The latter uses a Boltzmann distribution over structures, approximating the partition function by the set of generated alternative structures:

$$p(c \mid s) = \frac{p(c \mid s)}{\sum_{c \in C} p(c \mid s)} \approx \frac{\mathrm{e}^{-\sum_{i \sim j} \theta(x_i^c, x_j^c)}}{\sum_{a \in C_A \cup \{c_N\}} \mathrm{e}^{-\sum_{i \sim j} \theta(x_i^a, x_j^a)}} \tag{6}$$

where C_A is set of alternative structures and c_N the native structure. The contact energies θ are then obtained by optimizing the right hand side approximation to $p(c_N \mid s)$ for the training set. Further discussion of discriminative vs generative modeling for this problem is given elsewhere (Zhou and Schmidler, 2009).

The results of threading experiment are summarized in Table 3. The native structure is ranked among the pool of alternative structures according to the energy assigned by the

potential. At 0% the native structure ranks as best, at 5% native structure is assigned energy lower than 95% of alternative structures. We see that the pseudoposterior-based parameter estimation dramatically outperforms the quasi-chemical model (PFE) commonly used in the field. This shows the practical consequences of the inconsistency of the PFE estimators demonstrated in Section 4.4. Indeed, using a proper estimation scheme such as the MPLE or pseudoposterior, the generative model achieves nearly the same performance and the Dill model where discriminative performance is directly optimized. This will be surprising, good news for those protein modeling applications described in Section 2.2 which desire a generative model.

It is worth noting that there remains considerable room for improvement in the protein model described here. It is well known that potentials based on an distance-dependent, all-atom representation substantially improve fold recognition accuracy (Melo et al., 2002). However, our results suggest that such models should be fit by proper statistical estimators rather than the PFE commonly used. An advantage of the pseudoposterior approach is that it is easily extensible to such models.

6 Conclusion and Discussion

We have considered computational schemes for practical Bayesian estimation in Ising and Potts models, and generalizations relevant to protein modeling, containing an intractable normalization in the likelihood. Comparison of three alternative methods via simulation studies indicate that while importance sampling approximation of the ratio of normalizing constants, Møller's auxiliary variable method, and sampling from a pseudolikelihood-based posterior all work well in small Ising models, for realistic problems involving large graphs, higher-dimensional parameter vectors, and/or repeated samples, only the pseudoposterior method is practical. A critical limitation of both importance sampling and Møller's method is the large computational time required for generating exact samples using the bounding chain algorithm. Since the bounding chain algorithm is based on a heat-bath algorithm, exact sampling algorithms built on more sophisticated Markov chains may help alleviate this problem. However, in our hands Møller's algorithm also suffered from slow mixing due to the necessity of using a very narrow proposal distribution, due to the low acceptance rates arising from the use of a very high dimensional (equal to dimension of dataset) auxiliary variable. The latter problem has been addressed somewhat by Murray et al. (2006), who introduce multiple auxiliary variables. However, at present the first limitation (speed of

exact sampling) prohibits application of Møller's method or Murray's extension to Potts models on large graphs.

Surprisingly, the "pseudoposterior" method based on a pseudolikelihood approximation performs rather well in this setting. Sampling is fast and straightforward, and although pseudolikelihood estimates are known to exhibit significant bias when dependency in the graph is high (Geyer, 1992; Gile and Handcock, 2007), the asymptotic unbiasedness appears to be achieved relatively quickly when multiple independent samples of the graph are available as common in protein modeling problems. The resulting estimators significantly outperform the "pairing frequency estimators" commonly used in protein bioinformatics, both in recovering known parameters, and in a practical protein structure prediction test on real data. Indeed the PFE appear to be inconsistent, perhaps explaining previous observations that the pairing frequency estimates often do not appear to reflect the underlying energetics of protein structures (Thomas and Dill, 1996b).

Appendix: Pseudolikelihood optimization for the protein model

The log-pseudolikelihood function for the protein model is given by:

$$l_p = \sum_{i=1}^{n} \{ -(\sum_i \mu(x_i) + \sum_{i \sim j} \theta(x_i, x_j)) - \log(\sum_{x=1}^{20} \mathrm{e}^{-(\sum_i \mu(x) + \sum_{i \sim j} \theta(x, x_j))}) \}$$

Again this is log-concave and we use Newton-Raphson to maximize. The gradient vector and Hessian matrix are easily derived:

$$\frac{\partial l_p}{\partial \theta_{ab}} = \sum_{i=1}^{n} \{ -m_i^{ab} + \frac{\sum_{x=1}^{20} m_i^{ab} \mathrm{e}^{-E(x_i = x)}}{\sum_{x=1}^{20} \mathrm{e}^{-E(x_i = x)}} \}$$

$$\frac{\partial l_p}{\partial \mu_a} = \sum_{i=1}^{n} \{ -1_{\{x_i = a\}} + \frac{\mathrm{e}^{-E(x_i = a)}}{\sum_{x=1}^{20} \mathrm{e}^{-E(x_i = x)}} \}$$

$$\frac{\partial^2 l_p}{\partial\theta_{ab}\partial\theta_{cd}} = \sum_{i=1}^{n}\{\frac{[\sum_{x=1}^{20} m_i^{ab}\mathrm{e}^{-E(x_i=x)}][\sum_{x=1}^{20} m_i^{cd}\mathrm{e}^{-E(x_i=x)}]}{(\sum_{x=1}^{20}\mathrm{e}^{-E(x_i=x)})^2} - \frac{[\sum_{x=1}^{20} m_i^{ab}m_i^{cd}\mathrm{e}^{-E(x_i=x)}]}{\sum_{x=1}^{20}\mathrm{e}^{-E(x_i=x)}}\}$$

$$\frac{\partial^2 l_p}{\partial\mu_a\partial\mu_b} = \sum_{i=1}^{n}\{\frac{\mathrm{e}^{-E(x_i=a)}\mathrm{e}^{-E(x_i=b)}}{(\sum_{x=1}^{20}\mathrm{e}^{-E(x_i=x)})^2} - 1_{\{a=b\}}\frac{\mathrm{e}^{-E(x_i=a)}}{\sum_{x=1}^{20}\mathrm{e}^{-E(x_i=x)}}\}$$

$$\frac{\partial^2 l_p}{\partial\mu_a\partial\theta_{cd}} = \sum_{i=1}^{n}\{\frac{[\mathrm{e}^{-E(x_i=a)}][\sum_{x=1}^{20} m_i^{cd}\mathrm{e}^{-E(x_i=x)}]}{(\sum_{x=1}^{20}\mathrm{e}^{-E(x_i=x)})^2} - \frac{m_i^{cd}1_{\{x_i=a\}}\mathrm{e}^{-E(x_i=a)}}{\sum_{x=1}^{20}\mathrm{e}^{-E(x_i=x)}}\}$$

The identifiability constraints $\theta(1,1) = 0$ and $\mu(1) = 0$ are easily imposed. As explained in Section 5, the constraints $\sum\theta = 0$ and $\sum\mu = 0$ are used instead for the threading experiment, obtained by subtracting a constant.

References

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, 181:223–30.

Baddeley, A. and Turner, R. (2000). Practical maximum pseudolikelihood for spatial point patterns. *Australian and New Zealand Journal of Statistics*, 42:283–322.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall, Boca Raton, FL.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Stat. Soc. B*, 36:192–236.

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

Brush, S. (1967). History of the Lenz-Ising model. *Reviews of Modern Physics*, 39:883–893.

Comets, F. (1992). On consistency of a class of estimators for exponential families of Markov random fields on the lattice. *Ann. Statist.*, 20(1):455–468.

Fromer, M. and Yanover, C. (2008). A computational framework to empower probabilistic protein design. *Bioinformatics*, 24(13):214–222.

Gelman, A. and Meng, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Stat. Sci.*, 13(2):163–185.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattn. Anal. Mach. Intel.*, 6:721–741.

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics: Proc. 23rd Symp. Interface*, pages 156–163.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Stat. Sci.*, 7:473–511.

Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *J. Roy. Stat. Soc. B*, 54(3):657–699.

Geyer, C. J. and Thompson, E. A. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *J. Amer. Statist. Assoc.*, 90(431):909–920.

Gile, K. and Handcock, M. S. (2007). Comparison of maximum pseudo-likelihood and maximum likelihood estimation of exponential family random graph models. *Working Paper no. 74*.

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

Guyon, X. and Kunsch, H. (1992). Asymptotic comparison of estimators in the Ising model. *Stochastic Models, Statistical Methods, and Algorithms in Image Analysis. Lecture Notes in Statistics 74. Springer New York*, pages 177–198.

Hobohm, U. and Sander, C. (1994). Enlarged representative set of protein structures. *Prot. Sci.*, 3:522–524.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA*, 79(8):2554–2558.

Huber, M. (2004). Perfect sampling using bounding chains. *Ann. Appl. Prob.*, 14(2):734–753.

Jensen, J. L. and Kunsch, H. R. (1994). On asymptotic normality of pseudo likelihood estimates for pairwise interaction process. *Annals of the Institute of Statistical Mathematics*, 46(3):475–486.

Kleinman, C. L., Rodrigue, N., Bonnard, C., Philippe, H., and Lartillot, N. (2006). A maximum likelihood framework for protein design. *BMC Bioinformatics*, 7:326.

Kuhlman, B. and Baker, D. (2000). Native protein sequences are close to optimal for their structures. *Proc. Natl. Acad. Sci. USA*, 97(19):10383–10388.

Kuhlman, B., Dantas, G., Ireton, G., Varani, G., Stoddard, B. L., and Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. *Science*, 302(5649):1364–1368.

Lauritzen, S. L. and Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *J. Roy. Stat. Soc. B*, 50(2):157–224.

Maiorov, V. and Crippen, G. (1992). Contact potential that recognizes the correct folding of globular proteins. *J. Mol. Biol.*, 227(3):876–888.

Mase, S. (2000). Marked Gibbs processes and asymptotic normality of maximum pseudo-likelihood estimators. *Mathematische Nachrichten*, 209(1):151–169.

Melo, F., Sanchez, R., and Sali, A. (2002). Statistical potentials for fold assessment. *Protein Science*, 11(2):430–448.

Meng, X.-L. and Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, 6:831–860.

Meyerguz, L., Kempe, D., Kleinberg, J., and Elber, R. (2004). The evolutionary capacity of protein structures. In Bourne PE, G. D., editor, *RECOMB'04*, page 290 to 297. ACM Press, New York, NY, USA.

Micheletti, C., Seno, F., Maritan, A., and Banavar, J. (1998). Protein design in a lattice model of hydrophobic and polar amino acids. *Physical Review Letters*, 80(10):2237–2240.

Mirny, L. and Shakhnovich, E. I. (1996). How to derive a protein folding potential? A new approach to an old problem. *J. Mol. Biol.*, 264(5):1164–1179.

Miyazawa, S. and Jernigan, R. (1985). Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, 18(3):534–552.

Møller, J., Pettitt, A. N., Reeves, R., and Berthelsen, K. K. (2006). An efficient Markov Chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458.

Murray, I., Ghahramani, Z., and MacKay, D. J. C. (2006). MCMC for doubly-intractable distributions. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pages 359–366. AUAI Press.

Potts, R. and Domb, C. (1952). Some generalized order-disorder transformations. *Proceedings of the Cambridge Philosophical Society*, 48(1):106–109.

Propp, J. and Wilson, D. (1996). Exact sampling with coupled Markov Chains and applications to statistical mechanics. *Random Structures and Algorithms*, 9(1-2):223–252.

Reeves, R. and Pettitt, A. (2004). Efficient recursions for general factorisable models. *Biometrika*, 91(3):751–757.

Seno, F., Vendruscolo, M., Maritan, A., and Banavar, J. (1996). Optimal protein design procedure. *Physical Review Letters*, 77(9):1901–1904.

Shakhnovich, E. (1994). Proteins with selected sequences fold into unique native conformation. *Physical Review Letters*, 72:3907–3910.

Sippl, M. J. (1993). Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *Journal of Computer-Aided Molecular Design*, 7(4):473–501.

Thomas, P. D. and Dill, K. A. (1996a). An iterative method for extracting energy-like quantities from protein structures. *Proc. Natl. Acad. Sci. USA*, 93:11628–11633.

Thomas, P. D. and Dill, K. A. (1996b). Statistical potentials extracted from protein structures: How accurate are they? *J. Mol. Biol.*, 257(2):457–469.

Zhou, X. and Schmidler, S. C. (2009). Protein contact potentials: A statistical perspective. (in preparation).

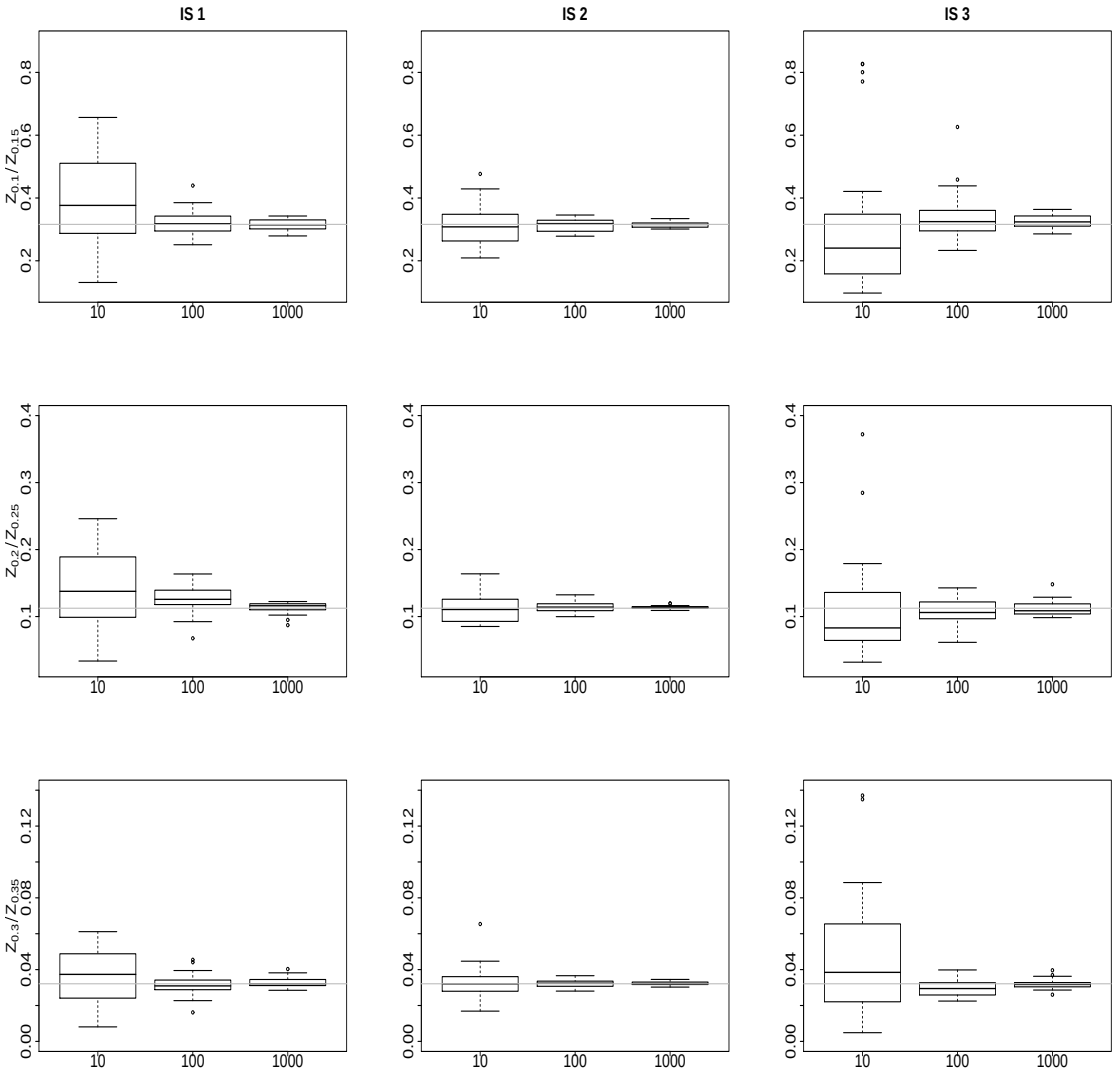


Figure 2: Comparison of the three different importance sampling methods of Section 3.1.1 for approximating the ratios $Z_{0.10}/Z_{0.15}$, $Z_{0.20}/Z_{0.25}$, and $Z_{0.30}/Z_{0.35}$ on a 10×10 Ising lattice, as sample size increases. Line shows exact value, computable for this size lattice via recursive methods (Reeves and Pettitt, 2004).

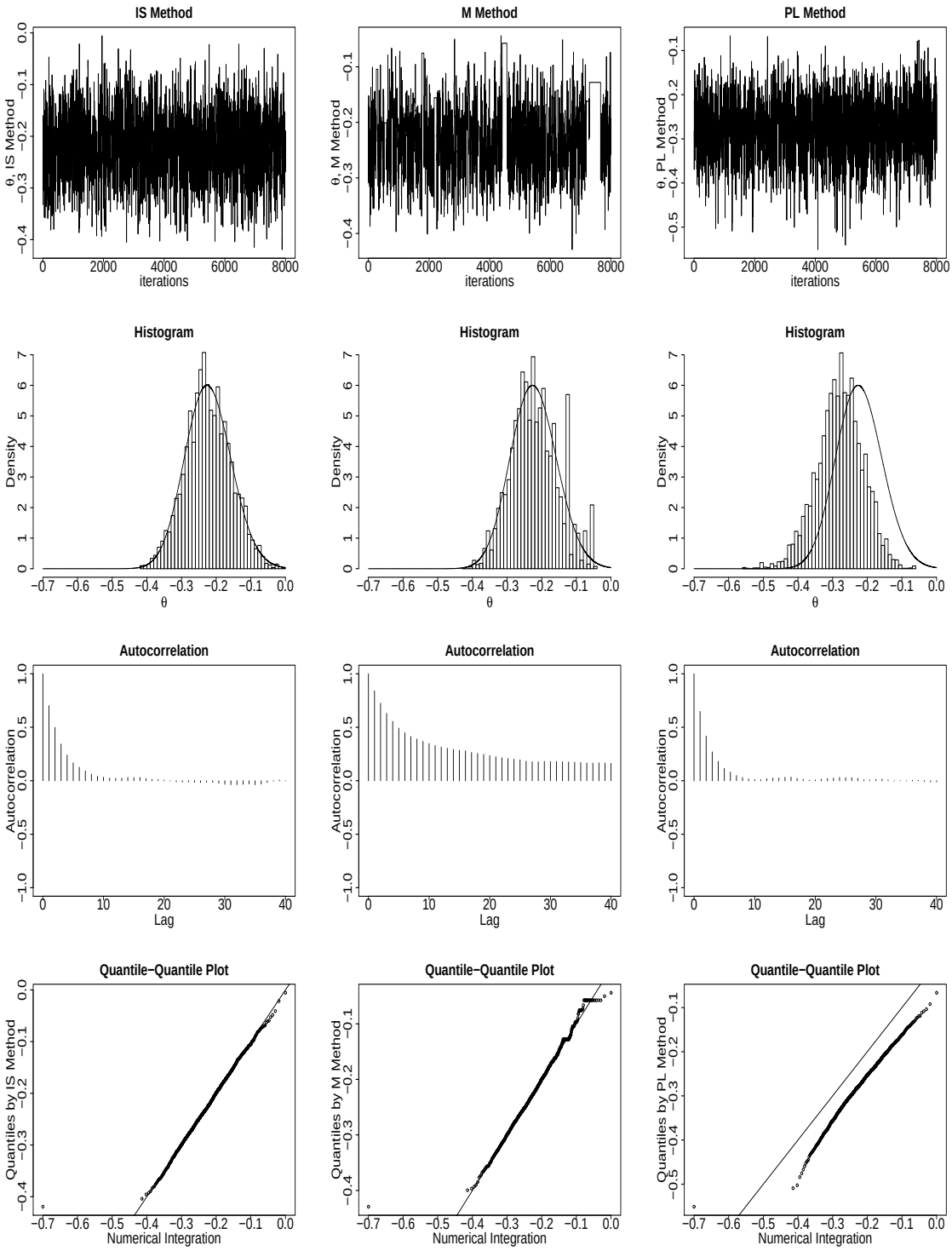


Figure 3: Comparison of three different parameter estimation methods for Ising model with $\theta = -0.3$ on a 10×10 lattice. Solid line on histograms is numerical integration using exact recursions.

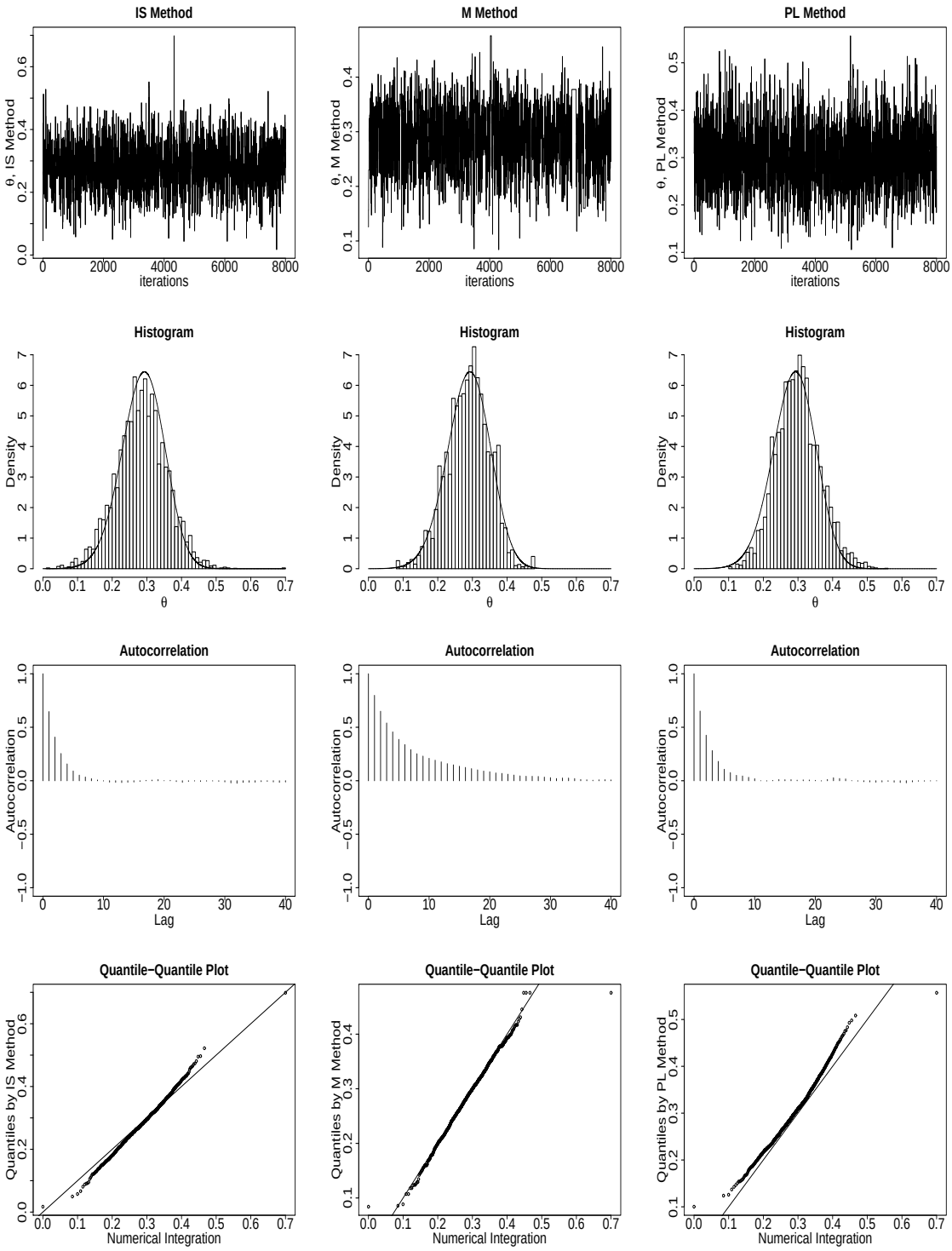
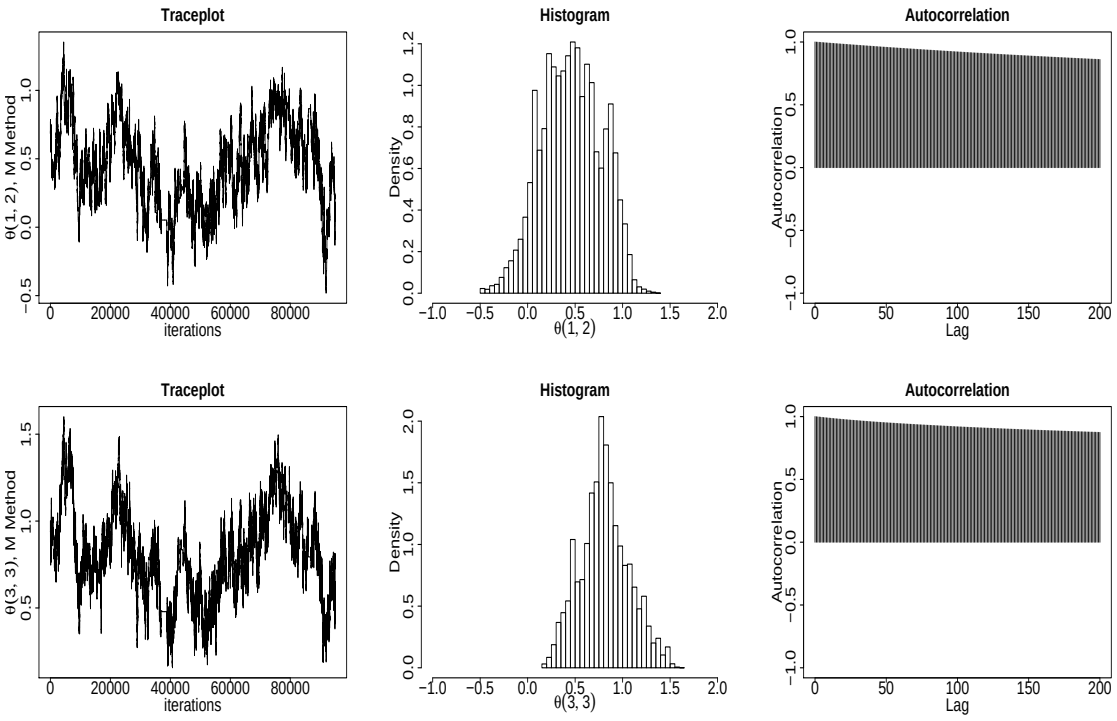
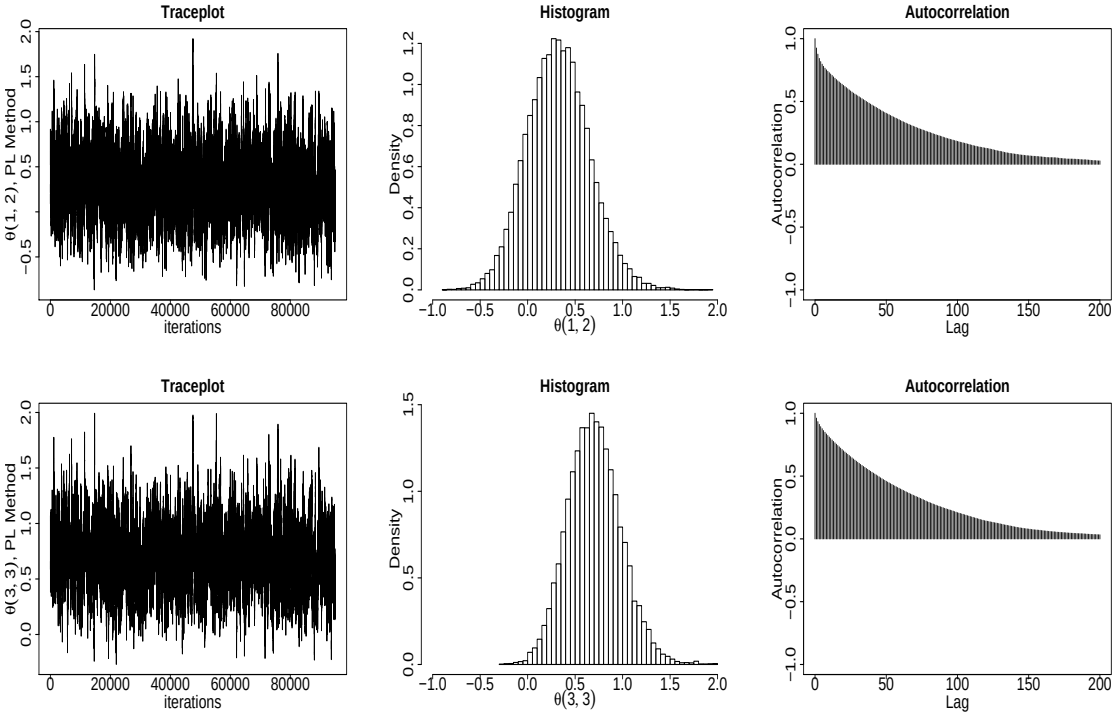


Figure 4: Comparison of three different parameter estimation methods for Ising model with $\theta = 0.3$ on a 10×10 lattice. Solid line on histograms is numerical integration using exact recursions.



(a)



(a)

Figure 5: Comparison of (a) Møller's and (b) pseudoposterior methods on two parameters $\theta(1,2)$ and $\theta(3,3)$ of a $k = 3$ color 10×10 lattice Potts model. Because the low acceptance rate requires the proposal σ^2 small, Møller's method exhibits high autocorrelation indicating slow mixing.

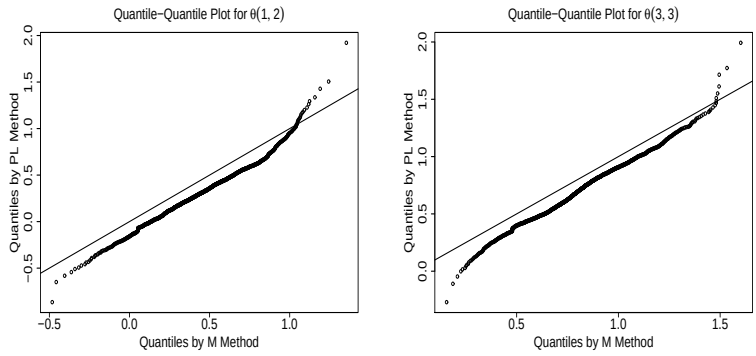


Figure 6: Quantile-Quantile Plot for $\theta(1,2)$ and $\theta(3,3)$ with M and PP methods.

θ	Ala	Arg	Asn	Asp	Cys	Glu	Gln	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	-0.075	0.43	0.058	0.14	-0.094	0.14	0.2	-0.0088	0.081	-0.0014	-0.11	0.058	0.12	0.081	0.44	0.14	0.028	0.028	0.033	-0.21
Arg	0.43	0.35	0.025	0.017	0.044	0.25	0.0076	-0.016	-0.084	-0.21	-0.1	0.56	-0.28	-0.3	0.53	0.29	-0.15	-0.08	-0.16	-0.088
Asn	0.058	0.025	-0.18	0.2	-0.31	0.083	0.09	0.023	-0.14	0.27	-0.07	0.24	-0.0350	.0053	0.34	0.51	0.038	-0.25	-0.35	0.036
Asp	0.14	0.017	0.2	0.51	0.1	0.44	0.48	0.19	-0.1	0.0098	0.33	0.14	0.19	0.23	0.72	0.092	0.16	-0.079	0.13	-0.01
Cys	-0.094	0.044	-0.31	0.1	-1.1	-0.24	0.16	-0.42	-0.37	-0.24	-0.24	0.067	-0.51	-0.48	0.12	-0.061	-0.082	-0.1	-0.52	-0.3
Glu	0.14	0.25	0.083	0.44	-0.24	0.039	0.33	-0.016	0.36	-0.22	-0.099	0.36	-0.42	-0.04	0.19	0.39	-0.052	0.35	-0.35	-0.1
Gln	0.2	0.0076	0.09	0.48	0.16	0.33	0.067	0.17	-0.098	0.044	-0.0047	-0.022	0.23	0.23	0.41	0.36	0.32	0.45	0.2	-0.067
Gly	-0.0088	-0.016	0.023	0.19	-0.42	-0.016	0.17	-0.05	-0.0079	0.07	-0.1	0.00031	-0.078	0.12	0.17	0.094	0.16	-0.076	-0.065	-0.035
His	0.081	-0.084	-0.14	-0.1	-0.37	0.36	-0.098	-0.0079	-0.21	0.074	-0.24	0.16	-0.41	-0.63	0.27	0.057	0.19	0.099	-0.27	-0.28
Ile	-0.0014	-0.21	0.27	0.0098	-0.24	-0.22	0.044	0.07	0.074	-0.44	-0.52	-0.027	-0.048	-0.46	0.18	0.077	-0.12	-0.46	-0.49	-0.36
Leu	-0.11	-0.1	-0.07	0.33	-0.24	-0.099	-0.0047	-0.1	-0.24	-0.52	-0.57	0.14	-0.13	-0.56	0.12	0.12	-0.12	-0.51	-0.31	-0.51
Lys	0.058	0.56	0.24	0.14	0.067	0.36	-0.022	0.00031	0.16	-0.027	0.14	0.56	-0.1	-0.12	0.67	0.43	0.19	0.085	-0.02	0.022
Met	0.12	-0.28	-0.035	0.19	-0.51	-0.42	0.23	-0.078	-0.41	-0.048	-0.13	-0.1	-0.077	-0.52	0.17	-0.069	0.11	0.31	-0.22	-0.29
Phe	0.081	-0.3	0.0053	0.23	-0.48	-0.04	0.23	0.12	-0.63	-0.46	-0.56	-0.12	-0.52	-0.57	-0.071	0.12	-0.11	-0.51	-0.22	-0.43
Pro	0.44	0.53	0.34	0.72	0.12	0.19	0.41	0.17	0.27	0.18	0.12	0.67	0.17	-0.071	0.73	0.66	0.26	0.18	-0.19	0.047
Ser	0.14	0.29	0.51	0.092	-0.061	0.39	0.36	0.094	0.057	0.077	0.12	0.43	-0.069	0.12	0.66	-0.099	0.058	-0.1	0.089	0.02
Thr	0.028	-0.15	0.038	0.16	-0.082	-0.052	0.32	0.16	0.19	-0.12	-0.12	0.19	0.11	-0.11	0.26	0.058	0.33	0.42	-0.059	0.095
Trp	0.028	-0.08	-0.25	-0.079	-0.1	0.35	0.45	-0.076	0.099	-0.46	-0.51	0.085	0.31	-0.51	0.18	-0.1	0.42	1.1	-0.56	0.068
Tyr	0.033	-0.16	-0.35	0.13	-0.52	-0.35	0.2	-0.065	-0.27	-0.49	-0.31	-0.02	-0.22	-0.22	-0.19	0.089	-0.059	-0.56	-0.37	-0.28
Val	-0.21	-0.088	0.036	-0.01	-0.3	-0.1	-0.067	-0.035	-0.28	-0.36	-0.51	0.022	-0.29	-0.43	0.047	0.02	0.095	0.068	-0.28	-0.53
μ	-1.2	-0.071	-0.42	-1.7	2.2	-0.28	-1.4	-0.94	0.97	1.1	0.94	-1.8	1.6	1.6	-2.1	-1.5	-1.1	1.6	1.4	0.95

Table 2: Pseudolikelihood Estimates for the Protein Model

Model	correct	<1%	1-5%	5-10%	10-20%	20-50%	>50%	total
Training Set								
Statistical Potentials (PFE)	3	16	3	4	5	5	1	37
Dill Estimates	36	1	0	0	0	0	0	37
Pseudoposterior	33	3	1	0	0	0	0	37
Test Set								
Statistical Potentials (PFE)	9	22	25	18	26	28	46	174
Dill Estimates	137	21	7	4	1	1	3	174
Pseudoposterior	134	21	7	2	6	2	2	174

Table 3: Threading Results on the Training and Test Sets with Different Models

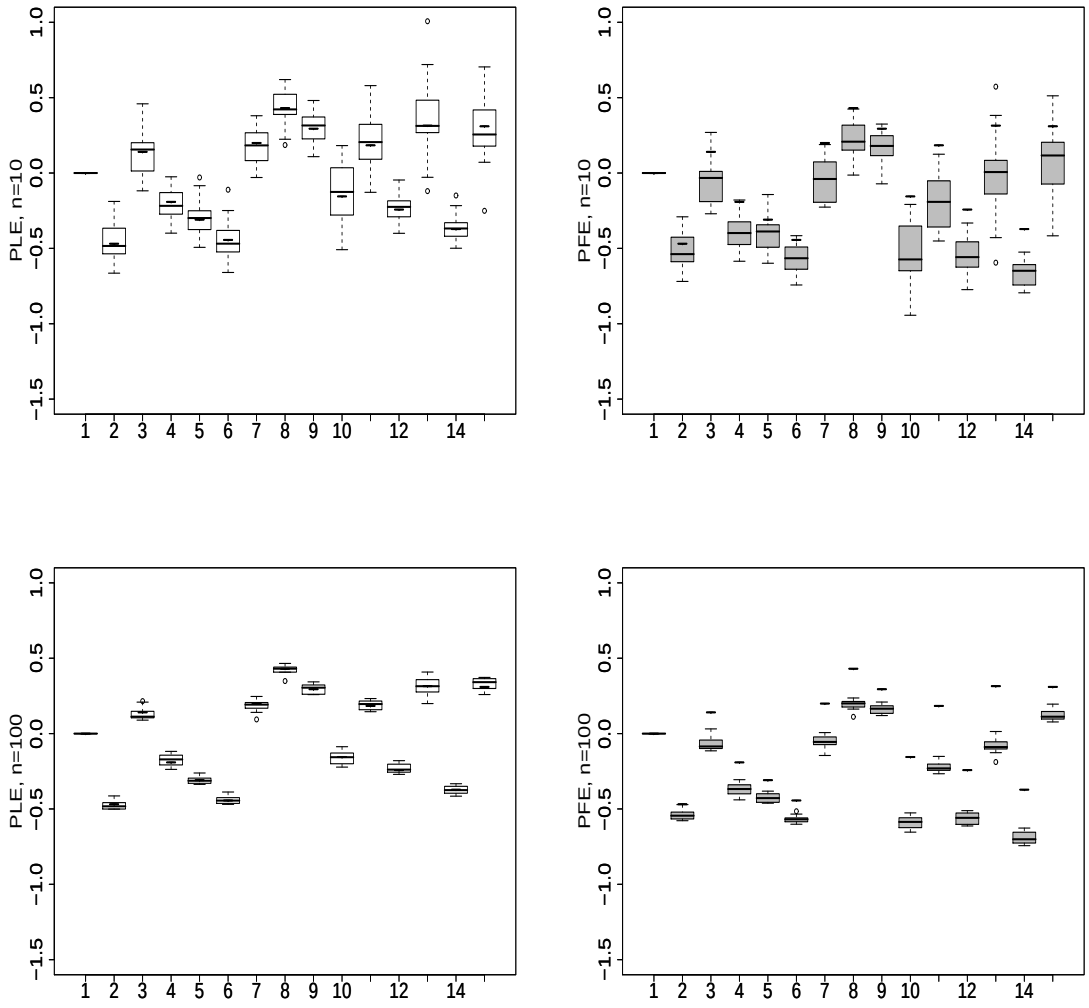
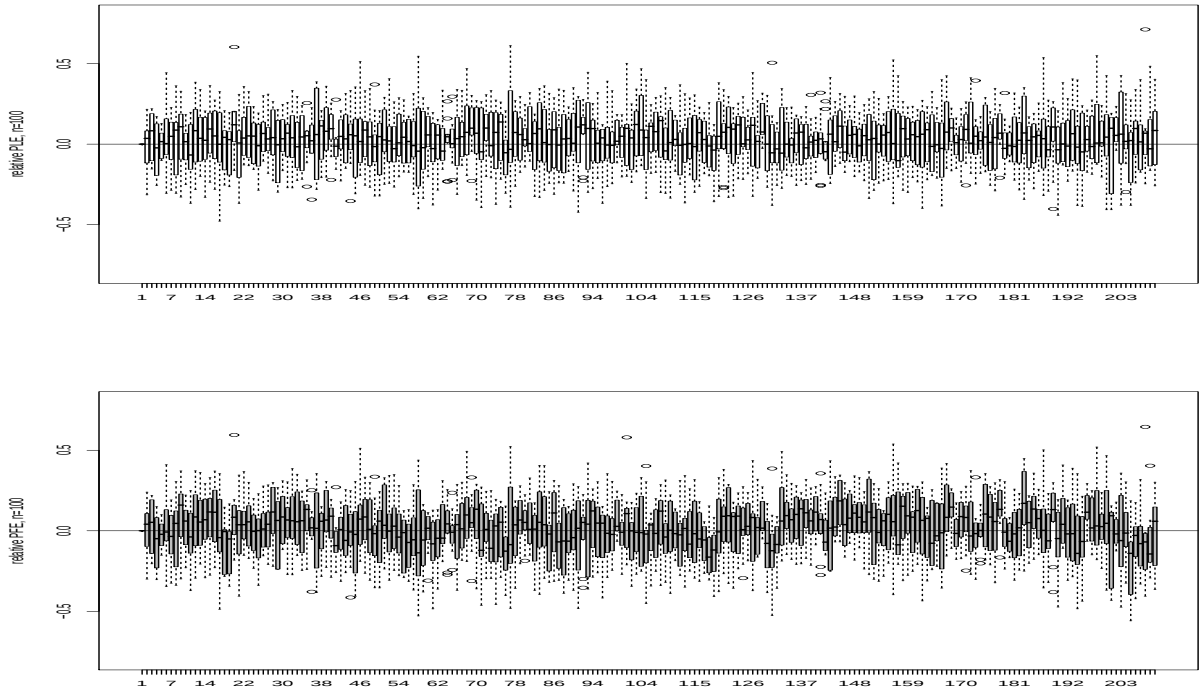
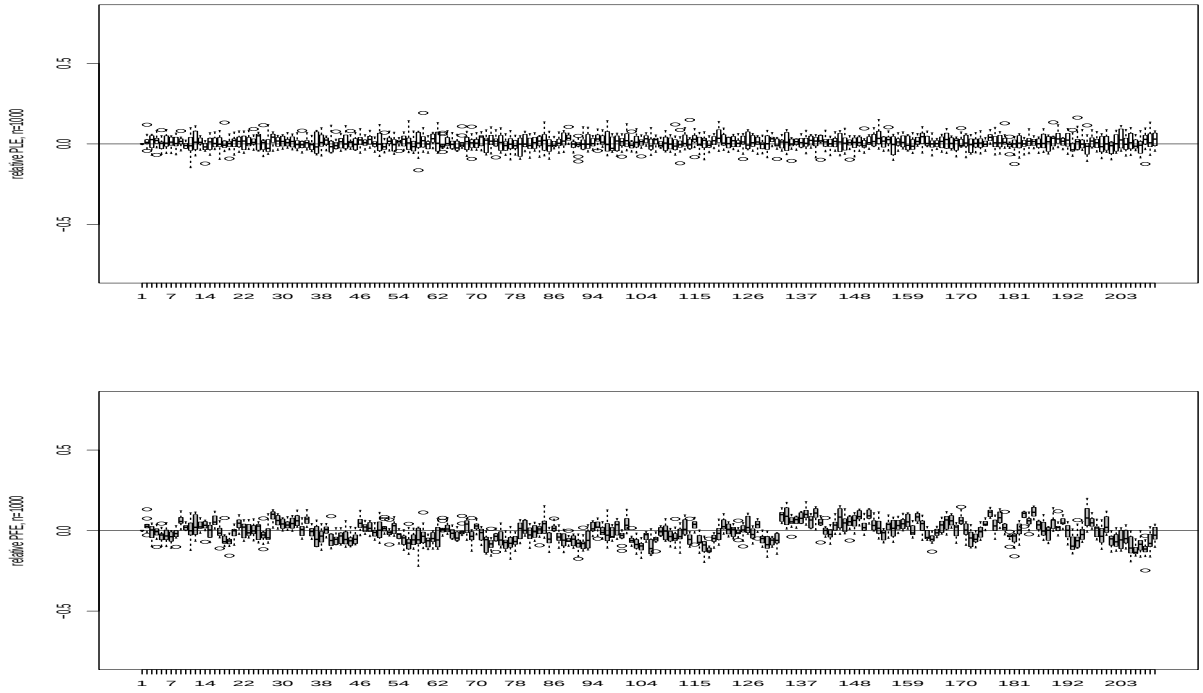


Figure 7: Comparison of MPLE and PFE estimators for $k = 5$ color 10×10 lattice Potts model, with sample sizes $n = 10$ and $n = 100$. Open boxplots give sampling distribution for pairing frequency estimates, grey boxplots represent pseudolikelihood estimates, and dark bars indicate true values.



(a)



(b)

Figure 8: Sampling distributions of errors for MPLE and PFE estimates of the 210 parameters of the $k = 20$ color Potts model. Shown are sample sizes (a) $n = 100$, and (b) $n = 1000$. At larger sample size, it can be seen that the PFE estimates converge to incorrect values.