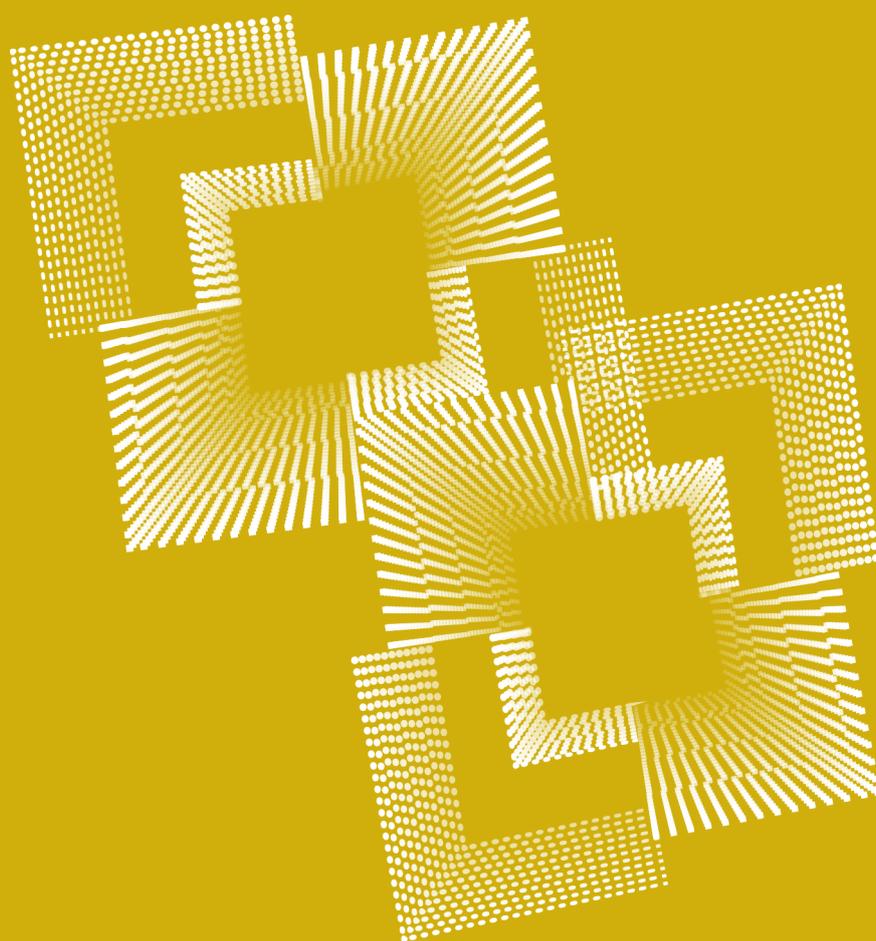# Pre-Processing for Approximate Bayesian computation in image analysis

M. T. Moores, C. C. Drovandi, K. Mengersen, C.P. Robert

CHAIRE HAVAS-DAUPHINE
ÉCONOMIE ET GESTION DES NOUVELLES DONNÉES

CONNECTING THE DOTS

Pre-processing for approximate Bayesian computation in image analysis

Matthew T. Moores[*][†]   Christopher C. Drovandi[*]

Kerrie Mengersen[*]   Christian P. Robert[‡][§][¶]

September 8, 2014

Abstract

Most of the existing algorithms for approximate Bayesian computation (ABC) assume that it is feasible to simulate pseudo-data from the model at each iteration. However, the computational cost of these simulations can be prohibitive for high dimensional data. An important example is the Potts model, which is commonly used in image analysis. Images encountered in real world applications can have millions of pixels, therefore scalability is a major concern. We apply ABC with a synthetic likelihood to the hidden Potts model with additive Gaussian noise. Using a pre-processing step, we fit a binding function to model the relationship between the model parameters and the synthetic likelihood parameters. Our numerical experiments demonstrate that the precomputed binding function dramatically improves the scalability of ABC, reducing the average runtime required for model fitting from 71 hours to only 7 minutes. We also illustrate the method by estimating the smoothing parameter for remotely sensed satellite imagery. Without precomputation, Bayesian inference is impractical for datasets of that scale.

1 Introduction

For many important problems, the computational cost of approximate Bayesian computation (ABC) is dominated by simulation of pseudo-data. This is particularly the case for the Potts model, for which an ABC algorithm was developed by Grelaud et al. [2009]. For real world applications in image analysis, the dimension of the state vector can correspond to millions of pixels. The

[*]Mathematical Sciences School, Queensland University of Technology, GPO Box 2434, Brisbane QLD 4001, Australia

[†]Current address: Department of Statistics, University of Warwick, Coventry, UK

[‡]CEREMADE, Université Paris Dauphine, 75775 Paris cedex 16, France

[§]CREST, INSEE, France

[¶]Department of Statistics, University of Warwick, Coventry, UK

distribution of these states is highly correlated, requiring algorithms such as Swendsen and Wang [1987] to simulate efficiently from the generative model. In his comparison of sequential Monte Carlo (SMC-ABC) with a particle Markov chain Monte Carlo (PMCMC) algorithm, Everitt [2012] found in both cases that the computational requirements were dominated by simulation of this state vector.

Adaptive ABC algorithms have been developed to reduce the number of iterations required through more efficient exploration of the posterior. Various particle-based methods have been proposed by Sisson et al. [2007], Beaumont et al. [2009], Toni et al. [2009], Drovandi and Pettitt [2011], Jasra et al. [2012], Filippi et al. [2013] and Sedki et al. [2013]. The SMC-ABC algorithm of Del Moral et al. [2012] uses multiple replicates of the summary statistics for each particle, which accounts for heteroskedasticity in the relationship between the statistics and the parameter values. This algorithm also has the advantage that computation of the importance weights is linear in the number of particles. Both the ABC tolerance ϵ and the Metropolis-Hastings (MH) proposal bandwidth σ_{MH}^2 are selected adaptively, which reduces the need to tune SMC-ABC for specific applications.

Regression adjustment for ABC was introduced by Beaumont et al. [2002], who post-processed the ABC output by fitting a local linear regression. By modelling the relationship between the simulated parameter values and the corresponding summary statistics, they improved the estimate of the posterior density even with large values of the ABC tolerance. Blum and François [2010] took a similar approach, except that they used a nonlinear, heteroskedastic regression. They then performed a second ABC run, using the estimate of the posterior to draw parameter values.

The idea proposed in this paper has a strong connection with indirect inference [e.g. Gouriéroux et al., 1993]. We assume that the underlying (intractable) distribution of the summary statistic of interest can be approximated well by an alternative parametric model with a tractable distribution. This so-called auxiliary model contains a different set of parameters. The indirect inference nature of our method requires learning the relationship between the parameters of the auxiliary model and the true model, often referred to as the mapping or binding function. Wood [2010] assume a normal parametric model with mean and covariance parameters. Our application involves a single summary statistic and we adopt the approach of Wood [2010] and use a normal distribution with a mean and a variance parameter. Wood [2010] learn the auxiliary parameters for each value of the true parameter proposed during a MCMC run by simulating independently a large collection (say of size M) of summary statistics from the true parameter and estimating the corresponding auxiliary parameters based on the sample moments of the generated summary statistics. We note that the resulting target distribution of the method depends on the choice of M [Drovandi et al., 2014]. Our approach differs in that we estimate the binding function prior to running our Bayesian algorithm via producing model simulations over a pre-defined grid (referred to here as the precomputation step). We attempt to recover the true mapping between the mean and variance auxiliary parameters with the true parameters (i.e. the limit as $M \to \infty$) by using non-

2

parametric regression techniques. An additional advantage is that the output of the precomputation can be reused for multiple datasets that share the same parameter space.

Despite superficial similarities with the "accurate ABC" method of Ratmann et al. [2014], where an auxiliary model is also constructed over a sufficient statistic, our approach cannot be seen as a special case of theirs as accurate ABC requires repeated observations of the summary statistics and consistency of the auxiliary model for its calibration step. We do not claim such proximity with the true posterior distribution and doubt it can be achieved for the models we consider below. Even though we are in the favourable case when the summary statistic is sufficient, most assumptions in Ratmann et al. [2014] do not apply to our setting. The construction of the binding function also highly differs in both motivations and complexity. We nonetheless agree that those different approaches of Wood [2010], Drovandi et al. [2011], Cabras et al. [2014], Ratmann et al. [2014], as well as ours, all relate in spirit to the original indirect inference method of Gouriéroux et al. [1993]. The reader is referred to Drovandi et al. [2014] for more details about the parametric auxiliary model approach adopted in this paper.

The rest of the paper is organised as follows. The ABC rejection sampler and SMC-ABC are reviewed in Sect. 2. Precomputation of the binding function is described in Sect. 3. In Sect. 4, we illustrate how this method can be applied to the hidden Potts model with additive Gaussian noise. Sect. 5 contains results from a simulation study as well as real imaging data from the Landsat 7 satellite. The article concludes with a discussion.

2 Approximate Bayesian Computation

The ABC rejection sampler introduced by Pritchard et al. [1999] draws values of the parameter from its prior distribution $\pi(\theta)$, then simulates pseudo-data $\mathbf{x}$ from the model. One or more summary statistics $s(\mathbf{x})$ are calculated from the pseudo-data and compared to the values of those statistics in the observed data, $s(\mathbf{y})$. If the difference between the statistics is within the ABC tolerance threshold, then the proposed parameter is accepted.

The simulation of pseudo-data from the generative model is typically the most computationally intensive step in this process. The ABC rejection sampler works best when prior information about the distribution of the parameter is available. The acceptance rate under a sparse or uninformative prior can be extremely low, requiring many pseudo-datasets to be generated for each parameter value that is accepted. The ABC tolerance is a tunable parameter, since a large tolerance means a higher acceptance rate but also increases the error in the estimate of the posterior distribution. If the summary statistic is sufficient for the parameter, then the distribution of the samples approaches the true posterior as ϵ approaches zero. However, the number of samples that are rejected also increases, to the point that almost none are accepted. Adaptive ABC methods have been developed to address this inefficiency.

2.1 Sequential Monte Carlo

The SMC-ABC algorithm of Del Moral et al. [2012] evolves a set of N parameter values, known as particles, through a sequence of T target distributions:

$$\pi_t(\theta, \mathbf{x}_{1\ldots M}|\mathbf{y}) \propto \frac{\sum_{m=1}^{M} \mathbb{I}\left(\Delta(\mathbf{x}_m) < \epsilon_t\right)}{M} \left(\prod_{m=1}^{M} p(\mathbf{x}_m|\theta)\right) \pi(\theta)$$

where ϵ_t is the ABC tolerance threshold such that $\epsilon_0 \geq \epsilon_1 \geq \cdots \geq \epsilon_T$, $\mathbf{x}_{1\ldots M}$ is a set of M replicates of the pseudo-data that are generated for each particle, $p(\mathbf{y}|\theta)$ is the likelihood, $\mathbb{I}(\cdot)$ is the indicator function, and $\Delta(\mathbf{x})$ is the distance between the summary statistics of the pseudo-data and the observed data, $\|s(\mathbf{x}) - s(\mathbf{y})\|$, for an appropriate norm. The algorithm involves four major stages: initialisation, adaptation, resampling, and mutation.

Initialisation The algorithm is initialised at $t = 0$ by drawing a population of particles $\theta_{i,t}$ from the prior, where $i \in 1 \ldots N$. Each particle is associated with M replicates of the summary statistics calculated from pseudo-data. The generation of multiple sets of pseudo-data $\mathbf{x}_{i,m,t}$ for each particle increases the computational cost relative to other ABC methods, but it better handles the situation where there is sizeable variability in the value of the summary statistic for a given parameter. This is the case for the hidden Potts model, as we explain in Sect. 4.

Adaptation At each iteration, the particles are assigned importance weights based on the following formula:

$$w_{i,t} \propto w_{i,t-1} \frac{\sum_{m=1}^{M} \mathbb{I}\left(\Delta(\mathbf{x}_{i,m,t-1}) < \epsilon_t\right)}{\sum_{m=1}^{M} \mathbb{I}\left(\Delta(\mathbf{x}_{i,m,t-1}) < \epsilon_{t-1}\right)} \tag{1}$$

These weights are normalised so that $\sum_{i=1}^{N} w_{i,t} = 1$. These weights gradually degenerate over successive iterations, which is measured by the effective sample size [ESS; Liu, 2001, pp. 34–36]:

$$ESS_t = \left(\sum_{i=1}^{N} w_{i,t}^2\right)^{-1} \tag{2}$$

The ABC tolerance ϵ_t is updated adaptively according to the desired rate $\alpha \in (0,1)$ of the reduction in the ESS:

$$ESS_t = \alpha ESS_{t-1} \tag{3}$$

This equation must be solved iteratively, e.g. by interval bisection, since ESS_t depends on the weights $w_{\cdot,t}$, which in turn depend on ϵ_t according to (1).

Resampling If ESS_t falls below a threshold value $N_{\min}$ then the particles are all resampled. The new population of N particles can either be drawn from a multinomial distribution with weights $\lambda_i = w_{i,t}$ or more complicated schemes can be used. Del Moral et al. employed the systematic resampling scheme of Kitagawa [1996]. Once the particles have been resampled, all of the importance weights are set to N^{-1} and thus the ESS is equal to N.

Mutation Finally, the particles with nonzero weight are updated using a random walk proposal $q_t(\theta'|\theta_{t-1})$. The bandwidth σ^2_{MH} can be chosen adaptively using an importance sampling approximation of the variance of θ under $\pi_{t-1}(\theta|\mathbf{y})$, as in Beaumont et al. [2009]. The pseudo-data is also updated using $q(\mathbf{x}'|\theta')$ and jointly accepted with probability $\min(1, \rho_i)$ according to the MH acceptance ratio:

$$\rho_i = \frac{\sum_{m=1}^{M} \mathbb{I}\left(\Delta(\mathbf{x}'_{i,m}) < \epsilon_t\right)}{\sum_{m=1}^{M} \mathbb{I}\left(\Delta(\mathbf{x}_{i,m,t-1}) < \epsilon_t\right)} \frac{q_t(\theta_{i,t-1}|\theta'_i)}{q_t(\theta'_i|\theta_{i,t-1})} \frac{\pi(\theta'_i)}{\pi(\theta_{i,t-1})} \tag{4}$$

3 Precomputation of the Summary Statistic

Model fitting with ABC can be decomposed into two separate stages: learning about the summary statistic, given the parameter $(s(\mathbf{x})|\theta)$; and choosing parameter values, given the summary statistic $(\theta|s(\mathbf{y}))$. The first stage is achieved by simulating pseudo-data from the generative model in a so-called precomputation step (discussed below), while the second is achieved via a Bayesian algorithm that utilises the output of the precomputation. In the case of latent models there are additional complications that will be discussed further in Sect. 4.2. To ease the exposition of this section, we assume the data $\mathbf{y}$ are observed without error.

The precomputation step involves simulating pseudo-data for fixed values of the parameter. Firstly we approximate the intractable distribution of the summary statistic, $f(s(\mathbf{y})|\theta)$, with an alternative parametric model that has a tractable likelihood function, $f_A(s(\mathbf{y})|\phi(\theta))$, which contains a different set of parameters ϕ. There is a strong connection with indirect inference (e.g. Gouriéroux et al. [1993]) as our method requires learning the mapping between θ and ϕ, $\phi(\theta)$, often referred to as the binding function.

Wood [2010] adopts this approach and uses the auxiliary likelihood within a MCMC algorithm. For each proposed value of θ, a set of M independent pseudo-datasets are generated from the true model, $\mathbf{x}_{1:M} = (\mathbf{x}_1, \ldots, \mathbf{x}_M)$, and the corresponding summary statistics are constructed, $s_{1:M} = (s(\mathbf{x}_1), \ldots, s(\mathbf{x}_M))$, which can be viewed as an independent and identically distributed sample from $s(\mathbf{x})|\theta$. The auxiliary model is then fit (using maximum likelihood or the method moments for example) to this sample in order to estimate the parameter ϕ. We denote this estimate of the mapping $\phi(\theta)$ as $\phi(\theta, s_{1:M})$. Drovandi et al. [2014] note that the target distribution of this method depends on the value of M, and

if the auxiliary model chosen is suitable, it is desirable to take M as large as possible.

The approach that we use to learn the mapping is different to that in Wood [2010]. Here we define a (not necessarily regular) grid over the parameter space Θ. For each θ within the grid a set of summary statistics, $s_{1:M}$, are generated from the true model and the corresponding auxiliary parameter is estimated, $\phi(\theta, s_{1:M})$. The next step is to use non-parametric regression techniques in order to smooth out the effect due to a finite choice of M and to obtain a direct approximation of the mapping, which we denote $\hat{\phi}(\theta)$. Here each component of ϕ is regressed on the model parameter θ. These non-parametric regression models can be used to predict the mapping for θ values not present in the grid. We refer to this procedure for estimating the mapping as the precomputation step. Our approach, in addition to mitigating the effect of M, has the advantage of being very useful when fitting the same model to multiple datasets because the output of the precomputation can be reused, thus amortising its computational cost. For alternative approaches to combining nonparametric regression models with ABC, see the regression adjustment method of Blum and François [2010] and the meta-modelling or emulation approach of [Wilkinson, 2014].

Our application involving the Potts model only has a single summary statistic and we follow Wood [2010] and assume a normal distribution for f_A with mean μ and variance σ^2. Therefore our precomputation step requires the generation of M pseudo-datasets for each value of β across a grid, recording the sample mean and standard deviation of the summary statistic (i.e. estimates of the auxiliary parameters μ and σ) and then applying two separate non-parametric regressions to estimate the mappings $\mu(\beta)$ and $\sigma(\beta)$. It is important to note that for reasonable size images it is not computationally feasible to generate perfect samples from the Potts model. Instead we use an MCMC algorithm to generate a set of M correlated pseudo-datasets and use these to calculate $s_{1:M}$ (see Sect. 4.1 for specific details of this for the Potts model).

There are several approximations induced by our method. The first is associated with replacing f with f_A. Despite the summary statistic of the Potts model being discrete, our choice of the normal distribution is appropriate due to the size of the images being analysed. The second source of approximation arises from the estimated mapping, $\hat{\phi}(\theta)$, due to the non-parametric regressions. However, the quality of the constructed mapping can be assessed visually and/or by standard data analytic techniques. For convenience in this paper we also introduce a third and seemingly unnecessary approximation effect. The natural implementation, used in Wood [2010] and Drovandi et al. [2014] for example, uses the auxiliary likelihood estimate $f_A(s(\mathbf{y})|\hat{\phi}(\theta))$ directly in a Bayesian algorithm and avoids any comparison of observed and simulated summary statistics (that is, specification of an ABC tolerance ϵ is not required). In this paper we use a more traditional ABC approach and draw pseudo summary statistics from $f_A(s(\cdot)|\hat{\phi}(\theta))$ for comparison with the observed data (and thus require ϵ). We adopt this approach here for two reasons. The first is for ease of implementation; having already developed an SMC-ABC algorithm, all that is required is

to replace the (expensive) simulation of pseudo-data from the model with the (cheap) summary statistic draw from the auxiliary model. The second motivation for our implementation approach is that a more direct comparison of the computational cost can be made between the SMC-ABC method with and without the precomputation step.

4 Hidden Potts Model

We illustrate our method using a hidden Potts model. The Potts model is a Markov random field with discrete states $z \in 1 \dots k$. It is defined in terms of its conditional probabilities:

$$p(z_i|z_{i \sim \ell}, \beta) = \frac{\exp\left\{\beta \sum_{i \sim \ell} \delta(z_i, z_\ell)\right\}}{\sum_{j=1}^{k} \exp\left\{\beta \sum_{i \sim \ell} \delta(j, z_\ell)\right\}} \tag{5}$$

where $i \in 1 \dots n$ are the nodes in the image lattice, also known as pixels, β is a scale parameter known as the inverse temperature, $i \sim \ell$ are the neighbouring pixels of i, and $\delta(u, v)$ is the Kronecker delta function. In this paper we use the first-order neighbourhood, so $i \sim \ell$ are the four pixels immediately adjacent to an internal node of the lattice. Pixels on the image boundary have less than four neighbours.

The inverse temperature parameter governs the strength of spatial association. A value of zero corresponds with spatial independence, while values greater than zero increase the probability of adjacent neighbours having the same state. The full conditional distribution of the inverse temperature is given by

$$p(\beta|\vec{z}) = \mathcal{C}(\beta)^{-1} \pi(\beta) \exp\left\{\beta \mathrm{S}(\mathbf{z})\right\} \tag{6}$$

where $\mathcal{C}(\beta)$ is an intractable normalising constant. It involves a sum over all k^n possible combinations of the labels $\mathbf{z} \in \mathcal{Z}$:

$$\mathcal{C}(\beta) = \sum_{\mathbf{z} \in \mathcal{Z}} \exp\left\{\beta \mathrm{S}(\mathbf{z})\right\} \tag{7}$$

A sufficient statistic is available for this model since it belongs to the exponential family, as noted by Grelaud et al. [2009]. If $\mathcal{E}$ is the set of all unique neighbour pairs, or edges in the image lattice, then the sufficient statistic is

$$\mathrm{S}(\mathbf{z}) = \sum_{i \sim \ell \in \mathcal{E}} \delta(z_i, z_\ell) \tag{8}$$

Thus, this statistic represents the total number of like neighbor pairs in the image. As β approaches infinity, all of the pixels in the image are almost surely assigned the same label, thus the expectation of $\mathrm{S}(\mathbf{z})$ approaches the total number of edges $|\mathcal{E}|$ asymptotically, while the variance approaches zero. When $\beta = 0$, the probability of any pair of neighbours being assigned the same label follows an independent Bernoulli distribution with $p = k^{-1}$, thus $\mathrm{S}(\mathbf{z})$ follows
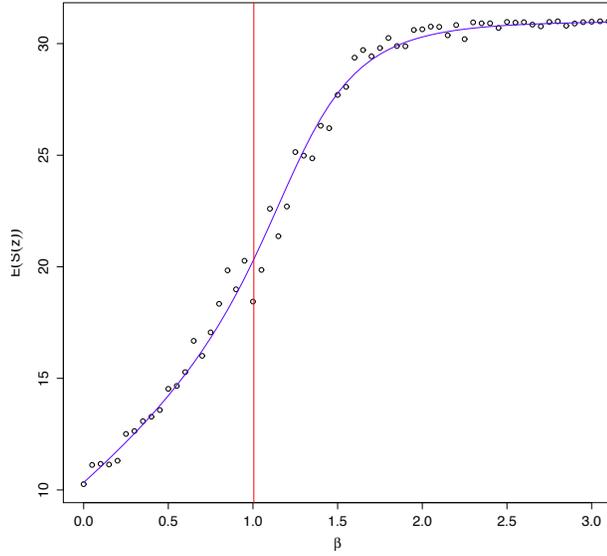
(a) Expectation of S($\vec{z}$)
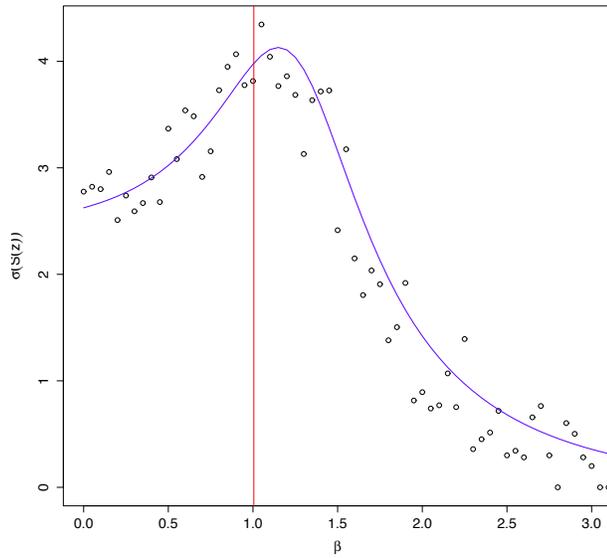
(b) Standard deviation of S($\vec{z}$)

Figure 1: Comparison of the exact distribution of the Potts model for a trivial image ($n = 20$, $k = 3$) with approximations using the Swendsen-Wang algorithm. The continuous, curved line is the exact value and the points are the approximations. The vertical line marks the critical value of β.

a Binomial distribution with expectation $|\mathcal{E}|/k$ and variance $|\mathcal{E}| k^{-1}(1 - k^{-1})$. The distribution of S($\mathbf{z}$) changes smoothly between these two extremes, as illustrated by Fig. 1, but its computation is intractable for nontrivial images. The expectations and standard deviations in Fig. 1 were calculated for $k = 3$ unique labels and $n = 20$ pixels, much less than required for any practical application.

The maximum variance of S($\mathbf{z}$)$|\beta$, which corresponds to the steepest gradient in the expectation, occurs near the critical temperature. This is the point at which the Potts model transitions from a disordered to an ordered state. The phase transition behaviour has analogies in physical systems, such as the Curie temperature in ferromagnetic materials. When $\beta > \beta_{\text{crit}}$, the values of the labels begin to exhibit long-range dependence and coalesce into clusters of similar values. Potts [1952] showed that this critical point can be calculated exactly for a two-dimensional regular lattice by $\beta_{\text{crit}} = \log\{1 + \sqrt{k}\}$, so $\beta_{\text{crit}} \approx 1.005$ for $k = 3$ and $\beta_{\text{crit}} \approx 1.238$ for $k = 6$. The nonlinearity and heteroskedasticity evident in Fig. 1 will need to be accounted for in our choice of binding function $\hat{\phi}(\beta)$.

4.1   Precomputation of $S(z)$

Since it is impossible to sample from $\beta|\mathbf{z}$ directly, we use ABC methods. This requires simulating pseudo-data from the Gibbs distribution $\pi(z_i|z_{i\sim\ell}, \beta)$ defined by (5). It is difficult to simulate from this distribution because neighbouring pixels are highly correlated with each other, particularly for $\beta > \beta_{\text{crit}}$, thus pixels can remain in the same state for many iterations. We use the algorithm of Swendsen and Wang [1987], which updates clusters of pixels simultaneously. The effect of the approximation error can be seen in Fig. 1 for a trivial image, where computation of the exact likelihood is feasible. Even though Swendsen-Wang is much less computationally intensive than perfect sampling, simulating pseudo-data remains expensive, which is why we approximate $f_A(\text{S}(\mathbf{z})|\hat{\phi}(\beta))$ offline using a pre-processing step.

To estimate the binding function for the Potts model, we use 1000 values of the inverse temperature, drawn from a truncated normal distribution $\beta \sim \mathcal{N}\left(\beta_{\text{crit}}, (\beta_{\text{crit}}/2)^2\right)\mathbb{I}(\beta > 0)$. This concentrates computational effort in the vicinity of the critical region, as shown in Fig. 2. The expectation and standard deviation have been approximated for a regular lattice with $n = 125 \times 125$ pixels and $k = 3$, which corresponds to the simulation study in section 5.1. For comparison, we show linear interpolation between 61 values of β on a regularly spaced grid, so that the approximation error due to a finite sample size can be seen. The simpler binding function provides a very good fit for the distribution of the expected values, but there is a larger approximation error in the estimate of the standard deviation. This is particularly evident at the critical point, where the variance of the sufficient statistic is much larger than estimated. We found that a binding function based on only 61 values did not provide sufficient accuracy for our purposes.

We perform 1000 iterations of Swendsen-Wang for each value of β, discarding the first 500 as burn-in. The remaining $M = 500$ iterations are used to compute
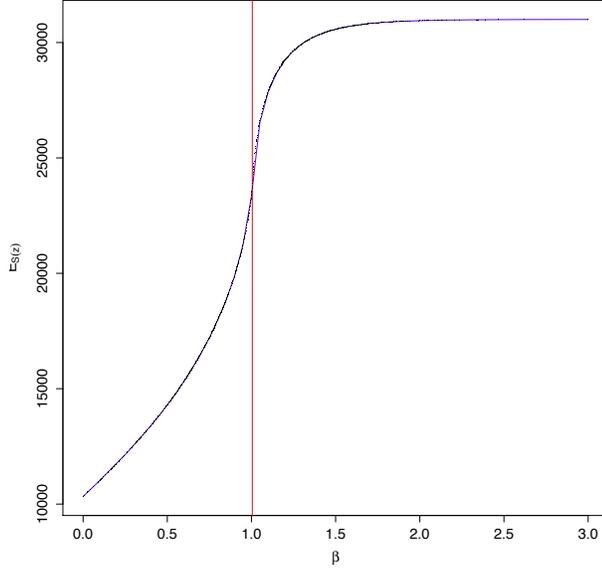
(a) Estimate of the expectation $\mathrm{E}\{\mathrm{S}(\mathbf{z})|\beta\}$
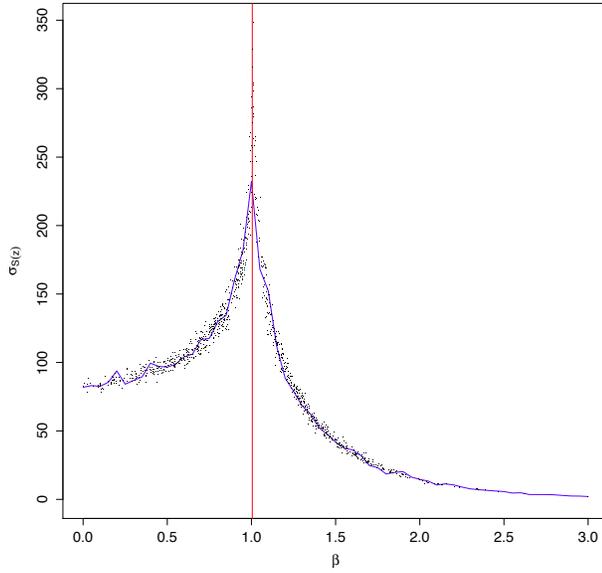
(b) Estimate of the standard deviation $\sigma\{\mathrm{S}(\mathbf{z})|\beta\}$

Figure 2: Approximation of the expectation and standard deviation of the sufficient statistic for the Potts model with $n = 125 \times 125$ and $k = 3$. The continuous line shows linear interpolation for 61 values of $\mathrm{S}(\vec{z})|\beta$, while the points are 987 values approximated using Swendsen-Wang.
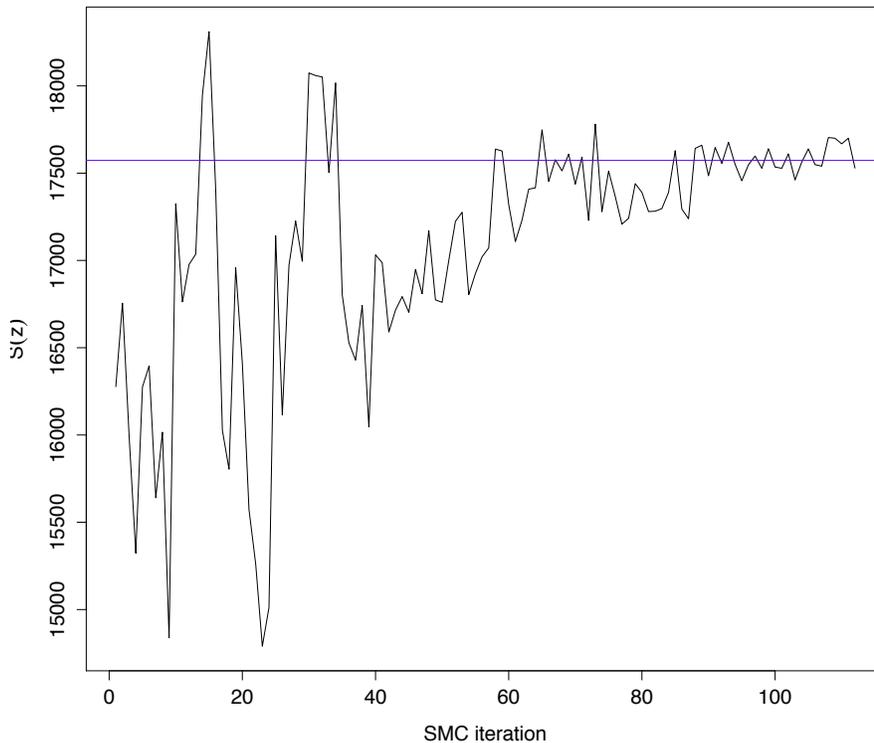
Figure 3: Change in the value of the sufficient statistic according to the current distribution of $\pi_t(\beta|\mathbf{z})$. The horizontal line shows the true value of S($\mathbf{z}$).

the expectation and variance of the sufficient statistic. It should be noted that this operation is embarrassingly parallel, since the computation for each value of β is completely independent. The results of this pre-processing step are stored in a matrix, which can then be used to fit the same model to multiple datasets. During model fitting, these pre-computed values are interpolated to obtain $\hat{\mu}(\beta')$ and $\hat{\sigma}(\beta')$, then the conditional distribution of the sufficient statistic is approximated by a Gaussian with these parameters.

4.2 Additive Gaussian noise

In the hidden Potts model, the observed data $\mathbf{y}$ are independently distributed conditional on the latent labels $\mathbf{z}$. Under the assumption of additive Gaussian

noise, the observation process is characterised by

$$y_i|z_i = j \quad \overset{iid}{\sim} \quad \mathcal{N}\left(\mu_j, \sigma_j^2\right) \tag{9}$$

Since each unique label value corresponds to the mean and variance of a Normal distribution, this model can be viewed as a spatially-correlated generalisation of the mixture of Gaussians.

When the parameters of these mixture components are unknown, they must be estimated as part of the model fitting procedure. The joint distribution of all random quantities is given by:

$$p(\mathbf{y}, \mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}, \beta) = p(\mathbf{y}|\mathbf{z}, \mu, \sigma^2)\pi(\mu)\pi(\sigma^2)p(\mathbf{z}|\beta)\pi(\beta)$$

where the vectors of noise parameters are $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)$ and $\boldsymbol{\sigma^2} = (\sigma_1^2, \ldots, \sigma_k^2)$.

The conditional distribution of the latent labels is dependent on both the current likelihood of the data as well as the distribution of the particles:

$$p(z_i|y_i, \boldsymbol{\mu}, \boldsymbol{\sigma^2}, \beta) = \frac{p(y_i; \mu_{z_i}, \sigma_{z_i}^2)}{\sum_{j=1}^k p(y_i; \mu_j, \sigma_j^2)} \; p(z_i|z_{i\sim\ell}, \beta) \tag{10}$$

where $p(y; \mu, \sigma^2)$ is the Gaussian likelihood and $p(z_i|z_{i\sim\ell}, \beta)$ is the Markov random field defined by (5).

This data augmentation approach can be problematic for SMC-ABC because it means that the summary statistic is a moving target, as shown in Fig. 3. The summary statistic in (8) is computed from the latent labels, therefore it only depends indirectly on the observed data through the noise parameters $\boldsymbol{\mu}, \boldsymbol{\sigma^2}$. Accurate posterior inference for β is only possible to the extent that the pixel labels and the noise parameters have also been estimated correctly. It can take many SMC iterations for all of the components of the model to converge. This is the third stage of ABC estimation for latent models that was mentioned in section 3.

One approach to overcoming this circular dependency would be to include the noise parameters μ_j, σ_j^2 in the state vector for each SMC particle. Updating these parameters would require generating latent labels from (10), which would severely limit the scalability of the algorithm. Simulating from the distribution of $\pi(\mathbf{z}|\mathbf{y}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}, \beta)$ is even more difficult than drawing pseudo-data from $\pi(\mathbf{z}|\beta)$ [Hurn, 1997, Higdon, 1998]. It is simply infeasible to do this for each particle individually. Apart from the multiplication of computational cost by serveral orders of magnitude, there is also the issue of memory utilisation when updating particles in parallel. A single copy of the state vector can require several megabytes of memory, depending on the size of the image. Maintaining a copy of $\mathbf{z}$ for each thread would be impractical for massively parallel implementation of the algorithm.

We propose a pragmatic alternative that preserves the scalability of our algorithm, while enabling estimation of all of the components of the hidden Potts model. We run a single MCMC chain in conjunction with our SMC-ABC algorithm, although it would also be possible to maintain a small number of parallel

12

chains. At the end of each SMC iteration, we draw a random sample from the current distribution of $\pi_t(\beta|\mathbf{z})$ according to the importance weights of the particles. These parameter values are used to update the pixel labels according to (10) by performing multiple iterations of chequerboard sampling [Winkler, 2003, chap. 8], one for each value of the parameter. After this sequence of updates, the new pixel labels reflect the distribution of β at the current SMC iteration. It is important to note that it requires several iterations to make substantial changes to the distribution of $\mathbf{z}$, thus it is the aggregate effect of all of these updates that transitions from $\mathbf{z}|\pi_{t-1}(\beta)$ to $\mathbf{z}|\pi_t(\beta)$. The new state vector is used to calculate S($\mathbf{z}$), as well as the sufficient statistics of the noise parameters $\bar{y}_j|\mathbf{z}$ and $s_j^2|\mathbf{z}$. We used 1000 MCMC iterations per iteration of SMC to produce the results in section 5.

One downside of this method is that the MCMC chain can become stuck in a low-probability region of the state space. This is the cause of the outliers that are evident in Fig. 7. Once the Markov chain crosses the phase transition boundary of the Potts model from a disordered to an ordered state, the correlations between neighbouring pixels make the probability of transitioning back extremely low, irrespective of the values of β that are used. To mitigate this problem, we initialize $\mathbf{z}, \boldsymbol{\mu}$ and $\boldsymbol{\sigma^2}$ at $\beta = 0$ during the initialisation phase of our algorithm. This increases the probability that the starting value of S($\mathbf{z}$) will be below the critical region.

5 Illustration

This section contains experimental results with synthetic data as well as real satellite imagery. In Sect. 5.1 we evaluate the performance and accuracy of our method for 20 images that have been simulated using the Swendsen-Wang algorithm for known values of β. We are thus able to compare the posterior distribution with the true parameter value. In Sect. 5.2 we demonstrate our method on an image obtained from the Landsat 7 satellite. This demonstrates that our pre-computation step enables inference with ABC for datasets of realistic size. We begin by calibrating our method against the SMC-ABC algorithm of Del Moral et al. [2012] as well as the approximate exchange algorithm [Murray et al., 2006, Cucala et al., 2009, Friel and Pettitt, 2011].

An R source package containing these algorithms is provided in Online Resource 1. Its computational engine is implemented in C++ using RcppArmadillo [Eddelbuettel and Sanderson, 2014] with OpenMP for parallelism. Reference implementations of SMC-ABC are available in the supplementary material accompanying Del Moral et al. [2012] and Everitt [2012]. Source code for the approximate exchange algorithm has been provided by Friel and Pettitt [2011] and Everitt [2012].

The elapsed times were recorded on 2.66GHz Intel Xeon processors. We used 8 parallel cores for fitting the model to each image and the precomputation was performed on a dual-CPU computer with 16 parallel cores. Memory usage varied depending on the number of pixels and the degree of parallelism. Approximate
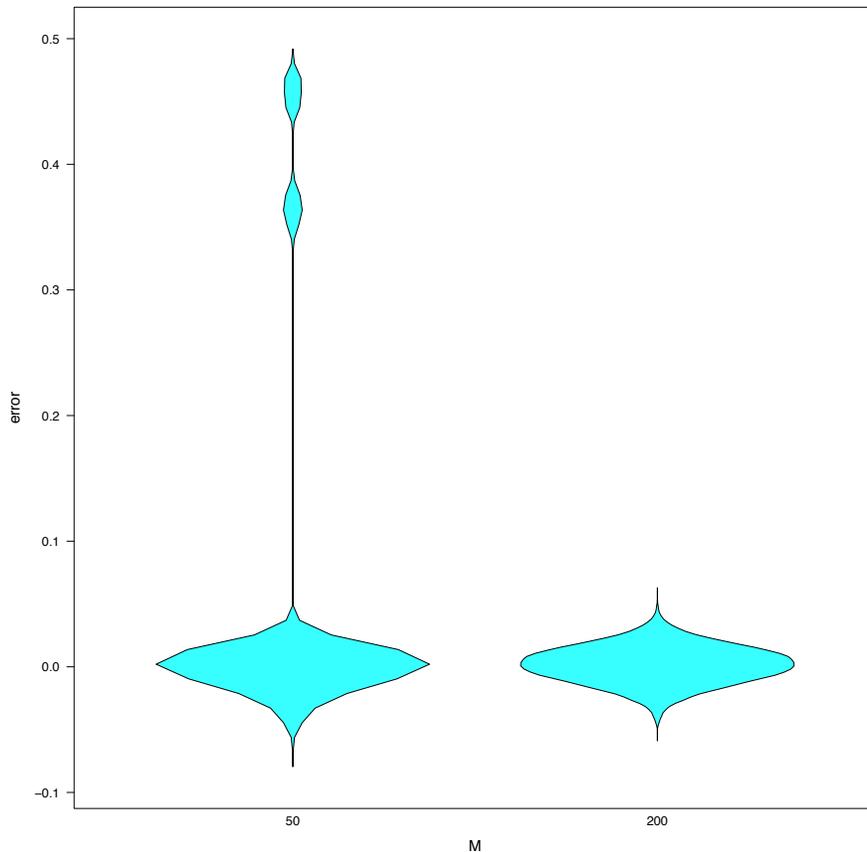
Figure 4: Distribution of posterior sampling error for β, comparing SMC-ABC using $M = 50$ replicates of the sufficient statistics with $M = 200$.

memory requirements for each computation are reported below.

We fit the model to each image using $N = 10,000$ SMC particles with $\alpha = 0.97$, using residual resampling [Douc et al., 2005]. We drew $M = 200$ summary statistics per particle from our precomputed binding function, but found that it was infeasible to simulate that much pseudo-data during model fitting since it took an average of 2 hours 45 minutes per SMC iteration. Instead, we used $M = 50$ for SMC-ABC with pseudo-data, which resulted in runtimes that were more reasonable. This made a difference to the accuracy of the posterior, as shown in Fig. 4.

The exchange algorithm of Murray et al. [2006] is an exact method for intractable likelihoods, but only when it is feasible to use perfect sampling to simulate from $\pi(\mathbf{z}|\beta)$. Even then, many studies such as McGrory et al. [2009]
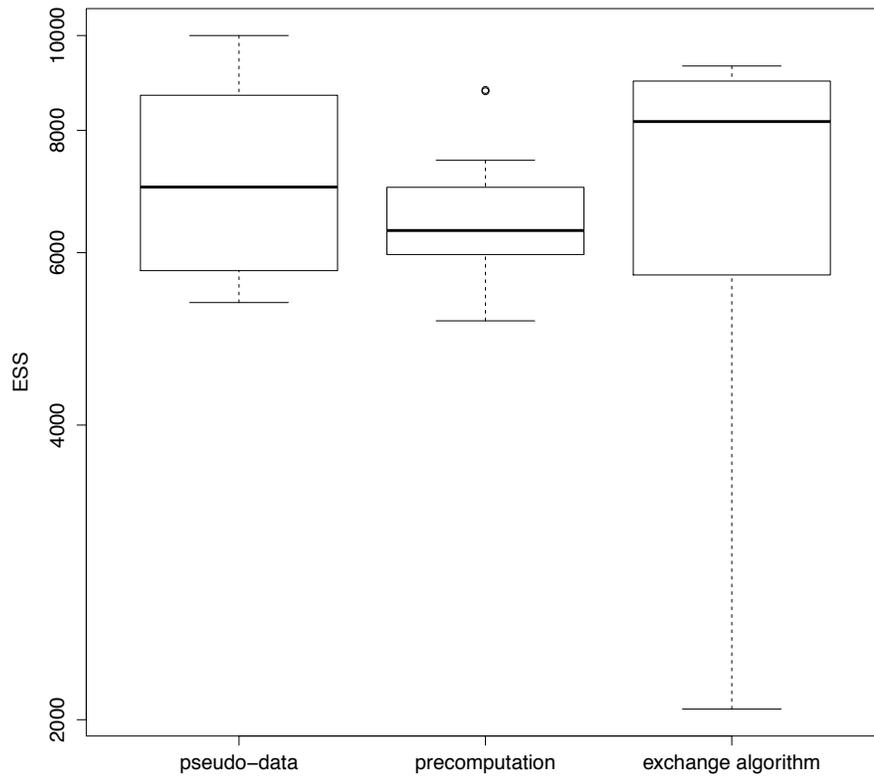
Figure 5: Distribution of effective sample size (ESS) for model fitting with the exchange algorithm in comparison to SMC-ABC with pseudo-data or precomputed $\hat{\phi}(\beta)$.

have shown that this algorithm is very computationally intensive and takes longer as the value of β increases. For this reason, Cucala et al. [2009] and Friel and Pettitt [2011] replaced the perfect sampling step with 500 iterations of Gibbs sampling, hence creating an approximate exchange algorithm. The effect of this approximation on the accuracy of posterior inference has been studied by Everitt [2012]. Since this is a MCMC method, the samples of $\pi(\beta|\mathbf{y})$ are correlated, reducing the effective sample size in comparison to the number of iterations. We found that 100,000 iterations were required to produce an ESS that was comparable with the SMC-ABC methods, as shown in Fig. 5.

5.1 Simulation Study

Since the inverse temperature cannot be directly observed, we have used a simulation study to evaluate the accuracy of our method where the true value of β is known. Following a similar methodology to that introduced by Cook et al. [2006], we first simulated 20 values of β from the prior, then generated 20 images from the model that corresponded to those parameter values. Each image had 125×125 pixels with $k = 3$ unique labels. We used a uniform prior on the interval $[0, 1.005]$ for β and natural conjugate priors $\pi(\mu_j) \sim \mathcal{N}(0, 100^2)$ and $\pi(\sigma_j^2) \sim \mathcal{IG}(1, 0.01)$ for the additive Gaussian noise.

Precomputation of the binding function took 1 hour 23 minutes for 987 values of β, using 1000 iterations of Swendsen-Wang for each. The resulting estimates of $\hat{\mu}(\beta)$ and $\hat{\sigma}(\beta)$ are illustrated in Fig. 2. Total CPU time for all 16 parallel threads was 21 hours 50 minutes, indicating over 98% utilisation of the available capacity. Memory usage was less than 1.3GB. Since the same mapping function was reused for all 20 images, this runtime could be amortised across the entire corpus. Thus, the cost of precomputation was less than 5 minutes per image.

Across the 20 simulated images, our algorithm took an average of 7 minutes per image for between 49 and 107 SMC iterations. Using pseudo-data to compute the sufficient statistic took an average of 71.4 hours for 45 to 111 SMC iterations. The approximate exchange algorithm took an average of 14.4 hours for 100,000 iterations. Fig. 6 illustrates that this two orders of magnitude difference in the distribution is consistent for both elapsed (wall clock) time and CPU time. This shows that the gain in performance is due to computational efficiency, not because of any increase in parallelism.

Fig. 7 shows that both SMC-ABC and the approximate exchange algorithm produced erroneous estimates of β for some of the images, with a large difference between the posterior distribution of the particles and the true value of the inverse temperature. This is most likely due to problems with the Markov chain undergoing phase transition, as explained in section 4.2. We were able to compensate for this problem by increasing the number of replicates of the summary statistic to $M = 200$. Thanks to precomputation of the binding function, this had little impact on the runtime of our algorithm.
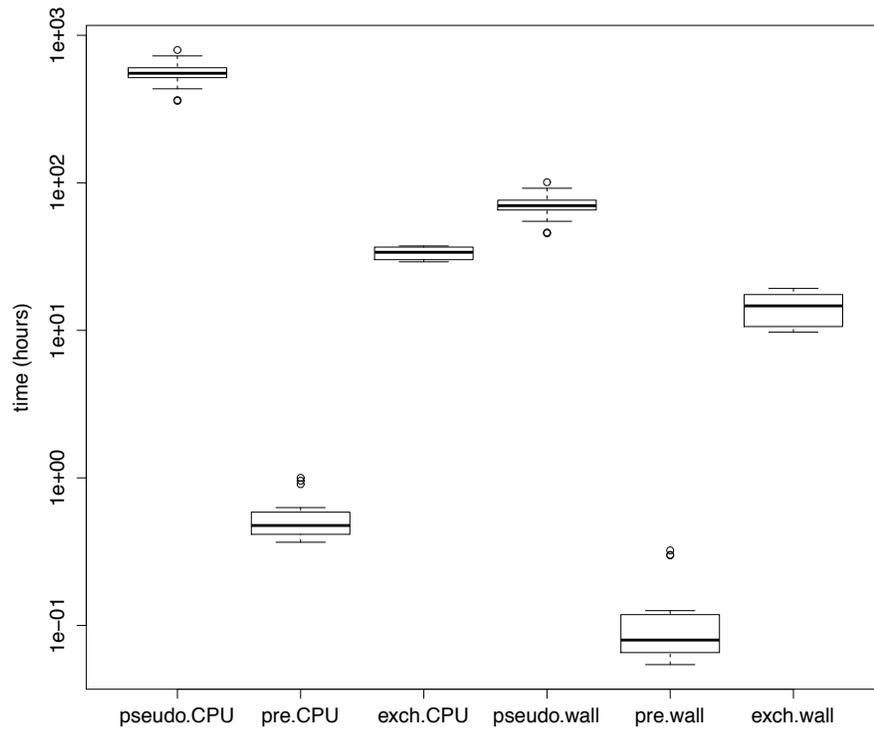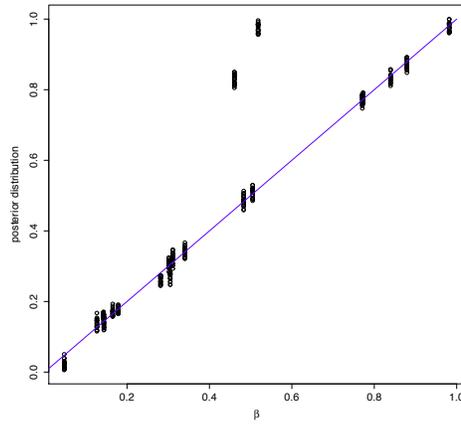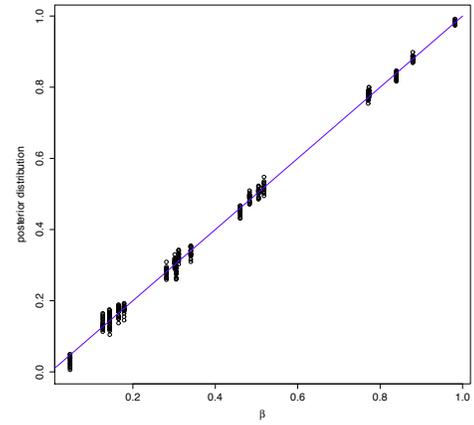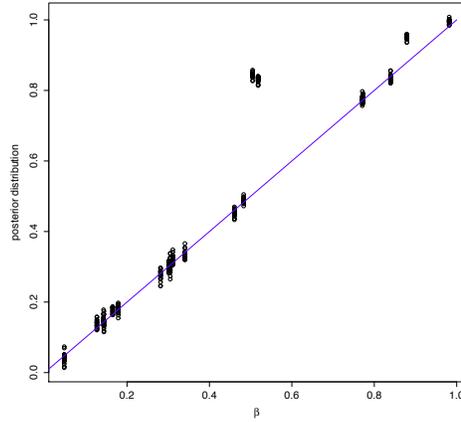
Figure 6: Distribution of CPU times (left) and elapsed (wall clock) times for model fitting with the exchange algorithm in comparison to SMC-ABC with pseudo-data or precomputed $\hat{\phi}(\beta)$.

(a) SMC-ABC with pseudo-data

(b) SMC-ABC with pre-computation

(c) approximate exchange algorithm

Figure 7: Results for the simulation study of 20 images. The x axis is the true value of β and the y axis shows the posterior samples from $\pi_t(\beta|\mathbf{z})$

5.2 Satellite Remote Sensing

We have also illustrated our method on real data, using a satellite image of Brisbane, Australia. The pixel values correspond to the Normalized Difference Vegetation Index (NDVI), which was calculated accordng to:

$$\text{NDVI} = \frac{NIR - VIS}{NIR + VIS} \tag{11}$$

In Landsat 7 images, the visible red (VIS) band corresponds to light wavelengths of $(0.63\ldots0.69)\mu m$ and the near-infrared (NIR) band corresponds to $(0.76\ldots0.90)\mu m$ [NASA, 2011].

The image was cropped to a region of interest that was approximately 40km east to west and 20km north to south, containing a total of 978,380 pixels. This region includes the city centre as well as suburbs and national parks to the south and west. By labelling the pixels, we aim to quantify the levels of vegetation present in the area and identify contiguous clusters of forest and parkland.

Precomputation of the mapping function took 13 hours 23 minutes for 987 values of β. Total CPU time for all 16 parallel threads was 75 hours 40 minutes, indicating over 98% utilisation of the available capacity. Memory usage was approximately 6.5GB.

We used weakly informative priors $\pi(\beta) \sim \mathcal{U}[0,3]$, $\pi(\mu_j) \sim \mathcal{N}(\bar{\bar{y}}, 5)$, and $\pi(\sigma_j^2) \sim \mathcal{IG}(1, 0.01)$ for the hidden Potts model with $k = 6$ unique labels. Model fitting took 5 hours 36 minutes using our algorithm. CPU time for 8 parallel threads was 39 hours, indicating 88% utilisation. The 95% posterior credible interval for β was $[1.243; 1.282]$.

Running the original SMC-ABC algorithm of Del Moral et al. [2012] on this dataset is clearly infeasible, due to the cost of simulating pseudo-data. It takes 89 hours to perform a single SMC iteration on our hardware. We also found that the exchange algorithm of Murray et al. [2006] was unable to scale to data of this dimension, even when using 500 iterations of Gibbs sampling for the auxiliary variable as recommended by Cucala et al. [2009]. It took 97 hours for 10,000 MCMC iterations. Discarding the first 5,000 as burn-in left an effective sample size of only 390 due to auto-correlation of the Markov chain.

6   Discussion

We have demonstrated that the scalability of ABC (and SMC-ABC in particular) can be dramatically improved using a precomputed binding function for indirect inference. We observed two orders of magnitude improvement in runtime in our simulation study, from an average of 71 hours down to only 7 minutes. This enables Bayesian inference for datasets of realistic scale, for which it was previously infeasible. An important example is imaging datasets with a million pixels or more, such as satellite remote sensing. We showed that our algorithm was able to estimate the smoothing parameter of a satellite image, while neither the approximate exchange algorithm nor the SMC-ABC algorithm of Del Moral et al. [2012] were scalable enough to be practical.

The computation of the binding function is embarrasingly parallel and therefore can make full use of modern computer architectures. Once the binding function has been computed, it can be reused to fit the same model to multiple datasets. This is an advantage in many applications such as satellite imaging, where there are a large number of images with approximately the same dimensions. In a longitudinal setting it would also be possible to update the binding function sequentially as each image is processed, to increase its resolution in the region of highest posterior density.

The other major issue that we have addressed is additive Gaussian noise, which is commonly encountered in real world imaging data. Everitt [2012] and Stoehr et al. [2014] have previously looked at ABC for latent models, where the state vector is not directly observed. However, their methods are only applicable where the observed data and the latent model are both discrete, sharing the same state space. The lack of identifiability induced by continuous observations creates a major problem for ABC, since the summary statistics become a moving target. We have introduced a pragmatic approach to reduce the tendency to become stuck in low-probability regions of the parameter space, while preserving the scalability of our method.

Our use of a precomputed binding function appears to be similar to that recently and independently developed in Cabras et al. [2014], who select a multivariate normal auxiliary model for the summary statistic and also use a precomputation step similar to above. It is important to note, however, that Cabras et al. [2014] assume that the covariance matrix of the auxiliary model is independent of θ. Such an assumption is severely violated in our Potts model application (see Fig. 1b) and unlikely to hold generally across models with intractable likelihoods. Furthermore, Cabras et al. [2014] assume throughout their paper that the summary statistic must be the same dimension as the parameter. We note here that this assumption is not required to use the preprocessing idea detailed in this paper. Finally, Cabras et al. [2014] use a regular grid over the parameter space. We suggest that for nonlinear, heteroskedastic auxiliary models it may be more appropriate to select a non-regular grid in order to obtain a good estimate of the mapping by including more points around regions where the gradient of the relationship is large.

The degree of complexity required for the binding function is dependent on the dimensionality of the parameter space, the number of summary statistics, and the properties of the relationship between them. These factors will also influence how much pseudo-data (in terms of the grid size and the number of replicate datasets) must be simulated in order to achieve a sufficiently good approximation of the likelihood. The nonlinear, heteroskedastic regression that was applied by Blum and François [2010] or Wilkinson [2014] would be a good choice in many cases, although simpler techniques for estimating the binding function could also be used. We assume that the summary statistics can be modelled as a continuous and smoothly-varying function of the parameters. The output of the precomputation step can be used to verify empirically that this assumption holds for the specific model under consideration.

Cabras et al. [2014] have shown that ABC with indirect inference can also

be applied to multivariate likelihoods, but only where the assumption of homoscedasticity is met. More research is needed to extend this method to Gibbs random fields such as the exponential random graph model (ERGM), which exhibits correlation between its summary statistics dependent on the value of θ.

The current methods for indirect inference would suffer from the curse of dimensionality if applied to models with a huge number of parameters and/or summary statistics, such as those encountered in genetics. Our method degrades very quickly as the number of parameters grows, because of a multi-grid requirement that grows as a power of the dimension of the parameter. We rely on asymptotic arguments as $M \to \infty$, therefore obtaining a sufficient number of simulations will become much more difficult as the dimension of the parameter space increases. The method that we have described is not universally applicable, nevertheless there exist a wide variety of models to which it could be successfully applied.

Acknowledgements

References

Mark A. Beaumont, Wenyang Zhang, and David J. Balding. Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

Mark A. Beaumont, Jean-Marie Cornuet, Jean-Michel Marin, and Christian P. Robert. Adaptive approximate Bayesian computation. *Biometrika*, 96(4): 983–990, 2009. doi: 10.1093/biomet/asp052.

Michael G. B. Blum and Olivier François. Non-linear regression models for approximate Bayesian computation. *Stat. Comput.*, 20(1):63–73, 2010. doi: 10.1007/s11222-009-9116-0.

Stefano Cabras, María Eugenia Castellanos, and Erlis Ruli. A quasi likelihood approximation of posterior distributions for likelihood-intractable complex models. *METRON*, 72(2):153–67, 2014. doi: 10.1007/s40300-014-0040-5.

Samantha R. Cook, Andrew Gelman, and Donald B. Rubin. Validation of software for Bayesian models using posterior quantiles. *J. Comput. Graph. Stat.*, 15(3):675–692, 2006. doi: 10.1198/106186006X136976.

Lionel Cucala, Jean-Michel Marin, Christian P. Robert, and D. Michael Titterington. A Bayesian reassessment of nearest-neighbor classification. *J. Am. Stat. Assoc.*, 104(485):263–273, 2009. doi: 10.1198/jasa.2009.0125.

Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Stat. Comput.*, 22(5):1009–1020, 2012. doi: 10.1007/s11222-011-9271-y.

Randal Douc, Olivier Cappé, and Éric Moulines. Comparison of resampling schemes for particle filtering. In *Proc. 4th Int. Symp. Image and Signal Processing and Analysis (ISPA)*, pages 64–69, Zagreb, Croatia, September 2005. IEEE. doi: 10.1109/ISPA.2005.195385.

Christopher C. Drovandi and Anthony N. Pettitt. Estimation of parameters for macroparasite population evolution using approximate Bayesian computation. *Biometrics*, 67(1):225–233, 2011. doi: 10.1111/j.1541-0420.2010.01410. x.

Christopher C. Drovandi, Anthony N. Pettitt, and Malcolm J. Faddy. Approximate Bayesian computation using indirect inference. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 60(3):317–337, 2011. doi: 10.1111/j.1467-9876.2010.00747.x.

Christopher C. Drovandi, Anthony N. Pettitt, and Anthony Lee. Bayesian indirect inference using a parametric auxiliary model. *Stat. Sci.*, 2014. To appear.

Dirk Eddelbuettel and Conrad Sanderson. RcppArmadillo: Accelerating R with high-performance C++linear algebra. *Comput. Stat. Data Anal.*, 71:1054–63, 2014. doi: 10.1016/j.csda.2013.02.005.

Richard G. Everitt. Bayesian parameter estimation for latent Markov random fields and social networks. *J. Comput. Graph. Stat.*, 21(4):940–960, 2012. doi: 10.1080/10618600.2012.687493.

Sarah Filippi, Chris P. Barnes, Julien Cornebise, and Michael P. H. Stumpf. On optimality of kernels for approximate Bayesian computation using sequential Monte Carlo. *Stat. Appl. Genet. Mol. Biol.*, 12(1):87–107, 2013. doi: 10. 1515/sagmb-2012-0069.

Nial Friel and Anthony N. Pettitt. Classification using distance nearest neighbours. *Stat. Comput.*, 21(3):431–437, 2011. doi: 10.1007/s11222-010-9179-y.

Christian Gouriéroux, Alain Monfort, and Eric Renault. Indirect inference. *J. Appl. Econometr.*, 8:S85–S118, 1993.

Aude Grelaud, Christian P. Robert, Jean-Michel Marin, François Rodolphe, and Jean-François Taly. ABC likelihood-free methods for model choice in Gibbs random fields. *Bayesian Analysis*, 4(2):317–336, 2009. doi: 10.1214/09-BA412.

David M. Higdon. Auxiliary variable methods for Markov chain Monte Carlo with applications. *J. Am. Stat. Assoc.*, 93(442):585–595, 1998. doi: 10.1080/01621459.1998.10473712.

Merrilee A. Hurn. Difficulties in the use of auxiliary variables in Markov chain Monte Carlo methods. *Stat. Comput.*, 7:35–44, 1997. doi: 10.1023/A:1018505328451.

Ajay Jasra, Sumeetpal S. Singh, James S. Martin, and Emma McCoy. Filtering via approximate Bayesian computation. *Stat. Comput.*, 22(6):1223–37, 2012. doi: 10.1007/s11222-010-9185-0.

Genshiro Kitagawa. Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *J. Comput. Graph. Stat.*, 5(1):1–25, 1996. doi: 10.1080/10618600.1996.10474692.

Jun S. Liu. *Monte Carlo Strategies in Scientific Computing.* Springer, New York, NY, 2001.

Clare A. McGrory, D. Michael Titterington, Robert Reeves, and Anthony N. Pettitt. Variational Bayes for estimating the parameters of a hidden Potts model. *Stat. Comput.*, 19(3):329–340, 2009. doi: 10.1007/s11222-008-9095-6.

Iain Murray, Zoubin Ghahramani, and David J. C. MacKay. MCMC for doubly-intractable distributions. In *Proc. 22nd Conf. UAI*, pages 359–366, Arlington, VA, 2006. AUAI Press.

NASA. Landsat 7 science data users handbook. Technical report, National Aeronautics and Space Administration, 2011. URL `http://landsathandbook.gsfc.nasa.gov/`.

Renfrey B. Potts. Some generalized order-disorder transformations. *Proc. Camb. Philos. Soc.*, 48:106–109, 1952. doi: 10.1017/S0305004100027419.

Jonathan K. Pritchard, Mark T. Seielstad, Anna Perez-Lezaun, and Marcus W. Feldman. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol. Biol. Evol.*, 16(12):1791–1798, 1999.

Oliver Ratmann, Anton Camacho, Adam Meijer, and Gé Donker. Statistical modelling of summary values leads to accurate approximate Bayesian computations. Technical report, 2014. arXiv:1305.4283.

Mohammed Sedki, Pierre Pudlo, Jean-Michel Marin, Christian P. Robert, and Jean-Marie Cornuet. Efficient learning in ABC algorithms. Technical report, 2013. arXiv:1210.1388.

Scott A. Sisson, Yanan Fan, and Mark M. Tanaka. Sequential Monte Carlo without likelihoods. *Proc. Natl Acad. Sci.*, 104(6):1760–1765, 2007. doi: 10. 1073/pnas.0607208104.

Julien Stoehr, Pierre Pudlo, and Lionel Cucala. Geometric summary statistics for ABC model choice between hidden Gibbs random fields. Technical report, 2014. arXiv:1402.1380.

Robert H. Swendsen and Jian-Sheng Wang. Nonuniversal critical dynamics in Monte Carlo simulations. *Phys. Rev. Lett.*, 58:86–88, 1987. doi: 10.1103/ PhysRevLett.58.86.

Tina Toni, David Welch, Natalja Strelkowa, Andreas Ipsen, and Michael P. H. Stumpf. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface*, 6(31):187–202, 2009. doi: 10.1098/rsif.2008.0172.

Richard David Wilkinson. Accelerating ABC methods using Gaussian processes. In *Proc. 17th Int. Conf. AISTATS*, volume 33 of *JMLR W&CP*, pages 1015–33, 2014.

Gerhard Winkler. *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods: A Mathematical Introduction.* Springer-Verlag, Berlin Heidelberg, 2nd edition, 2003.

Simon N. Wood. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102–4, 2010. doi: 10.1038/nature09319.
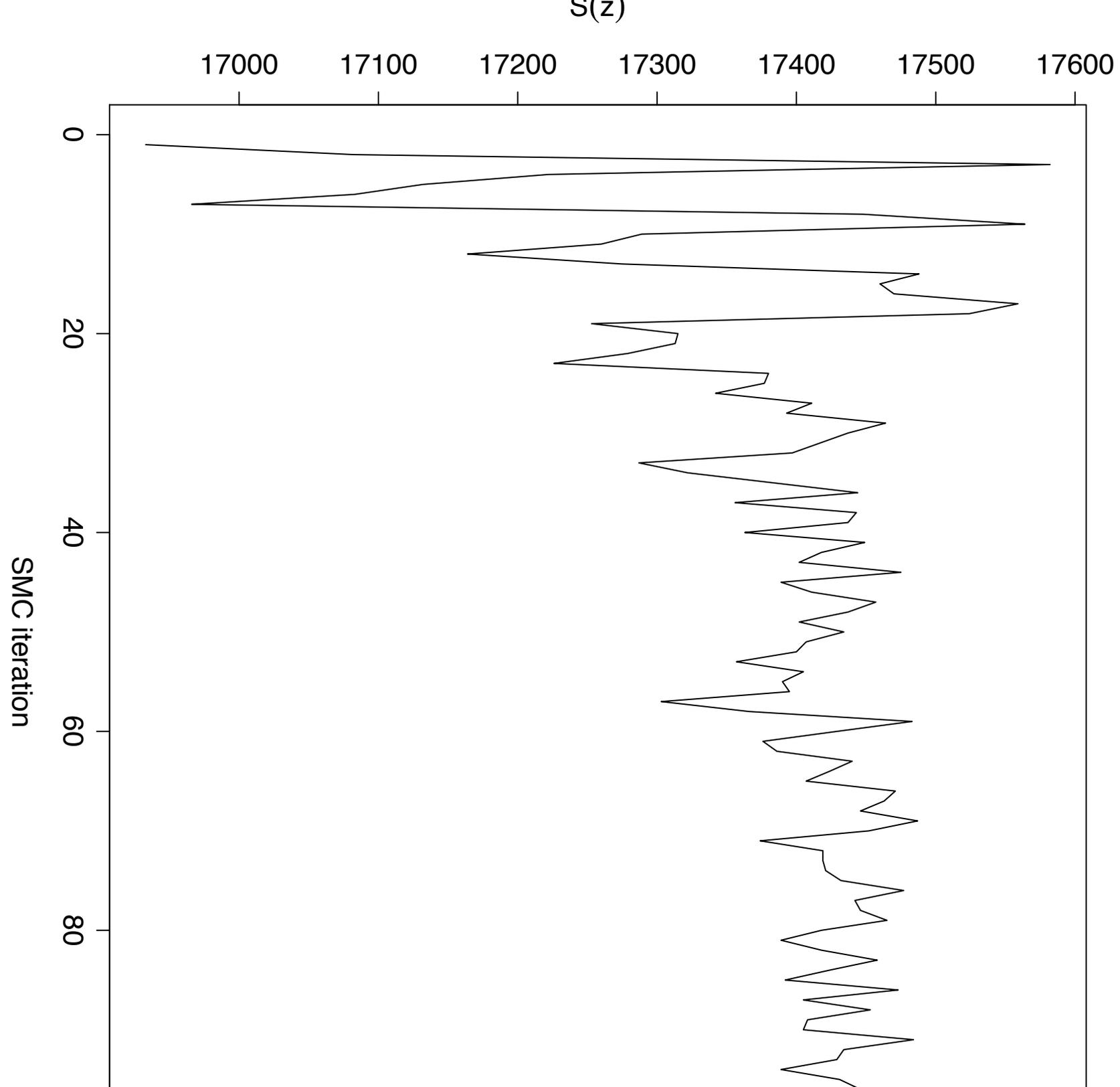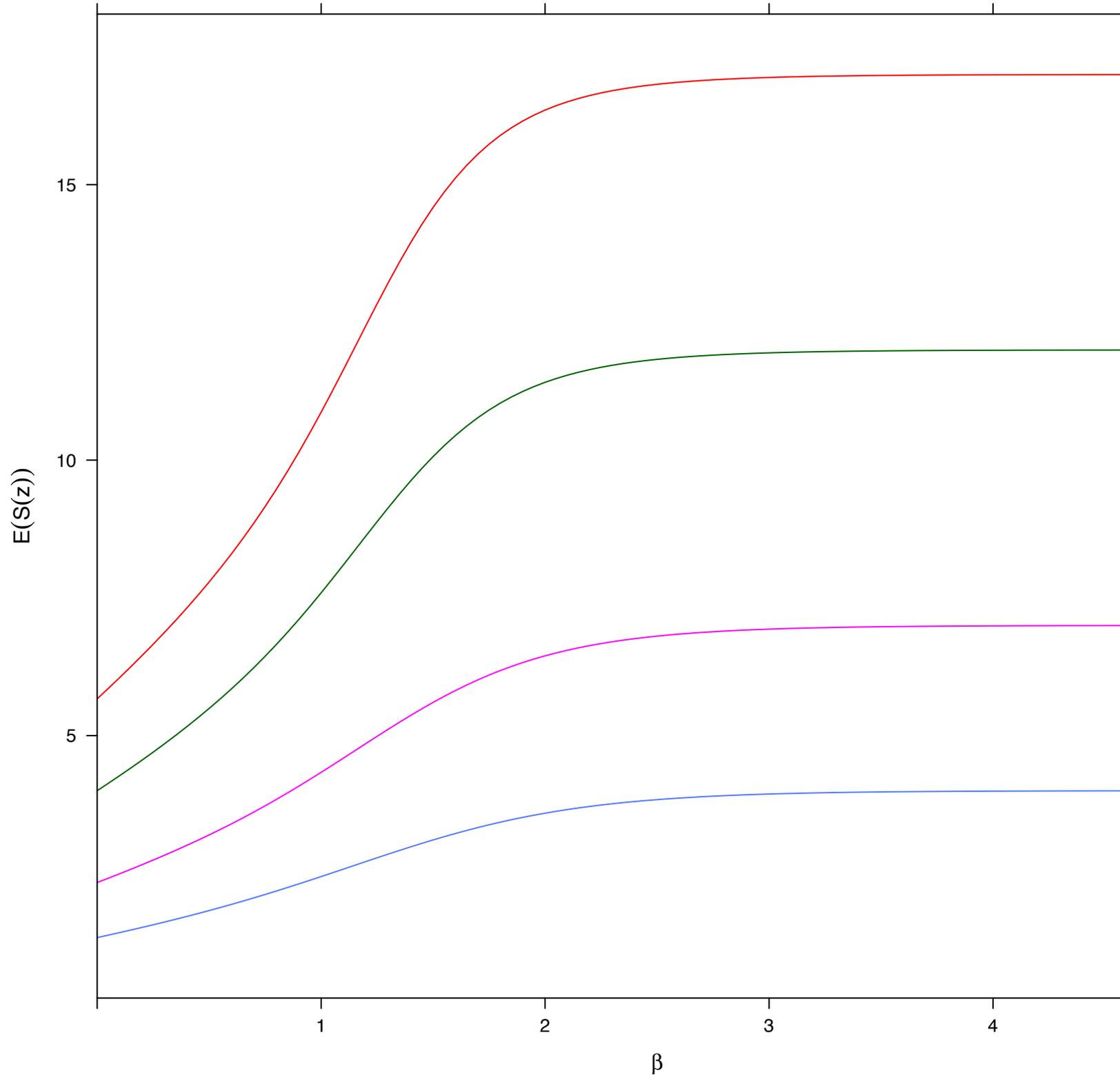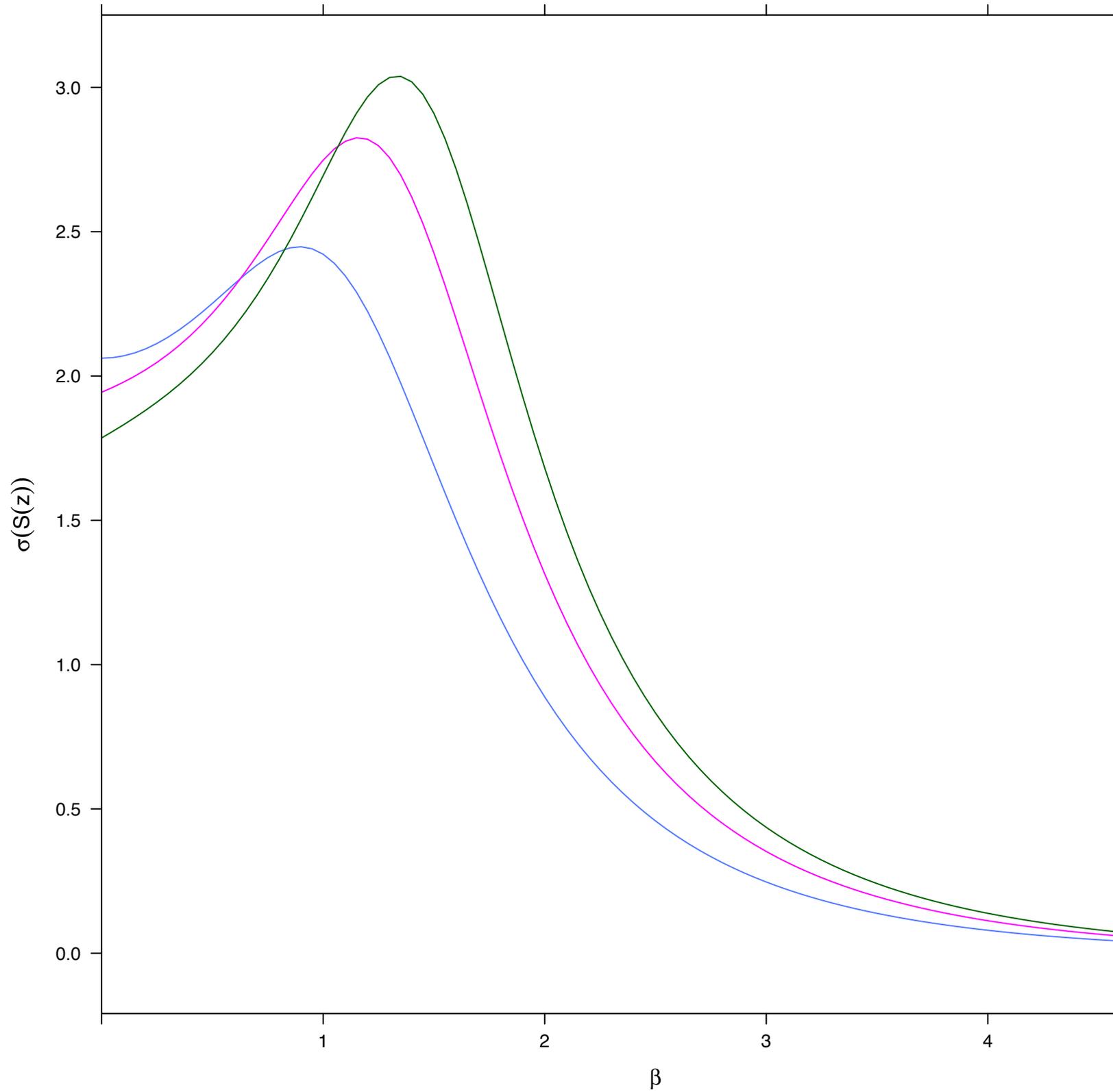
exact expectation of S(z) for n=12 and k=
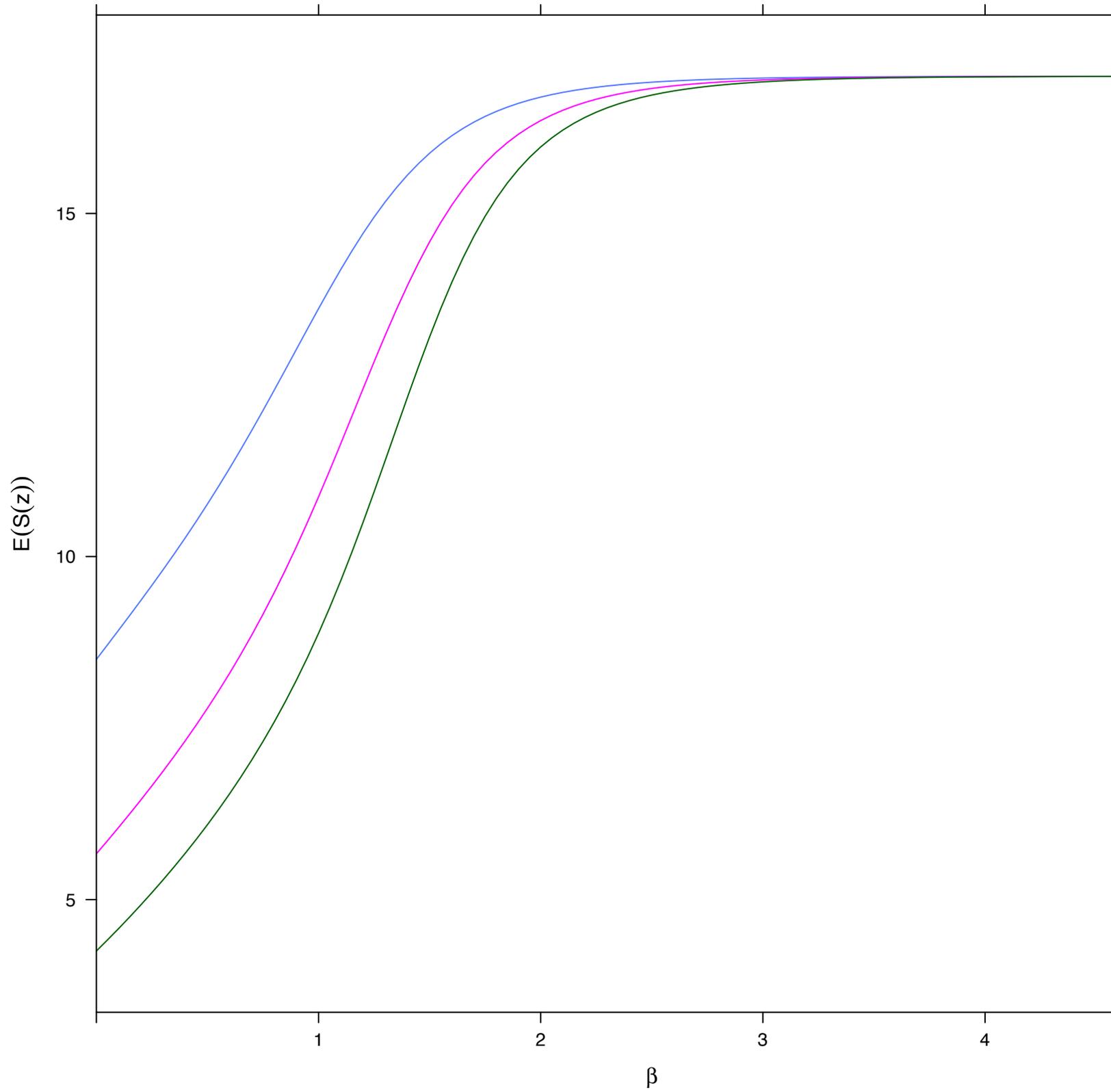
exact expectation of S(z) for k=3 and n=

exact standard deviation of S(z) for n=12 and k=

exact standard deviation of S(z) for k=3 and n=