Calibration of Boltzmann distribution priors in Bayesian data analysis

Martin Mechelke and Michael Habeck*

Max-Planck-Institute for Developmental Biology, Spemannstrasse 35, 72076 Tübingen, Germany and
Max-Planck-Institute for Intelligent Systems, Spemannstrasse 38, 72076 Tübingen, Germany

(Received 25 July 2012; revised manuscript received 24 September 2012; published 19 December 2012)

The Boltzmann distribution is commonly used as a prior probability in Bayesian data analysis. Examples include the Ising model in statistical image analysis and the canonical ensemble based on molecular dynamics force fields in protein structure calculation. These models involve a temperature or weighting factor that needs to be inferred from the data. Bayesian inference stipulates to determine the temperature based on the model evidence. This is challenging because the model evidence, a ratio of two high-dimensional normalization integrals, cannot be calculated analytically. We outline a replica-exchange Monte Carlo scheme that allows us to estimate the model evidence by use of multiple histogram reweighting. The method is illustrated for an Ising model and examples in protein structure determination.

DOI: 10.1103/PhysRevE.86.066705 PACS number(s): 05.10.−a, 02.50.Tt, 02.70.Uu, 87.15.A−

I. INTRODUCTION

Many complex data analysis problems involve a prior probability that is based on the Boltzmann distribution:

$$\pi(x|\beta) = \frac{1}{Z(\beta)}\, e^{-\beta E(x)}, \tag{1}$$

where microstate x is the parameter of interest (for example, an image or a protein structure) and E is an energy function that favors plausible configurations (without loss of generality $E > 0$); $Z(\beta)$ is the normalization constant or partition function. The hyperparameter $\beta \geqslant 0$ is the inverse temperature and determines how strongly the degrees of freedom x are coupled. Prior distributions of the above type occur, for example, in image analysis where Ising models or more generally Markov random fields are popular priors that capture spatial correlations [1]. Another example is protein structure determination from nuclear magnetic resonance (NMR) data. Here the prior distribution is the canonical ensemble based on an approximate potential energy [2,3].

In these data analysis problems, it is often unclear how much influence the prior probability should have. The weight of the prior is controlled by the inverse temperature β. If the system is at thermal equilibrium and the energy is an accurate description of the entire system, then $\beta = (k_B T)^{-1}$ where k_B is the Boltzmann constant and T the system's temperature. However, often there is no physical basis to choose the hyperparameter, for example, if we are working with nonphysical systems such as images or with approximate energy functions. Unless there is a physically justifiable reason to set β, it is a free parameter that needs to be estimated from the data.

The need to estimate hyperparameters has been discussed by MacKay in the context of Bayesian interpolation [4]. To infer the strength of the prior he applies the "evidence framework," which relies on a Gaussian approximation of the posterior distribution around the maximum *a posteriori* estimate. The evidence framework is a popular method to estimate hyperparameters in case the normalization integral of the prior is tractable. But it is not applicable to general Boltzmann priors (1) that cannot be normalized.

For intractable priors, the inverse temperature is typically chosen by visual inspection, trial and error, or in the best, cross-validation [5]. Geman and McClure [6] developed an expectation maximization algorithm to determine β iteratively in image restoration problems. Subsequent publications [7,8] applied the mean-field approximation to determine the hyperparameter. Methods that focus on temperature estimation for Ising and Potts model priors have been introduced, for example, in Ref. [9] and more recently by Kiwata [10]. These methods make assumptions about the functional form of the prior and the likelihood function and are therefore not generally applicable. An interesting and somewhat related method has been proposed by Atchadé *et al.* [11] (for a more detailed comparison see Sec. IV).

In this article, we develop a general algorithm that does not assume any particular prior distribution, likelihood function, or configuration space. Our framework allows us to estimate the full posterior distribution of the hyperparameter instead of just locating its maximum. We illustrate our method for data analysis problems in image restoration and protein structure determination.

II. METHOD

A. Inverse temperature calibration by maximization of the model evidence

A Bayesian approach to data analysis problems combines observed data D with the data-independent prior distribution (1). The likelihood function

$$L(x) = \Pr(D|x) \tag{2}$$

is the probability of observing the actual data assuming that the configuration of the system is known. Bayes's theorem [12] allows us to solve the inverse problem of inferring the system's configuration from observations by multiplying the prior probability and the likelihood function to obtain the posterior probability:

$$\Pr(x|\beta, D) = \frac{1}{\Pr(D|\beta)}\, L(x)\, \pi(x|\beta). \tag{3}$$

*michael.habeck@tuebingen.mpg.de

However, if we estimate x by maximizing or sampling from the posterior probability $\Pr(x|\beta,D)$, we assume that we know the correct value of β. As pointed out above, this is often unrealistic. The marginal likelihood or model evidence [13]

$$\Pr(D|\beta) = \int L(x)\,\pi(x|\beta)dx \tag{4}$$

measures how likely the observed data are for a particular value of β. The optimal β maximizes the evidence because it scales the Boltzmann distribution (1) such that those models which are most consistent with the data are preferred. If we multiply $\Pr(D|\beta)$ by a prior probability for β, we obtain the marginal posterior probability distribution of β, which can used, for example, to calculate the posterior mean of the hyperparameter.

Let us introduce the following quantity:

$$c_\lambda(\beta) = \int [L(x)]^\lambda\, e^{-\beta E(x)}\, dx. \tag{5}$$

Then $\Pr(D|\beta) = c_1(\beta)/c_0(\beta)$. The model evidence is the ratio of two high-dimensional integrals. The denominator is the partition function, which measures the volume of the subset of models that are consistent with the prior probability at a given temperature. With decreasing β this volume increases and approaches the total volume of configuration space. The nominator is an integral over the product of the likelihood function and the Boltzmann factor. This integral measures the volume of the subset of models that are both consistent with the data and the prior probability. The denominator prefers rigid models, the nominator prefers flexible models that fit the data well. Therefore the model evidence implements Occam's razor and trades model complexity against goodness-of-fit.

The representation of the model evidence in terms of $c_\lambda(\beta)$ suggests that it should be feasible to estimate model evidences by simulating the extended ensemble:

$$p(x|\lambda,\beta) = \frac{1}{c_\lambda(\beta)}\,[L(x)]^\lambda\, e^{-\beta E(x)}, \tag{6}$$

where we have two inverse temperatures, one for the prior probability and another for the data. For $\lambda = 1$, we obtain the posterior probability $p(x|\lambda = 1,\beta) = \Pr(x|D,\beta)$. For $\lambda = 0$, we obtain the prior distribution $p(x|\lambda = 0,\beta) = \pi(x|\beta)$. Values of λ between zero and one allow us to smoothly bridge between the prior and the posterior distribution. A similar family of distributions is used in bridge sampling [14]. The difference in our method is that bridge sampling does not allow for exchanges between different bridging distributions and that the simulation is sequential rather than parallel. Moreover, we do not apply bridge sampling estimators but combine samples from different bridging distributions with the help of the density of states (see Sec. II B).

The optimal β is obtained by maximizing the model evidence $\Pr(D|\beta)$, which we achieve by setting the derivative of its logarithm to zero:

$$\frac{\partial \log \Pr(D|\beta)}{\partial\beta} = \frac{c_1'(\beta)}{c_1(\beta)} - \frac{c_0'(\beta)}{c_0(\beta)} \stackrel{!}{=} 0.$$

We have

$$-\frac{\partial \log c_\lambda(\beta)}{\partial\beta} = -\frac{c_\lambda'(\beta)}{c_\lambda(\beta)} = \langle E\rangle_{x|\lambda,\beta},$$

where $\langle\cdot\rangle_{x|\lambda,\beta}$ denotes an average over the extended ensemble (6). Therefore the optimal prior weight $\hat\beta$ is determined by the equality [6,8]

$$\langle E\rangle_{x|\lambda=1,\beta=\hat\beta} = \langle E\rangle_{x|\lambda=0,\beta=\hat\beta}, \tag{7}$$

or in words $\langle E\rangle_{\text{"data"}} = \langle E\rangle_{\text{"no data"}}$. If we choose β according to Eq. (7), the expected interaction energy under the posterior distribution ($\lambda = 1$) is the same as the expected energy under the prior distribution ($\lambda = 0$). That is, for the Bayesian choice of β, incorporation of the data does not change our expectation about the interaction energy.

The expected interaction energy $U_\lambda(\beta) = \langle E\rangle_{x|\lambda,\beta}$ is monotonically decreasing in β. This is clear by taking its first derivative:

$$\frac{\partial U_\lambda(\beta)}{\partial\beta} = \langle [E - U_\lambda]^2\rangle_{x|\lambda,\beta} \geqslant 0, \tag{8}$$

which is proportional to the specific heat for $\lambda = 0$. Typically, we have $U_0(0) \geqslant U_1(0)$ and $U_0(\beta_{\max}) \leqslant U_1(\beta_{\max})$ for sufficiently large $\beta_{\max}$. At infinite temperature ($\beta = 0$), the configurations sampled from the prior are completely random, whereas those generated from $L(x)$ are clearly nonrandom and scatter about the maximum likelihood solution where the extent of scatter depends on the amount of data. For a reasonable choice of the energy function E the maximum likelihood solution will have a smaller energy than a random configuration. Therefore, $U_0(0) \geqslant U_1(0)$. At low temperatures ($\beta = \beta_{\max}$), configurations drawn from the prior will eventually collapse to the ground state. Configurations drawn from $L(x)\pi(x|\beta_{\max})$ still reflect the data to some degree and will seek a compromise between the maximum-likelihood and the ground-state configuration. Therefore, $U_0(\beta_{\max}) \leqslant U_1(\beta_{\max})$. Because $U_\lambda(\beta)$ is monotonically decreasing in β for all λ, both curves, $U_1(\beta)$ and $U_0(\beta)$, must at least cross once, and we have at least one β satisfying Eq. (7).

B. Replica-exchange Monte Carlo and multiple histogram reweighting

Practically, we calculate $\Pr(D|\beta)$ and $U_\lambda(\beta)$ for $\lambda = 0$ and $\lambda = 1$ by running two replica-exchange Monte Carlo (REMC) simulations [15,16] of the extended ensemble $p(x|\lambda,\beta)$ described in Eq. (6). REMC, also known as "parallel tempering" [17], is a variant of the Monte Carlo method by Metropolis *et al.* [18] and simulates the joint distribution $\prod_{r=1}^R p(x_r|\lambda_r,\beta_r)$ of multiple configurations at different inverse temperatures (λ_r,β_r). The replicas, $p(x_r|\lambda_r,\beta_r)$, are chosen such that they bridge between the target distribution, e.g., $p(x|\lambda = 1,\beta)$, and a flattened version, which is more suitable for sampling, e.g., $p(x|\lambda = 0,\beta = 0)$. Configurations are randomly exchanged between neighboring replicas. An exchange is accepted according to the Metropolis criterion, which allows configurations to diffuse between high- and low-temperature replicas and thereby escape metastable states and improve the mixing of the Markov chain. We need to run two REMC simulations (one for the prior, the other for the posterior expectation) because typically states from completely different regions in configuration space contribute to the expectation integral. However, once the expectation of the prior is estimated, it can be used to analyze different data sets [6].

The first REMC run simulates the ensemble $p(x|\lambda = 0,\beta)$ where $\beta: 0 \to \beta_{\max}$ is the only replica parameter. Samples from this simulation are used to estimate the expected interaction energy, $U_0(\beta)$, for the family of prior distributions obtained by varying β. The second REMC run simulates the ensemble $p(x|\lambda=1,\beta)$ where both λ and β are replica parameters. This simulation is the concatenation of two chains. In the first chain, we switch on the data, $\lambda: 0 \to 1$, while the prior stays switched off ($\beta = 0$). This chain is needed to improve the sampling, because it is typically hard to draw from the likelihood function even without any additional prior, and the system quickly gets stuck in a local mode. The additional chain allows the system to escape from subordinate modes by diffusion to a high-temperature heat bath. In the second chain, we keep the data fully switched on ($\lambda = 1$) and study the effect of the Boltzmann prior by increasing its temperature ($\beta: 0 \to \beta_{\max}$). Using this replica simulation, we can estimate the expected interaction energy, $U_1(\beta)$, for the family of posterior distributions obtained by varying the inverse temperature β. The maximum inverse temperature, $\beta_{\max}$, that is probed during REMC is an algorithmic parameter that must be chosen by the user. However, it also conceivable to use an iterative REMC scheme in which $\beta_{\max}$ is increased incrementally until a crossing of U_0 and U_1 is observed.

One of the ingenious aspects of the Metropolis algorithm is that it does not require the evaluation of normalization constants, but that its output can be used to estimate normalization constants. This feature is also used in our approach to estimate β. For both replica-exchange simulations, we combine states from all replicas by using multiple histogram reweighting [19], which estimates the density of states:

$$g_\lambda(E) = \int \delta(E - E(x))\,[L(x)]^\lambda\,dx, \tag{9}$$

where δ is the Dirac delta function. Application of the original version of multiple histogram reweighting, developed for discrete systems, requires binning of energies, which introduces a potential source of errors. The nonparametric version [20] eliminates the need to bin energies by estimating the density of states $\hat{g}_i$ associated with the energy of each of the sampled configurations x_i such that $g(E) \approx \sum_i \hat{g}_i\,\delta[E - E(x_i)]$. We reconstruct g_1 and g_0 from both REMC simulations. The model evidence as a function of β can then be estimated as

$$\Pr(D|\beta) = \frac{\int g_1(E)\,e^{-\beta E}\,dE}{\int g_0(E)\,e^{-\beta E}\,dE}. \tag{10}$$

The average interaction energies are obtained by evaluating

$$U_\lambda(\beta) = \frac{\int E\,g_\lambda(E)\,e^{-\beta E}\,dE}{\int g_\lambda(E)\,e^{-\beta E}\,dE}. \tag{11}$$

Knowledge of the density of states allows us to use configurations from all replicas, not just states from the prior and posterior, to obtain very accurate estimates of the expectations (11) and of the model evidence (10). The expectations are much more reliable than those obtained by Gibbs sampling [1] of the posterior and prior distributions as proposed in Ref. [6]. Furthermore, we obtain the model evidence for a whole β range and not just the location of its maximum. This allows us to make statements about how well β is determined by the data.

III. APPLICATIONS

A. Results for a Gaussian toy system

To explain our method of choosing β, let us first look at an analytically tractable system. We assume that the prior probability is a D-dimensional spherical Gaussian (i.e., a system of D uncoupled harmonic oscillators):

$$\pi(x|\beta) = \left(\frac{\beta}{2\pi}\right)^{\frac{D}{2}} e^{-\frac{\beta}{2}\|x\|^2} \tag{12}$$

with energy function $E(x) = \frac{1}{2}\|x\|^2 = \frac{1}{2}\sum_{i=1}^{D} x_i^2$. The observations $y_1, \ldots, y_N$ scatter about the true configuration with an amplitude of $1/\sqrt{\alpha}$ such that the likelihood function is

$$\begin{aligned} L(x) &= \left(\frac{\alpha}{2\pi}\right)^{\frac{ND}{2}} e^{-\frac{\alpha}{2}\sum_{i=1}^{N}\|y_i - x\|^2} \\ &= \left(\frac{\alpha}{2\pi}\right)^{\frac{ND}{2}} e^{-\frac{\alpha N}{2}(s^2 + \|x - \overline{y}\|^2)}, \end{aligned} \tag{13}$$

where $\overline{y} = \frac{1}{N}\sum_i y_i$ is the sample mean and $s^2 = \frac{1}{N}\sum_i \|y_i - \overline{y}\|^2$ the total sample variance. The posterior probability is

$$\Pr(x|\beta, D) = \left(\frac{\alpha N + \beta}{2\pi}\right)^{\frac{D}{2}} e^{-\frac{\alpha N + \beta}{2}\|x - \overline{x}\|^2}$$

with $\overline{x} = \frac{\alpha N}{\alpha N + \beta}\overline{y}$. It is also possible to calculate the model evidence analytically:

$$\Pr(D|\beta) = \left(\frac{\alpha}{2\pi}\right)^{\frac{ND}{2}} \left(\frac{\beta}{\alpha N + \beta}\right)^{\frac{D}{2}} e^{-\frac{\alpha N}{2}[s^2 + \frac{\beta}{\alpha N + \beta}\|\overline{y}\|^2]}.$$

Moreover, we have

$$U_\lambda(\beta) = \frac{1}{2}\left[\frac{D}{\alpha\lambda N + \beta} + \left(\frac{\alpha\lambda N}{\alpha\lambda N + \beta}\right)^2 \|\overline{y}\|^2\right].$$

Because the prior distribution (12) peaks at zero and the likelihood function (13) peaks at the sample mean $\overline{y}$, the mean squared deviation $\overline{b} = \|\overline{y}\|^2/D$ can be viewed as an empirical estimate of the bias that the prior imposes. Moreover, $v = 1/(\alpha N)$ is the width of the likelihood function (13). From $U_0(\hat{\beta}) = U_1(\hat{\beta})$ we obtain for the Bayesian prior weight

$$1 + v\hat{\beta} = \overline{b}\,\hat{\beta}, \tag{14}$$

which has a finite solution only if $\overline{b} > v$:

$$\hat{\beta} = (\overline{b} - v)^{-1}.$$

If $\overline{b} \leqslant v$ we can satisfy Eq. (14) only by letting $\hat{\beta} \to \infty$. We observe a kind of bias-variance tradeoff. The optimal weight decreases with increasing bias of the prior, which is a safety belt against putting too much weight on contributions that contradict the data [see Fig. 1(a)]. If the data are of a high quality (small v or likewise large α), the weight will tend to smaller values [see Fig. 1(b)], which makes sense, because the data suffice to characterize the true configuration, and we do not need to incorporate additional knowledge. An increase in the variance (large v) will lead to a higher weight, because differences between the prior and the likelihood can only be resolved to a lesser extent. For $\overline{b} \leqslant v$ the weight approaches infinity, because the possible bias introduced by the prior cannot be distinguished from the variability of the

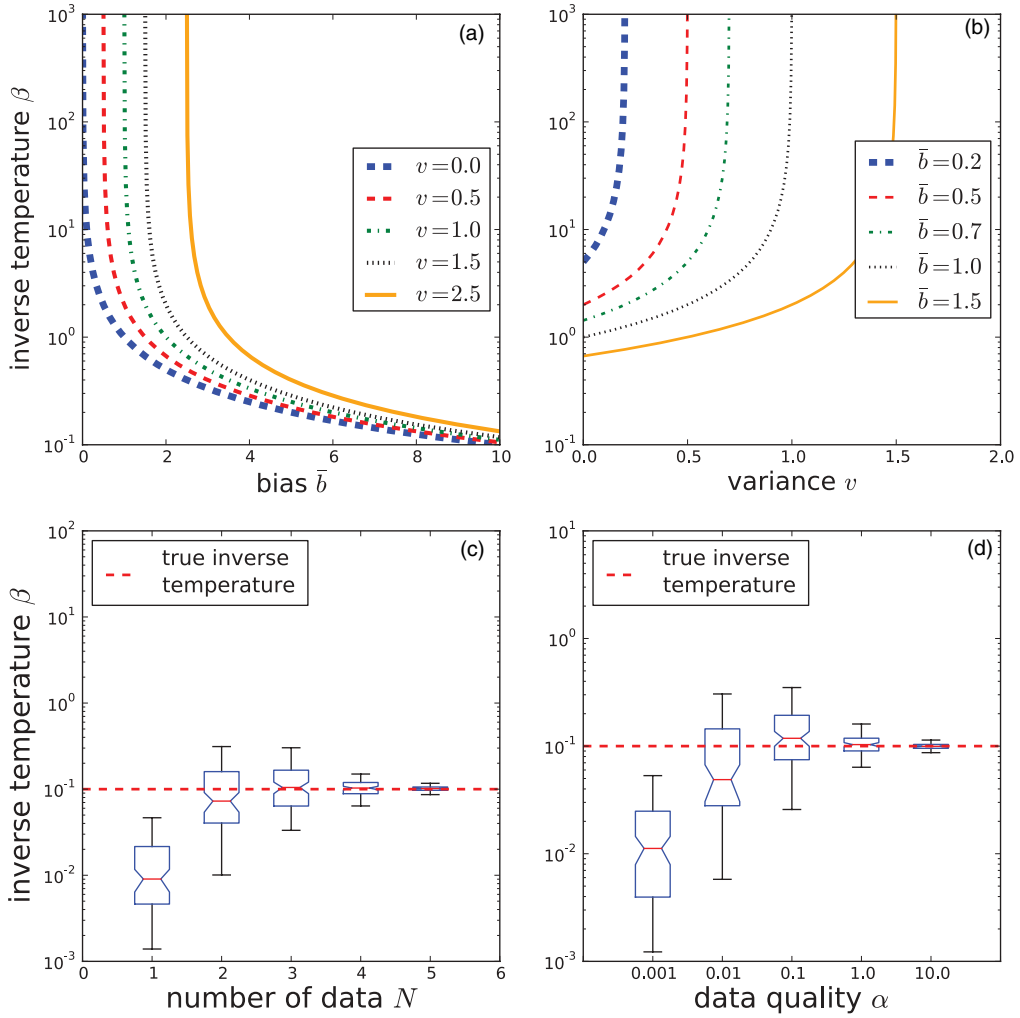

FIG. 1. (Color online) Optimal weighting of a Gaussian prior probability. (a) Dependence of the optimal inverse temperature on the empirical bias $\bar{b}$. (b) Optimal inverse temperature as a function of the variance v. (c) Estimation of the inverse temperature as a function of the number of data N. 100 repetitions were calculated for every value of N leading to the scatter shown as box plots. (d) Estimation of the inverse temperature as a function of the data quality α. Again, 100 repetitions were calculated for every α. In panels c and d, the true inverse temperature $\beta = 1/b$ is indicated by a red dashed line.

observations. The data cannot resolve differences between the prior distribution and the likelihood function, and therefore Bayesian model selection goes for the save option of trusting the prior. In a Bayesian treatment, one would also put a prior probability on β, which prevents β from approaching infinity and being negative. In our REMC scheme, the prior is implemented by probing β values in $[0,\beta_{\max}]$, and $\beta_{\max}$ is an algorithmic parameter.

It is less obvious to predict the effect of varying the number of data points, because changes in N both affect the variance v and, indirectly through $\overline{y}$, the empirical bias $\bar{b}$. If the data scatter about an unknown configuration μ, the sample mean $\overline{y}$ will follow the distribution

$$p(\overline{y}|\mu,\alpha,N) = \left(\frac{\alpha N}{2\pi}\right)^{D/2} e^{-\frac{\alpha N}{2}\|\overline{y}-\mu\|^2}$$

with $\langle\|\overline{y}\|^2\rangle = \|\mu\|^2 + D/(\alpha N)$ or $\langle\bar{b}\rangle = b + v$ where $b = \|\mu\|^2/D$ is the true bias. If we plug the expected empirical bias $\langle\bar{b}\rangle$ into Eq. (14), the optimal weight is $\hat{\beta} = 1/b$, which correctly approaches infinity if μ approaches zero. With increasing number of data points, the empirical $\bar{b}$ bias approaches the true bias b and the distribution of weights concentrates at $1/b$ [see Fig. 1(c)]. We observe the same effect if we increase the quality of the data α [see Fig. 1(d)].

B. Calibration of the Ising model in image reconstruction

We illustrate the method for an application in image reconstruction. Although this a toy model, it shows all the complexity of real-world applications. We use the two-dimensional Ising model on a $L \times L$ lattice with $L = 32$ as prior probability:

$$\pi(x|\beta) = \frac{1}{Z(\beta)} e^{\beta \sum_{i\sim j} x_i x_j},$$

where $\sum_{i\sim j}$ indicates a sum over the nearest neighbors on a square lattice and $x_i \in \{-1,1\}$. Here β controls how strongly the colors of neighboring pixels are coupled; for larger β we impose stronger spatial correlations. We use this prior

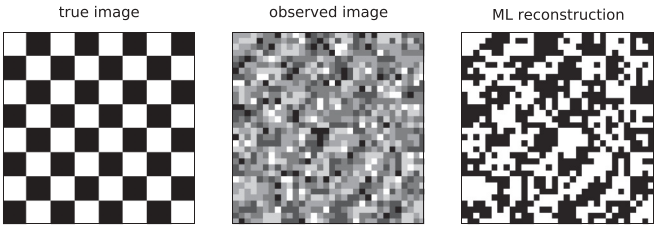

FIG. 2. Data for image reconstruction with an Ising model. Left: True image; middle: average observed image with $\theta = 0.65$ averaged over $N = 5$ observations; right: maximum likelihood reconstruction.

probability in the reconstruction of binary images. We will illustrate the method for images showing a checkerboard pattern.

Assume that we take noisy pictures of the checkerboard where each pixel is flipped with probability $1-\theta$, $\theta \in [0,1]$. The probability of observing $y_i \in \{-1,1\}$, given that the true color is x_i is

$$\Pr(y_i|x_i,\theta) = \frac{1}{2}\sqrt{\theta(1-\theta)}\left(\frac{\theta}{1-\theta}\right)^{x_i y_i/2}. \qquad (15)$$

The negative logarithm of the posterior probability therefore corresponds to the standard Hamiltonian of the Ising model (up to an additive constant) [9]:

$$-\log \Pr(x|\beta,D) = -\beta \sum_{i\sim j} x_i x_j - \sum_i h_i x_i,$$

where the data come in as an external magnetic field

$$h_i = N \log[\theta/(1-\theta)]\,\overline{y}_i/2$$

with $\overline{y}_i \in [-1,1]$ indicating the average intensity of the ith pixel over a total of N observed images. The maximum likelihood estimate, i.e., the binary image for which $\log \Pr(x|\beta,D)$ at $\beta = 0$ reaches its maximum, is $\hat{x}_i = \mathrm{sign}\{h_i\}$.

We generated $N = 5$ noisy images at $\theta = 0.65$. Figure 2 shows the true image, the observed average image, and the maximum likelihood reconstruction. We can calculate the inverse configurational temperature of an image x:

$$\beta(x) = \left.\frac{ds(E)}{dE}\right|_{E=E(x)}, \qquad (16)$$

where $s(E) = \log g_0(E)$ is the microcanonical entropy and $g_0(E)$ the density of states (9), which is exactly known for this system [21]. The true image has an inverse configurational temperature of $\beta(x_{\mathrm{true}}) = 0.38$. The ML estimate has a lower inverse configurational temperature of $\beta(\hat{x}) = 0.13$. We ran two replica-exchange Monte Carlo simulations ($\beta_{\max} = 2$), one ignoring the data, the other using the data, and estimated the densities of states, $g_0(E)$ and $g_1(E)$, using multiple histogram reweighting.

Figure 3(a) shows the internal energy curve $U(\beta)$ obtained with and without data. The curve obtained with data spans a narrower energy range, because the data impose additional ties and restrict the probable images. At $\beta = 0$, the curve obtained without data (i.e., internal energy curve of the Ising model) sets in at a higher value, because the spins are completely random. The curve shows a sigmoidal form and goes through a phase transition at $\beta_c \approx 0.44$. The Bayesian choice falls between the configurational temperatures of the true image and of the maximum likelihood reconstruction. The model evidence peaks at 0.26, and its width is 0.03 [Fig. 3(b)]. In Fig. 3(c) we show the accuracy of the reconstructed image as quantified by the Hamming distance between the mean posterior image and the true image. The accuracy is best for $\beta = 0.39$, which is very close to the inverse configurational temperature of the true image. The Bayesian choice ($\beta = 0.26$) results in posterior images that have a high accuracy compared to the maximum likelihood estimate ($\beta = 0$). The curve also illustrates the risk of putting too strong weight on the Boltzmann prior. For $\beta > 0.6$ the accuracy of the reconstructed image is worse than the maximum likelihood estimate.

Figure 4 shows the posterior means for all values of β that we probed. Too small β (top row) results in very grainy images, whereas too large β results in reconstructions showing large patches of same or similar color (bottom row). The Bayesian choice of β finds a compromise between these two limiting cases.

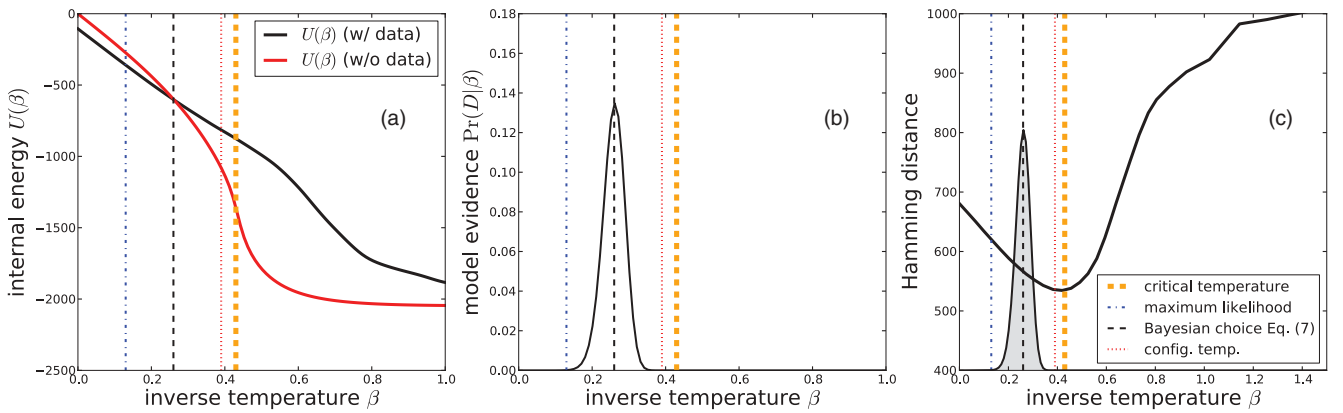

FIG. 3. (Color online) Estimation of the optimal temperature of the Ising model in image reconstruction. (a) Internal energy with and without data as a function of inverse temperature. Dashed vertical lines indicate the inverse configurational temperature of the true image (red) and of the maximum likelihood image (blue), the critical temperature (orange), and the optimal inverse temperature obtained from Eq. (7). (b) Model evidence as a function of β. (c) Hamming distance of posterior mean images and the true image.

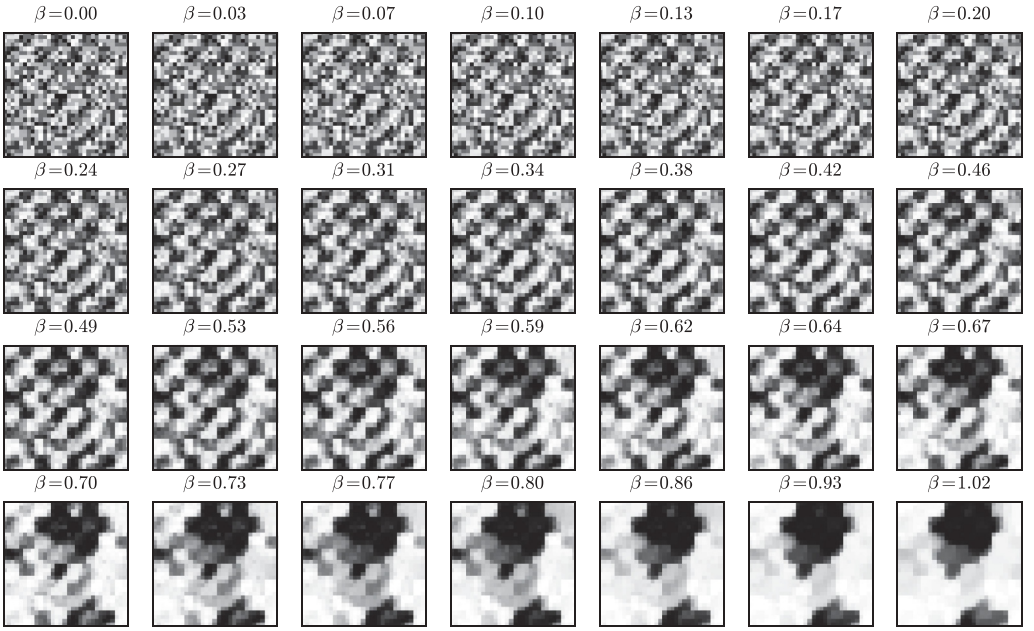

FIG. 4. Posterior mean images for different choices of β. β values between 0 and 2 were probed during REMC simulation. However, we show only the restored images for $\beta \leqslant 1$, because no significant changes are observed for larger β.

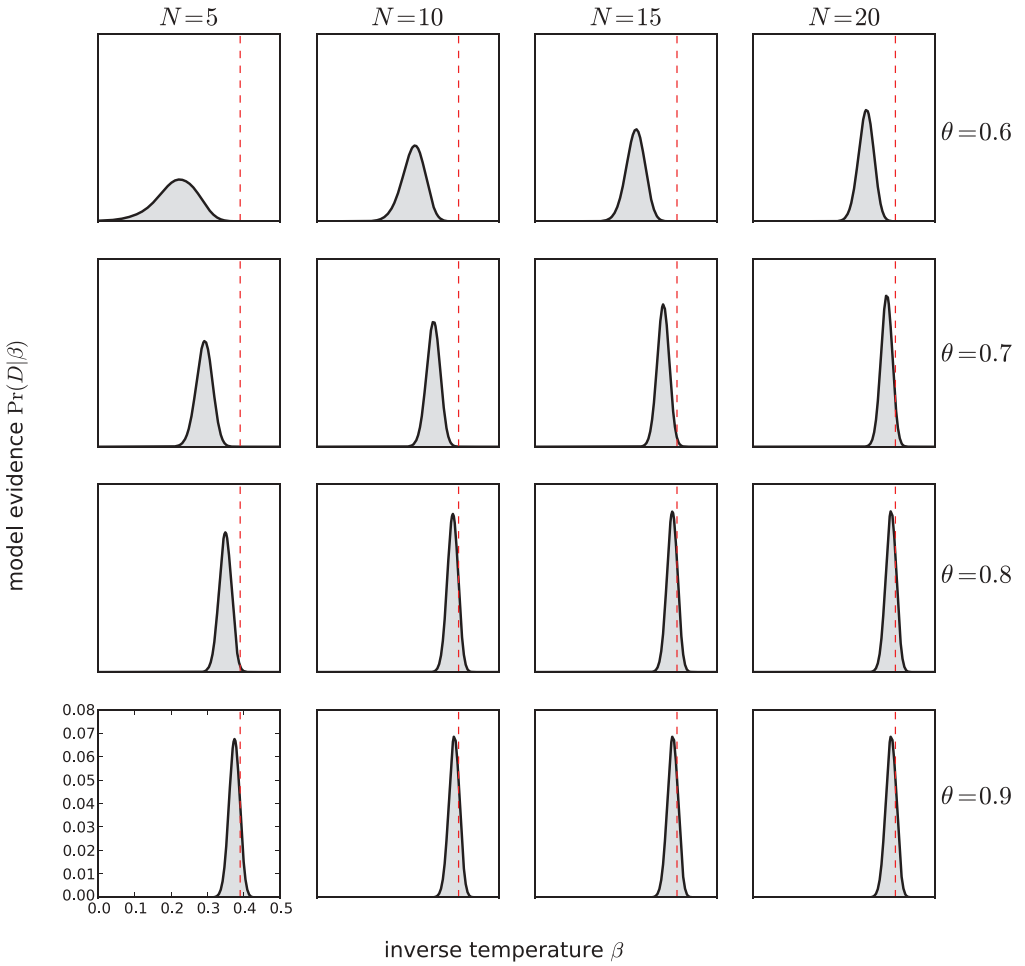

FIG. 5. (Color online) Model evidence for decreasing noise level $\theta = 0.6, 0.7, 0.8, 0.9$ and increasing amount of data $N = 5, 10, 15, 20$. The dashed vertical line indicates the configurational temperature of the true image.

To better understand the behavior of the prior temperature, we estimated β for varying noise level θ and varying number of data N. Figure 5 shows the results of probing all combinations of $N = 5,\ 10,\ 15,\ 20$ and $\theta = 0.6,\ 0.7,\ 0.8,\ 0.9$. We observe that the estimated coupling constant $\hat{\beta}$ approaches the configurational temperature of the true image as the number of data increases. For sparser data sets, the model evidence peaks at smaller β and becomes wider. This means that we should not try to compensate for the lack of data by cranking up the prior coupling. Rather, we should impose weaker prior ties and allow for more configurational flexibility. A similar behavior is observed with increasing noise level (decreasing θ). With smaller θ, β is estimated to be smaller than with larger θ. As the quality and amount of data deteriorates, Bayesian inference tells us to be cautious and introduce only weak prior correlations.

C. Optimal weighting of force fields in protein structure calculation

As a second application, we consider the calculation of a protein structure from NMR distance measurements [2]. Biomolecular structure calculation is based on minimalist force fields that are by far not as complex and realistic as modern molecular dynamics force fields [22]. Typically only van der Waals contributions are considered as noncovalent interactions; electrostatic and solvent interactions are ignored [23]. We use the van der Waals term from the Rosetta structure prediction software. This is a linearly ramped Lennard-Jones potential. Interactions at zero distance have a finite instead of an infinite energy; interactions between atoms that are more than 5.5 Å apart are set to zero [24].

What is the best weight for this force field? We first ran a calculation with a high-quality data set (PDB code 1D3Z) measured on the protein ubiquitin [25]. Ubiquitin has 76 amino acids and adopts a beta-barrel structure that is closed by an alpha helix. The data set comprises 1444 unique distance measurements; on average there are 19 distance restraints per amino acid. We ran two replica-exchange simulations to compute the model evidence as a function of the inverse temperature of the Lennard-Jones potential. Figure 6(a) shows the internal energy curves and the model evidence. The model evidence, $\Pr(D|\beta)$, peaks at a value of $\hat{\beta} = 0.87 \pm 0.05$. This peak falls well within the region of highest accuracy [as measured by the root mean square deviation (RMSD) from the crystal structure 1UBQ; Fig. 6(b)].

The second test case is a sparse data set measured on the Fyn-SH3 domain [2,26], a small beta-barrel domain. The data set comprises 154 distances out of which only 60 long-range distances define the fold. We estimated the inverse temperature for this challenging data set. The estimated

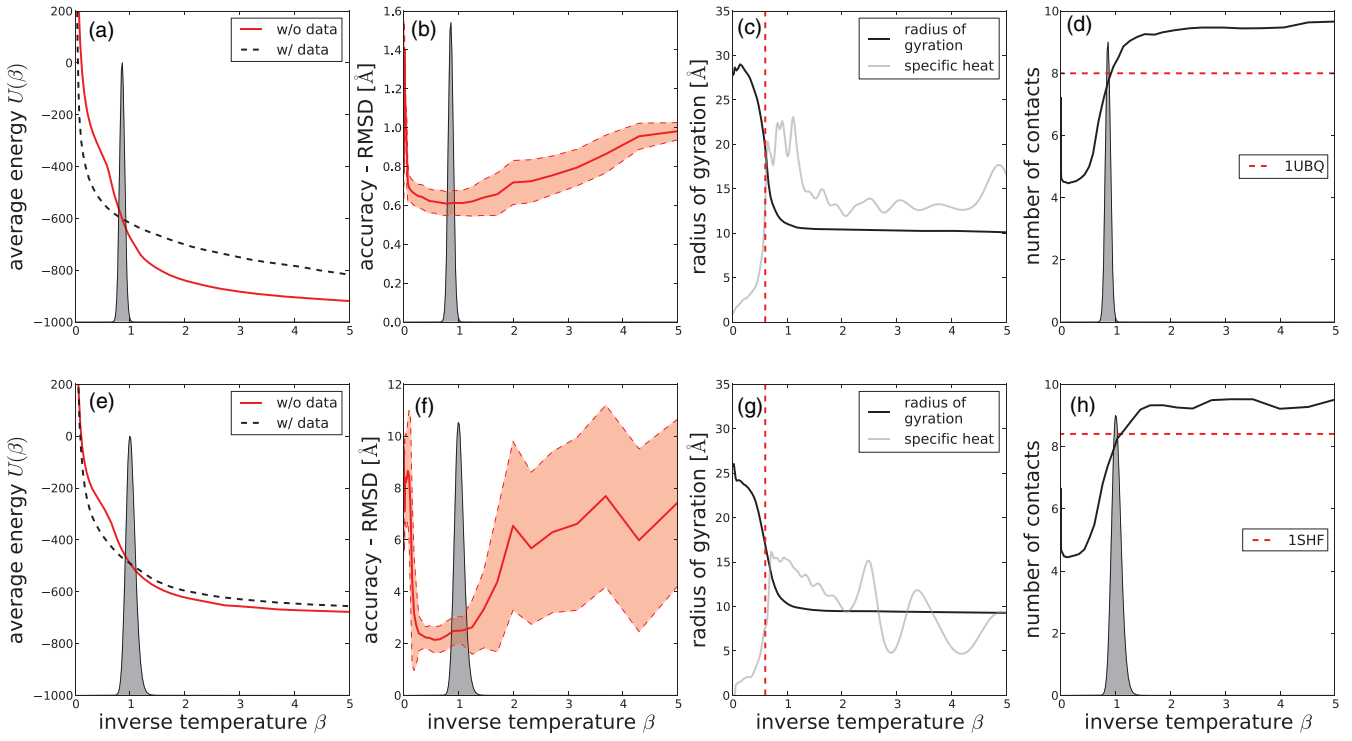

FIG. 6. (Color online) Calibration of the Boltzmann prior in protein structure calculation from NMR data. An approximate Lennard-Jones potential is weighted for high-quality (top row) and a sparse data set (bottom row). The model evidence (panels a, e) peaks at $\beta \approx 1$. Panels b and f show the accuracy of the structure ensemble obtained for different choices of the inverse temperature; the filled region indicates one standard deviation. Also shown is the compaction of structures in terms of the radius of gyration (panels c and g); the dashed vertical line marks the critical value $\beta = 0.6$ at which the compaction sets in. Another measure of compactness is the average number of nearest neighbor contacts (panels d and h); the dashed horizontal lines indicate the average number of contacts in the crystal structures 1UBQ and 1SHF, respectively. The radius of gyration and average number of contacts (black lines) are calculated over samples from the prior distribution, i.e., without using the data.

inverse temperature, $\hat{\beta} = 1.02 \pm 0.09$, is slightly larger than for ubiquitin [Fig. 6(e)]. Because we have fewer data, the model evidence is broader than the model evidence obtained with the ubiquitin data. Again, the model evidence peaks in the β region, which results in the most accurate structures [RMSD to crystal structure 1SHF; Fig. 6(f)]. Because the data are very sparse compared to the ubiquitin data, the range of β values for which the resulting ensemble shows an acceptable accuracy is much narrower. If we put too low or too strong weight on the Lennard-Jones potential, the ensemble comprises multiple conformers, e.g., mirror images of the correct structure. This results in an elevated, highly variable RMSD, which is indicated by a large standard deviation.

For both proteins, we observe a phase transition at approximately $\beta \approx 0.6$ when simulating the prior. The specific heat shows a first maximum, and the radius of gyration R_g drops from $R_g \approx 27/22$ Å to $R_g \approx 10/9$ Å (for ubiquitin/SH3 domain, respectively), indicating a sudden compaction of the structures at a critical inverse temperature [Fig. 6(c), 6(g)]. Likewise we can look at the average number of contacts in the structure ensembles generated with varying β [Fig. 6(d), 6(h)]. Contacts are defined between alpha carbons of the protein backbone. Whenever two alpha carbons are closer than 7.5 Å, we say that they form a contact. The average number of contacts an amino acid is engaged in is the number of CA-CA contacts that it forms with other residues. In Fig. 6(d) and 6(h) we show the average number of contacts as a function of the inverse temperature. Again we see a transition from elongated structures featuring a low contact density to compact structures that are densely packed. For too large inverse temperature, the structures are too dense with an average number of contacts of 9.7 for ubiquitin and 9.5 for the SH3 domain, whereas the corresponding crystal structure shows 8.0 (1UBQ) and 8.4 (1SHF) contacts on average. The Bayesian choice of the inverse temperature produces an ensemble of conformations whose degree of packing (as measured by our simple contact based measure) best corresponds to the packing of the crystal structure. This is observed for both proteins and argues that the optimal temperature of the Lennard-Jones potential is close to one both with sparse and with high-quality data. This is consistent with the value used in the Rosetta software. However, to make a general statement we need to look at more examples.

IV. CONCLUSION

We outline a Bayesian approach to adjust Boltzmann-type prior probabilities in data analysis problems. Our method can be used to calibrate the Boltzmann prior distribution when there is no physical basis that would determine the inverse temperature. We implement the method by using an extended replica-exchange Monte Carlo scheme in combination with histogram reweighting to obtain the density of states with and without data. Using the estimated densities of states we can compute the model evidence and locate its maximum and width. Tests with a Gaussian toy model and an Ising model show that the optimal coupling needs to be reduced as the number and quality of the data dwindles. An application to protein structure determination from NMR data suggests that the optimal weight of the Lennard-Jones potential as implemented in the Rosetta software is close to one.

The method of Atchadé *et al.* [11] bears some similarity to our approach. The method simulates two interleaved nonhomogeneous Markov chains and uses support points in β space that correspond to the β ladder of our replica-exchange simulations. Estimates of the partition function at each support point are updated as the algorithm progresses. The partition function estimates are smoothed to allow sampling from $p(\beta|D)$. Our approach combines samples from all chains by using multiple histogram reweighting. Although their method is in spirit similar to ours, there are many additional algorithmic parameters that need to be specified, among which are the learning rates for the partition function estimates and the kernel for parameter smoothing. Another difference is that Atchadé *et al.* do not use any exchange scheme, whereas our algorithm enhances the mixing by allowing exchange between the individual Markov chains.

An important question, especially for the design of energy functions in protein simulation, is how to estimate multiple inverse temperatures or weights for energies comprising multiple energy terms. To develop efficient algorithms that allow estimation of multiple weights or inverse temperatures will be the subject of future research.

ACKNOWLEDGMENTS

This work has been supported by Deutsche Forschungsgemeinschaft (DFG) grant HA 5918/1-1 and by the Max Planck Society.

[1] S. Geman and D. Geman, IEEE Trans. PAMI **6**, 721 (1984).
[2] W. Rieping, M. Habeck, and M. Nilges, Science **309**, 303 (2005).
[3] M. Habeck, M. Nilges, and W. Rieping, Phys. Rev. E **72**, 031912 (2005).
[4] D. J. C. MacKay, Neural Comput. **4**, 415 (1992).
[5] V. Johnson, W. Wong, X. Hu, and C. Chen, IEEE Trans. Pattern Anal. Mach. Intel. **13**, 413 (1991).
[6] S. Geman and D. E. McClure, Bull. Int. Stat. Inst. **LII-4**, 5 (1987).
[7] J. M. Pryce and A. D. Bruce, J. Phys. A **28**, 511 (1995).
[8] Z. Zhou, R. N. Leahy, and J. Qi, IEEE Trans. Image Process. **6**, 844 (1997).
[9] J. Inoue and K. Tanaka, Phys. Rev. E **65**, 016125 (2001).
[10] H. Kiwata, Physica A **391**, 2215 (2012).
[11] Y. F. Atchadé, N. Lartillot, and C. P. Robert, arXiv:0804.3152 (2008).
[12] E. T. Jaynes, *Probability Theory: The Logic of Science* (Cambridge University Press, Cambridge, 2003).
[13] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms* (Cambridge University Press, Cambridge, 2003).
[14] A. Gelman and X. Meng, Stat. Sci. **13**, 163 (1998).
[15] R. H. Swendsen and J.-S. Wang, Phys. Rev. Lett. **57**, 2607 (1986).
[16] M. Habeck, M. Nilges, and W. Rieping, Phys. Rev. Lett. **94**, 0181051 (2005).
[17] C. J. Geyer, in *Computing Science and Statistics: Proceedings of*

the 23rd Symposium on the Interface, edited by E. M. Keramidas (Interface Foundation of North America, Fairfax Station, VA, 1991), pp. 156–163.

[18] N. Metropolis, M. Rosenbluth, A. Rosenbluth, A. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1957).

[19] A. M. Ferrenberg and R. H. Swendsen, Phys. Rev. Lett. **63**, 1195 (1989).

[20] M. Habeck, in *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, Vol. 22, edited by N. Lawrence and M. Girolami (2012), pp. 486–494, http://jmlr.csail.mit.edu/proceedings/papers/v22/.

[21] P. D. Beale, Phys. Rev. Lett. **76**, 78 (1996).

[22] A. T. Brünger and M. Nilges, Q. Rev. Biophys. **26**, 49 (1993).

[23] J. P. Linge and M. Nilges, J. Biomol. NMR **13**, 51 (1999).

[24] B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, and D. Baker, Science **302**, 1364 (2003).

[25] G. Cornilescu, J. L. Marquardt, M. Ottiger, and A. Bax, J. Am. Chem. Soc. **120**, 6836 (1998).

[26] T. K. Mal, S. J. Matthews, H. Kovacs, I. D. Campbell, and J. Boyd, J. Biomol. NMR **12**, 259 (1998).