Genomics Proteomics Bioinformatics 12 (2014) 39–47

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

ORIGINAL RESEARCH

Bayesian Peak Picking for NMR Spectra

Yichen Cheng [1], Xin Gao [2], Faming Liang [1,*]

[1] *Department of Statistics, Texas A&M University, College Station, TX 77843, USA*
[2] *Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia*

Received 1 June 2013; accepted 29 July 2013
Available online 31 October 2013

KEYWORDS

Markov chain Monte Carlo;
Nuclear magnetic resonance;
Peak picking

Abstract Protein structure determination is a very important topic in structural genomics, which helps people to understand varieties of biological functions such as protein-protein interactions, protein–DNA interactions and so on. Nowadays, nuclear magnetic resonance (NMR) has often been used to determine the three-dimensional structures of protein *in vivo*. This study aims to automate the peak picking step, the most important and tricky step in NMR structure determination. We propose to model the NMR spectrum by a mixture of bivariate Gaussian densities and use the stochastic approximation Monte Carlo algorithm as the computational tool to solve the problem. Under the Bayesian framework, the peak picking problem is casted as a variable selection problem. The proposed method can automatically distinguish true peaks from false ones without preprocessing the data. To the best of our knowledge, this is the first effort in the literature that tackles the peak picking problem for NMR spectrum data using Bayesian method.

Introduction

Determination of structure-function relationships has been a long-standing research topic in structural genomics. Nowadays, nuclear magnetic resonance (NMR) has often been used to determine the three-dimensional structures of proteins, especially for the small proteins that are partially disordered, exist in multiple stable conformations in solution, show weak interactions with ligands, or do not crystallize readily. The NMR protein structure determination commonly involves a series of steps, such as peak picking, chemical shift assignment, nuclear Overhauser effect (NOE) assignment and structural calculation [1]. Among them, peak picking is the most important and tricky step and it is also the prerequisite for all the followed steps (see *e.g.*, [2,3]). As shown in **Figure 1** using protein TM1112 as an example, a typical NMR spectrum contains many peaks. We show 3D plot of protein TM1112 in panel A and show contour plot in panel B for the same protein. Here, H dimension corresponds to chemical shift in hydrogen dimension and N dimension corresponds to chemical shift in nitrogen dimension. Each peak, which is often referred to as a signal, represents a group of nuclei that can be coupled through bonds (scalar coupling) or space (spin–spin coupling). Peak picking step extracts the frequencies of each peak, which correspond to the chemical shift values of the corresponding nuclei. Such chemical shift values are then assigned to the corresponding atoms of the protein by considering the inter- and

* Corresponding author.
E-mail: fliang@stat.tamu.edu (Liang F).

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

http://dx.doi.org/10.1016/j.gpb.2013.07.003

intra-residue information that different spectra contain. The assignment is used to interpret NOE peaks, which provide distance constraints for the structural calculation step. However, the peak picking step is usually very time-consuming. Typically, it costs an experienced spectroscopist weeks or even months to accomplish the task. To automate this step, a variety of methods have been proposed, including neural networks [4], singular value decomposition [5,6], wavelet-based smoothing [7], among others.

The existing methods select peaks based on the intensities or the volumes of the peaks, and often fail for complex spectra. For example, they often fail to identify peaks with low intensity and overlapping peaks, and fail to distinguish false peaks with high intensities/volumes from true ones. In addition, they require a preprocessing step of data smoothing to remove noise. In this paper, we propose a Bayesian method to tackle this problem. We model the spectrum by a mixture of bivariate Gaussian densities and use the stochastic approximation Monte Carlo (SAMC) algorithm to estimate the positions and intensities of the peaks. Under the Bayesian framework, we cast the peak picking problem as a variable selection problem. Therefore, sophisticated Bayesian variable selection methods can be applied to seek for high-quality solutions to this problem.

The rest of this paper is structured as follows. We will first introduce the Bayesian model for NMR spectrum data. Next, we describe in detail the SAMC algorithm for peak picking. Following that, we give the results for both simulation studies and real NMR data, which show the benefit of the proposed method. We then conclude the paper with a brief discussion.

A Bayesian model for NMR spectra

For simplicity, this section describes only the model for the NMR spectra in two-dimensional (2D) space. The 2D NMR experiments, such as ^{15}N-HSQC, are among the most frequently used spectra for protein structure determination. Extension of the proposed method to higher-dimensional spaces is straightforward.

Suppose that the NMR spectrum consists of a total of n ($= L \times W$) grid points. Let $g(i,j)$ denote the intensity of the spectrum at the grid point (i,j) for $i = 1,\ldots,L$ and $j = 1,\ldots,W$. Then we model $g(i,j)$ as a mixture of bivariate Gaussian densities:

$$g(i,j) = \sum_{k=1}^{m} a_k \phi_k\left(i,j|\mu_{k1},\mu_{k2},\tau_{k1}^2,\tau_{k2}^2\right) + \epsilon_{ij}, \quad i = 1,\ldots,L \text{ and } j = 1,\ldots,W, \tag{1}$$

where $\phi_k(\cdot)$ is the k^{th} component of the mixture density function with mean $(\mu_{k1},\ \mu_{k2})'$ and covariance matrix diag $\left(\tau_{k1}^2,\tau_{k2}^2\right)$, a_k is the volume (or amplitude) of the k^{th} component, and ϵ_{ij} is the error term, which is assumed to be normally distributed with mean 0 and variance σ^2. We use M to denote a model and use $m = |M|$ to denote its size, *i.e.*, the number of components included in the mixture density function.

By lining up all the n grid points, the model (1) can be written in the matrix–vector form as follows:

$$Y = \Phi a + \epsilon, \tag{2}$$

where

$$Y = \begin{pmatrix} g(1,1) \\ \vdots \\ g(1,W) \\ \vdots \\ g(L,1) \\ \vdots \\ g(L,W) \end{pmatrix}, \quad \Phi = \begin{pmatrix} \phi_1(1,1) & \cdots & \phi_m(1,1) \\ \vdots & & \\ \phi_1(1,W) & \cdots & \phi_m(1,W) \\ \vdots & & \\ \phi_1(L,1) & \cdots & \phi_m(L,1) \\ \vdots & & \\ \phi_1(L,W) & \cdots & \phi_m(L,W) \end{pmatrix}, a = \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1W} \\ \vdots \\ \epsilon_{L1} \\ \vdots \\ \epsilon_{LW} \end{pmatrix}.$$

Here Y is an n-vector representing the spectrum intensity for each grid point; Φ is an $n \times m$ matrix that carries the information of m Gaussian density functions, each column of Φ corresponds to one Gaussian density component, and ϕ_k (i,j) is defined as in (1) but with parameters omitted; a is a m-vector consisting of the volumes of each component; and ϵ is an n-vector representing the random error.

Let $\vartheta = (\vartheta_1,\ldots,\vartheta_n)$, where $\vartheta_i = \left(\mu_{i1},\mu_{i2},\log\left(\tau_{i1}^2\right),\log\left(\tau_{i2}^2\right)\right)$. Then the likelihood function of the model (1) is given by

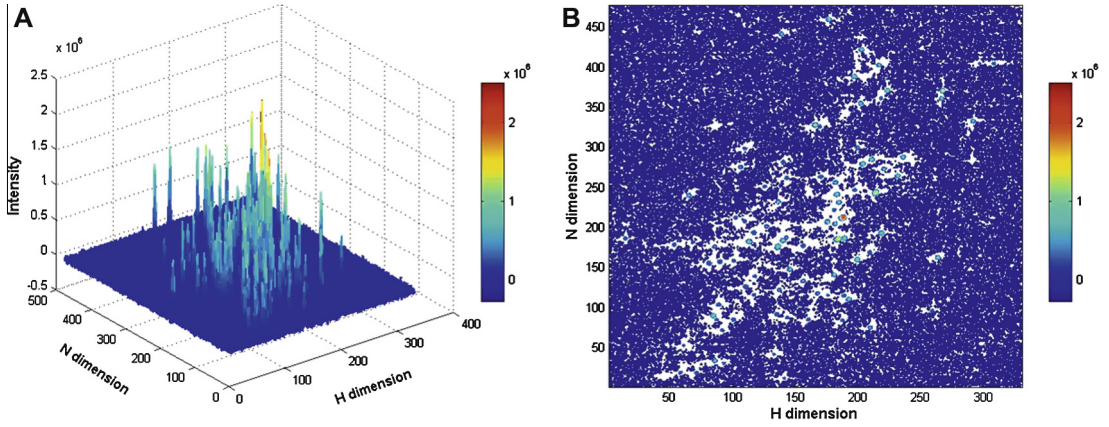

Figure 1 Illustration of 2D NMR spectrum data using protein TM1112 as an example

A. A 3D plot of 2D NMR spectrum data for protein TM1112. The Z axis is for the intensity of the spectrum. **B.** A contour plot of the same spectrum data. Here, H dimension corresponds to chemical shift in hydrogen dimension and N dimension corresponds to chemical shift in nitrogen dimension. One unit in H dimension represents 0.0148 ppm and one unit in D dimension represents 0.0873 ppm.

$$f(Y|\vartheta,a,\sigma^2,m)=\frac{1}{(2\pi)^{n/2}|\sigma^2 I_m|^{n/2}}\exp\left\{-\frac{1}{2}(Y-\Phi a)^T(\sigma^2 I_n)^{-1}(Y-\Phi a)\right\},$$

where I_m denotes an $m\times m$ identity matrix.

To conduct Bayesian analysis for the model (1), we consider the following prior distributions for the unknown parameters:

$$\begin{aligned}&a\sim\mathcal{N}(0,\sigma^2 V),\\&\mu_{i1}\sim U(0,L),\quad \mu_{i2}\sim U(0,W),\\&\tau_{i1}^2\sim IG(\alpha,\beta),\tau_{i2}^2\sim IG(\alpha,\beta),\\&\tfrac{\nu}{\sigma^2}\sim\chi_\nu^2,\end{aligned}$$

where $IG(\cdot,\cdot)$ denotes an inverse gamma distribution, $U(\cdot,\cdot)$ denotes a uniform distribution, and v, V are hyperparameters to be specified by the user. In this paper, we set $V=(\Phi'\Phi)^{-1}$; that is, we specify a Zellner's g-prior for the regression coefficients a with $g=1$. Following [8], we set $v=1$ and $\alpha=\beta=0.05$. The latter leads to vague priors for τ_{i1}'s and τ_{i2}'s. Since, for a given spectrum, the peak positions are always bounded, we let μ_{ij}'s be subject to the uniform priors.

Furthermore, we assume the prior distribution of m follows a truncated Poisson distribution with mean λ; that is,

$$P(|M|=m)=\frac{1}{C}\frac{\lambda^m}{m!}e^{-\lambda},\ m\in\{1,\ldots,m_{max}\},$$

where $C=\sum_{i=1}^{m_{max}}\frac{\lambda^m}{m!}e^{-\lambda}$, and λ and m_{max} are hyperparameters to be specified by the user. In practice, one may set λ to a small number to avoid finding too many false peaks. In this paper, we set $\lambda=1$ in all computations which yield good results. Our numerical results indicate that the choice of m_{max} is not crucial for peak picking, as long as it is not too small, *e.g.*, smaller than the number of true peaks. In this paper, we set m_{max} to 10 for the simulation studies, and a relatively small number, *e.g.*, two times of the number of amino acids, for a given protein.

Integrating out a and σ^2 gives us the posterior

$$\begin{aligned}f(\vartheta,|M|=m|\boldsymbol{Y})\propto&\frac{\lambda^m}{m!}\frac{1}{L^m W^m}\prod_{i=1}^{m}\left\{\frac{\beta^\alpha}{\Gamma(\alpha)}(\tau_{i1}^2)^{-\alpha-1}\exp\left(-\frac{\beta}{\tau_{i1}^2}\right)\right\}\\&\times\prod_{i=1}^{m}\left\{\frac{\beta^\alpha}{\Gamma(\alpha)}(\tau_{i2}^2)^{-\alpha-1}\exp\left(-\frac{\beta}{\tau_{i2}^2}\right)\right\}P(\boldsymbol{Y}|\vartheta,m),\end{aligned}\tag{3}$$

where

$$P(\boldsymbol{Y}|\vartheta,m)=\frac{\Gamma\left(\frac{v+n}{2}\right)(v)^{v/2}}{\pi^{n/2}\Gamma\left(\frac{v}{2}\right)|I_n+\Phi V\Phi^T|^{1/2}}\times\{v+\boldsymbol{Y}^T(I+\Phi V\Phi^T)^{-1}\boldsymbol{Y}\}^{-(v+n)/2}.$$

Note that the intensity for a true peak should be positive for the 2D NMR spectrum considered here. However, in our model, no any constraints are imposed concerning the value of a. This allows us to integrate out a from the posterior and, as a consequence, this accelerates the convergence of the simulation of the posterior. The marginal posterior distribution of a is normal with mean $(\Phi^T\Phi+V^{-1})^{-1}\Phi^T\boldsymbol{Y}$ and covariance matrix $\sigma^2(\Phi^T\Phi+V^{-1})^{-1}$. Hence, a can be estimated based on its expectation $(\Phi^T\Phi+V^{-1})^{-1}\Phi^T\boldsymbol{Y}$ conditional on the samples of ϑ and m obtained at each iteration.

Bayesian peak picking

The Bayesian peak picking problem is to determine the number of peaks, m, and the peak positions $(\mu_{11},\ \mu_{12}),\ldots,(\mu_{m1},\ \mu_{m2})$ through simulating from the posterior (Eq. (3)). However, it is not known how many peaks there are for a given NMR spectrum, although the intensities at the grid points around the peaks are relatively high. Based on this observation, we propose following algorithm for Bayesian peak picking.

For an $L\times W$ grid NMR spectrum, we first select N poles as "peak candidates". This can be done by selecting N poles with the highest intensities, or, if we have the results from some other methods, we can set them to be part of the peak candidates as well. In this paper, we have tried both. Let $\{(P_{1,1},P_{1,2}),\ldots,(P_{N,1},\ P_{N,2})\}$ denote the pool of candidate peaks, which gives all candidate components for the model (Eq. (1)). Then the peak picking problem is casted as a Bayesian variable selection problem, selecting appropriate components from the pool of candidate peaks.

For the solution of the Bayesian variable selection problem, we apply the stochastic approximation Monte Carlo (SAMC) algorithm [9] to estimate both the number and positions of the peaks through simulating from the posterior distribution (Eq. (3)). SAMC is an adaptive Markov chain Monte Carlo (MCMC) algorithm which possesses the self-adjusting mechanism and is immune to local trap problems. At each step, SAMC updates the set of selected peaks by either adding a peak (birth move), deleting a peak (death move), or refining the position of a selected peak (position update). Let $\mathbb{P}_{\mathrm{I}}^t$ denote the peaks included in the model at iteration t and let $\mathbb{P}_{\mathrm{R}}^t$ denote the remaining peaks that are not included in the current sample. Hence, $\mathbb{P}_{\mathrm{I}}^t\cup\mathbb{P}_{\mathrm{R}}^t=\{(P_{1,1},P_{1,2}),\ldots,(P_{N,1},P_{N,2})\}$ The birth move creates a new peak by randomly selecting one from the set $\mathbb{P}_{\mathrm{R}}^t$ and proposing a peak position based on the selected peak. The death move removes one peak from the set $\mathbb{P}_{\mathrm{I}}^t$. The position update refines the position of a randomly-selected peak, which does not change the dimension of the model (Eq. (1)).

A brief review of the SAMC algorithm

Let $f(x)=c\psi(x)$, $x\in\chi$, denote a distribution that we are working with, where c denotes a constant and X denotes the sample space of the distribution. Let $U(x)=-\log(\psi(x))$ denote the energy function of the distribution. SAMC works on a partitioned sample space. For example, the sample space can be partitioned into κ disjoint subregions according to the energy function: $E_1=\{x{:}U(x)<u_1\}$, $E_2=\{x{:}u_1\leqslant U(x)<u_2\},\ldots,$ $E_{k-1}=\{x{:}u_{\kappa-2}\leqslant U(x)<u_{\kappa-1}\}$ and $E_\kappa=\{x{:}U(x)\geqslant u_{\kappa-1}\}$, where $u_1,u_2,\ldots,u_{\kappa-1}$ are prespecified numbers. SAMC algorithm aims to sample from the following distribution:

$$P_\theta(x)\propto\sum_i^\kappa\frac{\psi(x)}{e^{\theta_i}}I(x\in E_i),\tag{4}$$

where $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_k)$ and $\theta_i=\log\int_{E_i}\psi(x)dx$, and $I(\cdot)$ is the indicator function. It is easy to see that sampling from (Eq. (4)) will lead to a "random walk" in the space of energy, if the sample space is partitioned according to the energy function and each subregion is treated as a "point".

However, $\boldsymbol{\theta}$ is usually unknown. SAMC provides an automatic mechanism to estimate $\boldsymbol{\theta}$ in simulations from $f(x)$. As shown in [10], SAMC is essentially a dynamic importance sampling algorithm. Let θ_{ti} denote the estimate of $\log\int_{E_i}\psi(x)dx$ at

iteration t, and define $\boldsymbol{\theta}_t = (\theta_{t1}, \ldots, \theta_{tk})$. Then one iteration of the SAMC algorithm can be described as follows.

(1) Conditioned on the current sample $x^{(t)}$, simulate a sample $x^{(t+1)}$ according to a Markov transition kernel, which admits the following distribution as the invariant distribution:

$$P_{\theta_t}(x) \propto \sum_{i}^{\kappa} \frac{\psi(x)}{e^{\theta_{ti}}} I(x \in E_i) \tag{5}$$

(2) Set $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \gamma_{t+1}(\boldsymbol{e}_{t+1} - 1/\kappa)$, where $e_{t+1} = (e_{t+1,1}, \ldots, e_{t+1,m})$, $e_{t+1,i} = 1$ if $x^{(t+1)} \in E_i$ and 0 otherwise, and γ_{t+1} is called the gain factor.

The gain factor sequence $\{\gamma_t\}$ is positive and non-decreasing, and satisfies the conditions $\sum \gamma_t = \infty$ and $\sum \gamma_t^{\xi} < \infty$ for some $\xi \in (1,2)$. In this paper, we set $\gamma_t = \frac{\delta t_0}{\max(t_0, t)}$, $t_0 = 5000$, $\delta = 0.5$.

When the dimension of x is high or when the sampling space X is too large, SAMC may take long time to converge. For this reason, we adopt a variant of SAMC, annealing stochastic approximation Monte Carlo [11], for simulating from the posterior (Eq. (3)). Annealing SAMC shrinks the sample space at each iteration according to the current sample. To be precise, at each iteration, annealing SAMC draws samples from the distribution

$$P_{\theta_t}(x) \propto \sum_{i}^{\prod (U_{min}^{(t)} + \aleph)} \frac{f(x)}{\exp(\theta_{t1})} I(x \in E_i) \tag{6}$$

where $U_{min}^{(t)}$ is the best value of $U(x)$ obtained by iteration t, $\aleph > 0$ is a user-defined parameter that determines the broadness of the sample space at each iteration, and $\Pi(u)$ denotes the index of subregions based on the energy function; if $u_{i-1} < u < u_i$, then $\Pi(u) = i$. Clearly, if $\aleph$ is large, say $\aleph \geqslant 20$, then it follows from the principle Occam's razor [12] that the samples simulated using annealing SAMC can still be used for Bayesian inference. In this paper, we set $\aleph = 1000$.

SAMC for Bayesian peak picking

In this section, we use M^* to denote the proposed model, use $M^{(t)}$ to denote the current model, use $\boldsymbol{\vartheta}^*$ to denote the parameter vector proposed for the model M^*, and use $\boldsymbol{\vartheta}^{(t)}$ to denote the parameter vector of the current model. At each iteration, SAMC randomly chooses to make one of the following moves with equal probability: position update, birth move and death move.

Position update

In this move, we randomly choose one component from the current model, say, the i-th component $\boldsymbol{\vartheta}^{(t)} = \left(\mu_{i1}^{(t)}, \mu_{i2}^{(t)}, \log\left(\tau_{i1}^2\right)^{(t)}, \log\left(\tau_{i2}^2\right)^{(t)}\right)$, then we propose to replace it by $\boldsymbol{\vartheta}^* = \left(\mu_{i1}^*, \mu_{i2}^*, \log\left(\tau_{i1}^{*2}\right), \log\left(\tau_{i2}^{*2}\right)\right)$, which is generated by one of the following with equal probability:

$$\begin{aligned} &\vartheta_{i,j}^* = \vartheta_{i,j}^{(t)} + un \times S, \text{ for one } j \text{ randomly drawn from } \{1,2,3,4\} \\ &\boldsymbol{\vartheta}_i^* = \boldsymbol{\vartheta}_i^{(t)} + un \times S \times \boldsymbol{e}, \end{aligned} \tag{7}$$

where un is a random variable generated from the standard normal distribution, S is called the step size, and $\boldsymbol{e}$ is a vector randomly drawn from a unit sphere of dimension 4.

The proposal is accepted with probability

$$\alpha = \min\left\{1, \frac{\exp\{\theta_{J(\boldsymbol{\vartheta}^{(t)})}\} P(\boldsymbol{\vartheta}^*, |M^*| \| Y) T(\boldsymbol{\vartheta}^* \to \boldsymbol{\vartheta}^{(t)})}{\exp\{\theta_{J(\boldsymbol{\vartheta}^*)}\} P(\boldsymbol{\vartheta}^{(t)}, |M^{(t)}| \| Y) T(\boldsymbol{\vartheta}^{(t)} \to \boldsymbol{\vartheta}^*)}\right\} \tag{8}$$

where $J(\boldsymbol{\vartheta})$ denotes the index of the subregion that the corresponding model belongs to and $T(\boldsymbol{\vartheta}^{(t)} \to \boldsymbol{\vartheta}^*)$ denotes the proposal distribution that is determined by Eq. (7).

Birth move

This move is to randomly choose a pole from the list of unselected peak candidates to add to the current model. For example, the peak $\{P_{i,1}, P_{i,2}\}$ is chosen, then the related parameters are proposed as follows:

$$\mu_{i1}^* = P_{i,1} + un_1 \times S, \tag{9}$$

$$\mu_{i2}^* = P_{i,2} + un_2 \times S, \tag{10}$$

$$\log\left(\tau_{i1}^{*2}\right) = U(\log(L_3), \log(U_3)), \tag{11}$$

$$\log\left(\tau_{i2}^{*2}\right) = U(\log(L_4), \log(U_4)), \tag{12}$$

where $un1$ and $un2$ are random samples drawn from the standard normal distribution. The acceptance probability of the move is given by

$$\alpha = \min\left\{1, \frac{Q(|M^*| \to |M^{(t)}|)}{Q(|M^{(t)}| \to |M^*|)} \frac{P(Death|M^*)}{P(Birth|M^{(t)})} R_{PU}\right\}, \tag{13}$$

where $R_{PU} = \frac{\exp\left\{\theta_{J(\vartheta^{(t)})}\right\} P(\vartheta^*, |M^*| \| Y) T(\vartheta^* \to \vartheta^{(t)})}{\exp\{\theta_{J(\vartheta^*)}\} P(\vartheta^{(t)}, |M^{(t)}| \| Y) T(\vartheta^{(t)} \to \vartheta^*)}$ is the acceptance rate for position update move. where $J(\boldsymbol{\vartheta})$ denotes the index of the subregion that the corresponding model belongs to; $Q(M^* \to M^{(t)})/Q(M^{(t)} \to M^*) = \left|\text{`P}_\text{R}^\text{t}\right| / \left(\text{`P}_\text{I}^\text{t} + 1\right)$ accounts for the probability of adding a pole/component to the current model; $\mathrm{T}(\cdot \to \cdot)$ denotes the proposal distribution determined by Eqs. (9)–(12); $P(Birth|M^{(t)}) = 1/3$ if $1 < |M^{(t)}| < m_{max}$, $P(Birth|M^{(t)}) = 2/3$ if $|M^{(t)}| = 1$, and $P(Birth|M^{(t)}) = 0$ if $|M^{(t)}| = m_{max}$; and $P(Death|M^*)$ if $1 < |M^*| < m_{max}$, $P(Death|M^*) = 0$ if $|M^*| = 1$, and $P(Death|M^*) = 2/3$ if $|M^*| = m_{max}$.

Death move

This move is to randomly delete one component from the model (Eq. (1)). The acceptance probability of this move is given by

$$\alpha = \min\left\{1, \frac{Q(|M^*| \to |M^{(t)}|)}{Q(|M^{(t)}| \to |M^*|)} \frac{P(Birth|M^*)}{P(Death|M^{(t)})} R_{PU}\right\} \tag{14}$$

where $R_{PU} = \frac{\exp\left\{\theta_{J(\vartheta^{(t)})}\right\} P(\vartheta^*, |M^*| \| Y) T(\vartheta^* \to \vartheta^{(t)})}{\exp\left\{\theta_{J(\vartheta^*)}\right\} P(\vartheta^{(t)}, |M^{(t)}| \| Y) T(\vartheta^{(t)} \to \vartheta^*)}$ is the acceptance rate for position update move. where $J(\boldsymbol{\vartheta})$ denotes the index of the subregion that the corresponding model belongs to; $Q(M^* \to M^{(t)})/Q(M^{(t)} \to M^*) = \left|\text{`P}_\text{I}^\text{t}\right| / \left(\text{`P}_\text{R}^\text{t} + 1\right)$ accounts for the probability of removing a component from the current model; $T(\cdot \to \cdot)$ denotes the proposed distribution determined by Eqs. (9)–(12); $P(Birth|M^*) = 1/3$ if $1 < |M^*| < m_{max}$, $P(Birth|M^*) = 2/3$ if $|M^*| = 1$, and $P(Birth|M^*) = 0$ if $|M^*| = m_{max}$; and $P(Death|M^{(t)}) = 1/3$ if $1 < |M^{(t)}| < m_{max}$, $P(Death|M^{(t)}) = 0$ if $|M^{(t)}| = 1$, and $P(Death|M^{(t)}) = 2/3$ if $|M^{(t)}| = m_{max}$.

Peak identification

At the end of the SAMC run, the peaks can be identified according to the marginal inclusion probability, that is, the posterior probability of each pole. Since SAMC is essentially a dynamic importance sampling algorithm [10], the marginal inclusion probability for a given pole can be estimated by

$$\widehat{I}_i = \frac{\sum_{t=t_1+1}^{t_1+t_2} I_i^t \exp(\theta_{J(\vartheta^{(t)})})}{\sum_{t=t_1+1}^{t_1+t_2} \exp(\theta_{J(\vartheta^{(t)})})}, \quad i = 1, 2, \ldots, N, \tag{15}$$

where t_1 denotes the number of burn-in iterations, t_2 denotes the number of iterations used for posterior calculation, and I_i^t is an indicator variable which is 1 if the i-th candidate peak is included in the model $M^{(t)}$ and 0 otherwise. In this paper, we set $t_1 = t_2 = 50{,}000$ for the simulation study and $t_1 = t_2 = 250{,}000$ for the real data examples. Alternatively, the peaks can be identified based on the maximum *a posteriori* (MAP) model. In our examples, the peaks identified by these two methods tend to be identical.

If a pole is identified as a peak, the related parameters can be estimated by

$$\hat{\vartheta}_{i,j} = \frac{\sum_{t=t_1+1}^{t_1+t_2} \vartheta_{i,j}^{(t)} \exp(\theta_{J(\vartheta^{(t)})})}{\sum_{t=t_1+1}^{t_1+t_2} \exp(\theta_{J(\vartheta^{(t)})})}. \tag{16}$$

It follows from the theory of SAMC, both $\widehat{I}_i$ and $\hat{\vartheta}_{i,j}$ are consistent.

Post-processing of simulation results

When applying the proposed method to NMR spectrum data, several issues need to be taken care for post-processing the simulation results. (1) As aforementioned, we did not restrict the peak intensity parameter vector a to be positive for the reason of computational efficiency. If the simulated model contains some components of negative intensities, we can directly eliminate them from the model. Those components capture the outrageous noise of the data, and removing them corresponds to a denoising step employed by other methods. (2) It is believed that the spreads of true peaks are relatively small as compared to the range of the spectrum. In model (Eq. (1)), the spreads of components are measured by τ_{i1} and τ_{i2} for $i = 1, 2, \ldots, N$. Hence, for a component, say component i, if τ_{i1} or τ_{i2} is large, then it is reasonable to treat it as an overall trend rather than a peak. This suggests us to remove it from the model. In our study, we found that it is good enough to set the threshold for τ_{i1} and τ_{i2} to be $\sqrt{L}/2$ and $\sqrt{W}/2$, respectively; that is, removing the peaks with $\tau_{i1} > \sqrt{L}/2$ or $\tau_{i2} > \sqrt{W}/2$. (3) In the practice of NMR peak picking, the tolerance limit for N dimension is 0.5 and that for H dimension is 0.05. Hence, if the simulated model contains two components that are close to each other in the sense that the difference between their locations is within the tolerance range, then we will combine them into a single peak.

Numerical results

Simulation study

In the simulation study, we generated an image of size 50×50 with 5 peaks. The volumes of the 5 peaks are 452293.9, 532729.6, 719234.05, 403184 and 215974.5, respectively. Their intensities are 14353.41, 15907.05, 18044.68, 43738.34 and 23187.57, respectively. Extra noises are added to the image. To study the sensitivity of our method to the noise, two situations are considered. (1) The noise follows a normal distribution with mean 0 and standard deviation 4000 and (2) the noise follows a normal distribution with mean 0 and standard deviation 4000; in addition, some extra negative spikes are put around the point (10, 20) with the volume 100,000.

Figure 2 shows the example for situation 1. The image with noises added is shown in Figure 2A, for which the true peaks are hard to detect using naked eyes, whereas the recovered image by SAMC, and the pure image without noises added are shown in Figure 2B and C. The comparison of the recovered image and the pure image shows that we have successfully denoised the image and recovered the locations and shapes of the peaks. As shown in **Figure 3**, the results for situation 2 is similar.

Table 1 shows the peak position estimation by our method for the simulated example with the noise as simulated in situation 1. It is easy to see that the estimation is rather accurate. In this table, we also include the marginal inclusion probability of candidate poles. The poles corresponding to the true peaks have a marginal inclusion probability of 1 and

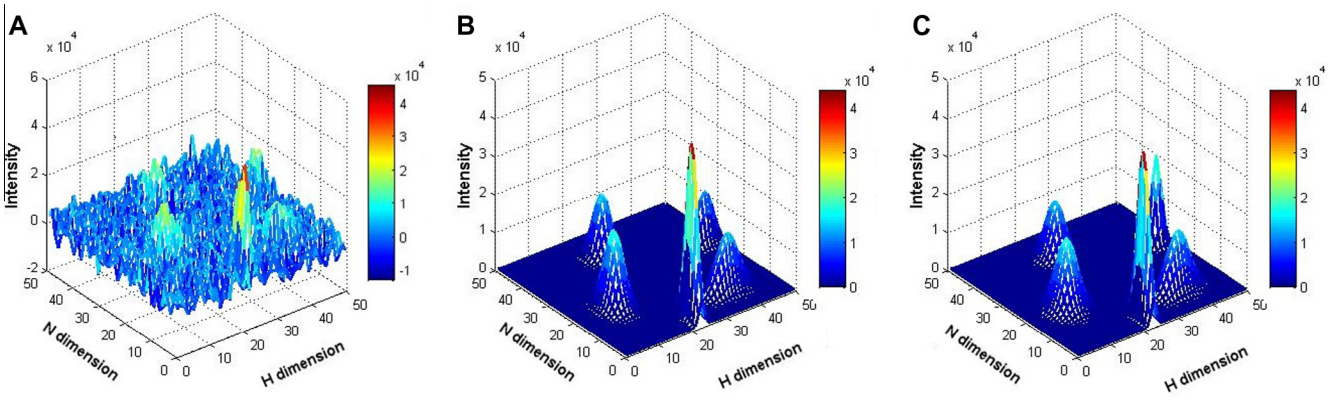

Figure 2 A simulated image of 5 peaks with the noise simulated as in situation 1

A. The image with noises added, for which the true peaks are hard to detect using naked eyes. **B.** The recovered image by SAMC. **C.** The pure image without noises added.

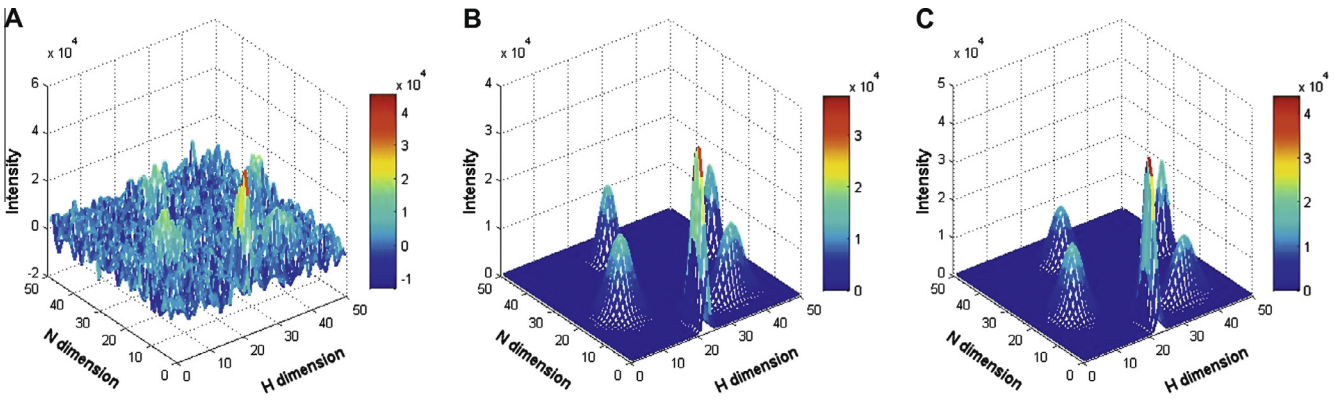

Figure 3 A simulated image of 5 peaks with the noise simulated as in situation 2

A. The image with noises added, for which the true peaks are hard to detect using naked eyes. **B.** The recovered image by SAMC. **C.** The pure image without noises added.

Table 1 Peak position estimation for the simulated example in situation 1

Peak	True position		Estimated position		MIP
	μ_1	μ_2	$\widehat{\mu_1}$	$\widehat{\mu_2}$	
1	40	24	40.80	24.14	1
2	10	37	9.70	36.94	1
3	20	12	20.16	11.91	1
4	5	23	4.84	22.94	1
5	30	46	30.23	46.01	1

Note: (μ_1, μ_2) is the location of the peaks and $(\widehat{\mu_1}, \widehat{\mu_2})$ is the estimation using Bayesian peak picking method. MIP refers to the marginal inclusion probability of the corresponding pole.

all others have a marginal inclusion probability of 0. This implies that our method has converged to true peaks.

NMR peak picking

We have applied the proposed method to six proteins along with a comparison with an existing method. We used 2D ^{15}N-HSQC spectra for the experiment. For the N dimension, a peak is considered correct if its distance from the truth is less than 0.5. For the H dimension, a peak is considered correct if the distance is less than 0.05. In the 2D space, a peak is considered correct if both the N and H dimensions are within the tolerance ranges when compared to the true peak.

Let N_T denote the number of true peaks in a given spectrum, let N_P denote the number of peaks being picked and let T_P denote the number of true peaks being picked. Then the recall rate is defined as T_P/N_T, which is the identification rate of a true peak; and the precision is defined as T_P/N_P, which is the proportion of true peaks among the identified peaks. Figure 4 shows the results of our method for protein

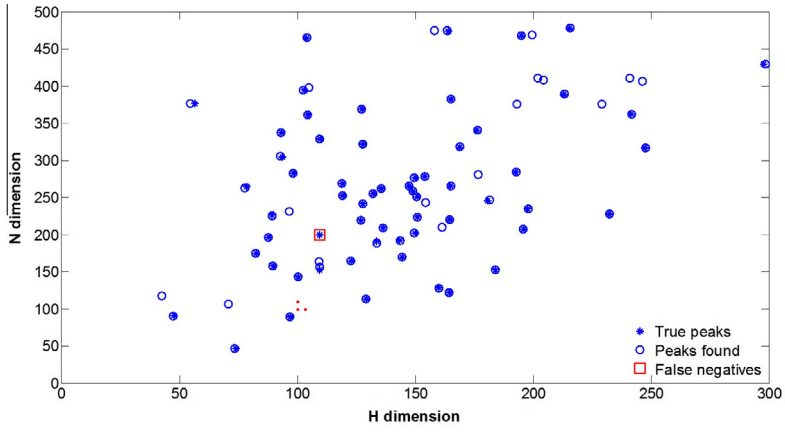

Figure 4 Results of Bayesian peak picking method for protein HACS1

One unit in H dimension represents 0.0148 ppm and one unit in D dimension represents 0.0873 ppm.

SAM domain, SH3 domain and nuclear localization signals 1(HACS1), where the asterisk (*) denotes the true peaks and the circle denotes the identified peaks using the proposed method. The only peak that was not identified by our method is the one around the grid point (109,200). The contour plot given in **Figure 5** shows that the intensity around the grid point (109,200) is very low.

Figure 6 gives the peak picking results using our method for protein coilin. **Figure 7** shows the contour plot of the NMR spectrum for coilin. It is obvious that in the region of $[240,320] \times [110,140]$, there are lots of peaks with very high intensities. However, there are no true peaks residing in this regions. Results show that our method is able to exclude a big portion of false peaks in that suspicious region, although not all of them.

Table 2 summarizes the results of our method for 6 proteins along with a comparison with PICKY [6], a newly developed powerful peak picking method. Table 2 reports the recall and precision and F-score [13] values for PICKY and the proposed method. On average, the proposed method is 1.0% more accurate in recall and 3.9% more accurate in precision for these 6 proteins.

Taking a closer look at Table 2, we can see that the proposed method has made improvements over PICKY under different situations. Our method has made the most significant improvements over PICKY on proteins vancomycin resistance associated regulator (VraR) and HACS1. For these two proteins, PICKY gives high recall rates but low precision values. Compared to PICKY, our method works well in eliminating false peaks. However, our method does not improve the results of PICKY for *Thermotoga maritima* enzyme protein TM1112, for which PICKY already did a good job. From this example, we find that our method can fail to identify overlapping peaks as other existing methods do. Table 2 reported the results with the candidate poles selected according to the intensities and according to the preliminary results of PICKY. Overall,

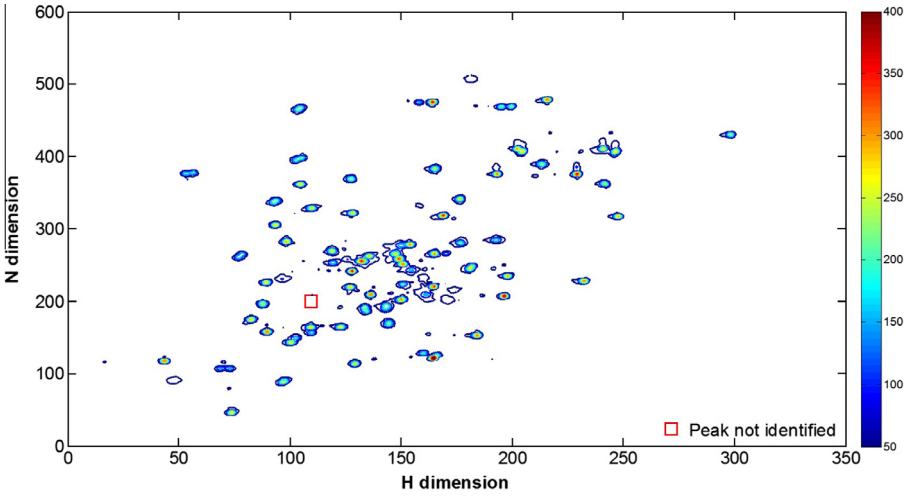

Figure 5 Contour plot for the ^{15}N-HSQC spectrum of protein HACS1
One unit in H dimension represents 0.0148 ppm and one unit in D dimension represents 0.0873 ppm.

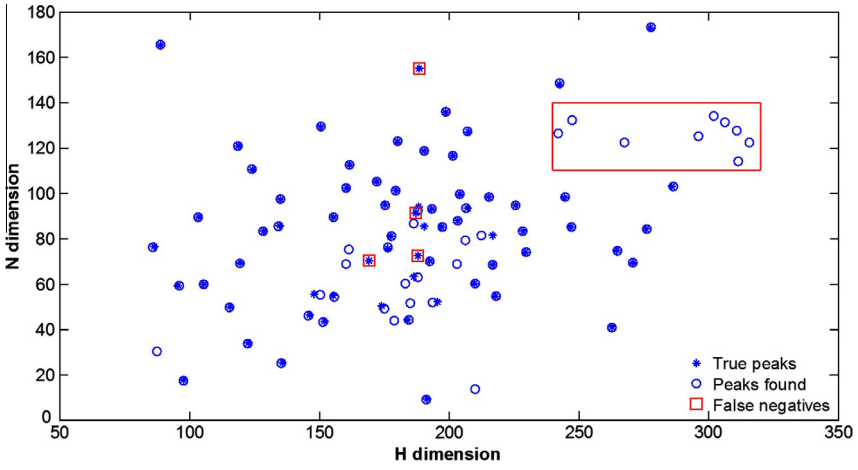

Figure 6 Results of Bayesian peak picking method for protein coilin
One unit in H dimension represents 0.0118 ppm and one unit in D dimension represents 0.1508 ppm. Asterisk (*) denotes the true peaks and the circle denotes the identified peaks using the proposed method.

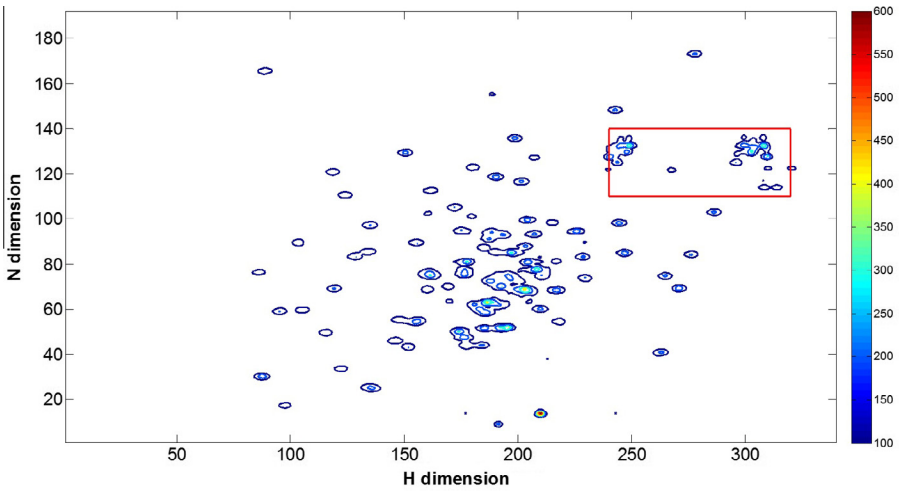

Figure 7 Contour plot for the ^{15}N-HSQC spectrum of protein coilin
One unit in H dimension represents 0.0118 ppm and one unit in D dimension represents 0.1508 ppm. Red square box marks the true peaks that are not detected by our method.

Table 2 Numerical results for the 6 proteins tested

Protein name	Protein length	PICKY			SAMC1			SAMC2		
		Recall	Precision	F-score	Recall	Precision	F-score	Recall	Precision	F-score
TM1112	89	96	89	92.4	94	89	91.4	95	85	89.7
RP3384	64	94	86	89.8	91	83	86.8	93	91	92.0
ATC1776	101	78	82	80.0	83	84	83.5	87	76	81.1
Coilin	98	97	70	81.3	94	77	84.7	94	80	86.4
VraR	72	87	93	89.9	93	98	95.4	91	98	94.4
HACS1	74	95	67	78.6	98	81	88.7	98	81	88.7
Average	—	91.2	81.2	85.3	92.2	85.3	88.4	93.0	85.2	88.7

Note: SAMC1, results of SAMC with peak candidates selected by intensities in a descending order; SAMC2, results of SAMC with peak candidates from the results of PICKY.

our method does not perform differently under the two aforementioned settings, since the self adjustment mechanism of SAMC makes the simulation less dependent on the starting point.

Discussion

In this paper, we proposed a Bayesian method to tackle the problem of NMR peak picking. Our numerical results indicate that the proposed method tends to produce more accurate results than the existing methods. To the best of our knowledge, this is the first effort in the literature that tackles the NMR peak picking problem using a Bayesian method. Our method has a few advantages over the existing methods. (1) Through choosing appropriate prior distributions, our method automatically penalizes the models with too many or too few peaks. (2) Our method can automatically distinguish true peaks from false ones without preprocessing the data. While the existing methods need to first remove the noise by setting a threshold at a risk of signal deletion. (3) Our method has the ability to estimate the spread and volume of each peak during the process of peak picking. This helps to reconstruct the denoised spectrum as compared to the existing methods which just give the peak positions.

A drawback of our method is that it is computationally intensive. This difficulty can be alleviated through parallel computing. We can partition the spectrum into multiple subregions and then process each of the subregions in parallel. For instance, for TM1112, we partition the spectrum into 6 subregions, and the run of SAMC takes only a few hours for each subregion. This is acceptable to most NMR laboratories.

Our method can be improved in various ways. For example, we can improve the fitting of the model to the spectra by replacing the Gaussian density function with a skew Gaussian density function, as the latter has a much more flexible density shape than the former. Other different prior distributions can also be tried for the model parameters, *e.g.*, the mixture *g*-prior [14], which can leads to the consistency of variable selection.

Authors' contributions

YC and FL conceived and designed the method. XG collected the data and YC analyzed the data. YC, XG and FL wrote the paper. All authors read and approved the final manuscript.

Competing interests

The authors declared that no competing interests exist.

Acknowledgements

This study was partially supported by grants from the National Science Foundation of USA (Grant No. DMS-1007457 and DMS-1106494) and the award (KUS-C1-016-04) made by King Abdullah University of Science and Technology (KAUST) to FL.

References

[1] Wüthrich K. NMR of Proteins and Nucleic Acids. New York: Wiley; 1986.
[2] Gao X. Mathematical approaches to the NMR peak-picking problem. J Appl Comput Math 2012;1:1.
[3] Gao X. Recent advances in computational methods for nuclear magnetic resonance data processing. Genomics Proteomics Bioinformatics 2013;11:29–33.
[4] Corne S, Jognson P, Fisher J. An artificial neural network for classifying cross peaks in two dimensional NMR spectra. J Magn Reson 1992;100:256–66.
[5] Koradi R, Billeter M, Engeli M, Güntert P, Wüthrich K. Automated peak picking and peak integration in macromolecular NMR spectra using AUTOPSY. J Magn Reson 1998;135:288–97.
[6] Alipanahi B, Gao X, Karakov E, Donaldson L, Li M. PICKY: a novel SVD-based NMR spectra peak picking method. Bioinformatics 2009;25:i268–75.
[7] Liu Z, Abbas A, Jing BY, Gao X. WaVPeak: picking NMR peaks through wavelet-based smoothing and volume-based filtering. Bioinformatics 2012;28:914–20.
[8] Raftery A, Madigan D, Hoeting J. Bayesian model averaging for linear regression models. J Am Stat Assoc 1997;92:179–91.
[9] Liang F, Liu C, Carroll R. Stochastic approximation in Monte Carlo computation. J Am Stat Assoc 2007;102:305–20.
[10] Liang F. On the use of stochastic approximation Monte Carlo for Monte Carlo integration. Stat Probab Lett 2009;79:581–7.
[11] Liang F. Annealing stochastic approximation Monte Carlo for neural network training. Mach Learn 2007;68:201–33.
[12] Madigan D, Raftery A. Model selection and accounting for model uncertainty in graphical models using Occam's window. J Am Stat Assoc 1994;89:1535–46.
[13] Huang YJ, Powers R, Montelione GT. Protein NMR recall, precision, and F-measure scores (RPF scores): structure quality assessment measures based on information retrieval statistics. J Am Chem Soc 2005;127:1665–74.
[14] Liang F, Paulo R, Molina G, Clyde M, Berger J. Mixtures of g priors for Bayesian variable selection. J Am Stat Assoc 2008;103:410–23.