Gibbs sampling

From Wikipedia, the free encyclopedia

In statistics, Gibbs sampling or a Gibbs sampler is a Markov chain Monte Carlo (MCMC) algorithm for obtaining a sequence of observations which are approximated from a specified multivariate probability distribution, when direct sampling is difficult. This sequence can be used to approximate the joint distribution (e.g., to generate a histogram of the distribution); to approximate the marginal distribution of one of the variables, or some subset of the variables (for example, the unknown parameters or latent variables); or to compute an integral (such as the expected value of one of the variables). Typically, some of the variables correspond to observations whose values are known, and hence do not need to be sampled.

Gibbs sampling is commonly used as a means of statistical inference, especially Bayesian inference. It is a randomized algorithm (i.e. an algorithm that makes use of random numbers), and is an alternative to deterministic algorithms for statistical inference such as the expectation-maximization algorithm (EM).

As with other MCMC algorithms, Gibbs sampling generates a Markov chain of samples, each of which is correlated with nearby samples. As a result, care must be taken if independent samples are desired (typically by thinning the resulting chain of samples by only taking every nth value, e.g. every 100th value). In addition, samples from the beginning of the chain (the burn-in period) may not accurately represent the desired distribution.

Contents

- 1 Introduction
- 2 Implementation
  - 2.1 Relation of conditional distribution and joint distribution
- 3 Inference
- 4 Mathematical background
- 5 Variations and extensions
  - 5.1 Blocked Gibbs sampler

- 8 Notes
- 9 References
- 10 External links

Introduction

Gibbs sampling is named after the physicist Josiah Willard Gibbs, in reference to an analogy between the sampling algorithm and statistical physics. The algorithm was described by brothers Stuart and Donald Geman in 1984, some eight decades after the death of Gibbs.[1]

In its basic version, Gibbs sampling is a special case of the Metropolis–Hastings algorithm. However, in its extended versions (see below), it can be considered a general framework for sampling from a large set of variables by sampling each variable (or in some cases, each group of variables) in turn, and can incorporate the Metropolis–Hastings algorithm (or more sophisticated methods such as slice sampling, adaptive rejection sampling and adaptive rejection Metropolis algorithms[2][3][4]) to implement one or more of the sampling steps.

Gibbs sampling is applicable when the joint distribution is not known explicitly or is difficult to sample from directly, but the conditional distribution of each variable is known and is easy (or at least, easier) to sample from. The Gibbs sampling algorithm generates an instance from the distribution of each variable in turn, conditional on the current values of the other variables. It can be shown (see, for example, Gelman et al. 1995) that the sequence of samples constitutes a Markov chain, and the stationary distribution of that Markov chain is just the sought-after joint distribution.

Gibbs sampling is particularly well-adapted to sampling the posterior distribution of a Bayesian network, since Bayesian networks are typically specified as a collection of conditional distributions.

Implementation

1. We begin with some initial value $\mathbf{x}^{(i)}$.
2. We want the next sample. Call this next sample $\mathbf{x}^{(i+1)}$. Since $\mathbf{x}^{(i+1)} = (x_1^{(i+1)}, x_2^{(i+1)}, \ldots, x_n^{(i+1)})$ is a vector, we sample each component of the vector, $x_j^{(i+1)}$, from the distribution of that component conditioned on all other components sampled so far. But there is a catch: we condition on $\mathbf{x}^{(i+1)}$'s components up to $x_{j-1}^{(i+1)}$, and thereafter condition on $\mathbf{x}^{(i)}$'s components, starting from $x_{j+1}^{(i)}$ to $x_n^{(i)}$. To achieve this, we sample the components in order, starting from the first component. More formally, to sample x_j^{i+1}, we update it according to the distribution specified by $p(x_j^{(i+1)} | x_1^{(i+1)}, \ldots, x_{j-1}^{(i+1)}, x_{j+1}^{(i)}, \ldots, x_n^{(i)})$. Note that we use the value that the $j+1$th component had in the ith sample, not the $i+1$th sample.
3. Repeat the above step k times.

If such sampling is performed, these important facts hold:

- The samples approximate the joint distribution of all variables.
- The marginal distribution of any subset of variables can be approximated by simply considering the samples for that subset of variables, ignoring the rest.
- The expected value of any variable can be approximated by averaging over all the samples.

When performing the sampling:

- The initial values of the variables can be determined randomly or by some other algorithm such as expectation-maximization.
- It is not actually necessary to determine an initial value for the first variable sampled.
- It is common to ignore some number of samples at the beginning (the so-called burn-in period), and then consider only every nth sample when averaging values to compute an expectation. For example, the first 1,000 samples might be ignored, and then every 100th sample averaged, throwing away all the rest. The reason for this is that (1) successive samples are not independent of each other but form a Markov chain with some amount of correlation; (2) the stationary distribution of the Markov chain is the desired joint distribution over the variables but it may

autocorrelation between samples, rather than moving around quickly, as is desired). Other techniques that may reduce autocorrelation are collapsed Gibbs sampling, blocked Gibbs sampling, and ordered overrelaxation; see below.

Relation of conditional distribution and joint distribution

Furthermore, the conditional distribution of one variable given all others is proportional to the joint distribution:

$$p(x_j|x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_n) = \frac{p(x_1,\ldots,x_n)}{p(x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_n)} \propto p(x_1,\ldots,x_n)$$

"Proportional to" in this case means that the denominator is not a function of x_j and thus is the same for all values of x_j; it forms part of the normalization constant for the distribution over x_j. In practice, to determine the nature of the conditional distribution of a factor x_j, it is easiest to factor the joint distribution according to the individual conditional distributions defined by the graphical model over the variables, ignore all factors that are not functions of x_j (all of which, together with the denominator above, constitute the normalization constant), and then reinstate the normalization constant at the end, as necessary. In practice, this means doing one of three things:

1. If the distribution is discrete, the individual probabilities of all possible values of x_j are computed, and then summed to find the normalization constant.
2. If the distribution is continuous and of a known form, the normalization constant will also be known.
3. In other cases, the normalization constant can usually be ignored, as most sampling methods do not require it.

Inference

Gibbs sampling is commonly used for statistical inference (e.g. determining the best value of a parameter, such as determining the number of people likely to shop at a particular store on a given day, the candidate a voter

The most likely value of a desired parameter (the mode) could then simply be selected by choosing the sample value that occurs most commonly; this is essentially equivalent to maximum a posteriori estimation of a parameter. (Since the parameters are usually continuous, it is often necessary to "bin" the sampled values into one of a finite number of ranges or "bins" in order to get a meaningful estimate of the mode.) More commonly, however, the expected value (mean or average) of the sampled values is chosen; this is a Bayes estimator that takes advantage of the additional data about the entire distribution that is available from Bayesian sampling, whereas a maximization algorithm such as expectation maximization (EM) is capable of only returning a single point from the distribution. For example, for a unimodal distribution the mean (expected value) is usually similar to the mode (most common value), but if the distribution is skewed in one direction, the mean will be moved in that direction, which effectively accounts for the extra probability mass in that direction. (Note, however, that if a distribution is multimodal, the expected value may not return a meaningful point, and any of the modes is typically a better choice.)

Although some of the variables typically correspond to parameters of interest, others are uninteresting ("nuisance") variables introduced into the model to properly express the relationships among variables. Although the sampled values represent the joint distribution over all variables, the nuisance variables can simply be ignored when computing expected values or modes; this is equivalent to marginalizing over the nuisance variables. When a value for multiple variables is desired, the expected value is simply computed over each variable separately. (When computing the mode, however, all variables must be considered together.)

Supervised learning, unsupervised learning and semi-supervised learning (aka learning with missing values) can all be handled by simply fixing the values of all variables whose values are known, and sampling from the remainder.

For observed data, there will be one variable for each observation—rather than, for example, one variable corresponding to the sample mean or sample variance of a set of observations. In fact, there generally will be no variables at all corresponding to concepts such as "sample mean" or "sample variance". Instead, in such a case there will be variables representing the unknown true mean and true variance, and the determination of sample values for these variables results automatically from the operation of the Gibbs

to add additional variables and take advantage of conjugacy. However, logistic regression cannot be handled this way. One possibility is to approximate the logistic function with a mixture (typically 7–9) of normal distributions. More commonly, however, Metropolis–Hastings is used instead of Gibbs sampling.

Mathematical background

Suppose that a sample X is taken from a distribution depending on a parameter vector $\theta \in \Theta$ of length d, with prior distribution $g(\theta_1, \ldots, \theta_d)$. It may be that d is very large and that numerical integration to find the marginal densities of the θ_i would be computationally expensive. Then an alternative method of calculating the marginal densities is to create a Markov chain on the space Θ by repeating these two steps:

1. Pick a random index $1 \leq j \leq d$
2. Pick a new value for θ_j according to $g(\theta_1, \ldots, \theta_{j-1}, \cdot, \theta_{j+1}, \ldots, \theta_d)$

These steps define a reversible Markov chain with the desired invariant distribution g. This can be proved as follows. Define $x \sim_j y$ if $x_i = y_i$ for all $i \neq j$ and let p_{xy} denote the probability of a jump from $x \in \Theta$ to $y \in \Theta$. Then, the transition probabilities are

$$p_{xy} = \begin{cases} \frac{1}{d} \frac{g(y)}{\sum_{z \in \Theta : z \sim_j x} g(z)} & x \sim_j y \\ 0 & \text{otherwise} \end{cases}$$

So

$$g(x)p_{xy} = \frac{1}{d} \frac{g(x)g(y)}{\sum_{z \in \Theta : z \sim_j x} g(z)} = \frac{1}{d} \frac{g(y)g(x)}{\sum_{z \in \Theta : z \sim_j y} g(z)} = g(y)p_{yx}$$

since $x \sim_j y$ is an equivalence relation. Thus the detailed balance equations are satisfied, implying the chain is reversible and it has invariant distribution g.

In practice, the suffix j is not chosen at random, and the chain cycles

Numerous variations of the basic Gibbs sampler exist. The goal of these variations is to reduce the autocorrelation between samples sufficiently to overcome any added computational costs.

Blocked Gibbs sampler

- A blocked Gibbs sampler groups two or more variables together and samples from their joint distribution conditioned on all other variables, rather than sampling from each one individually. For example, in a hidden Markov model, a blocked Gibbs sampler might sample from all the latent variables making up the Markov chain in one go, using the forward-backward algorithm.

Collapsed Gibbs sampler

- A collapsed Gibbs sampler integrates out (marginalizes over) one or more variables when sampling for some other variable. For example, imagine that a model consists of three variables A, B, and C. A simple Gibbs sampler would sample from p(A|B,C), then p(B|A,C), then p(C|A,B). A collapsed Gibbs sampler might replace the sampling step for A with a sample taken from the marginal distribution p(A|C), with variable B integrated out in this case. Alternatively, variable B could be collapsed out entirely, alternately sampling from p(A|C) and p(C|A) and not sampling over B at all. The distribution over a variable A that arises when collapsing a parent variable B is called a compound distribution; sampling from this distribution is generally tractable when B is the conjugate prior for A, particularly when A and B are members of the exponential family. For more information, see the article on compound distributions or Liu (1994).[5]

Implementing a collapsed Gibbs sampler

Collapsing Dirichlet distributions

In hierarchical Bayesian models with categorical variables, such as latent Dirichlet allocation and various other models used in natural language processing, it is quite common to collapse out the Dirichlet distributions that are typically used as prior distributions over the categorical

1. Collapsing out a Dirichlet prior node affects only the parent and children nodes of the prior. Since the parent is often a constant, it is typically only the children that we need to worry about.
2. Collapsing out a Dirichlet prior introduces dependencies among all the categorical children dependent on that prior — but no extra dependencies among any other categorical children. (This is important to keep in mind, for example, when there are multiple Dirichlet priors related by the same hyperprior. Each Dirichlet prior can be independently collapsed and affects only its direct children.)
3. After collapsing, the conditional distribution of one dependent children on the others assumes a very simple form: The probability of seeing a given value is proportional to the sum of the corresponding hyperprior for this value, and the count of all of the other dependent nodes assuming the same value. Nodes not dependent on the same prior must not be counted. Note that the same rule applies in other iterative inference methods, such as variational Bayes or expectation maximization; however, if the method involves keeping partial counts, then the partial counts for the value in question must be summed across all the other dependent nodes. Sometimes this summed up partial count is termed the expected count or similar. Note also that the probability is proportional to the resulting value; the actual probability must be determined by normalizing across all the possible values that the categorical variable can take (i.e. adding up the computed result for each possible value of the categorical variable, and dividing all the computed results by this sum).
4. If a given categorical node has dependent children (e.g. when it is a latent variable in a mixture model), the value computed in the previous step (expected count plus prior, or whatever is computed) must be multiplied by the actual conditional probabilities (not a computed value that is proportional to the probability!) of all children given their parents. See the article on the Dirichlet-multinomial distribution for a detailed discussion.
5. In the case where the group membership of the nodes dependent on a given Dirichlet prior may change dynamically depending on some other variable (e.g. a categorical variable indexed by another latent categorical variable, as in a topic model), the same expected counts are still computed, but need to be done carefully so that the correct set of variables is included. See the article on the Dirichlet-multinomial distribution for more discussion, including in the context of a topic model

inverse-gamma-distributed variance out of a network with a single Gaussian child will yield a Student's t-distribution. (For that matter, collapsing both the mean and variance of a single Gaussian child will still yield a Student's t-distribution, provided both are conjugate, i.e. Gaussian mean, inverse-gamma variance.)

If there are multiple child nodes, they will all become dependent, as in the Dirichlet-categorical case. The resulting joint distribution will have a closed form that resembles in some ways the compound distribution, although it will have a product of a number of factors, one for each child node, in it.

In addition, and most importantly, the resulting conditional distribution of one of the child nodes given the others (and also given the parents of the collapsed node(s), but not given the children of the child nodes) will have the same density as the posterior predictive distribution of all the remaining child nodes. Furthermore, the posterior predictive distribution has the same density as the basic compound distribution of a single node, although with different parameters. The general formula is given in the article on compound distributions.

For example, given a Bayes network with a set of conditionally independent identically distributed Gaussian-distributed nodes with conjugate prior distributions placed on the mean and variance, the conditional distribution of one node given the others after compounding out both the mean and variance will be a Student's t-distribution. Similarly, the result of compounding out the gamma prior of a number of Poisson-distributed nodes causes the conditional distribution of one node given the others to assume a negative binomial distribution.

In these cases where compounding produces a well-known distribution, efficient sampling procedures often exist, and using them will often (although not necessarily) be more efficient than not collapsing, and instead sampling both prior and child nodes separately. However, in the case where the compound distribution is not well-known, it may not be easy to sample from, since it generally will not belong to the exponential family and typically will not be log-concave (which would make it easy to sample using adaptive rejection sampling, since a closed form always exists).

joint distribution. If the child nodes of the collapsed nodes are continuous, this distribution will generally not be of a known form, and may well be difficult to sample from despite the fact that a closed form can be written, for the same reasons as described above for non-well-known compound distributions. However, in the particular case that the child nodes are discrete, sampling is feasible, regardless of whether the children of these child nodes are continuous or discrete. In fact, the principle involved here is described in fair detail in the article on the Dirichlet-multinomial distribution.

Gibbs sampler with ordered overrelaxation

- A Gibbs sampler with ordered overrelaxation samples a given odd number of candidate values for $x_j^{(i)}$ at any given step and sorts them, along with the single value for $x_j^{(i-1)}$ according to some well-defined ordering. If $x_j^{(i-1)}$ is the s^{th} smallest in the sorted list then the $x_j^{(i)}$ is selected as the s^{th} largest in the sorted list. For more information, see Neal (1995).[6]

Samplers-within-Gibbs and other extensions

It is also possible to extend Gibbs sampling in various ways. For example, in the case of variables whose conditional distribution is not easy to sample from, a single iteration of slice sampling or the Metropolis-Hastings algorithm can be used to sample from the variables in question. A more efficient alternative is the application of the adaptive rejection sampling (ARS) methods for sampling from the full-conditional densities.[7][8][9][10][11] When the ARS techniques cannot be applied, the adaptive rejection Metropolis sampling algorithms are often employed.[2][3][4] Furthermore, other alternatives can be found in literature.[12][13]

It is also possible to incorporate variables that are not random variables, but whose value is deterministically computed from other variables. Generalized linear models, e.g. logistic regression (aka "maximum entropy models") can be incorporated in this fashion (BUGS for example allows

(0, 0) and (1, 1) each have probability ½, but the other two vectors (0, 1) and (1, 0) have probability zero. Gibbs sampling will become trapped in one of the two high-probability vectors, and will never reach the other one. More generally, for any distribution over high-dimensional, real-valued vectors, if two particular elements of the vector are perfectly correlated (or perfectly anti-correlated), those two elements will become stuck, and Gibbs sampling will never be able to change them.

The second problem can happen even when all states have nonzero probability and there is only a single island of high-probability states. For example, consider a probability distribution over 100-bit vectors, where the all-zeros vector occurs with probability ½, and all other vectors are equally probable, and so have a probability of $\frac{1}{2(2^{100}-1)}$ each. If you want to estimate the probability of the zero vector, it would be sufficient to take 100 or 1000 samples from the true distribution. That would very likely give an answer very close to ½. But you would probably have to take more than 2^{100} samples from Gibbs sampling to get the same result. No computer could do this in a lifetime.

This problem occurs no matter how long the burn-in period is. This is because in the true distribution, the zero vector occurs half the time, and those occurrences are randomly mixed in with the nonzero vectors. Even a small sample will see both zero and nonzero vectors. But Gibbs sampling will alternate between returning only the zero vector for long periods (about 2^{99} in a row), then only nonzero vectors for long periods (about 2^{99} in a row). Thus convergence to the true distribution is extremely slow, requiring much more than 2^{99} steps; taking this many steps is not computationally feasible in a reasonable time period. The slow convergence here can be seen as a consequence of the curse of dimensionality.

Note that a problem like this can be solved by block sampling the entire 100-bit vector at once. (This assumes that the 100-bit vector is part of a larger set of variables. If this vector is the only thing being sampled, then block sampling is equivalent to not doing Gibbs sampling at all, which by hypothesis would be difficult.)

Software

Church is free software for performing Gibbs inference over arbitrary distributions that are specified as probabilistic programs.

PyMC is an open source Python library for Bayesian learning of general Probabilistic Graphical Model with advanced features and easy to use interface.[14]

IA2RMS (http://a2rms.sourceforge.net) is a Matlab code of the Independent Doubly Adaptive Rejection Metropolis Sampling method for drawing from the full-conditional densities.[4]

Notes

1. Geman, S.; Geman, D. (1984). "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images". IEEE Transactions on Pattern Analysis and Machine Intelligence. 6 (6): 721–741. doi:10.1109/TPAMI.1984.4767596.
2. Gilks, W. R.; Best, N. G.; Tan, K. K. C. (1995-01-01). "Adaptive Rejection Metropolis Sampling within Gibbs Sampling". Journal of the Royal Statistical Society. Series C (Applied Statistics). 44 (4): 455–472. JSTOR 2986138.
3. Meyer, Renate; Cai, Bo; Perron, François (2008-03-15). "Adaptive rejection Metropolis sampling using Lagrange interpolation polynomials of degree 2". Computational Statistics & Data Analysis. 52 (7): 3408–3423. doi:10.1016/j.csda.2008.01.005.
4. Martino, L.; Read, J.; Luengo, D. (2015-06-01). "Independent Doubly Adaptive Rejection Metropolis Sampling Within Gibbs Sampling". IEEE Transactions on Signal Processing. 63 (12): 3123–3138. doi:10.1109/TSP.2015.2420537. ISSN 1053-587X.
5. Liu, Jun S. (September 1994). "The Collapsed Gibbs Sampler in Bayesian Computations with Applications to a Gene Regulation Problem". Journal of the American Statistical Association. 89 (427): 958–966. doi:10.2307/2290921. JSTOR 2290921.
6. Neal, Radford M. (1995). Suppressing Random Walks in Markov Chain Monte Carlo Using Ordered Overrelaxation (Technical report). University of Toronto, Department of Statistics. 9508.
7. Gilks, W. R.; Wild, P. (1992-01-01). "Adaptive Rejection Sampling for Gibbs Sampling". Journal of the Royal Statistical Society. Series C (Applied Statistics). 41 (2): 337–348. doi:10.2307/2347565.
8. Hörmann, Wolfgang (1995-06-01). "A Rejection Technique for Sampling from T-concave Distributions". ACM Trans. Math. Softw. 21 (2): 182–193. doi:10.1145/203082.203089. ISSN 0098-3500.
9. Martino, Luca; Míguez, Joaquín (2010-08-25). "A generalization of the adaptive

12. Ritter, Christian; Tanner, Martin A. (1992-09-01). "Facilitating the Gibbs Sampler: The Gibbs Stopper and the Griddy-Gibbs Sampler". Journal of the American Statistical Association. 87 (419): 861–868. doi:10.1080/01621459.1992.10475289. ISSN 0162-1459.
13. Martino, L.; Yang, H.; Luengo, D.; Kanniainen, J.; Corander, J. (2015-12-01). "A fast universal self-tuned sampler within Gibbs sampling". Digital Signal Processing. Special Issue in Honour of William J. (Bill) Fitzgerald. 47: 68–83. doi:10.1016/j.dsp.2015.04.005.
14. "PyMC on GitHub".

References

- Bishop, Christopher M. (2006), Pattern Recognition and Machine Learning, Springer, ISBN 0-387-31073-8
- Bolstad, William M. (2010), Understanding Computational Bayesian Statistics, John Wiley ISBN 978-0-470-04609-8
- Casella, G.; George, E. I. (1992). "Explaining the Gibbs Sampler". The American Statistician. 46 (3): 167. doi:10.2307/2685208. JSTOR 2685208. (Contains a basic summary and many references.)
- Gelfand, Alan E.; Smith, Adrian F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities", Journal of the American Statistical Association, 85 (410): 398–409, doi:10.2307/2289776, JSTOR 2289776, MR 1141740
- Gelman, A., Carlin J. B., Stern H. S., Dunson D., Vehtari A., Rubin D. B. (2013), Bayesian Data Analysis, third edition. London: Chapman & Hall.
- Levin, David A.; Peres, Yuval; Wilmer, Elizabeth L. (2008), "Markov Chains and Mixing Times (http://www.uoregon.edu/~dlevin/MARKOV/)", American Mathematical Society.
- Robert, C. P.; Casella, G. (2004), Monte Carlo Statistical Methods (second edition), Springer-Verlag.

External links

- The OpenBUGS Project (http://www.openbugs.net/) — Bayesian inference Using Gibbs Sampling
- A practical application of Gibbs sampling in genomics (http://ccmbweb.ccv.brown.edu/gibbs/gibbs.html)
- PyMC (https://github.com/pymc-devs/pymc) — Markov Chain Monte Carlo in

Categories: Markov chain Monte Carlo
