SOFTWARE HIGHLIGHT

Software for Graphical models: a review

by Kevin Murphy
murphyk@cs.ubc.ca

Graphical models (GMs) are a way to represent conditional independence assumptions by using graphs. Specifically, nodes represent random variables and lack of edges represent conditional independencies. The graph is a useful visual representation of complex stochastic systems. The graphical structure is also the basis of efficient inference algorithms.

There are many different kinds of graphical models, but the two most popular ones are based on directed acylic graphs (also called "Bayesian networks") and on undirected graphs (also called "Markov random fields"). In this article, we review some of the more popular and/or recent software packages for dealing with graphical models. A more extensive comparison can be found at http://www.cs.ubc.ca/~murphyk/Software/bnsoft.html.

BUGS

BUGS (Bayesian inference using Gibbs Sampling)[1] assumes the model is specified in the form of a DAG (directed acyclic graph), and uses Gibbs sampling for inference. A large number of different conditional distributions (node types) are supported. Internally, various algorithms (such as adaptive rejection sampling and slice sampling) are used to sample from the full conditionals. The software is easy to use, especially since it has recently become possible to call it directly from R (using R2WinBUGS[2]) and Matlab (using MatBUGS[3]), thus bypassing the rather cumbersome GUI.

Unfortunately, single site Gibbs sampling can be very slow to "mix", resulting in unreliable posterior inferences. In addition, Gibbs sampling cannot be used to find posterior modes, and cannot easily be used to compute the marginal likelihood, which is useful for model selection (although BUGS does return the DIC score of a model).

BUGS is freely available as an executable file. The most recent version, called WinBUGS, only runs on Windows (although one can run it on Linux systems using Wine). Recently, an open-source alternative called OpenBUGS[4] has been created, but it is not nearly as mature as WinBUGS. OpenBUGS is written in a language called "Component Pascal".

JAGS

JAGS (Just Another Gibbs Sampler)[5] is very similar in functionality to BUGS. The main difference is that it is fully open source, and works easily on multiple platforms (Windows, unix, etc). The principle advantage over OpenBUGS is that it is written in Java, which is a more widely known language than Component Pascal. In addition, it seems to have a simpler design than OpenBUGS.

VIBES

VIBES (Variational Inference for Bayesian Networks)[6] is open-source Java, and is designed to be similar to BUGS in functionality, but it uses the variational mean field algorithm for inference. This is potentially much faster, but less accurate than Gibbs sampling. In addition, it is limited to the conjugate exponential family. Note that VIBES is no longer being supported; its author is developing a replacement called Infer.NET (see below).

Infer.NET

Infer.NET[7] is a software package developed at Microsoft Research in Cambridge. They anticipate an initial public release in Spring 2008. The code will not be open source but will be freely available

[1] http://www.mrc-bsu.cam.ac.uk/bugs
[2] http://cran.r-project.org/src/contrib/Descriptions/R2WinBUGS.html
[3] http://www.cs.ubc.ca/~murphyk/Software/MATBUGS/matbugs.html
[4] http://mathstat.helsinki.fi/openbugs/
[5] http://www-ice.iarc.fr/~martyn/software/jags/
[6] http://vibes.sourceforge.net/
[7] http://research.microsoft.com/mlp/ml/Infer/Infer.htm
[8] http://research.microsoft.com/mlp/ml/Infer/Csoft.htm

for academic use. The model is specified using a new programming language called Csoft[8], which allows one to combine stochastic code with standard C# code. Thus one can easily specify graphical models of various kinds. Various Bayesian inference algorithms are supported, including Gibbs sampling, variational mean field, and expectation propagation. The package is designed to generate model-specific code, and to run very fast, even on large models.

BNT

BNT (Bayes Net Toolbox)[9] is open-source Matlab, and supports many different models and inference algorithms. In particular, it supports DAG models, "dynamic Bayesian networks" (which are DAG models unrolled in time) and influence/ decision diagrams. It also has undocumented and partial support for undirected models.

In terms of inference, BNT, like many other GM packages, can only perform Bayesian inference on discrete or Gaussian random variables. Hence parameter inference is performed using point estimation techniques such as EM or gradient descent. However, conditional on the parameters, inference of the remaining variables can often be performed exactly, using the junction tree algorithm, which includes well-known algorithms, such as the forwards-backwards algorithm, as special cases. If exact methods are too slow, a variety of different approximate inference algorithms are supported, such as "loopy belief propagation".

Hugin

Hugin[10] is a commercial package with functionality similar to BNT. It was one of the first packages for DAG models (including influence diagrams), and it is arguably the most mature. However, there are now a large number of other packages, such as Genie, MSBNx, Netica, PNL, etc. with very similar functionality (see `http://www.cs.ubc.ca/~murphyk/Software/bnsoft.html` for a comparison). These packages focus on exact inference in discrete-state (or conditionally Gaussian) models, using the junction tree or variable elimination algorithm. Some of them also support parameter estimation using EM. These packages are aimed at the business/ data-mining market, and hence they often put more emphasis on user interface and I/O issues than on core functionality.

gR

gR (graphical models in R)[11] is a collection of packages rather than a single package. The main package is `gRbase`, which is a way of defining data and models. There is also the `dynamicGraph` package, for visualizing and editing graphs. There are no Bayesian inference alorithms implemented in R. However, R interfaces to several existing model-fitting packages are provided, including `CoCo`, and `mimR`, both of which are designed for fitting contingency tables, which can be represented as undirected GMs.

Blaise

Blaise[12] is a Java software package that supports efficient Monte Carlo inference (including MCMC and sequential Monte Carlo samplers) in a large class of probabilistic models, including directed graphical models and non-parametric Bayesian models. The plan is to release a first version to the public, under a restricted open source license, in Spring 2008.

Gaussian graphical models

Gaussian graphical models are an important special case of graphical models that support efficient Bayesian inference using techniques from sparse linear algebra. GMRFsim[13] supports inference in undirected GGMs, and GDAGSim[14] supports inference in directed GGMs. Both of these can be used to perform block sampling inside an MCMC sampler. The ggm[15] R package can be used to fit undirected GGMs parameters using point estimation techniques.

Model selection

In addition to inference about states and parameters, there is much interest (especially in the

[9] `http://www.cs.ubc.ca/~murphyk/Software/BNT/bnt.html`
[10] `http://www.hugin.com`
[11] `http://www.r-project.org/gR`
[12] `http://publications.csail.mit.edu/abstracts/abstracts07/bonawitz/bonawitz.html`
[13] `http://www.math.ntnu.no/~hrue/GMRFsim`
[14] `http://www.staff.ncl.ac.uk/d.j.wilkinson/software/gdagsim`
[15] `http://cran.r-project.org/src/contrib/Descriptions/ggm.html`

systems biology community) in inference about the graph structure itself. The model selection problem is very difficult, because the space of all graphs on n nodes has size $O(2^{n^2})$. There are basically three main approaches to this: greedy search (and variants), MCMC model averaging, and constraint-based methods. Most of the work has focused on learning DAG models, although the WinMine[16] package learns dependency networks.

There are many packages that perform greedy search in DAG space: BNT, DAGlearn[17], Banjo[18], Deal[19], etc. BNT supports simple hill-climbing. DAGlearn uses L1-penalized logistic regression to reduce the search space. Banjo uses simulated annealing. Deal uses hill-climbing, but can handle conditionally Gaussian models (the other packages assume discrete data).

There are very few publically available packages that perform Bayesian model averaging in the space of DAGs. BNT implements a Metropolis Hastings method with a simple local proposal. BDAGL[20] uses a more sophisticated proposal based on dynamic programming. The GGM package[21] does model averaging in the space of undirected Gaussian GMs, using MCMC and stochastic search techniques. The HdBCS (high dimensional Bayesian covariance selection)[22] package is similar to the GGM package, but searches in the space of DAGs and then converts the result to an undirected GGM.

The constraint-based approach to structure learning, in which one eliminates edges if certain conditional independencies are detected in the data (using some hypothesis testing procedure), is generally faster but more error-prone than the above Bayesian techniques. BNT implements some of the simpler algorithms. Tetrad[23] is a more elaborate package.

Traditionally, the constraint-based approach has been the method of choice for people interested in learning "causal" models from observational data. However, this approach can of course be used to fit "acausal" models, too. For example, the SIN[24] R package uses conditional independency tests to learn the structure of undirected Gaussian GMs. The GeneNet[25] R uses an FDR approach to threshold the partial correlation coefficients to induce a sparse GGM. The glasso[26] R package uses L1 regularization to estimate a sparse precision matrix.

[16] http://research.microsoft.com/~dmax/WinMine/tooldoc.htm
[17] http://www.cs.ubc.ca/~murphyk/Software/DAGlearn/index.html
[18] http://www.cs.duke.edu/~amink/software/banjo
[19] http://www.math.auc.dk/novo/deal
[20] http://www.cs.ubc.ca/~murphyk/Software/BDAGL/
[21] http://xpress.isds.duke.edu:8080/softwarelinks/ggm.html
[22] http://www.stat.duke.edu/~adobra/hdbcs.html
[23] http://www.phil.cmu.edu/projects/tetrad/tetrad4.html
[24] http://cran.r-project.org/src/contrib/Descriptions/SIN.html
[25] http://www.strimmerlab.org/software/genenet/index.html
[26] http://www-stat.stanford.edu/~tibs/glasso/index.html