Bayesian Research at the NASA Ames Research Center, Computational Sciences Division

Robin D. Morris, rdm@email.arc.nasa.gov

February 7, 2003

[I am writing this in the week following the break-up of the space shuttle Columbia on re-entry. Our thoughts are with the families of Rick Husband, Michael Anderson, Laurel Clark, David Brown, William McCool, Kalpana Chawla and Ilan Ramon.]

NASA Ames Research Center is one of NASA's oldest centers, having started out as part of the National Advisory Committee on Aeronautics, (NACA). The site, about 40 miles south of San Francisco, still houses many wind tunnels and other aviation related departments. In recent years, with the growing realization that space exploration is heavily dependent on computing and data analysis, its focus has turned more towards Information Technology. The Computational Sciences Division has expanded rapidly as a result. In this article, I will give a brief overview of some of the past and present projects with a Bayesian content. Much more than is described here goes on with the Division. The web pages at `http://ic.arc.nasa.gov` give more information on these, and the other Division projects.

AUTOCLASS: Bayesian research at Ames began in 1985. The first major project, lead by Peter Cheeseman, was AUTOCLASS, a system for performing unsupervised classification of data, where the number and description of the natural classes of the data is not known. AUTOCLASS handles missing data, mixed real and discrete attributes, and estimates the posterior probability over a range of model structures. It is one of the earliest examples of a restricted class of Bayes Net system. AUTOCLASS has proved extremely useful in practice, and has found subtly different classes that were unknown to the investigators, as well as many previously known classes (but unknown to AutoClass). AUTOCLASS is publicly available.

IND: Another early project, lead by Wray Buntine, was the IND system, which was concerned with Bayesian software for supervised classification using decision trees. A tree is "grown" from data using a recursive partitioning algorithm to create a tree which (hopefully) has good prediction of classes on new data. As well as reimplementing parts of some of the standard Decision Tree algorithms (e.g. C4) and offering experimental control suites, IND also introduced Bayesian and MML methods and more sophisticated search in growing trees. These produce more accurate class probability estimates that are important in applications like diagnosis

The approach used in IND has subsequently been adapted to learning Bayesian networks from data, to learning n-grams for language modeling, and to a classification model known as Alternating Decision Trees . The data structures and algorithms have been quiet influential. Moreover, rumor has it that Breiman, an influential Bayesian antagonist, was motived by INDs apparent successes to develop the Bagging approach to classification trees that subsequently became the empirical champion in the field.

IND has seen widespread use in empirical and applied studies, and is publicly available.

AUTOBAYES: An ongoing project of general applicability in Bayesian analysis is the AutoBayes project.

AutoBayes is an automatic program synthesis system for the machine learning domain under development by the Automated Software Engineering group since 1999. From the outside, AutoBayes is essentially a compiler for a modeling language similar to the BUGS language; inside, however, it employs sophisticated code generation methods and is one of the more complex synthesis systems produced by the Automated Software Engineer-

ing community. AutoBayes takes as its input a statistical model, extracts a Bayesian network from it, and then generates a program which solves the learning task specified in the model. Unlike BUGS, however, AutoBayes is not restricted to a single generic algorithm (i.e., Gibbs sampling) but can generate different algorithms which are specialized for the model.

AutoBayes contains a comprehensive schema library. A schema contains two parts, a definition of when it is applicable, and a code template. During synthesis, AutoBayes finds the schemas that are applicable, then instantiates a code fragment in a model-specific way (e.g. an EM schema is instantiated if a sub-problem is recognised as a finite mixture model). These code fragments can spawn new, simpler, synthesis tasks, which are solved recursively; the recursion terminates if subproblems can be solved either numerically or symbolically. An important aspect here is the interaction of the schemas with the symbolic subsystem (i.e., a simple Mathematica-like symbolic-algebraic kernel) which allows the identification and efficient solution of tractable subproblems, even if they are embedded in the original model.

This divide-and-conquer approach allows AutoBayes to synthesize larger programs in a bottom-up fashion, using both schemas and symbolic solutions as building blocks. After synthesis, the code is optimized and translated into a C/C++ program which can be run standalone or linked dynamically into the Matlab or Octave environments.

AutoBayes has been used to generate code for a spectrum of models, ranging from textbook examples (e.g., normal models with various priors) to "almost state-of-the-art" machine learning algorithms; it has also been applied successfully to some data analysis problems within NASA.

Autonomy

A major research area in the Computational Sciences Division is to provide autonomous capabilities for spacecraft and rovers - communication bandwidth to space exploration vehicles and onboard storage are limited; the spacecraft collect vastly more data than can be returned, and cannot be controlled in real time[1]. The need for autonomy, both for spacecraft operations and scientific discovery, is obvious. One great success in this area was the *Remote Agent* system on the Deep Space One spacecraft. This was a "traditional" AI system, based around planning and scheduling, modeling the state of the spacecraft, and a smart executive module. Currently, research is underway to address some of the limitations of that system, and many of the approaches being pursued are Bayesian.

Diagnosis: Diagnosis is the problem of detecting and identifying any faults or unexpected events that occur in a system from observations of that system. Bayesian belief updating methods are being applied to this problem, maintaining a belief distribution over the state of the system, and updating the distribution based on a model of the evolution of the system and on new observations as they arrive. The models used are probabilistic hybrid automata – they contain a mixture of discrete states and continuous variables. The evolution of the system is governed by a transition function which gives the probability of a transition from one discrete state to another, and a set of differential equations which model the behavior of the continuous variables, and are dependent on the discrete state.

Optimal approaches to this problem are computationally infeasible, particularly on-board a spacecraft or planetary rover[2] . Particle filters can be used to track the state in reasonable computation time. However, diagnosis problems present some interesting challenges for particle filter algorithms, particularly because the fault states have very low probability of occurring. Several variants of particle filters have been developed, tuned to solving diagnosis problems.

Scheduling: New research is applying Bayesian techniques to scheduling problems. The domain here is one in which a number of tasks must be scheduled, given a set of constraints on when the tasks must be performed. Completing certain tasks, or subsets of the tasks results in a numerical reward. There is uncertainty about the duration of the individual tasks so the problem becomes one of building the schedule that maximizing the expected reward obtainable.

Autonomous Exploration: This project investigated the application of Bayesian statistics to the problem of autonomous geological exploration with a robotic vehicle. It

[1] Communication with spacecraft on Mars typically occurs twice per day.

[2] The computational capacity of the Mars rovers scheduled for launch later this year is equivalent to a 25MHz PowerPC

concentrated on the sub-problem of classifying rock types while addressing the issues associated with operating onboard a mobile robot. The Bayesian paradigm was used in a natural way to solve the more general robotic problems of autonomously profiling an area and allocating scarce sensor resources. Major considerations are the need to use of multiple sensors and the ability of a robotic vehicle to acquire data from different locations. Needless sensor use must be curtailed if possible, such as when an object is sufficiently well identified given sensor data acquired so far. Furthermore, by investigating rocks in many locations, the robot has the opportunity to profile the environment. Different rock samples are statistically dependent on each other. These dependencies can be exploited to substantially improve classification accuracy.

The classification system was been implemented onboard the Nomad robot developed at Carnegie Mellon University, and applied to the task of recognizing meteorites amongst terrestrial rocks in Antarctica. In January 2000 A.D., Nomad was deployed to Antarctica where it made the first autonomous robotic identification of a meteorite.

Data Analysis

NASA has been described as a *data collection agency* – each mission returns huge quantities of data, and Earth observing satellites return data at such a rate that it is difficult to archive, let alone analyze. Naturally, therefore, there are a number of data analysis projects within the Division.

Planetary Nebula Modeling: Stars like our sun end their lives as swollen red giants surrounded by cool extended atmospheres. The nuclear reactions in their cores create carbon, nitrogen and oxygen, which are transported by convection to the outer envelope of the stellar atmosphere. As the star finally collapses to become a white dwarf, this envelope is expelled from the star to form a planetary nebula (PN) rich in organic molecules. The physics, dynamics, and chemistry of these nebulae are poorly understood and have implications not only for our understanding of the stellar life cycle but also for organic astrochemistry and the creation of prebiotic molecules in interstellar space.

This project is working toward generating three-dimensional models of planetary nebulae, which include the size, orientation, shape, expansion rate and mass distribution of the nebula, as well as the distance from earth. Such a reconstruction of a PN is a challenging problem for several reasons. First, the data consist of images obtained over time from the Hubble Space Telescope and long-slit spectra obtained from Kitt Peak National Observatory and Cerro Tololo Inter-American Observatory. These images are of course taken from a single viewpoint in space, which amounts to a very challenging tomographic reconstruction. Second, that there are two disparate data types requires that we utilize a method that allows these data to be used together to obtain a solution. Bayesian model estimation is applied using a parameterized physical model that incorporates much prior information about the known physics of the PN. By modeling the nebula in three-dimensions it is possible reconcile the observed tangential expansion observed as an angular size change of the object with the radial expansion velocity determined from the Doppler shift in the spectral lines thus providing accurate estimates of the objects expansion velocity, dynamical age, and distance from earth.

Event analysis for GLAST: The Gamma Ray Large Area Space Telescope is a project to map the incidence of gamma rays from the entire sky. It is an orbiting telescope, scheduled for launch in 2006. It works by converting an incident gamma ray into an electron-positron pair in one of a stack of tungsten layers, and then detecting the positions where these charged particles cross layers of silicon microstrip detectors. However, the analysis is complicated by numerous secondary processes – the electron and positron are scattered each time they traverse the layers, and can also knock out further electrons, which cause the microstrips to fire as they cross them. We are studying the feasibility of using a detailed model of the physics of the detector to define importance sampling distributions to enable a particle filter type approach to be used to estimate, for each event, the direction from which the gamma ray came, and its energy.

Analysis of hyper-spectral solar flux data: This effort aims at developing a Bayesian framework for analyzing hyper-spectral data on solar radiation in the atmosphere, collected with a custom-built NASA radiometer in various field campaigns around the world. This framework is expected to allow efficient and accurate determination, from heterogeneous data, of the chemical composition and the physical state of the atmosphere, thus sig-

nificantly enhancing our understanding of, as well as our capability to model and predict, the Earth system. Specific goals include: retrieval of cloud physical parameters for understanding their evolution and for assessing their impact on weather and global climate; identification of composition, size, shape, and distribution of aerosols for evaluating their effects on solar radiation budget; quantification of the influence of tropospheric ozone and carbon-based trace gases on radiative forcing.

The main thrust of present research is toward developing forward physical models - one for the atmosphere and one for the instrument - suitable for use as likelihood functions within a Bayesian parameter estimation scheme.

Computer Vision: The low-level vision problem is conceived as the construction of a 3-D surface model of the local world, where the model is represented as a triangulated mesh with reflectance parameters associated with each triangle. In addition to inferring the 3-D mesh, the lighting and camera parameters must also be inferred. This is an extremely hard inference problem, because an observed image depends on the 3-D model geometry and reflectance as well as the camera and lighting parameters. The likelihood function is essentially the computer graphics problem: given all the model information, what would the image look like. Bayes theorem inverts this function and allows the 3-D model to be inferred given the images. The forward model is well understood; the physics of light scattering and camera optics is well known.

The simplified 2-D problem was first investigated. If the camera and lighting are essentially constant, there is no significant parallax, and it is impossible to separate out the effects of surface geometry from surface reflectance. Instead, model "images" at super-resolution are reconstructed from multiple images of the same area. Super-resolution is possible because each image is an independent sample of the unknown surface. This program yielded spectacular improvements in resolution, enabling features to be seen that were completely invisible in the individual images. Currently, this research is being extended to full 3-D surface reconstruction from images with different camera views and lighting. Although straight forward in principle, it is extremely difficult in practice because there are typically millions of model parameters to be inferred, and because simultaneous estimation of camera, lighting and 3-D surface model from images is greatly complicated by their mutual dependence. For this simultaneous inference to work, it is necessary to use techniques such as feature matching between images to "bootstrap" the inference procedure. Once such heuristic information initializes the joint model search sufficiently close to the global maximum, standard gradient methods to find the MAP estimate (and its associated covariance matrix) seem to work well in practice. This is ongoing research that should give a full Bayesian foundation to the problem of low-level computer vision.

Separation of Neural Signals: The electric potentials and magnetic fields generated by ensembles of synchronously active neurons in response to external stimuli provide information essential to understanding the processes underlying cognitive and sensorimotor activity. Interpreting the recordings of these potentials and fields can be problematic as each detector records signals that have been simultaneously generated by various regions throughout the brain. Separating these signals into a set of components each originating from a synchronous ensemble has proven to be a very difficult problem.

The differential variability component analysis (dVCA) algorithm relies on a more physiologically realistic source model that accounts for variability of response amplitude and latency across multiple experimental trials. Rather than making any unrealistic assumptions of independence of components, this algorithm utilizes the differential variability of the evoked waveforms to aid their characterization. By applying the Bayesian methodology to this new source model, we derive an algorithm that uses EEG data simultaneously recorded from multiple electrodes to identify multiple components each representing synchronous neuronal activity from an ensemble of neurons displaying a distinct trial-to-trial variability pattern. In addition, this algorithm estimates the single-trial amplitude and latency of each component active during any particular evoked response.

Analysis of Earth Observing Data: A couple of projects involved in the analysis of Earth Observing Data are the following.

One project is looking at using naive Bayes classifiers applied to MODIS (Moderate Resolution Imaging Spectroradiometer) data for generating a cloud mask product. The current methods of generating the cloud mask products from MODIS data at the DAACs (Distributed Active Archive Centers) are too slow to allow for the product to be included in the broadcast stream, and so are not used in

other data products, limiting their accuracy. The goal is to use naive Bayes to produce a quick product which could be sent out along with the data.

A second project is looking at the uncertainty present in the data products themselves, many of which are derived from the raw satellite observations. The derivation of these data products from the observations and other data is often via some empirically determined relationships (e.g. the production of Leaf Area Index maps from Normalised Difference Vegetation Index maps). The Earth Science community then uses these derived quantities with little appreciation of the range of uncertainty present, and the effect of that uncertainty on predictions made using these derived data products. In this project we are analyzing the relationships used to generate certain data products, with a view to quantifying the uncertainty, and making it available together with the data product.

Novel Interfaces: This builds on work done on Monte-Carlo methods for mixture modeling. In particular a Bayesian approach to the parameterization of Gaussian mixture models, looking at the case where the distributions change over time. This work will be applied to "virtual keyboards", where electrical signals from the muscles in the users forearm are captured by dry electrodes on the skin, and decoded to recognize the gestures associated with pressing a particular key. It will also used to enable a "virtual joystick", used to fly a high-fidelity aircraft simulator. These models are being developed to augment HMM-based models to improve performance.