MODERN STATISTICAL METHODS: A FIRST INTRODUCTION

HÅVARD RUE AND ODDVAR K. HUSBY
DEPARTMENT OF MATHEMATICAL SCIENCES
NTNU, NORWAY

APRIL 16, 1998

SUMMARY

The statistical community has during the last 10 years experienced a significant transformation stimulated by technological developments in computing, theoretical developments in stochastic modeling and simulation based inference. Armed with these tools, the modern statistician can now easily handle many problems which were practically unsolvable just few years ago. In this introductory paper we discuss three different aspects of these developments, and show how these can be merged into a powerful tool for analyzing statistical models.

ADDRESSES: H. Rue and O. K. Husby, Department of Mathematical Sciences, The Norwegian University for Science and Technology, N-7034 Trondheim, Norway.
E-MAIL: hrue@math.ntnu.no and okh@math.ntnu.no

1 Introduction

The advances made since the late 1980's in the development of computational power, and in the theoretical developments of building statistical models and simulation methods, have had a significant impact on the statistical community. Problems practically unsolvable just ten years ago, are now routinely solved in the first exercises in one of the many *modern computational statistics*-courses given around the world. Armed with these modern tools, the statistician can now handle better than ever the challenges raised by modern science and technology. Despite the challenges in *applying* these modern tools to analyze datasets and model real worlds problems, the basic ideas and principles are easy to understand (as with the rest of *statistics*). In this paper we wish to give the more mathematically oriented reader a basic understanding of these new tools, and make him/her able to explore a part of the field him/her-self through the popular public domain software BUGS.

We do not intend to give a complete review or exposition of all new developments nor to give all the standard references, but try to explain the main ideas and give some basic references to important WWW-pages and introductory books.

We will focus on the following three aspects of modern computational statistics.

Graphical models are stochastic models that are defined via a graph or a network. In most cases, these models have a quite complicated global structure made up of simple local structures/building-blocks. The graph for the model gives both an easy interpretation of the model, e.g. that smoking influences the risk for cancer, and provides a representation for the model well suited to be analyzed using *Markov chain Monte Carlo* methods.

Markov chain Monte Carlo (MCMC) methods make use of the underlying graph structure to make computational efficient simulations from the stochastic model, in most cases conditioned on the observed data. We can then answer questions concerning properties of our model like parameter estimation, uncertainty bounds for the parameters, prediction for future observations and so on. The probability density for unknown parameters or predictions are often estimated using *nonparametric* methods, as the required densities seldom are known to be for example Gaussian or exponential.

Nonparametric methods do not assume any parametric probability density functions or other functional relationship, but merely let the "data speak for them self". This is especially useful when the amount of data is large, a case that in our computerized world occurs more often than before. Nonparametric techniques cover regression, likelihood estimation and probability density estimation among others.

We will restrict our focus on the appropriate subsets of graphical models, MCMC methods, and nonparametric methods, which are needed to construct a complete tool to build and analyze stochastic models. We can then handle and analyze statistical problems ranging from easy to complicated, all solved equally easy in the same framework using the same methodology and often the same computer program!

The plan on this introductory paper is as follows. In Section 2 we introduce, one by one, graphical models, MCMC and nonparametric methods. In Section 3 we merge these techniques into a complete and powerful tool for analyzing stochastic models, and demonstrate its use on a classic problem.

2 Three Aspects of Modern Computational Statistics

In this section we present appropriate subsets of three chosen aspects of modern computational statistics, which are needed to make a complete tool to analyze a huge variety of models/problems all within the same framework and using the same techniques.

We will now define some basic notation. Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ denote continuous (for simplicity) random variables, for which their joint (probability) density function is $f(x_1, \ldots, x_n)$ or $f(\boldsymbol{x})$. Let further $V = \{1, \ldots, n\}$, $\boldsymbol{x}_A = \{x_i : i \in A\}$ where $A \subseteq V$, and $\boldsymbol{x}_{-A} = \{x_i : i \notin A, i \in V\}$. The density for any subset of the $\boldsymbol{x}$'s, $\boldsymbol{x}_A$ say, can be found from $f(\boldsymbol{x})$ by integrating out $\boldsymbol{x}_{-A}$, as

$$f(\boldsymbol{x}_A) = \int_{\boldsymbol{x}_{-A}} f(\boldsymbol{x})\, d\boldsymbol{x}_{-A}.$$

We use $f(\cdot)$ as a generic notation for the (probability) density for those x_i's that appears as the arguments: $f(x_1, x_2)$ is the joint density for (x_1, x_2), $f(x_3, x_4 \mid x_1, x_2)$ is the conditional density for (x_3, x_4) given (x_1, x_2),

$$f(x_3, x_4 \mid x_1, x_2) = \frac{f(x_1, x_2, x_3, x_4)}{f(x_1, x_2)}.$$

2.1 Graphical Models

The graphs enters the scene as a way of expressing the *causal* relation between random variables. Consider Figure 1a where the single-mother x_1 gets two children x_2 and x_3. The children are (thought of as) independent replicates of the mother's genes. Consider the nodes (or vertices) in the graph as the random variables (x_1, x_2, x_3) and the *directed* edges as a way of specifying the *causal* relation or dependency in time between (x_1, x_2, x_3). As the mother is present before the children and the children are independent replicates of the mother, the natural way of partitioning the joint density $f(x_1, x_2, x_3)$, is

$$f(x_1, x_2, x_3) = f(x_2 \mid x_1) f(x_3 \mid x_1) f(x_1).$$

Thus, starting from the graph Figure 1a, we take advantage of the causal relations between the variables to express the partitioning of $f(\boldsymbol{x})$.

The missing father appears in Figure 1b, where the mother x_1 and father x_2 bear a child x_3. The child must appear after its parents, so the graph Figure 1b displays the natural causal structure between (x_1, x_2, x_3), and thus the partitioning of the joint density is

$$f(x_1, x_2, x_3) = f(x_3 \mid x_1, x_2) f(x_1) f(x_2).$$

Let us generalize the situation in Figure 1a-b. We represent the causal relationship between the variables $x_1, \ldots, x_n$ using a directed acyclic graph (DAG): the edges are *directed* and shows the direction of the causal relationship, and the graph is *acyclic*, meaning that no internal cycles in the graph are allowed. Define pa(i) as the parents to node i (i.e. x_i): the set of indices to those variables x_j having a directed edge pointing to node i. We can now interpret the graph as a family tree like the one in Figure 2a, with ancestors, grandparents, parents and children. We have a generalized family-tree as each node are allowed to have one or several parents, a parent and child can have a common child, etc. If we

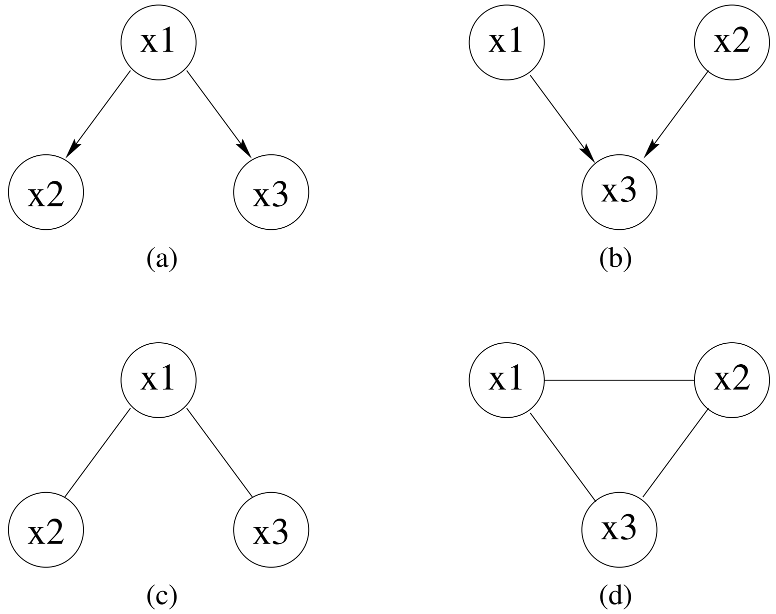


FIGURE 1: Panel (a) and (b) displays graphically the causal structure of (x_1, x_2, x_3) where in (a), $f(x_1, x_2, x_3) = f(x_2 \mid x_1)f(x_3 \mid x_1)f(x_1)$, and in (b), $f(x_1, x_2, x_3) = f(x_3 \mid x_1, x_2)f(x_1)f(x_2)$. Panel (c) and (d) displays the moral-graph of (a) and (b), respectively.

make use of the causal relationship the directed edges between the nodes represent, the joint density $f(\boldsymbol{x})$ can be partitioned as

$$f(\boldsymbol{x}) = \prod_{i \in V} f(x_i | \boldsymbol{x}_{\mathrm{pa}(i)}) \tag{1}$$

where $f(\cdot|\boldsymbol{x}_{\emptyset}) \equiv f(\cdot)$. We can thus build a rather complicated global model $f(\boldsymbol{x})$ for $\boldsymbol{x}$ by using only simple local building blocks, the conditional densities $f(x_i|\boldsymbol{x}_{\mathrm{pa}(i)})$.

We will now discuss the (conditional) dependency structure of $\boldsymbol{x}$ with the causal structure described by a DAG and a density as in (1). To fix the ideas, consider the situation in Figure 1b. Assume you one day see a child with two obvious characteristics C_1 and C_2, what will be your guess about which parent having which characteristic? Due to heritage, either the father has C_1 and the mother C_2, the father has both C_1 and C_2 and the mother none, etc. In statistical words, knowing the child makes our knowledge of the parents dependent. Assume now a situation like in Figure 1a where the mother has characteristics C_1 and C_2, will then our knowledge of the two children become dependent? The answer is in this case no, as the children are independent replicates of the mother. The important statistical concept these examples demonstrate is *conditional independence*: x_2 and x_3 is conditional independent given x_1 iff

$$f(x_2, x_3 \mid x_1) = f(x_2 \mid x_1)\; f(x_3 \mid x_1).$$

Compare this to the well-known definition of independence, x_2 and x_3 is independent iff $f(x_2, x_3) = f(x_2)f(x_3)$. The situation is somewhat more complicated in Figure 2a: which variables does x_i depend on if we condition on all the remaining variables $\boldsymbol{x}_{-i}$? The answer can be derived from the DAG by *moralizing* it: marry the unmarried parents by adding an undirected edge between them and then

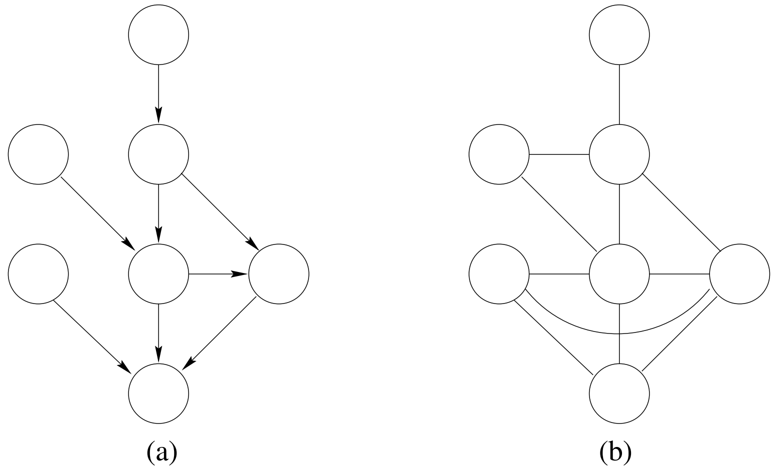


FIGURE 2: Panel (a) displays a DAG for $\boldsymbol{x}$ and panel (b) displays the conditional independence graph of (a) which is found by marring unmarried parents and removing the arrows of edges.

remove the direction on the remaining edges. The resulting graph is called the moral-graph or the conditional independence graph. Figure 1c-d and Figure 2b show the moral-graph for Figure 1a-b and Figure 2a. To answer our question, we need to define the *boundary* of node i, $\text{bd}(i)$: the set of all nodes j having an edge from j to i in the moral-graph. Using (1) it is easy to show that $\boldsymbol{x}$ has a *local Markov property*,

$$f(x_i \mid \boldsymbol{x}_{-i}) = f(x_i \mid \boldsymbol{x}_{\text{bd}(i)}) \propto \prod_{j \in V \,:\, i \in \{j\} \cup \text{pa}(j)} f(x_j | \boldsymbol{x}_{\text{pa}(j)}). \tag{2}$$

The intuitive answers we derived from Figure 1a-b is now immediate clear from Figure 1c-d.

FURTHER READING. The books by Whittaker (1990) and Lauritzen (1996) discuss general graphical models at an intermediate and more advanced level, respectively, and Cox & Wermuth (1996) discuss graphical models with a focus on considerations arising in observational studies in the social and medical sciences including epidemiology. For a commercial successful example of graph-based models, we mention probabilistic (or normative) expert systems. They have become increasingly popular as a way of representing knowledge and to guide decisions through graphs, refer to `http://www.hugin.dk` for online examples and further references.

2.2 MARKOV CHAIN MONTE CARLO METHODS

Markov chain Monte Carlo is an iterative technique for simulating samples from a high-dimensional probability density. To avoid presenting the algorithms within the framework of Markov chains, we will use an alternative view which gives the understanding of the most popular algorithm as successive fixed-point iterations, where the fixed-"point" is the required density $f(\boldsymbol{x})$.

Assume $\boldsymbol{x}$ has density $f(\boldsymbol{x})$, and that we want to generate M samples from $f(\boldsymbol{x})$,

$$\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(M)} \;\sim\; f(\boldsymbol{x}).$$

The samples $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(M)}$ can be used to estimate any property of $\boldsymbol{x}$, e.g.

$$\mathrm{E}(x_i) \approx \frac{1}{M}\sum_{k=1}^{M} x_i^{(k)}, \qquad \mathrm{Prob}(x_i > 0) = \mathrm{E}(1[x_i > 0]) \approx \frac{1}{M}\sum_{k=1}^{M} 1[x_i^{(k)} > 0] \tag{3}$$

where $1[B]$ is one if B is true and zero otherwise, and so on. We can use Monte Carlo methods to replace troublesome high-dimensional integrals by simple averages. Note that the estimates in (3) are unbiased (have the correct expectation) even if the $\boldsymbol{x}^{(k)}$'s are dependent. The advantages of having independent samples from $f(\boldsymbol{x})$, is that we have control over the variance (or error) in the approximations. The error is much harder to control if the $\boldsymbol{x}^{(k)}$'s are dependent.

Sampling from a known univariate density $f(x_1)$ is usually straightforward, but how can we sample from the density $f(\boldsymbol{x})$ when the dimension of $\boldsymbol{x}$ is larger than one? The general solution is to rewrite $f(\boldsymbol{x})$ as

$$f(\boldsymbol{x}) = f(x_1)f(x_2 \mid x_1)f(x_3 \mid x_1, x_2)\cdots$$

and use the algorithm

$$\begin{aligned} x_1' &\sim f(x_1) \\ x_2' &\sim f(x_2 \mid x_1') \\ x_3' &\sim f(x_3 \mid x_1', x_2'), \quad \text{and so on.} \end{aligned}$$

After n-steps $\boldsymbol{x}'$ will be a sample from $f(\boldsymbol{x})$ and we can repeat this algorithm M times to obtain M independent samples. However, for near all interesting cases this approach is practically impossible: densities like $f(x_1)$, $f(x_2 \mid x_1)$ and so on, are very hard to obtain because we have to integrate out the remaining x_j's, and usually n is large! However, conditional densities like $f(x_i \mid \boldsymbol{x}_{-i})$ are usually easy to obtain, which intuitively suggests the following algorithm. Let $\tilde{\boldsymbol{x}}$ be the an reasonable start-value, and sample

$$\begin{aligned} x_1' &\sim f(x_1 \mid \tilde{x}_2, \tilde{x}_3, \tilde{x}_4, \dots, \tilde{x}_n) \\ x_2' &\sim f(x_2 \mid x_1', \tilde{x}_2, \tilde{x}_3, \dots, \tilde{x}_n) \\ x_3' &\sim f(x_3 \mid x_1', x_2', \tilde{x}_3 \dots, \tilde{x}_n) \\ &\vdots \\ x_n' &\sim f(x_n \mid x_1', x_2', x_3' \dots, x_{n-1}'). \end{aligned} \tag{4}$$

Then we use the current value $\boldsymbol{x}'$ as the start-value $\tilde{\boldsymbol{x}}$ for the next iteration and we repeat the n-steps. We iterate (4) until we (think) it has converged, and then *each following iterate* $\boldsymbol{x}'$ will be a sample from the required distribution $f(\boldsymbol{x})$. However, these samples will be dependent on each other. Algorithm (4) is the most popular MCMC-algorithm and is called the *Gibbs-sampler*. Note the similarities with the algorithm in (4) and a fixed-point iteration. The difference is that the fixed-"point" is the required probability density $f(\boldsymbol{x})$ instead of a solution to some equations. The algorithm converges in *distribution* under some conditions to $f(\boldsymbol{x})$.

We can verify that *if* we start with $\tilde{\boldsymbol{x}}$ from the correct density f, then also $\boldsymbol{x}'$ will be a sample from f, i.e. f is the fixed-"point" density for the iteration scheme (4). Assume $\tilde{\boldsymbol{x}}$ has the correct density f and we sample x_1' from $f(x_1 \mid \tilde{\boldsymbol{x}}_{-1})$. The density for $(x_1', \tilde{\boldsymbol{x}}_{-1})$ then becomes $f(x_1' \mid \tilde{\boldsymbol{x}}_{-1})f(\tilde{\boldsymbol{x}}_{-1})$ which is equal to $f(x'_1, \tilde{\boldsymbol{x}}_{-1})$, i.e. the density f is maintained after the first single step in (4), and then also through n-steps.

The Gibbs-sampler algorithm simplify if $f(\boldsymbol{x})$ has the representation (1). The local Markov property says that $f(x_i \mid \boldsymbol{x}_{-i})$ only depends on $\boldsymbol{x}_{\text{bd}(i)}$ and only a few conditional densities appear in the right hand side of (1). If $\boldsymbol{x}$ has a *local structure*, meaning that each node x_i has only a few nodes in its boundary, then it is easy to compute the conditional density and hence easy to sample x'_i from the conditional density given the rest.

FURTHER READING. The book edited by Gilks, Richardson & Spiegelhalter (1996) gives a throughout introduction with relevant references to more general MCMC-techniques, theory, applications and practical aspects. The Web-page `http://www.stats.bris.ac.uk/MCMC/` contains pointers to recent developments in the field.

2.3 NONPARAMETRIC METHODS

Nonparametric techniques in statistics have become increasingly popular during the last years, along with the theoretical advances and the development of practical algorithms. The main idea is to replace parameterized densities and functional relationships, for example Gaussian distributions and linear relationships, with the density or relationship the "data tells us". The data-driven estimates are often obtained using some kind of smoothing-technique. This approach requires relatively large data sets, which occurs more and more often in our computerized world.

We will only make use of a simple nonparametric technique for estimating the underlying density for a sample $x^{(1)}, \dots, x^{(M)}$ from an unknown continuous density g. The classic approach is to make a discrete histogram of the samples. This requires the critical choices of bin limits and the number of bins, and we can obtain histograms with different characteristics, and thus also interpretations, by varying the bin limits and the number of bins. Further, the histogram is not continuous, as the underlying density g is assumed to be. The nonparametric approach to this problem is to make a "continuous histogram", as

$$\widehat{g}(x) = \frac{1}{Nh} \sum_{k=1}^{M} \phi\left(\frac{x - x^{(k)}}{h}\right), \quad \phi(z) = \exp(-z^2/2)/\sqrt{2\pi}. \tag{5}$$

Other choices also exists for the kernel $\phi(z)$. Here, h is a (positive) smoothing-parameter which determines how smooth our estimate should be. There exists a variety of methods to estimate h from the data. The interpretation of (5) is as follows. When constructing a histogram, $x^{(k)}$ is either in a specific bin or not, while in (5) $x^{(k)}$ is a member of a *continuum* of bins centered at x, with weight $\phi((x - x^{(k)})/h)/h$. The weight is maximum at $x^{(k)}$ and decays with increasing normalized distance $|x - x^{(k)}|/h$ from $x^{(k)}$. Thus, h can be interpreted as a scaling-parameter that determines how quickly the contribution from $x^{(k)}$ should decay.

FURTHER READING. The books by Silverman (1986), Green & Silverman (1994) and Wand & Jones (1995), are written on a pleasant level and give a nice introduction to the world of nonparametric statistics.

3 THE COMPLETE TOOL

In this section we will collect the three subjects in Section 2 to construct a complete tool to analyze data and stochastic models. To keep the focus on ideas, we will concentrate on a simple example for which comparable textbook solutions are available. However, the ideas presented in this section are general and apply equally well and easy to simple as well as hard problems. The FURTHER READING section below contains references to more challenging problems and larger case-studies.

3.1 MAKING A COMPLETE TOOL

Assume the data $y_1, \dots, y_N$ are N independent samples from a Gaussian distribution with unknown mean μ and variance σ^2. What are the estimates of μ and σ^2 and what are their densities? The (frequentistic) answers are available in any textbook in statistics as

$$\widehat{\mu} = \frac{1}{N}\sum_{k=1}^{N} y_k, \qquad \widehat{\mu} \sim \mathrm{N}(\mu, \sigma^2/N),$$

and

$$\widehat{\sigma^2} = \frac{1}{N-1}\sum_{k=1}^{N} (y_k - \widehat{\mu})^2, \qquad \frac{(N-1)\widehat{\sigma^2}}{\sigma^2} \sim \chi^2_{N-1}. \tag{6}$$

Here, "N(mean, variance)" denotes the Gaussian distribution with given mean and variance, and χ^2_{N-1} is the chi-squared distribution with $N-1$ degrees of freedom.

An alternative solution to the above classical problem using modern computational methods, proceeds in several steps.

1. First we identify all variables in the problem, this includes both unknown parameters and observed data. We have two unknown parameters, μ and σ^2, and N observed data points, $y_1, \dots, y_N$. This makes a total of $N+2$ variables, which we denote as $\boldsymbol{x} = (x_1, \dots, x_n)$ where $n = N+2$.
2. The next step is to specify the DAG for $\boldsymbol{x}$. Note that y_1 is a child of its parents μ and σ^2, and so are $y_2, \dots, y_N$. Figure 3 displays the DAG for our example when $N = 5$.
3. Further, we need to specify the joint density for all vertices in the graph, $f(\boldsymbol{x})$. As the DAG for our example is available, the partitioning of $f(\boldsymbol{x})$ is given in (1),

$$f(\boldsymbol{x}) = f(\mu, \sigma^2, y_1, \dots, y_N) = \left[\prod_{i=1}^{N} f(y_i \mid \mu, \sigma^2)\right] f(\mu) f(\sigma^2). \tag{7}$$

   Here, $f(y_i \mid \mu, \sigma^2)$ is known to be Gaussian with mean μ and variance σ^2. We need to specify $f(\mu)$ and $f(\sigma^2)$ which is the density for μ and σ^2 *before* we see the data. These densities are called the *a-priori* distributions, and we have silently entered the Bayesian arena. Common (uninformative) choices for $f(\mu)$ is a Gaussian density with mean zero and a large value of the variance, and the standard choice for σ^{-2} is a gamma distribution with small parameters (see Section 3.2). With these choices, our estimates will mainly depend on the contribution from the data.

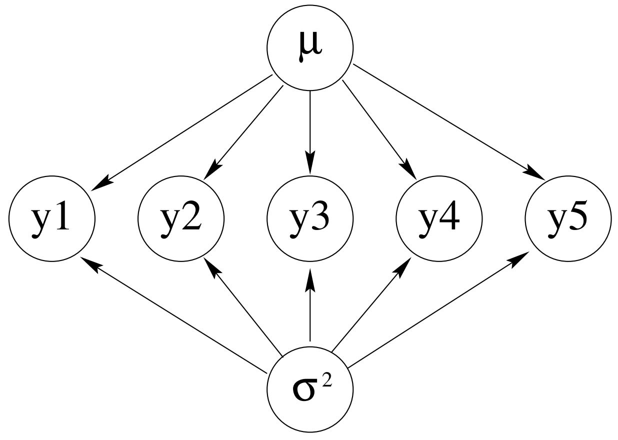


FIGURE 3: (a) The DAG for the classic problem in Section 3 when $N = 5$.

4. How do the observed data enter our model? Recall from (7) that if we know the joint density of $\boldsymbol{x}$ then we also know the conditional density for the unknown parameters conditional on the observed ones,

$$f(\mu, \sigma^2 \mid y_1 \dots, y_N) = \frac{f(\mu, \sigma^2, y_1, \dots, y_N)}{f(y_1, \dots, y_N)} \tag{8}$$

$$\propto f(\mu, \sigma^2, y_1, \dots, y_N). \tag{9}$$

The statement (9) is true, as the denominator in (8) does not depend on (μ, σ^2). By generating samples from the *conditional* density (8), we obtain samples from the joint density of (μ, σ^2) *after* the information in the observed data have been accounted for.

The Gibbs sampler algorithm (4) operates on the full density and not the conditional one, so we need to modify it to generate samples from (8) instead. The modification is in fact trivial: *start the Gibbs sampler iteration with the observed values in the nodes $y_1, \dots, y_N$, and update only those variables that are not observed, while computing the conditional densities using f as before*. A similar proof of the one in Section 2.2 will show (using (9)) that the required conditional density (8) is now the fixed-"point" in the Gibbs-sampler iteration (4). We need the moral-graph found from the DAG to find those variables really needed when calculating the conditional densities in (2).

5. We are now ready (in principle) to run the Gibbs-sampler algorithm and to obtain samples from $f(\mu, \sigma^2 \mid y_1 \dots, y_N)$ after the algorithm has converged. The nonparametric method in Section 2.3 can then be used to estimate the marginal density of μ and σ^2 conditioned on the data. Further, we can estimate the mean, variance, quantiles etc. similar to (6).

3.2 EXAMPLE OF USAGE

To help us with all the practical matter our modern computational approach in Section 3.1 require, we will make use of the public domain software BUGS (see the FURTHER READING section below). The input BUGS needs, is the DAG for the problem and all the conditional densities in (1), the observed data and (optionally) initial values for the unobserved ones.

For our problem, the definition of the DAG in the language of BUGS is as follows.

```
File: classic.bug

     model classic;
     const
        N = 25;
     var
        mu, tau, variance,  y[N];
     data in "classic.dat";
     inits in "classic.init";
     {
         for (i in 1:N) {
             y[i] ~ dnorm(mu, tau);
         }
         mu   ~ dnorm(0.0, 1.0E-4);
         tau  ~ dgamma(1.0E-3,1.0E-3);
         variance <- 1/tau;
     }
```

Here, we recognize all our variables in the `var` statement, further `dgamma` denotes the gamma-distribution and `dnorm` the Gaussian one. The only exception is the `tau`-node; in the Bayesian world the natural specification of variability is the precision, or inverse variance, $\tau = 1/\sigma^2$. The second argument in `dnorm` is the precision. We get around this feature by creating deterministic relation between the node `variance` and `tau` by the "<-" assignment.

We simulated an artificial data-set by sampling $N = 25$ independent samples from a Gaussian distribution with mean zero and variance one. We specify these values in the file `classic.dat` which is read by the BUGS program.

```
list(y=c(0.463,  0.622, -0.293, -0.564, -0.232,  0.809, -0.527,
          0.620, -2.201, -0.826,  0.328,  0.252,  0.349, -0.424,
          0.618, -0.184, -1.248, -1.651, -0.280, -0.065,  0.277,
         -0.106, -0.856,  0.301, -0.653)))
```

These observed values will now be kept fixed during the iterations of (4), and we sample therefore from the conditional density given in (8).

It is optional to specify initial values in the file `classic.init` for the unobserved nodes, which we do for illustrative purposes.

```
list(mu=0,tau=1,seed=1998)
```

The variable `seed` is a special specification to BUGS to set the seed for the internal pseudo-random generator, which makes it possible to repeat our simulation exactly.

We do the analysis by asking BUGS to perform the following commands.

```
compile("classic.bug")
update(0)
monitor(mu)
```

```
monitor(variance)
update(10000)
stats(mu)
stats(variance)
q()
```

The first command compile the specification of the DAG contained in the file `classic.bug`. The second command specifies the number of iterations of (4) needed to reach the fixed-"point". We set for illustrative purposes this value to zero, to see if we can detect the transient-phase before the iteration has converged from the output. We then ask BUGS to monitor what is happening in the nodes `mu` and `variance`, do $10\,000$ more iterations of (4), and then write out the empirical statistics of the two nodes during $10\,000$ iterations. A brief summary of the results we obtain, are

```
              mean      2.5% :  97.5%  CI
mu         -2.174E-1 -5.181E-1   9.469E-2
variance    6.049E-1  3.393E-1   1.080E+0
```

The empirical mean of μ is -0.22 with a 95% interval of $[-0.52, 0.09]$, and the empirical mean of σ^2 is 0.60 with a 95% interval of $[0.34, 1.08]$. We can compare these estimates with standard textbook-answers, and we obtain $\widehat{\mu} = -0.22$ with a 95% interval $[-0.53, 0.09]$ for μ, and $\widehat{\sigma^2} = 0.55$ with a 95% interval $[0.34, 1.07]$ for σ^2. (Note that even if the number of iterations goes to infinity, the values obtained from the BUGS program and the textbook answers will not agree in general, although they will be very similar. This is because the textbook answers are computed under a frequentistic framework, and our new results are computed in the Bayesian framework. The interpretation of the results are also somewhat different.)

Figure 4 shows a more detailed analysis of the values of $\mu^{(k)}$ and $(\sigma^2)^{(k)}$ after each full sweep of (4). The top-left panel, displays a trace-plot of the first 1000 values $\mu^{(1)}, \ldots, \mu^{(1000)}$ and the top-right panel shows the estimated density using (5). The bottom panels shows similar results for $(\sigma^2)^{(1)}, \ldots, (\sigma^2)^{(1000)}$. The Gibbs-sampling algorithm seems to have converged quickly for this problem, as there is no sign of any transient phase before the trace-plots settle down (in distribution).

To demonstrate more power of this modern approach, assume we want to estimate the density of a *new* observation y_{new}. What is the density of the predicted value y_{new}? Such question can be answered by adding the following line to `classic.bug`,

```
ynew ~ dnorm(mu, tau);
```

where `ynew` is an unobserved variable. The extended DAG is displayed in Figure 5a. BUGS will now sample from the density

$$f(\mu, \sigma^2, y_{\text{new}} \mid y_1, \ldots, y_N) \tag{10}$$

from which we in principle can obtain our required prediction density as

$$f(y_{\text{new}} \mid y_1, \ldots, y_N) = \int_\mu \int_{\sigma^2} f(\mu, \sigma^2, y_{\text{new}} \mid y_1, \ldots, y_N)\, d\mu\, d\sigma^2. \tag{11}$$

Using our new tool, the answer is embarrassingly simple. Since we obtain samples from the *joint* conditional density (10), the subset $y_{\text{new}}^{(1)}, \ldots, y_{\text{new}}^{(M)}$ of the samples will be distributed according to

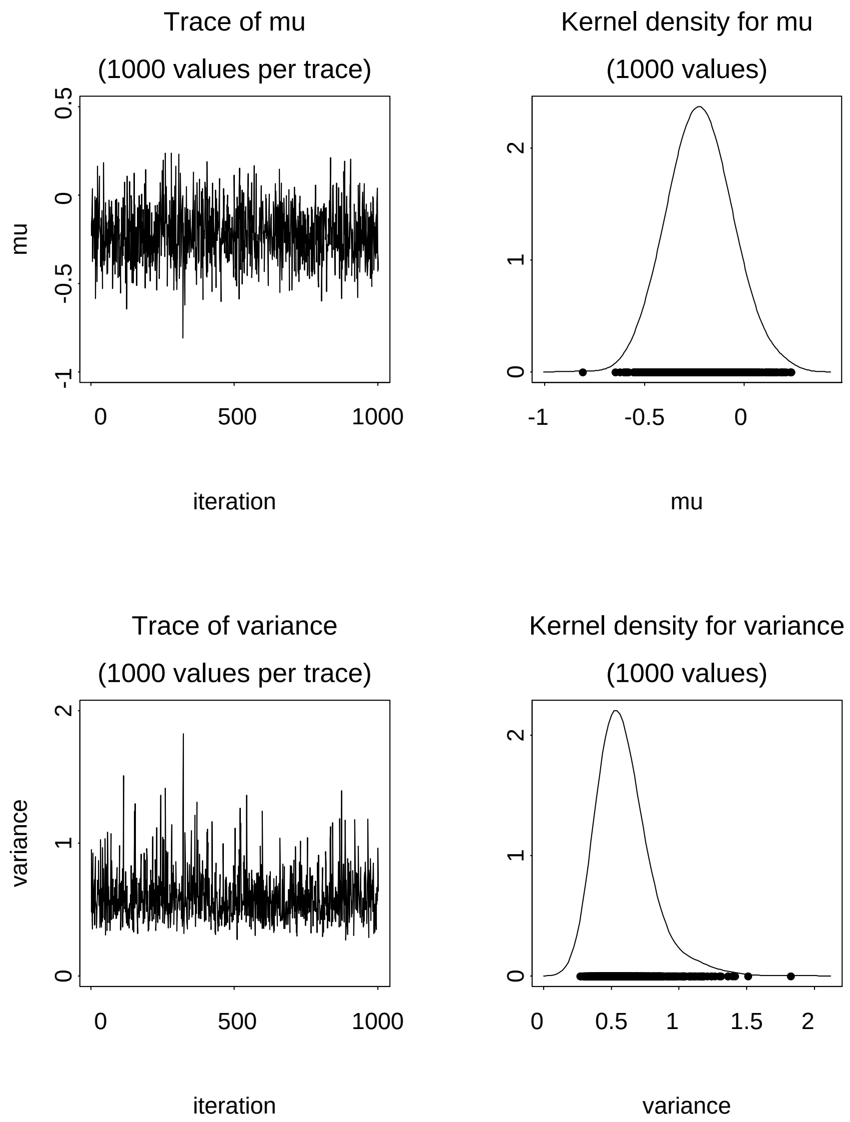
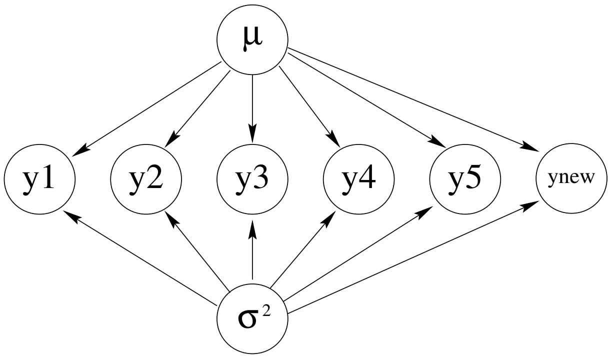
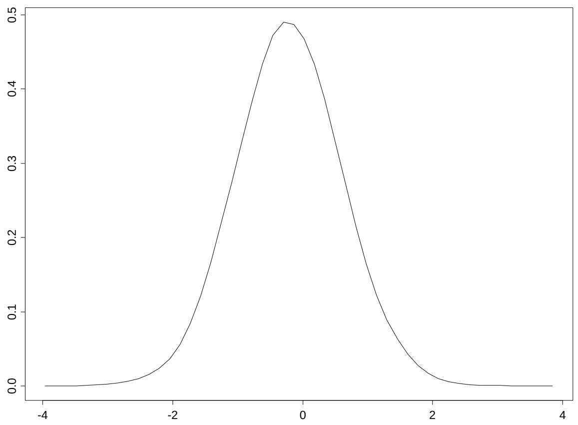


FIGURE 4: The output from the BUGS program applied to our classic problem in Section 3.2. The top panels shows the trace plot and density estimate of the variable μ, and the bottom panels shows the trace plot and density estimate of the variable σ^2, all for the first 1000 iterations of the Gibbs-sampler.



(a)



(b)

FIGURE 5: (a) the DAG for the prediction problem with $N = 5$, (b) the estimated predictive density for a new observation.

the marginal density (11). Thus, including the lines `monitor(ynew)` and `stats(ynew)` in the `classic.cmd` file, give us the answer. We obtain for y_{new}, the mean -0.21 and a 95% predictive interval $[-1.73, 1.36]$, and the density in Figure 5b. Since the y_i's are Gaussian, it is possible to compute predictive densities analytically. However, our modern solution applies equally well and easy to non-Gaussian examples for which analytical solutions is not practically obtainable. A further look at the referred examples and case-studies in the FURTHER READING-section should hopefully convince the reader about the power of modern computational statistics.

FURTHER READING. The best way to learn more about the presented tool, is to visit the www-address `http://www.mrc-bsu.cam.ac.uk/bugs/`. Here you will find pointers to relevant literature, free binary versions of the BUGS program, documentation and a lot of examples of usage. Particular useful are the two large sets of worked-out examples and case-studies using BUGS on *non-trivial* problems. It is fairly easy to follow these examples which demonstrates some of the power these modern techniques can handle.

REFERENCES

COX, D. R. & WERMUTH, N. (1996). *Multivariate dependencies: models, analysis and interpretation*, Monographs on Statistics and Applied Probability, London: Chapman & Hall.

GILKS, W. R., RICHARDSON, S. & SPIEGELHALTER, D. J. (1996). *Markov Chain Monte Carlo in Practice*, London: Chapman & Hall.

GREEN, P. J. & SILVERMAN, B. (1994). *Nonparametric regression and generalized linear models: a roughness penalty approach*, Monographs on Statistics and Applied Probability, London: Chapman & Hall.

LAURITZEN, S. L. (1996). *Graphical models*, Oxford statistical science series, Oxford: Clarendon Press.

SILVERMAN, B. (1986). *Density estimation for statistics and data analysis*, Monographs on Statistics and Applied Probability, London: Chapman & Hall.

WAND, M. P. & JONES, M. C. (1995). *Kernel smoothing*, Monographs on Statistics and Applied Probability, London: Chapman & Hall.

WHITTAKER, J. (1990). *Graphical models in applied multivariate statistics*, Wiley series in probability and mathematical statistics. Probability and mathematical statistics, Chichester: Wiley.