THREE STATISTICAL TECHNOLOGIES WITH HIGH POTENTIAL IN BIOLOGICAL IMAGING AND MODELING

Moshe Fridman* and J. Michael Steele**

Department of Statistics
University of Pennsylvania
Philadelphia PA 19104-6302

ABSTRACT

The three technologies that are surveyed here are (1) wavelet approximations, (2) hidden Markov models, and (3) the *Markov chain Renaissance*. The intention of the article is to provide an introduction to the benefits these technologies offer and to explain as far as possible the sources of their effectiveness. We also hope to suggest some useful relationships between these technologies and issues of importance on the agenda of biological and medical research.

INTRODUCTION

The purpose of this article is to review some of the most significant recent progress in statistical theory and to focus attention to the extent that is possible on the assistance these advances offer to biological and medical technology.

The first topic we engage is the theory and application of *wavelets*, which is possibly the most far-reaching development in all of applied mathematics over the last ten years. The emerging technology has important implications for all domains of signal processing, or wherever one works to reconstruct a sound, an image, or a more elaborate object such as a three-dimensional representation of a human organ. The roots of the theory of wavelets can be traced to sophisticated questions of harmonic analysis, but the explosive development would never have taken place if the basic ideas were not simple, easy to implement on computers, and demonstrably superior to earlier technologies in some important instances. The features of simplicity and broad impact are common to all of material of this review.

The second topic we take up is the technology of *hidden Markov models*. These models offer a natural tool for dealing with one of the fundamental problems in stochastic modeling: many naturally generated stochastic processes exhibit temporal heterogeneity that is driven by an underlying (but unobservable) change in the signal generating system. Because of substantial changes in computational technology, we now find that the range of uses for the methods of the hidden Markov model (HMM) are much more substantial than had previously been imagined.

*Research partially supported by NSF DMS92-11634
**Research partially supported by ARO Grant DAAL03-91-G-0110 and NSF DMS92-11634

Computational Approaches in Molecular Radiation Biology
Edited by M.N. Varma and A. Chatterjee, Plenum Press, New York, 1994

The source of the strength of the HMM seems to be due to its ability to acknowledge the relationships between changing regimes where on a short term basis one could adequately model the observed data by a homogeneous process. A second source of the strength of HMM's are their exceptional ability to incorporate structural features of the phenomena under study into the structural features of the model. Often the topology of the HMM (the number of states, the transition matrix structure, and the observed sequence distribution) is designed to incorporate as many features of the observed sequence as the underlying science can justify. Although such modeling is not *a priori* effective, history has done much to support the practice. The HMM has been applied with telling success in a variety of scientific contexts.

The third topic we explore is actually a broader development that nowadays often goes under the banner of the *Markov Chain Renaissance.* There are two particular subjects in this domain that help focus our review: *simulated annealing* and the *Gibbs sampler.* Each of these topics has been subjected to intensive study over the last ten years. Together with some closely related developments – like the uses of Markov chain simulation methods in the theory of algorithms and the theory of random walks on finite structure like groups – simulated annealing and Gibbs sampling have lead to a rebirth of research interest in discrete time, finite state space Markov chains, whence the notion of a *Renaissance.*

The method of simulated annealing offers an approach to optimization problems that are particularly common in computationally intensive statistical problems such as those provided by image analysis and related inverse problems. Still, the method is exceptionally general and also gives insights into many problems of combinatorial optimization. Similarly, Gibbs sampling is a broadly applicable tool for the analysis and understanding of multivariate distributions that had been viewed as computationally intractable. The most notable successes in the application of the Gibbs sampler have been in making possible a numbers of natural Bayesian procedures, including some that are of importance in imaging.

All of the work review here has the capacity for further theoretical development and more extensive application. The collection of applications of these tools to issues in biological and medical technology is already extensive, but the natural expectation is that we have seen only a small fraction of the important possibilities that lie ahead.

WAVELETS

The purpose of this section is to introduce the central ideas and the main benefits that come from wavelet analysis. Although these notions are *per force* mathematical, our intention is always to keep one eye trained on suggestions for the empirical sciences, especially for those that are dependent on effective imaging and image storage. The development recalled here of wavelet theory substantially follows the notation and conceptualization of Daubechies (1992) and Chui (1992a) both of which offer a down to earth initiation to the practical study of wavelets. The much anticipated volume of Meyer (1993) also can be expected to offer an excellent starting point. The recent survey given in Strang (1993) offers many useful insights on the relationship of wavelets to Fourier and fast Fourier transforms.

The mathematical high road to the study of wavelets if offered by the three volume treatise of Meyer (1990) and Coifman and Meyer (1991). There are also two recent edited volumes by Chui (1992b) and Ruskai, M.B. *et al* (1992) that focus substantially on

applications. In fact, any electronic bibliographic search of the recent scientific literature tends to yield an almost oppressively large number of current contributions. Although for the moment wavelet analysis may seem exotic, the likely expectation is that wavelet application will evolve to become part of the tool kit one is expected to bring to any problem in signal analysis.

Scale and Location

The scientific benefits from wavelet analysis come principally from their ability to help us focus on an object at many different scales of resolution. The implementation of this idea is illustrated most clearly by consideration of the continuous wavelet transform, though eventually one finds that the exciting applications almost all evolve from the discrete cousins.

The continuous wavelet transform begins with a "suitable" function ψ and then creates the scale location family

$$\psi_{ab}(x) = |a|^{-\frac{1}{2}}\psi(\frac{x-b}{a}),$$

where the scaling normalization $|a|^{-1/2}$ has been chosen so that all of ψ_{ab} have the same L^2 norm. The features of ψ that qualify it for suitability are not cut in stone; but the *mother wavelets* ψ that have proved to be most useful are typically smooth, have compact support, and have the property that m of its moments vanish for some $m \geq 1$. A big part of the art of wavelet analysis remains the choice (or construction) of the mother wavelet that is most suitable for the task at hand.

The wavelet transform is given by the mapping $f \mapsto \langle f, \psi_{ab}\rangle$, or more explicitly by

$$T^{WAV}(a,b) = |a|^{-\frac{1}{2}} \int_{-\infty}^{\infty} f(x)\psi(\frac{x-b}{a})dx.$$

Just as one has an inversion formula for the Fourier transform, there is a formula that lets us invert the wavelet transform:

$$f(x) = \frac{1}{C_\psi}\int_{-\infty}^{\infty}\int_{0}^{\infty} T^{WAV}(a,b)\psi_{ab}(x)\frac{dadb}{a^2},$$

and where the normalizing constant C_ψ is given by

$$C_\psi = \int_{-\infty}^{\infty}\frac{1}{|u|}|\hat{\psi}(u)|^2 du < \infty$$

where $\hat{\psi}$ is the Fourier transform of ψ.

Source of Power

Even if one agrees that the source of technological power in the wavelet transform is the fact that the wavelet coefficients $\langle \psi_{ab}, f\rangle$ weigh f over different scales "a" and locations "b", there are still mysteries to be resolved concerning the sources of efficiency of wavelet transformations and representations. Still, three key points emerge:

(1) Wavelets are responsive to the empirical fact that some natural phenomena are related to scale. This fact is observable in the natural occurrence of scaled similarities (as in snowflakes) and also in more diverse phenomena, such as ranges of validity of approximations, averaging methods of differential equations, and renormalization methods of statistical physics.

(2) Wavelets can take advantage of sharper localization and smoothness properties than traditional Fourier (or windowed Fourier) methods. The availability of both scale and location parameters provides an extra degree of freedom that provides some relief from constrictive phenomena of Fourier analysis that say in essence that f and $\hat{f}$ can not both be too concentrated; for example, f and $\hat{f}$ cannot both have compact support.

(3) Wavelets also seem to have aspects of redundancy that provide useful robustness properties. This redundancy turns out to be useful in several respects, and, in particular, it permits substantial data compression for images by permitting the wavelet transform to be stored with limited precision. This fact is one that suggest there may be an important role for HDTV image compression and transmission.

In the next few subsections, we will take a look at some of the mathematics that underlies the technological effectiveness of wavelet representations. Each of these subsections calls of the lectures of Daubechies (1992) for notation, organization, and insight. The first of these details some simple and calculations that show rather generally how approximation properties of some discrete series are usefully abstracted through the language of *frames* that goes back to Duffin and Schaeffer (1952).

Frames and Practical Inversion

Any set $\{\psi_j : j \in J\}$ of elements of a Hilbert space $\mathcal{H}$ is called a *frame* provided there are constants $0 < A, B < \infty$ such that for all $f \in \mathcal{H}$ we have

$$A||f||^2 \leq \sum_{j \in J} |\langle f, \psi_j \rangle|^2 \leq B||f||^2.$$

Certainly, if $\{\psi_j : j \in J\}$ is a complete orthonormal basis for $\mathcal{H}$, then $\{\psi_j : j \in J\}$ is a frame with $A = B = 1$, but the reason for bothering with this notion is that there are frames that exhibit a much different character than bases. One example to keep in mind consists of three vectors in $\mathbb{R}^2$ with 120 degree angles; for such a set $\{\psi_1, \psi_2, \psi_3\}$ one can check that in the required inequality we have $A = B = \frac{3}{2}$. This frame is certainly not a basis, and it also helps us see that when $A = B > 1$ the frame carries a measure of redundancy.

Given any frame, we define a frame operator $F : \mathcal{H} \mapsto \ell^2$ by taking

$$(Ff)_j = c_j \equiv \langle f, \psi_j \rangle,$$

and we note that the adjoint operator $F^* : \ell^2 \mapsto \mathcal{H}$ defined by the relationship $\langle Ff, c \rangle_{\ell^2} = \langle f, F^*c \rangle_{\mathcal{H}}$ can be given explicitly by the formula

$$f^*c = \sum_{j \in J} c_j \psi_j.$$

One can check using the definitions that F^*F is invertible and that we have the bounds $A\,Id \leq F^*F \leq B\,Id$ and $B^{-1}Id \leq (F^*F)^{-1} \leq A^{-1}Id$.

Now, if $\{\psi_j : j \in J\}$ is any frame and F is the associated frame operator, we can define a new frame $\{\tilde{\psi}_j : j \in J\}$ by taking

$$\tilde{\psi}_j = (F^*F)^{-1}\psi_j.$$

This *dual frame* turns out to be the key to a practical inversion formula, since one can check just from the definitions that the frame operator $\tilde{F}$ associated with $\{\tilde{\psi}_j : j \in J\}$ satisfies

$$\tilde{F}F = Id = F\tilde{F},$$

or, in long-hand, we have the Frame Resolution of the Identity:

$$\sum_{j\in J}\langle f, \psi_j\rangle\tilde{\psi}_j = f = \sum_{j\in J}\langle f, \tilde{\psi}_j\rangle\psi_j.$$

This last formula offers many suggestions of how to approximate f, and despite the abstract simplicity of the result, the technological implications are substantial–even in a world where closely related identities have been known for more than a hundred years.

Reconstruction Calculations

The bound $A\,Id \leq F^*F \leq B\,Id$ suggests that one may think of F^*F as a crude approximation of $\frac{1}{2}(A+B)Id$. When we replace $\tilde{\psi}_j = (F^*F)^{-1}\psi_j$ by the approximation $2(A+B)^{-1}\psi_j$ in the first equality in our resolution of the identity equation and work out the remainder term we find the *First Frame Approximation*:

$$f = \frac{2}{A+B}\sum_{j\in J}\langle f, \psi_j\rangle\psi_j + Rf,$$

where we have introduced a remainder term Rf that is explicitly defined by

$$R = Id - \frac{2}{A+B}F^*F.$$

Now, the norm of R is less than 1, so we do have an approximation of some sort, but one needs to judge how good the approximation may be and to explore the ways in which it might be improved. The first step is to consider the operator inequalities

$$-\frac{B-A}{B+A}\,Id \leq R \leq \frac{B-A}{B+A}\,Id,$$

from which a traditional argument gives the norm bounds $||R|| \leq (B-A)/(A+B) = r/(2+r)$ where $r = -1 + B/A$.

The fact that R is linear with norm less than one carries the seed of self-improvement. By the definition of R we have

$$(F^*F)^{-1} = \frac{2}{A+B}(Id - R)^{-1}$$

so we get a series representation for $\tilde{\psi}_j$ by

$$\tilde{\psi}_j = (F^*F)^{-1}\psi_j = \frac{2}{A+B}\sum_{k=0}^{\infty}R^k\psi_j.$$

This suggest how we can do better than just approximate $\tilde{\psi}_j$ by $2\psi_j/(A+B)$; we can take as many terms of the geometric sum as we like. Specifically, we can create an $N+1$ term approximation by taking

$$\tilde{\psi}_j^N = \frac{2}{A+B}\sum_{k=0}^{N} R^k\psi_j = \tilde{\psi}_j - \frac{2}{A+B}\sum_{k=N+1}^{\infty} R^k\psi_j = [Id - R^{N+1}]\tilde{\psi}_j.$$

There is a simple, instructive computation of the error that is made by using $\tilde{\psi}_j^N$ in place of $\tilde{\psi}_j$ in the resolution of the identity. First we note

$$\Delta_N = f - \sum_{j\in J}\langle f,\psi_j\rangle\tilde{\psi}_j^N = \sum_{j\in J}\langle f,\psi_j\rangle(\tilde{\psi}_j - \tilde{\psi}_j^N),$$

so expressing $\tilde{\psi}_j - \tilde{\psi}_j^N$ in terms of R gives

$$\Delta_N = \sum_{j\in J}\langle f,\psi_j\rangle R^{N+1}\tilde{\psi}_j$$

$$= R^{N+1}\sum_{j\in J}\langle f,\psi_j\rangle\tilde{\psi}_j = R^{N+1}f,$$

where the last step invokes the resolution of the identity. Finally, taking norms we see $||\Delta_n|| \leq ||R||^{N+1}||f||$, so using the approximation $\tilde{\psi}_j^N$ in lieu of $\tilde{\psi}_j$ gives us an approximation that conveys geometrically fast.

An important feature of the $\tilde{\psi}_j^N$ is that when we write out $\tilde{\psi}_j^N$ in an expansion in terms of ψ_ℓ we can get a simple recursion for the coefficients. In particular, if we write

$$\tilde{\psi}_j^N = \sum_{\ell\in J}\alpha_{j\ell}^N\psi_\ell$$

we find

$$\alpha_{j\ell}^N = \frac{2}{A+B}\delta_{\ell j} + \alpha_{j\ell}^{N-1} - \frac{2}{A+B}\sum_{m\in J}\alpha_{jm}^{N-1}\langle\psi_m,\psi_\ell\rangle.$$

As Daubechies (1992, page 63) points out, this expression may look "daunting" but in practice many of the $\langle\psi_m,\psi_\ell\rangle$ are negligibly small.

Closing the Loop

Our excursion into the theory of frames serves to give a sense of the organizing role that Hilbert space methods give to the theory of wavelets, but eventually one has to leave the soft analysis of Hilbert space to engage the hard analysis that provides us with mother wavelets ψ for which the doubly indexed set of functions given by

$$\psi_{m,n}(x) = a^{-m/2}\psi(a^{-m}x - nb)$$

provide us with honest frames with useful scientific approximation properties.

The easiest example to write down is the *Sombrero Function* given by the second derivative of the Gaussian hump $e^{x^2/2}$:

$$\psi(x) = \frac{2}{\sqrt{3}}\pi^{-1/4}(1-x^2)e^{-x^2/2}.$$

For $a = 2$ and $b = 1$ this function provides a frame that has $A = 3.223$ and $B = 3.596$.

This particular frame is not as nice as one would hope; specifically, ψ does not have compact support. Still, even this function yields a frame that is quite useful, since the tails of the Gaussian decay so rapidly. The world's catalog of good mother wavelets and associated frames is increasing rapidly, and some useful general principals are starting to emerge about how one "designs" a good wavelet.

Designer Wavelets

Mallat (1989) and Meyer (1990) used the idea of incrementing the information needed to represent a picture at one level of resolution to a more refined level of resolution to articulate a set of concrete mathematical ideas that are evolving as instrumental in the design of wavelet approximations. The emerging theory goes under the name of multiresolution analysis. The abstract set up requires a set of approximation closed subspaces V_j of $L^2(\mathbb{R})$ that satisfy the four nesting and self-similarity conditions:

$$\cdots V_2 \subset V_1 \subset V_0 \subset V_{-1} \subset V_{-2} \subset \cdots,$$

$$\overline{\cup_{j\in\mathbb{Z}} V_j} = L^2(\mathbb{R}),$$

$$\overline{\cap_{j\in\mathbb{Z}} V_j} = \{0\},$$

and finally the condition that adds teeth to this abstract structure by tying the V_j's all together

$$f \in V_j \Longleftrightarrow f(2^j \cdot) \in V_0.$$

The moral principal of multiresolution analysis is that whenever one finds a collection of subspaces satisfying the four preceding conditions, then there is an orthonormal wavelet basis $\psi_{j,k} = 2^{-j/2}\psi(2^{-j}x - k)$ such that the projection operators P_j onto V_j satisfy the key identity

$$P_{j-1} = P_j + \sum_{k\in\mathbb{Z}} \langle f, \psi_{j,k}\rangle \psi_{j,k}.$$

Beyond mere morality, there is a process that is quite often successful in finding a mother wavelet that accommodates the multiresolution. This process is not simple enough to recall here but it is simple enough for one to take seriously in the context of any scientific problem for which there are logical subspaces V_j satisfying the conditions detailed above. This is also the path by which the first basis of compactly supported wavelets was developed in Daubechies (1988), the work that perhaps most pointedly initiated the current flurry of wavelet activity.

Last Wavelet Steps

The preceding review has been reasonably complete in detailing the most basic definitions and properties of wavelet analysis. With luck there is also some suggestion of where one is likely to find effective applications, though the most compelling process is to review those collections of applications that have been edited and the current applications as they appear. Still, there is one more conceptual step that even an introduction like this

should address, and that is wavelet constructions for higher dimensions. The bottom line is that there are at least two trustworthy constructions, one based on tensor products and one based on the use of lattices in $\mathbb{R}^d$ for $d > 1$. For further discussion of these points one should consult Gröchenig (1991), Gröchenig and Madych (1992), and Kovačević, J. and Vetterli, M. (1991),

HIDDEN MARKOV MODELS

We have already described the key qualitative feature of HMM's — they accommodate the relationships between differing regimes of homogeneity in processes that are only locally homogeneous. This feature is evident in almost all of the important HMM applications including (1) speech recognition (cf. the survey and extensive bibliography of Juang and Rabiner (1991)), (2) the study of DNA composition where one uses sequences of bases in a DNA molecule to identify types of molecule segments (Churchill (1989)), (3) hypothesis testing in the study of different haematopoiesis theories, where data on counts of specific bone marrow cells are used to determine the number of unobserved active stem cells from which all blood cells develop, and information on the number of active stem cells is essential for the determination of a correct haematopoiesis model (Guttorp *et al.* (1990)), (4) modeling of ion channels (or large proteins that span cell membranes), where data on electrical single channel currents in neuronal membranes are obtained to study the multiple conductance levels of the ion channel (Chung *et al.* (1990), Fredkin and Rice (1992)), and (5) electrocardiology, where sequences of electrocardiac wave patterns that represent different states of the heart are recognized from EKG data (Coast *et al.* (1991)). The variety of these applications surely suggests that HMM's are a flexible tool, but to understand how HMM's actually serve in these contexts we naturally need to engage the mathematical details of the model. As a first step, we need to lay out some basic definitions.

Model Definition

Let $A = (a_{ij})$ be an $N \times N$ Markov transition matrix, and let $\Pi = (\pi_i), 1 \leq i \leq N$ denote an arbitrary probability distribution. If we view Π as an initial distribution, we can define a discrete-time finite-state Markov chain $\{Q_t\}_{t=1}^{\infty}$ by (Π, A). We further denote the N possible states of the Markov chain by $\{S_1, S_2, \ldots, S_N\}$, and for each $1 \leq i \leq N$, we associate a probability density function $b_i(\cdot)$.

For the triple $(A, B = \{b_i\}_{i=1}^{N}, \Pi)$ we observe a stochastic process $\{O_t\}_{t=1}^{\infty}$. Given that $Q_t = S_i$, the random variable O_t has density $b_i(\cdot)$ and are assumed independent of $\{O_s : s \neq t\}$.

The sequence $\{O_t\}$ is referred to as the *observed sequence or observation sequence*, while the sequence $\{Q_t\}$ is unobserved and referred to as the *hidden, or unobserved sequence*. The probability density function for the sequence $O = (O_1, \ldots, O_T)$ is given by

$$P_{(A,B,\Pi)}(O_1 = o_1, \ldots, O_T = o_T) = \sum_{1 \leq q_1, \ldots, q_T \leq N} P_{(A,B,\Pi)}(O, Q) =$$

$$\sum_{1 \leq q_1, \ldots, q_T \leq N} \pi_{q_1} b_{q_1}(o_1) a_{q_1 q_2} b_{q_2}(o_2) \cdots a_{q_{T-1} q_T} b_{q_T}(o_T).$$

One often has to distinguish the situation where the $b_j(\cdot)$ are probability mass functions, and in such cases we refer to the *discrete value model.* Typically, one assumes that there is a set of parameters of interest that determine the distributions assumed for the hidden and observable processes. Here we will assume that there is an $n \geq 1$ and an open subset Λ of the Euclidean n space such that for each $\lambda \in \Lambda$ we have a one-one correspondence $\lambda \leftrightarrow (A(\lambda), B(\lambda), \Pi(\lambda))$. The set Λ is defined to be the parameter space of the model.

After having defined and parameterized the model, we turn to the three fundamental questions that arise when applying HMM's.

1. Probability **Evaluation** Problem. Given the observation sequence $O = (O_1, O_2, \ldots, O_T)$ and a model $\lambda = (A, B, \Pi)$, how can we compute $P_\lambda(O)$ in the discrete value model? Correspondingly, how do we compute the likelihood function in the continuous value model?
   We have obtained an expression for the likelihood $P_\lambda(O)$, but the expression is only of theoretical value for it contains far too many summands to be evaluated numerically.

2. Parameter **Estimation** Problem. Given the sequence O, how can we estimate the model parameters $\lambda = (A, B, \Pi)$? Moreover, if we proceed by maximum likelihood, how do we calculate λ to maximize $P_\lambda(O)$?

3. State sequence **Identification** Problem. Given the observation sequence O and the parameter set λ, how can we compute a state sequence $q = (q_1, q_2, \ldots, q_T)$ that has maximal conditional probability $P_\lambda(Q \mid O)$?

Maximum likelihood is the main estimation principle that has been used in the solutions to the parameter estimation and the best sequence of states determination problems, but there are other estimation principals that have been successfully used for these problems, such as the *state optimized likelihood criterion* in Juang and Rabiner (1990). Here, the usual likelihood objective function is replaced by the state optimized likelihood function

$$\max_q \sum_q \pi_{q_1} b_1(o_1) \prod_{t=2}^{T} a_{q_{t-1}q_t} b_{q_t}(o_t).$$

However, we shall not elaborate on alternative methods for maximum likelihood estimation.

Except in some trivial cases, one cannot provide a closed-form solution for the maximum likelihood estimates of the parameters or hidden sequence associated with the hidden Markov model. Hence, one has to call on algorithms that were developed to address the three problems mentioned above. Each of these algorithms offers computational challenges that must be met for the method to be effective.

In addition to the computational difficulties in applying HMM's, there exist a number of theoretical and implementational HMM problems. We next list a number of HMM limitations and some suggested solutions.

Identifiability Identifiability has been studied extensively in connection with mixture distributions. A closely related problem is the identifiability of the HMM parameter. Identifiability makes the estimation of λ an unambiguous problem, and it is a

necessary condition for existence of a consistent estimate for λ. One of the subtle features of the HMM is that the parameterization is unidentifiable. The most obvious way to see this is to consider a permutations of the Markov chain state indices, i.e., a relabeling of the states. As a consequence, all the results on estimation of the HMM parameters and their properties, apply only up to equivalence classes of parameters that define the same distribution for the observation sequence. For the discrete value model, Petrie (1969) discusses the identifiability problem, and in the continuous value model there is a detailed discussion in Leroux (1992).

For both models, if A is irreducible and aperiodic, so that it has a unique stationary distribution, and the observed sequence densities are distinct, then the only ambiguity remaining in parameter values originates from the symmetry of the likelihood function with respect to permutations in state labels, but not from any other changes in parameter values.

Local Optimality All the methods for HMM likelihood maximization can only offer a local maximum point of the likelihood. The particular local maximum point obtained will frequently depend upon the location of the initialization used in these iterative methods. For an interesting approach to the maximum likelihood estimation of HMM parameters using the simulated annealing method, see Paul (1985).

Model Dimensionality The number of states N and the number of distinct symbols M for the discrete observation distribution case have to be chosen *a priori* to appropriately specify an HMM. Misspecification of these parameters implies incorrect dimensions of the model's parameter space, λ. This could potentially lead to incorrect solutions of the estimation and state identification problems stated above.

Duration Models One of the inherent limitations of HMM's is the constraint that the Markov model imposes on the amount of time that the unobserved process can stay in a given state. The probability of staying in state S_i for d observations is $p_i(d) = (a_{ii})^{d-1}(1 - a_{ii})$, that is the time to exit state S_i is geometrically distributed with success probability $(1 - a_{ii})$. This geometric state duration distribution is inappropriate for many applications. Several alternatives for implementing different state duration models leading to *hidden semi-Markov models* (HSMM) have been proposed, and a review of the alternative models can be found in Rabiner (1989).

Another model that has been considered by Levinson (1986) in an attempt to alleviate some of these problems, is to use a parametric state duration density instead of the nonparametric $p_i(d)$ used in the HSMM. In particular, the Normal family and the Gamma family have been considered. For these models, reestimation formulas have been derived and successfully used in applications.

A variety of other implementational issues have been discussed in the literature. Some of the most important ones can be found in Rabiner (1989).

HMM Development and Computational Complexity

Early contributions to identifiability and statistical inference problems for functions of finite Markov chains were given in papers by Blackwell and Koopmans (1957), Gilbert

(1959), and Baum and Petrie (1966), but the landmark papers by Baum and Eagon (1967), Petrie (1969), and Baum, Petrie, Soules and Weiss (1970) seem to be the first to introduce solutions to the first two fundamental problems in HMM's mentioned above. These landmark papers contain a set of theorems that lead to iterative reestimation procedure for the maximum likelihood estimation of the HMM parameters.

Although the early work in this field tended to restrict attention to the case where the observed sequence of signals is a sequence of univariate random variables with log-concave densities $b_i(\cdot)$, $1 \leq i \leq N$, the work by Liporace (1982), Juang (1985), and Juang, Levinson and Sondhi (1986) offers extensions to multivariate stochastic observations of Markov chains with densities $b_i(\cdot)$ that are mixtures of log-concave and elliptically symmetric densities. As for the estimation of the hidden state sequence, an algorithm based on dynamic programming methods called the *Viterbi algorithm* was introduced by Viterbi (1967) and has been further studied in Forney (1973).

The computational and analytical complexity inherent in HMM's originates from the complicated nature of its likelihood function. In this section we briefly present three algorithms that offer computationally feasible solutions to the three fundamental questions posed above. We shall only focus on the main ideas of the algorithms. The interested reader can find an excellent exposition of the procedures in Rabiner (1989).

The Evaluation Problem

The evaluation problem is a question in computational efficiency. A naive evaluation of $P_\lambda(O)$ is an infeasible computation that involves N^T possible state sequences. Instead we can invoke a simple but powerful procedure called the *Forward-Backward* procedure, that evaluates $P_\lambda(O)$ in an order of $N^2 \times T$ operations.

The key idea is that instead of proceeding T steps at a time separately with each possible state sequence realization as in the naive evaluation, the Forward-Backward evaluation proceeds one step at a time simultaneously with all possible state sequence realizations. The latter process allows for a recursive calculation of partial sequence probabilities that make this procedure so useful.

Formally, given a model λ and an observed sequence $(o_1, \ldots, o_T)$, we define *forward* and *backward* probabilities as follows:

$$\alpha_t(i) \stackrel{def}{=} P_\lambda(O_1 = o_1, O_2 = o_2, \ldots O_t = o_t, Q_t = S_i),$$

$$\beta_t(i) \stackrel{def}{=} P_\lambda(O_{t+1} = o_{t+1}, O_{t+2} = o_{t+2}, \ldots O_T = o_T | Q_t = S_i).$$

Recursive formulas for the forward and backward probabilities are readily calculated using Bayes' Rule, the Markov property of the state sequence Q, and the conditional independence of $\{O_t, 1 \leq t \leq T\}$ given Q. Namely, we have

$$\alpha_t(i) = \left[\sum_{j=1}^{N} \alpha_{t-1}(j) a_{ji}\right] b_i(o_t), \quad \text{with} \quad \alpha_1(i) = \pi_i b_i(o_1),$$

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij}\, b_j(o_{t+1})\, \beta_{t+1}(j), \quad \text{with} \quad \beta_T(i) = 1,$$

where $2 \leq t \leq T$, $1 \leq i \leq N$.

Finally, we obtain the following formulae,

$$P_\lambda(O) = \sum_{i=1}^{N} \alpha_T(i),$$

$$P_\lambda(O, Q_t = S_i) = \alpha_t(i)\beta_t(i), \quad \text{for any } 1 \leq t \leq T,$$

$$P_\lambda(O, Q_t = S_i, Q_{t+1} = S_j) = \alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j).$$

The first formula can serve as a tool for likelihood based model comparisons. Under a given model λ, the last two formulae provide a method to obtain *a posteriori* estimates

$$\hat{\pi}_i = \frac{\alpha_1(i)\beta_1(i)}{\sum_{j=1}^{N} \alpha_T(j)},$$

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} \alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{t=1}^{T-1} \alpha_t(i)\beta_t(i)}.$$

for the initial and transition probabilities. Also, note that we can find the most likely state at time t by $\hat{q}_t = \arg\max_i P_\lambda(Q_t = S_i|O)$. The sequence $(\hat{q}_1, \hat{q}_2, \ldots, \hat{q}_T)$ is often called the *Maximal Aposteriori Probability*, or MAP, estimate of $(q_1, q_2, \ldots, q_T)$

The Estimation Problem

The most widely used optimization technique for the maximum likelihood estimation of the HMM parameters is known as the *Baum-Welch* algorithm. The key observation, noted originally by Baum and Eagon (1967) in the case of a discrete value model, is that there is a transformation $\tau : \Lambda \to \Lambda$ of the parameter space such that the transformed parameter $\tau(\lambda)$ is guaranteed to increase the likelihood $L(\cdot)$. There exist several ways to arrive at the form of the transformation τ, such as standard constrained optimization techniques or the *a posteriori* approach illustrated above, but the one that proved to be the most useful is based on an auxiliary function that is closely related to the Kullback-Leibler number introduced in Kullback and Leibler (1951). Baum *et al.* (1970) derive the explicit form of the transformation τ for the continuous value model with log-concave density functions $b_i(\cdot)$, $1 \leq i \leq N$, and prove that fixed point solutions of τ are locally optimal points of the likelihood function.

Formally, define the auxiliary function

$$Q(\lambda, \lambda') = \sum_{q} P_\lambda(O, Q = q) \ln P_{\lambda'}(O, Q = q),$$

where the summation is over all feasible paths q through the state product space. Let

$$\tau : \lambda \to \hat{\lambda} \stackrel{def}{=} \arg\max_{\lambda'} Q(\lambda, \lambda').$$

The Baum-Welch algorithm begins with a feasible initial estimate of the parameter values $\lambda = \lambda_0$, to which the transformation τ is applied to obtain a new estimator $\hat{\lambda}$. The process is iterated by replacing the old values in λ by the newly obtained values $\hat{\lambda}$, until a fixed point of τ is approximated. For this procedure we have the following important result established in Baum *et al.* (1970).

Theorem *Under the above assumption on the densities $b_j(\cdot)$, we have for all $\lambda \in \Lambda$ that $L(\tau(\lambda)) \geq L(\lambda)$. Moreover, equality can hold if and only if λ is a critical point of L, or equivalently, λ is a fixed point of τ.*

The significance of this result can be brought out in several ways:

1. The explicit form of the transformation τ is given by a set of so called *reestimation formulas* that express the new value for each parameter as a function of its old value. These formulas are obtained by differentiating $Q(\lambda, \lambda')$ with respect to each one of the primed parameters and equating the derivatives to zero. Part of the usefulness of the Baum-Welch approach is the form of the auxiliary function that greatly facilitates the manipulation of the primed parameters. also, note that only first derivatives are required by the Baum-Welch algorithm.

2. The reestimation formulas involve probability expressions of the sort handled by the Forward-Backward procedure. Hence, to reduce the computations to a feasible order, one usually invokes the Forward-Backward procedure within each iteration of the Baum-Welch algorithm.

The Identification Problem

We are often interested in uncovering the true state sequence that led to a given observation sequence O. Although the probability measure $P_\lambda(O)$ does not explicitly involve a specific state sequence realization it can often provide useful insight into the structure of the mechanism that generates the observations.

When the HMM has state transitions with zero probabilities and we choose to maximize separate state subsequences, the optimal state sequence resulting from this process may not even be a valid state sequence. The Viterbi algorithm overcomes this problem by maximizing over the entire state sequence. Prior to discussing the algorithm, we describe a trellis structure that captures the features of the problem.

Consider an $N \times T$ vertex trellis structure whose vertices are arranged in T columns corresponding to time slots 1 through T and N rows representing the N states of the Markov chain. Directed edges connect between all possible pairs of edges with positive transition probabilities. Clearly, this construction has the property that for every possible state sequence there is a unique corresponding path through the trellis and vice versa. For a given model λ, we attach weights to the edges and initial vertices in a way that our problem becomes to find the longest path through the trellis. Denote the jth vertex in the tth time slot by S_j^t. The edge of the trellis connecting vertex i at time slot $t-1$ with vertex j at time slot t have weight $a_{ij}b_j(o_t)$, $1 \leq i, j \leq N$, $2 \leq t \leq T$ and the initial weight for vertex S_i^1 is $\pi_i b_i(o_1)$, $1 \leq i \leq N$. We will calculate the weights of partial state sequences, as we move along the time slots, by multiplying the weights along the edges

of the path. For a particular complete path q, the weights product along the path's edges results in

$$\pi_{q_1} b_{q_1}(o_1) \prod_{t=2}^{T} a_{q_{t-1}q_t} b_{q_t}(o_t) = P_\lambda(Q = q, O).$$

Let $q(S_j^t)$ denote any path segment from time 1 to t, ending at state S_j, $1 \leq j \leq N$. Let $\hat{q}(S_j^t)$ denote the longest such path segments, also called the *survivors*. Assume for simplicity that $\hat{q}(S_j^t)$ is uniquely defined for any (j,t), or else choose one such path arbitrarily. Then for any time $t > 1$ there are N survivors in all, one for each possible terminating state (vertex) of the partial path.

The main observation is that the longest complete path must begin with one of these survivors. If it did not, we could have found a longer path segment from time 1 to t which would be a contradiction. Thus, at any time t, we need to remember only the N survivors $\hat{q}(S_j^t)$, $1 \leq j \leq N$ and their lengths. To get from time t to $t+1$, we need only extend all time-t survivors by one time unit. This is done by selecting for each time-$(t+1)$ state S_k^{t+1}, $1 \leq k \leq N$, the time-t survivor that is longest when extended . The length of the new survivors is recalculated by multiplying the length of its last edge times the total length of the corresponding old survivors. The algorithm proceeds indefinitely, advancing one time unit at a time, without the number of survivors ever exceeding N. As was the case for the Forward-Backward procedure, computations are on the order of $N^2 \times T$ operations.

Recent Advances in HMM's

The popularity of HMM's continues to grow rapidly both in applied and theoretical work. On the theoretical side an important part of the research focuses on inferential properties of likelihood methods. Leroux (1992) established under mild conditions the consistency of the maximum likelihood estimate for the continuous value model, and thus complemented the pioneering work by Baum and Petrie (1966) and Petrie (1969) where the consistency and asymptotic normality of the maximum likelihood estimates had been established under the discrete value model. Another important step was taken in Bickel and Ritov (1993) where the log-likelihood for continuous value HMM's is shown to obey the local asymptotic normality conditions of LeCam as a consequence of which asymptotically efficient analogs of the maximum likelihood estimates can be constructed and the information bound that gives their asymptotic variance can be estimated.

Aggoun and Elliot (1992) consider the case of a continuous time Markov chain observed in Gaussian noise. Finite dimensional normalized and unnormalized predictors are obtained for the state of the chain, for the number of jumps from one state to another, and for the occupation time in any state.

In the hidden sequence estimation area, some simplified estimation procedures for both the filtration and interpolation problems in a two-state HMM are proposed and analyzed in Khasminskii, Lazareva and Stapleton (1993). These estimates have been designed to perform well for the case where $a_{01} = \epsilon\lambda, a_{10} = \epsilon\mu; \epsilon \to 0$ and $0 \leq \lambda, \mu \leq 1$, thus creating a similarity to the change point detection problem. Kogan (1991) gives conditions under which the estimated chain, as given by the Viterbi algorithm, has a lower recognition error rate than the alternative MAP estimated chain, for the two-state HMM case with symmetric transition matrix.

An extension of the basic HMM paradigm to the regression setting is proposed by Fridman (1993). The *hidden Markov model regression* offers a way to extend the benefits of HMM's to problems that are naturally studied through regression analysis. In general terms, it is assumed in HMM regression that the regression parameter values depend on the Hidden Markov chain state. As a result, given that the state at time t is S_i, we have that $Y_t = X_t'\beta_i + \sigma_i\epsilon_t$, where the error terms ϵ_t are i.i.d. $N(0,1)$. There is a connection to the *switching regression model* introduced in Quandt (1972) and Quandt and Ramsey (1978) is a special case of HMM regression for a Markov transition matrix with the property $a_{ij} = a_j$.

MARKOV CHAIN RENAISSANCE

The subject of discrete time discrete space Markov chains have receive greatly increased attention over the last five years because of several developments in the theory of algorithms. One of these developments concerns simulated annealing, which we introduce in the next section. Two other potentiating developments were the Gibbs sampler and the invention of Markov chain methods for making uniform selections from large discrete sets, like the set of all matchings in a graph. Only the first of these is engaged in this survey.

An Algorithm for All Problems

A bewildering variety of substantial computational problems can be cast in to the framework of determining the minimum of a real-valued function on a finite set, $f : S \to \mathbb{R}$. For example, if $\{x_1, x_2, ..., x_n\}$ denotes a set of n points in the plane and S denotes the set of permutations of the n-set $\{1, 2, ..., n\}$, then by taking

$$f(\sigma) = \sum |x_{\sigma(k)} - x_{\sigma(k+1)}|$$

we see that the determination of an element of the set of minimizers of f,

$$S^* = \{s^* \in S : f(s^*) = \min\{f(s) : s \in S\},$$

is the same as solving the famous *traveling salesman problem*, or *TSP* . Naturally, it is no surprise that the *TSP* is only one of a hord of problems that can be put into the specified form —the form is so general as to seem to impose virtually no restriction. Rather, the surprise comes from the fact that we can still say something useful, even at so great a level of generality.

Naturally, one has to cut some slack. The *TSP*, like many of the interesting problems of combinatorial optimization, can be regarded as computationally intractable. It is an element of the class of *NP-Complete* problems, and, as a consequence, it is extremely unlikely that one can ever obtain a good algorithm for the *TSP*, if one regards a good algorithm as one that can provide an optimal solution in an amount of time that grows only polynomially in the size of the input. Because of this natural barrier of intractability, many problems like the *TSP* have been studied in the context of approximate solutions.

It is remarkable that one can give anything like a general recipe for making progress toward the general problem of determining an element of S^*. We can in fact provide a

sequence of such recipes. We begin with one which is not quite practical, but it still offers some genuine insight, and —even in its naive form –it is cleverer and more practical than the often useless idea of exhaustive search.

A skeletal version of the recipe is simple. First, we introduce a probability measure μ_t on S by defining

$$\mu_t(s) = exp(-tf(s))/Z(t)$$

where $Z(t)$ is chosen just in order to normalize everything properly, i.e.

$$Z(t) = \sum_{s \in S} exp(-tf(s)).$$

Second, we select a "large" t. Third, and finally, we just choose an element of S at random according to the measure μ_t.

This is a phony recipe in a couple of ways; but it has something to it, and its faults can be substantially ameliorated. But even before seeing how it can be made honest, we should get an idea why it might work. The essential point is that as $t \to \infty$ the measure μ_t becomes increasingly concentrated on the states for which s is small. Formally, it is not hard to prove that

$$\lim_{t \to \infty} \mu_t(S^*) = 1$$

so ultimately $\mu(s)$ is concentrated on the best choices for s. This limit result would have to be supplemented by more precise continuity statements if we were to try use it to justify our recipe along the lines presented thus far, but a wiser course —and one the development of the theory has actually taken— is to improve our recipe at least a couple more times before aiming at the convergence theorems that offer an honest justification.

An Honest Version

So, how do we honestly pick an element of S according to the measure μ_t? The denominator $Z(t)$ in the definition of μ_t can have many billions of summands in even an "easy" problem, so naive methods of drawing a sample according to μ_t are meaningless. Luckily, there is a brilliant trick due to Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (1953) that provides the key. As an incomplete historical note, it is interesting to record that the fifth of these authors is the Edward Teller who is an acknowledged father of the hydrogen bomb.

The essence of the trick is that one can make progress by viewing μ_t as the stationary distribution of a Markov chain with state space S. For this to buy us anything, we need to be able to simulate the steps of the chain, but we have lots of room to maneuver in this aim, because, after all, we get to invent the chain.

To restrict our search for a good Markov chain to particularly tractable situations, we will impose a graph structure on S, and we will only consider Markov chains were one always moves to a neighbor in the graph (or else stays still, as is sometimes useful in a Markov chain to guarantee aperiodicity). In many applications, the set S has a graph structure is naturally at hand, but in any event the restriction is modest since all it really means is that for each $s \in S$ we have a set $N(s) \subset S - \{s\}$ that we call the neighbors of s.

As we start our hunt for chains that have μ_t as their stationary measure, we might also pay particular attention to those that are reversible, since the stationarity equations then can be replaced by the much simpler *total balance* equation. As a first step, we may as well also restrict our attention to chains for which the graph on S is *regular*, i.e. $|N(s)| = N$, some constant, for all $s \in S$. By making sure that the cardinality N is feasibly small, we will be able to simulate the behavior of any Markov chain that only makes transitions to neighboring states.

To make matters precise, we need transition matrix p_{ij}, where $p_{ij} = 0$ except for $j \in N(i)$ and for which μ_t is the stationary measure. Hunting, as we are, in the context of reversible chains, we want our transition probabilities to satisfy the *total balance* condition:

$$\mu_t(i)p_{ij} = \mu_t(j)p_{ji}$$

Determination of a suitable p_{ij} is now pretty easy, but instead of continuing with more heuristic groping, it seems better to look at one good answer and its associated story.

Before writing down the transition probabilities, it is useful to indicate first how to simulate the behavior of the required Markov chain $\{X_n, n \geq 0\}$. The story goes like this: (1) if the current state X_n is equal to i, first choose a neighbor j of i at random, (2) if the value of f at j improves on the value of f at i, then move to the chosen state i, but (3) if the value of f would not be improved by the move (so that $f(j) \geq f(i)$), then make the move with probability $\exp(-t(f(j) - f(i)))$ while otherwise staying at the state i. Formally, for $i \neq j$ we have

$$P(X_{n+1} = j \mid X_n = i) = \begin{cases} 1/N & \text{if } f(j) < f(i) \\ (1/N)\exp(-t(f(j) - f(i)) & \text{otherwise} \end{cases}$$

The transition probabilities in case $i = j$ are just those needed to pick up the leftovers:

$$P(X_{n+1} = i \mid X_n = i) = 1 - \sum_{j:j \neq i} P(X_{n+1} = j \mid X_n = i).$$

A useful point to check at this juncture, is that a chain with the transition function given above does satisfy the total balance condition, so to make a choice from S according to the measure μ_t all we need to do is start at an arbitrary point $X_0 = s$ and run the Markov chain for a "long time" after which X_n will be a realization from a distribution that is approximately equal to μ_t.

To see how simulation of X_n can be practical even when direct simulation of μ_t is not, just reconsider the *TSP*. In an n-city problem the cardinality of S is $(n-1)!$, the number of cyclic permutations, but we can introduce a natural graph on these permutations where the degree of each vertex is bounded by n. The graph is defined by saying that two permutations are considered adjacent if we can go from one to the other by a an "interchange operation" given by picking two non-intersecting edges of the cycle, breaking the cycle at each of these edges, and building a new cycle by reconnecting the two disconnected components of the cycle in the opposite way from their initial connection.

At this point we have in hand a method for "solving" all problems in combinatorial optimization—but, of course, it solves some problems better than others. After the next section, we will review the performance of the method, but even as it sits, it has some victories. One of these that is entertaining to code is the famous "Eight Queens" problem: How can one place eight queens on a chessboard in such a way that no pair of queens are

mutually attacking? This problem is one that is often assigned in programming courses where backtracking algorithms are studied, but it is also a nice one to study with our naive sampling method. Once one chooses an appropriate reduction of the problem (say, to "rook good" configurations of queens) it is not hard to find an appropriate graph on the set of configurations (pick a pair of queens and switch the two in the only way that preserves "rook-goodness").

A More Honest Version

The only scurrilous part of our recipe that remains is that of our choice of t and our choice of how long to run the Markov chain. To many people it is unsatisfying to say that the choices are simply up to them, and luckily the search for a more adaptive procedure turns out to be a source of insight, and there is a fortuitous charm in combining the two problems into one.

Specifically, we consider a sequence of values t_n such that $t_n \to \infty$ as $n \to \infty$, and now let $\{X_n : n \geq 1\}$ evolve as before except on the nth step we use t_n in place of t. Letting the t_n grow to infinity, provides us a way to combine the issues of picking large t and a suitably large n. The issue now is to obtain the conditions on the sequence $\{t_n : n \geq 1\}$ that suffice to provide us some proper assurance that the method is effective.

A central result in this direction is due to Hajek (1988). In addition to answering the basic question as asked, it also provides us with some special insight. A key notion in Hajek's theorem is that of a *height*. Specifically, we say a state s *communicates with* S^* *at height* h if $h(s)$ is the smallest number such that there is a path from s to some state of $t \in S^*$ for which each state u on the path satisfies

$$f(u) \leq f(s) + h(s).$$

Theorem If h^* is the largest of the set of heights $\{h(s) : s \in S\}$, then we have

$$\lim_{n \to \infty} P(X_n \in S^*) = 1,$$

if and only if

$$\sum_{n=1}^{\infty} \exp(-t_n h^*) = \infty.$$

One consequence of this result is that the choice $t_n = log(n)/h$ is sufficient to provide convergence if and only if h satisfies $h \geq h^*$. As a practical matter one seldom knows h^*, but still taking $t_n = log(n)/h$ for a speculatively chosen h is a common occurrence, or at least it it is common among those bother with changing t as n evolves. Many people have found that in their application the change provided by $t = log(n)/h$ is too slow to be worth the trouble and hence they fall back on the naive process of just picking a "big" t. This approach makes one miss out on a lot of engaging theory; but, as a matter of practice, it can work out well, and picking t directly is not substantially more *ad hoc* than picking h.

Origins and Aspects of Metaphor

There are three ideas that were stirred together to make the theory just described. The first idea is that we might get a small value of $f(s)$ if we pick a point at random

according to μ_t. The second idea— an old but wonderful one—is that a simple Markov chain can be use to help us make the selection process practical. The third idea is that we can link the processes of letting X_n "run for a long time" and of "picking a big t" by letting X_n evolve as a time inhomogeneous.

The final algorithm that combines all of these ideas goes under the engaging name of *Simulated Annealing*. It was introduce independently by Kirkpatrick, Gelett and Vecchi (1983) and Cerny (1985), and it was first described in the context of a physical analogy that has continued to exert considerable influence. It is traditional nowadays in the discussion of simulated annealing to a call on the language of statistical mechanics: (1) μ_t is called the Gibbs distribution, (2) $Z(t)$, the partition function, (3) $T = K/t$, temperature, and (4) the sequence $T_n = K/t_n$, the cooling schedule. There is a certain beauty in this way of describing the optimization process, and the metaphor brings with it a rich collection of intuitions and experience from statistical mechanics. Still, there is some merit in taking a bare-bones look without introducing language which - though apt and properly evocative - can make simple ideas look more mystical than they otherwise might. A second benefit of the less fortified description is that it offers us a different set of opportunities future for development than those offered by the physical insights. Finally, the usual presentation does not pick apart the three steps as thoroughly as we have done here, so the bare-bones route also suggests opportunities for innovation that precede even the proto-simulated annealing algorithm.

How Well Does One Do?

The jury is out on many aspects of the simulated annealing algorithm, and it is unlikely that a definitive understanding of its merits will come about anytime soon. One difficult issue is that there is no *single* simulated annealing algorithm. One has a whole family of algorithms for every problem one might study because of the need to choose (1) the graph structure to be used (2) one candidate from among many f that would yield the given S^* of interest, and (3) some specific values for $\{t_n\}$, or the equivalent.

It seems almost impossible to make all of these selections in a way that would lead to a definitive evaluation, but some reliable experience has evolved. In particular, the papers of Johnson, Aragon, McGeoch, and Schevon (1990,1991,1992) provide many useful comparisons.

Metropolis-Hastings Algorithm

The method sketched above for generation of an observation from the Gibbs distribution has an extension that has the benefits of being less *ad hoc* , more general, and providing a connection to the final topic of this review: the Gibbs sampler. The extension is due to Hastings (1970) and it tells us how an arbitrary Markov chain can be modified to provide one with a specified stationary distribution. Quite simply, to provide a Markov chain $\{X_n : 1 \leq n < \infty\}$ that has stationary distribution π on the state space $\mathcal{S}$, the Metropolis-Hasting Algorithm begins with an arbitrary transition function $q(x, y)$ and

makes modifications. In particular, a new transition function defined by

$$p(x,y) = \begin{cases} q(x,y)\alpha(x,y) & \text{if } x \neq y \\ 1 - \sum_z q(x,z)\alpha(x,z) & \text{if } x = y \end{cases}$$

where α is defined by

$$\alpha(x,y) = \begin{cases} \min\{\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\} & \text{if } \pi(x)q(x,y) > 0 \\ 1 & \text{if } \pi(x)q(x,y) = 0. \end{cases}$$

The Gibbs Sampler

The Gibbs sampler shares with simulated annealing the feature of being a Monte Carlo integration method that proceeds by a Markovian updating scheme. The product that is delivered by the Gibbs sampler is an observation from a multivariate distribution, and the raw material that is required for the algorithms is a collection of conditional distributions from which one can easily draw observations.

Suppose that we have a collection of k real, possibly vector-valued, random variables $U_1, \ldots, U_k$ whose full conditional densities, denoted by $f_{U_i}(\cdot|U_j; j \neq i), 1 \leq i \leq k$, have a simple known form. By that we mean that the full conditional densities are available for sampling, given values of the appropriate conditioning random variable.

Our interest is to simulate an observation from the joint density $f_{(U_1,\ldots,U_k)}$ with the eventual intention of gaining insight into the joint density or some other quantity that can be estimated using such observations, such as an estimate of the marginals f_{U_i}.

The idea is to generate a sample of k-tuples from the joint density using only the available full conditional densities. By simulating a large enough sample, any population characteristic can be approximated to a desired degree of accuracy. Before we formally describe the method, we introduce some simplifying notation from Gelfand and Smith (1990). Densities are denoted generally by brackets, so joint, conditional, and marginal forms appear as $[X,Y], [X|Y]$, and $[X]$ correspondingly. Multiplication of densities is denoted by $*$ and marginalization by forms such as $[X|Y] = \int_{Z,W}[X|Y,Z,W] * [Z|W,Y] * [W|Y]$. We assume that the joint density exists and is strictly positive over the product sample space. Besag (1974) shows that this condition ensures that knowledge of the full conditional densities uniquely defines the full joint density.

The Gibbs sampler algorithm generates an approximation to an observation from $f_{(U_1,\ldots,U_k)}$ by iterations of a k-step process, each complete pass of which corresponds to a step of a Markov chain. Specifically, given an arbitrary set of values $(U_1^{(0)}, U_2^{(0)}, \ldots, U_k^{(0)})$ we draw

$$\begin{aligned} U_1^{(1)} &\sim [U_1|U_2^{(0)}, U_3^{(0)}, \ldots, U_k^{(0)}] \\ U_2^{(1)} &\sim [U_2|U_1^{(1)}, U_3^{(0)}, \ldots, U_k^{(0)}] \\ &\vdots \\ U_k^{(1)} &\sim [U_k|U_1^{(1)}, U_2^{(1)}, \ldots, U_{k-1}^{(1)}]. \end{aligned}$$

Upon completion of this first iteration of the algorithm, we obtain the vector $U^{(1)} = (U_1^{(1)}, U_2^{(1)}, \ldots, U_k^{(1)})$. Next, we generate $U^{(2)}$ using conditioning values taken from $U^{(1)}$. After i such iterations we arrive at $(U_1^{(i)}, U_2^{(i)}, \ldots, U_k^{(i)})$.

The great usefulness of the algorithm comes from the following theoretical results, established by Geman and Geman (1984). Under mild conditions,

$$(U_1^{(i)}, U_2^{(i)}, \ldots, U_k^{(i)}) \xrightarrow{d} [U_1, \ldots, U_k]$$

as $i \to \infty$, and hence for each $1 \leq j \leq k$, $U_j^{(i)} \xrightarrow{d} [U_j]$. The rate of convergence here (in the sup norm) is geometric in i. Furthermore, for any measurable function G of $U_1, \ldots, U_k$ whose expectation exists, an ergodic theorem holds, namely

$$\lim_i i^{-1} \sum_{l=1}^{i} G(U_1^{(l)}, U_2^{(l)}, \ldots, U_k^{(l)}) \xrightarrow{a.s.} E(G(U_1, \ldots, U_k)).$$

As a result, Gibbs sampling through m replications of the aformentioned i iterations produces m i.i.d. k-tuples of the form $(U_{1r}^{(i)}, \ldots, U_{kr}^{(i)})$, $1 \leq r \leq m$, with the desired joint density. Gelfand and Smith (1990) recommend a density estimate for $[U_j]$, $1 \leq j \leq k$ having the form,

$$[\hat{U}_j]_i = m^{-1} \sum_{r=1}^{m} [U_j | U_t = U_{tr}^{(i)}; t \neq j].$$

The the Gibbs sample can be viewed from many directions, but one fruitful perspective is to consider it as an adaptation of the Metropolis-Hastings algorithm (Metropolis *et al.* (1953), and Hastings (1970)). The Gibbs sample seems to have been first formally developed by Geman and Geman (1984) in the context of image reconstruction, though as with most good ideas the roots of the Gibbs sampler can be traced to many suggestive sources.

In the statistical framework, Tanner and Wong (1987) used a similar technique in their substitution sampling approach to missing data problems. Gelfand and Smith (1990) showed the applicability of the Gibbs sampler to general parametric Bayesian computations and a variety of other conventional statistical problems. The Gibbs sampler approach, along with several other computer-intensive statistical methods, are reshaping many traditional methods in statistics.

Further reviews of uses of Monte Carlo Markov Chain (MCMC) methods for Bayesian computations and inference can be found in Besag and Green (1993), and in Smith and Roberts (1993). Tierney (1991) gives an outline of MCMC methods for exploring posterior distributions. Geyer (1991) explores the use of MCMC in likelihood based inference. A Gibbs sampler approach to generalized linear models with random effects is given in Zeger and Karim (1991). Hierarchical Bayesian analysis of changepoint problems are approached using the Gibbs sampler in Carlin, Gelfand and Smith (1992). The simulation tool of Gibbs sampling is employed in Albert and Chib (1993) to generate marginal posterior distributions for all parameters of interest in an AR model subject to Markov mean and variance shifts that is closely related to HMM's. This method is expedient since the conditional posterior distributions of the states given the parameters, and the parameters given the states, all have form amenable to Monte Carlo sampling. A variety of applications of the Gibbs sampler in Medicine are reviewed in Gilks *et al.* (1993).

An important question that has been recently given a considerable amount of attention in the literature is what is the "best" way to extract information from a Gibbs sampler sequence? More specifically, the two issues at steak are,

Convergence What is a "long enough" run of a Gibbs sampler?

Sampling How can we design an efficient Gibbs sequence, or sequences, sampling strategy?

As was the case in choosing a cooling scheme for the simulated annealing procedure, different ways of extracting information from a Gibbs sequence have been suggested, and it seem unlikely that there exist one optimal solution to this problem. An excellent bibliography on the Gibbs sampler, and the convergence rate and output sampling problem can be found in the special discussion paper of Smith and Roberts (1993).

CONCLUDING REMARKS

We have reviewed the theory and applications of wavelets, of hidden Markov models, of simulated annealing, and of the Gibbs sampler. This is almost a litany of the major steps in statistical science over the last ten years, so perhaps one has done about as well as possible just to get a taste of the possibilities for applications. Still, because the ideas behind these developments are fundamentally simple, perhaps also just enough detail has been given so that the central mathematical facts might be honestly understood.

The work reviewed here is far from done, and the best is surely years ahead of us. Pointers have been given throughout the review to many articles and books that develop our topics with much greater detail, though the most compelling work is not to be found in the books but rather in the marriage of the most basic parts of these technologies to problems of importance in science.

REFERENCES

Wavelets

Chui, C.K. (1992a), *An Introduction to Wavelets.* Academic Press, New York.

Chui, C.K.(1992b), *ed.*, *Wavelets: A Tutorial in Theory and Applications.* Academic Press, New York.

Coifman, R.R. and Meyer, Y. (1991), "Remarques sur l'analyse de Fourier á fenêtre," *C.R. Acad. Sci. Paris Sér. I*, 259-261.

Daubechies, I. (1988), "Orthonormal bases of compactly supported wavelets," *Comm. Pure Appl. Math.*, **41**, 909-996.

Daubechies, I., (1992) *Ten Lectures on Wavelets*, SIAM Publications, Philadelphia PA.

Duffin, R.J. and Schaeffer, A.C. (1952), "A class of nonharmonic Fourier series," *Trans. Amer. Math. Soc.*, **72**, 341-366.

Gröchenig, G.K. (1991), "Describing functions: atomic decompositions versus frames," *Monatsh. Math.*, **112**, 1-42.

Gröchenig, K. and Madych, W.R. (1992), "Multiresolution analysis, Haar bases and self-similar tilings of $I\!R^n$," *IEEE Trans. Inform. Theory*, **38**, 556-568.

Kovačević, J. and Vetterli, M. (1991), "Nonseparable multidimensional perfect reconstruction filter banks and wavelet bases for $I\!R^n$," *IEEE Trans. Inform. Theory*, **38**, 533-535.

Mallat, S. (1989), "Multiresolution approximation and wavelets," *Trans. Amer. Math. Soc.*, **315**, 69-88.

Meyer, Y. (1990), *Ondelettes ét opéateurs, I: Ondelettes, II: Opéateurs de Calderón-Zygmund, III: Opéateurs multilinéaires.* Hermann, Paris.

Meyer, Y. (1993), *Wavelets: Algorithms and Applications.* SIAM Publications, Philadelphia, PA.

Strang, G. (1993), " Wavelet Transforms versus Fourier Transforms", *Bulletin of the American Mathematical Society* **28**,2, 288-305.

Ruskai, M.B., Beylkin, G., Coifman, R.R., Daubechies, I., Mallat, S. Meyer, Y. and Raphael, L. (1992), *eds.*, *Wavelets and their Applications.* Jones and Bartlett, Boston.

HMM

Aggoun, L. and Elliot, R.J. (1992), "Finite dimensional predictors for hidden Markov chains," *System Control Lett.*, **19**, 335-340.

Baum, L.E. and Eagon, J.A. (1967), "An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model of Ecology," *Bull. Amer. Math. Soc.*, **73**, 360-363.

Baum, L.E. and Petrie, T. (1966), "Statistical inference for probabilistic functions of finite state Markov chains," *Ann. Math. Statist.*, **37**, 1554-1563.

Baum, L.E., Petrie, T., Soules, G. and Weiss, N. (1970), "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Statist.*, **41**, 164-171.

Bickel, P.J. and Ritov, Y. (1993), "Inference in hidden Markov models I Local asymptotic normality in the stationary case," Technical Report, Department of Statistics, University of California, Berkeley.

Blackwell, D. and Koopmans, L. (1957), "On the identifiability problem for functions of finite Markov chains," *Ann. Math. Statist.*, **28**, 1011-1015.

Chung, S.H., Moore, J.B., Xia, L., Premkumar, L.S. and Gages, P.W. (1990), "Characterization of single channel currents using digital signal processing techniques based on hidden Markov models," *Phil. Trans. Roy. Soc. Lond. Ser. B*, **329**, 265-285.

Churchill, G.A. (1989), "Stochastic models for heterogeneous DNA sequences," *Bull. Math. Biol.*, **51**, 79-94.

Coast, D.A., Cano, G.G. and Briller, S.A. (1991), "Use of hidden Markov models for Electrocardiographic signal analysis," *J. of Electrocardiology*, **23**, 184-191.

Forney, G.D. (1973), "The Viterbi algorithm," *Proc. IEEE*, **61**, 268-278.

Fredkin, D.R. and Rice, J.A. (1992), "Bayesian restoration of single channel patch clamp recordings," *Biometrics*, **48**, 427-448.

Fridman, M. (1993), "Hidden Markov model regression," *Unpublished Thesis*, University of Pennsylvania.

Gilbert, E.J. (1959), "On the identifiability problem for functions of finite Markov chains," *Ann. Math. Statist.*, **30**, 688-697.

Guttorp, P., Newton, M.A. and Abkowitz, J.L. (1990), "A stochastic model for haematopoiesi in cats," *IMA J. Math. Appl. Med. Biol.*, **7**, 125-143.

Juang, B.H. (1985), "Maximum Likelihood estimation for mixture multivariate stochastic observations of Markov chains," *AT & T Tech. J.*, **64**, 1235-1249.

Juang, B.H., Levinson, S.E. and Sondhi, M.M. (1986), "Maximum likelihood estimation for multivariate mixture observations of Markov chains," *IEEE Trans. Inform. Theory*, **IT-32**, 307-309.

Juang, B.H. and Rabiner, L.R. (1990), "The Segmental K-Means algorithm for estimating parameters of hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Process*, **ASSP-38**, 1639-1641.

Juang, B.H. and Rabiner, L.R. (1991), "Hidden Markov models for speech recognition," *Technometrics*, **33**, 251-272.

Khasminskii, R.Z., Lazareva, B.V. and Stapleton, J. (1993), "Some procedures for state estimation of a hidden Markov chain with two states," Technical Report, Department of Statistics and Probability, Michigan State University.

Kogan, J.A. (1991), "Optional reconstruction of Markov sequences through indirect observations," *6th USSR-Japan Symp. on Probab. and Math. Statist.*, Kiev.

Kullback, S. and Leibler, R.A. (1951), "On information and sufficiency," *Ann. Math. Statist.*, **22**, 79-86.

Leroux, B.G. (1992), "Maximum-likelihood estimation for hidden Markov models," *Stochastic Process. Appl.*, **40**, 127-143.

Levinson, S.E. (1986), "Continuously variable duration hidden Markov models for automatic speech recognition," *Computer, Speech and Language*, **1**, 29-45.

Liporace, L.A. (1982), "Maximum Likelihood estimation for multivariate observations of Markov sources," *IEEE Trans. Inform. Theory*, **IT-28**, 729-734.

Paul, D.B. (1985), "Training of HMM recognizers by simulated annealing," *Proc. ICASSP*, New York: IEEE, 13-16.

Petrie, T. (1969), "Probabilistic functions of finite state Markov chains," *Ann. Math. Statist.*, **40**, 97-115.

Quandt, R.E. (1972), "A new approach to estimating switching regressions," *J. Amer. Statist. Assoc.*, **67**, 306-310.

Quandt, R.E. and Ramsey, J.B. (1978), "Estimating mixtures of Normal distributions and switching regressions," *J. Amer. Statist. Assoc.*, **73**, 730-738.

Rabiner, L.R. (1989), "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, **77**, 257-285.

Viterbi, A.D. (1967), "Error bounds for convolutional codes and an asymptotically optimal decoding algorithm," *IEEE Trans. Inform. Theory*, **IT-13**, 260-269.

Markov Renaissance

Albert, J.H. and Chib, S. (1993), "Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts," *J. Bus. & Econ. Statist.*, **11**, 1-16.

Besag, J. (1974), "Spatial interaction and the statistical analysis of lattice systems (with discussion)," *J. Roy. Statist. Soc. Ser. B*, **36**, 192-236.

Besag, J. and Green, P.J. (1993), "Spatial statistics and Bayesian computations," *J. Roy. Statist. Soc. Ser. B*, **55**, 25-37.

Cerny, V. (1985), "A thermodynamic approach to the traveling salesman problem: An efficient simulation", *J. Optim. Theory Appl.*, **45**, 41-51.

Gelfand, A.E. and Smith, A.F.M. (1990), "Sampling-based approaches to calculating marginal densities," *J. Amer. Statist. Assoc.*, **85**, 398-409.

Geman, S. and Geman, D. (1984), "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattn Anal. Mach. Intell.*, **6**, 721-741.

Geyer, C.J. (1991), "Markov chain Monte Carlo maximum likelihood," *Computer Science and Statistics: Proc. 23rd Symp. Interface*, Fairfax Station: Interface Foundation, 156-163.

Gilks, W.R., Clayton, D.G., Spiegelhalter, D.J., Best, N.G., McNeil, A.J., Sharples, L.D. and Kirby, A.J. (1993), "Modelling complexity: Applications of Gibbs sampling in medicine," *J. Roy. Statist. Soc. Ser. B*, **55**, 39-52.

Johnson, D.S., Aragon, C., McGeoch, L. and Schevon, C. (1990), "Optimization by simulated annealing: An experimental evaluation, Part I: Graph partitioning," *Oper. Res.*, **37**, 865-892.

Johnson, D.S., Aragon, C., McGeoch, L. and Schevon, C. (1991), "Optimization by simulated annealing: An experimental evaluation, Part II: Graph coloring and number partitioning," *Oper. Res.*, **39**, 378-406.

Johnson, D.S., Aragon, C., McGeoch, L. and Schevon, C. (1992), "Optimization by simulated annealing: An experimental evaluation, Part III: The traveling salesman problem," in preparation.

Hajek, B. (1988), "Cooling schedules for optimal annealing," *Math. Oper. Res.*, **13**, 311-329.

Hastings, W.K. (1970), "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, **57**, 97-109.

Kirkpatrick, S., Gelett, C.D. and Vecchi, M.P (1983), "Optimization by simulated annealing," *Science*, **220**, 621-630.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953), "Equation of state calculation by fast computing machines," *J. Chem. Phys.*, **21**, 1087-1092.

Smith, A.F.M. and Roberts, G.O. (1993), "Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussion)," *J. Roy. Statist. Soc. Ser. B*, **55**, 3-23.

Tanner, M.A and Wong, W. (1987), "The calculation of posterior distributions by data augmentation (with discussion)," *J. Amer. Statist. Assoc.*, **82**, 528-550.

Tierney, L. (1991), "Exploring posterior distributions using Markov chains," *Computer Science and Statistics: Proc. 23rd Symp. Interface*, Fairfax Station: Interface Foundation, 563-570.

Zeger, S. and Rizaul Karim, M. (1991), Generalized linear models with random effects: A Gibbs sampling approach," *J. Amer. Statist. Assoc.,* **86**, 79-86.