The Annals of Applied Probability
1995, Vol. 5, No. 3, 577–602

HIDDEN MARKOV RANDOM FIELDS[1]

BY HANS KÜNSCH, STUART GEMAN AND ATHANASIOS KEHAGIAS

ETH Zentrum, Brown University and Brown University

A noninvertible function of a first-order Markov process or of a nearest-neighbor Markov random field is called a hidden Markov model. Hidden Markov models are generally not Markovian. In fact, they may have complex and long range interactions, which is largely the reason for their utility. Applications include signal and image processing, speech recognition and biological modeling. We show that hidden Markov models are dense among essentially all finite-state discrete-time stationary processes and finite-state lattice-based stationary random fields. This leads to a nearly universal parameterization of stationary processes and stationary random fields, and to a consistent nonparametric estimator. We show the results of attempts to fit simple speech and texture patterns.

1. Introduction. If $X = X_1, X_2, \ldots$ is a Markov process and $Y = Y_1, Y_2, \ldots$ is a deterministic or stochastic function of X, then Y is called a *hidden Markov model* (HMM), or sometimes a *hidden Markov process*. Usually, the dependency of Y_t on X is more-or-less local, as when $Y_t = f(X_t)$ for some function f or $Y_t = g(X_t, X_{t+1}, \eta_t)$ for some function g and an iid process $\{\eta_t\}$, independent of X. In any case, Y itself is generally not Markov, and may in fact have a complicated dependency structure. Nevertheless, the *conditional* distribution of X given Y may remain simple, as in the above two examples where X given Y is still first-order Markov. The combination of a rich marginal structure for Y and a simple posterior structure for X makes hidden Markov processes a common modeling tool.

EXAMPLE 1. *Filtering* (cf. [34]). Although the general (nonlinear) filter problem falls within this framework, let us specialize to the linear case: X (known as the *state* process) is not only Markov, but satisfies a simple linear (stochastic) difference equation

$$X_{t+1} = aX_t + \omega_t,$$

where $\{\omega_t\}$ is iid. The *observation* process Y is a HMM, linearly related to X, as in

$$Y_t = bX_t + \omega_t',$$

Received December 1993; revised November 1994.
[1] Supported by Army Research Office Contract DAAL03-92-G-0115 to the Center for Intelligent Control Systems, NSF Grant DMS-88-13699 and Office of Naval Research Contract N00014-91-J-1021.
AMS 1991 *subject classifications.* Primary 60G60; secondary 62M05.
Key words and phrases. Hidden Markov models, Markov random fields, speech models, textures.

where $\{\omega_t'\}$ is another iid noise process, independent of $\{\omega_t\}$. The object is to estimate the state X_t from the observations $\{Y_s\}$, $s \in [0, T]$. This is termed *smoothing* if $0 \leq t < T$, *filtering* if $t = T$ and *prediction* if $t > T$. In any case, the fact that X given Y is still Markov is central to obtaining practical estimation formulas. Beyond this, linearity is exploited to derive efficient recursive estimators (e.g., the Kalman filter) for a host of "on-line" applications in tracking and control.

EXAMPLE 2. *Speech recognition* (see, e.g., [1] and [43]). Here X is a Markov chain with finite (but very large) state space. In principle, the state of X_t represents all of the information relevant to predicting utterances of a speaker at times $\tau > t$. In practice, this information is modeled by representing, jointly, the word (and, sometimes, word pair), phoneme and part of the phoneme (e.g., beginning, middle or end) being articulated at time t. The transition matrix for X is built hierarchically, by successively modeling the variations in pronunciation of parts of phonemes, phonemes and words, as well as (some of) the constraints and regularities in word sequences (syntax). Observations are of the acoustic signal, or some transformation or simplification, and are represented by Y. A stochastic model for Y_t given X_t is developed (or estimated more-or-less nonparametrically). The result is a HMM for the observable acoustic signal (or its transformation) Y, and the object is to estimate X (especially the word sequence) given Y. The posterior is Markov, which is fortunate since this simple dependency structure admits dynamic-programming-like computational tools for the calculation (or at least approximation) of an optimal estimator for X, as well as for computing expectations of various sufficient statistics involved in the estimation of the model parameters. This HMM setup, or some of its variations, is the basis for the most successful speech recognition systems.

EXAMPLE 3. *Ion channel kinetics* (see [3], [2], [24] and [36]). Nerve cells can propagate electrical activity without attenuation over long distances. Lossless conduction involves an active process of opening and closing selective membrane ion channels, and thereby exchanging selected ions between inter- and intracellular spaces. Experiments can be devised to measure the changing conductance of one or a small number of channels in response to various chemical or electrical stimuli. These experiments reveal that ion channels typically move through only a few effective states, being, for example, simple "open" or "closed" with essentially no intermediate levels of conductance. The actual molecular basis for these measurable states is more complicated and is often modeled as a Markov process with multiple states. The observable conductance is then a function of this process, through which, for example, certain of the molecular states manifest themselves as an open channel and others as a closed channel. Thus the observable conductance is a HMM. Purported mechanisms for channel kinetics can be tested by using observed channel conductances to infer the structure and transition probabilities of the (hidden) molecular Markov process. In these applications, the time parameter is generally continuous.

EXAMPLE 4. *Amino acid sequence analysis.* Hundreds or thousands of amino acids strung linearly together constitute a protein. Typically, there are only 20 distinct types of amino acids found, but there are of course a very large number of possible *sequences*. The particular sequence of amino acids that constitutes a protein is known as its "primary" structure. The determination of primary structure is known as sequencing, a process that has been increasingly automated; the result is a large existing data bank of primary structures. The *function* of a protein is largely determined by the folded three-dimensional (or "tertiary") structure that the amino acid chain assumes in vivo. Tertiary structure can sometimes be determined by experimental and imaging techniques, but the process is laborious and the number of sequenced proteins far exceeds the number of proteins with known tertiary structure. Hence, a fundamental problem in biology is the prediction of tertiary structure from primary structure.

One general approach is to search through sequences with known tertiary structures in order to find a "good match" to a sequence with unknown tertiary structure. Similar sequences tend to have similar structure, and in fact there are broad categories of structure that most proteins (or portions thereof) fall into. In an effort to exploit these structural categories, Krogh, Brown, Mian, Sjölander and Haussler [38] built probabilistic models for amino acid sequences conditional on structural classes. These models are built up from known structure-sequence pairs, and then are used to infer a likely structural class for a novel amino acid sequence. Thus, for example, a stochastic model is built for the sequence of amino acids constituting a typical globin (protein that transports oxygen and carbon dioxide). A new amino acid sequence can be evaluated under the globin model to determine its fit, and thereby to predict whether or not it will exhibit a globin-like tertiary structure. Preliminary tests have been highly successful.

The actual models constructed by Krogh, Brown, Mian, Sjölander and Haussler are HMM's with the amino acids constituting the observables and a Markov process, with carefully constructed state space and restricted transitions, constituting the hidden process. (A very similar approach is taken by Churchill [16] in constructing HMM's for the sequence of bases constituting a DNA molecule.) Transition probabilities are estimated from existing data bases, as are state-dependent distributions on the 20 available amino acids. Here again the conditional Markov structure of the unobserved (in fact, virtual) process is heavily exploited to develop computationally feasible estimation and inference algorithms (involving various dynamic-programming-like procedures).

EXAMPLE 5. *Texture models.* This is just a proposal, but it serves to introduce a generalization that will be a primary focus of our theoretical development. Consider a digitized image of a textured pattern such as cloth, wood or sand. The image can be thought of as a realization of a stochastic process $\{Y_t\}$, $t \in \Lambda = \{(i, j)\colon 1 \le i \le N, 1 \le j \le M\}$, where $N = M = 512$, for example, and Y_t is the grey level observed at picture element (or pixel) t. Many authors (e.g., [19], [33] and [21]) have proposed modeling $\{Y_t\}$, condi-

tioned on the texture type and the imaging parameters (distance to camera, orientation, discretization, etc.), as a Markov random field. Since there is usually an organization to the texture that essentially rules out nearest-neighbor models, this approach demands that one either pick, more-or-less arbitrarily, a neighborhood structure or attempt to estimate the neighborhood structure. In either case, there is then the requirement of choosing (or estimating) parameters that determine the associated clique functionals.

A different approach to obtaining the necessary structure would be to employ a *hidden* Markov random field, using a simple nearest-neighbor process for the underlying Markov structure. Thus $Y_t = f(X_t)$, f a fixed "hiding function," where X_t is a nearest-neighbor Markov random field. As in the one-dimensional examples discussed previously, Y will not generally be Markov, although the conditional distribution on X, given Y, is still a nearest-neighbor Markov random field. Is it possible to introduce sufficiently rich structure into the Y process to capture the regularity/variability of real textures through this mechanism? We will return to this shortly.

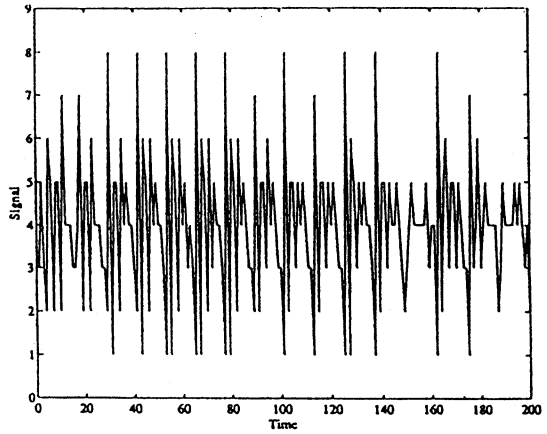
The last example, especially, raises the issue of generality: How general is the class of processes that can be well approximated by a hidden Markov model? To be concrete, we shall restrict ourselves to nearest-neighbor processes (which is to say, first-order Markov when working in one dimension) and we will only allow instantaneous and deterministic "hiding" functions: $Y_t = f(X_t)$. [In one dimension, many variations are popular: Y_t might depend, randomly or deterministically, on X_t or, simultaneously, on X_t and X_{t-1}. Restricting to *finite* state spaces, it is not difficult to show that these four classes are *equivalent*, in the sense that the set of achievable distributions, for the observable process Y, is identical in each case (see [5] and [35]). One constructs an explicit distribution-preserving transformation from a HMM of one type to a HMM of another type.] Furthermore, X_t (and hence also Y_t) will always have finite state space. So, for example, consider a stationary process $Z_t \in \{0, 1\}$, $t = 1, 2, \ldots$, which we shall try to model (or "fit") with a HMM of the form $Y_t = f(X_t)$, where X_t is first-order Markov, $X_t \in \{0, 1, \ldots, N\}$, f: $\{0, 1, \ldots, N\} \to \{0, 1\}$. By varying N, f and the transition probability matrix for X, how close can we get (how similar to Z can we make Y)?

The answer depends very much on the measure of similarity. Ornstein and Weiss [40], for example, study related questions under a strong notion of similarity: Given two discrete-state stationary processes Y and Z, $d(Y, Z) \leq \varepsilon$ if there exists a stationary process $\Psi = \{\Psi_t\} = \{(Y'_t, Z'_t)\}$ such that:

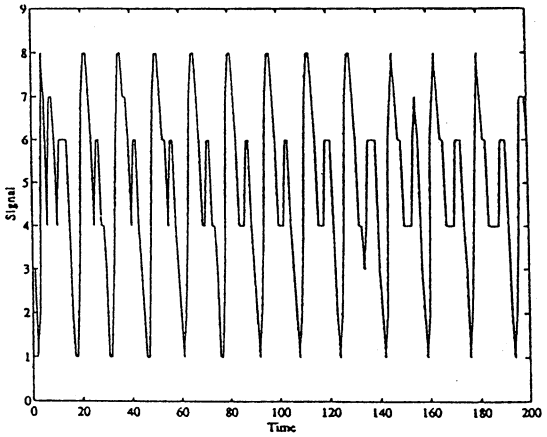
1. Y' and Z' have the same distributions as Y and Z, respectively.
2. $P(Y'_1 \neq Z'_1) \leq \varepsilon$.

The Ornstein–Weiss *distance*, d, between Y and Z is the *infimum* over all such ε. The results of Ornstein and Weiss indicate that the class of Z which can be arbitrarily well approximated by HMM's Y, relative to d, is highly restricted.

3.2.1. *Speech waveforms.* Segments of two phonemes were extracted from a single utterance of the word "one"; see Figure 1. The acoustic signal was sampled at 10 kHz (one sample every 0.1 ms) and 4096 amplitude levels, although each amplitude was later rounded to one of eight equally spaced values, and the samples themselves were subsampled to one observation at every 0.5 ms. Panel (a) shows a 100-ms segment from the phoneme /a/, which follows the initial /u/ and precedes the final /n/ in the pronunciation of "one." There are 200 data points, each having one of eight values. Panel (b) shows an analogous segment from the final /n/ of the same utterance. The nearly periodic waveforms are characteristic of so-called voiced phonemes, and derive ultimately from more-or-less periodic oscillations of the vocal chords.



(a)



(b)

FIG. 1. (a) 100-*millisecond segment from the phoneme* /a/. (b) 100-*ms segment from the phoneme* /n/.

We treated each signal as a sample $z_1, z_2, \ldots, z_{200}$, from an eight-valued stationary process, which we attempted to fit with a series of hidden Markov models of increasing size. In each case, we employed the "hiding function" $f(x) = 1 + x \bmod 8$, and computed approximate maximum likelihood $N \times N$ transition probability matrices, for $N = 10, 20, 30, 40, 50$ and 60. Maximum likelihood computations were made via the Baum reestimation formula [4, 6], which is an instance of the EM procedure [18]. Estimates were only approximately maximum likelihood since this is an iterative hill-climbing algorithm; it can approach a local maximum and, as a practical matter, it must be terminated short of convergence. We began each run (one run for each value of N) with a randomly generated transition probability matrix and continued until there were only negligible changes in the transition matrix.

The results are most easily judged by viewing samples from the resulting HMM's. Figures 2 (for the /a/ sequence) and 3 (for the /n/ sequence) show random samples from $Y_t = f(X_t^N)$, where $\{X_t^N\}_{t=1}^{200}$ is first-order Markov on $\{0, 1, \ldots, N-1\}$ with the estimated $N \times N$ transition probability matrix, and $X_1^N = 1$ ($N = 10, 20, 30, 40, 50$ and 60). In both sets of experiments, one gets the impression that the fit generally improves with increasing N, although there is the suggestion of some deterioration at $N = 60$. Since there are only 200 (highly correlated) samples, it may be that, at $N = 60$, the familiar problem of over-fitting has been encountered. There may, as well, be computational problems with the iteration procedure, perhaps related to local maxima. In any case, it would be interesting to perform similar experiments with larger data sets; essentially infinite amounts of data are easily available.

It may also be interesting to splice together such signals, as a novel approach to speech synthesis. In this regard, one would need to fit, as well, the nonstationary speech units associated with various consonants. Because we are after a signal of only finite duration, it is not impossible, and perhaps not unreasonable to speculate, that exactly the same models would be effective for fitting consonants.

An obvious alternative approach would be to fit each signal with an Nth-order Markov process. However, even at the modest eight-level discretization used in our experiments, this would involve estimating $7 \cdot 8^N$ parameters, which evidently places a severe restriction on the process order. It may be true, in contrast, that the hidden process provides an efficient coding of the nearly periodic structure by dedicating single or multiple states to positions within the cycle, although we have performed no systematic experiments to test this conjecture.

3.2.2. *Binary textures.* The experiments with two-dimensional processes were more difficult and less successful. We adopted the modest goal of fitting some simple binary textures. These were derived from real textures, borrowed from the well-used Brodatz collection [14], by simply thresholding grey-level pictures. A suitable threshold produces substantial islands of "ones" positioned among a sea of "zeros." The shape and pattern of the islands, of course, depends upon the texture. Figure 4 has two examples: straw and paper. In each, there are $80 \times 60 = 4800$ pixels; "ones" are depicted with dots and "zeros" with stars.

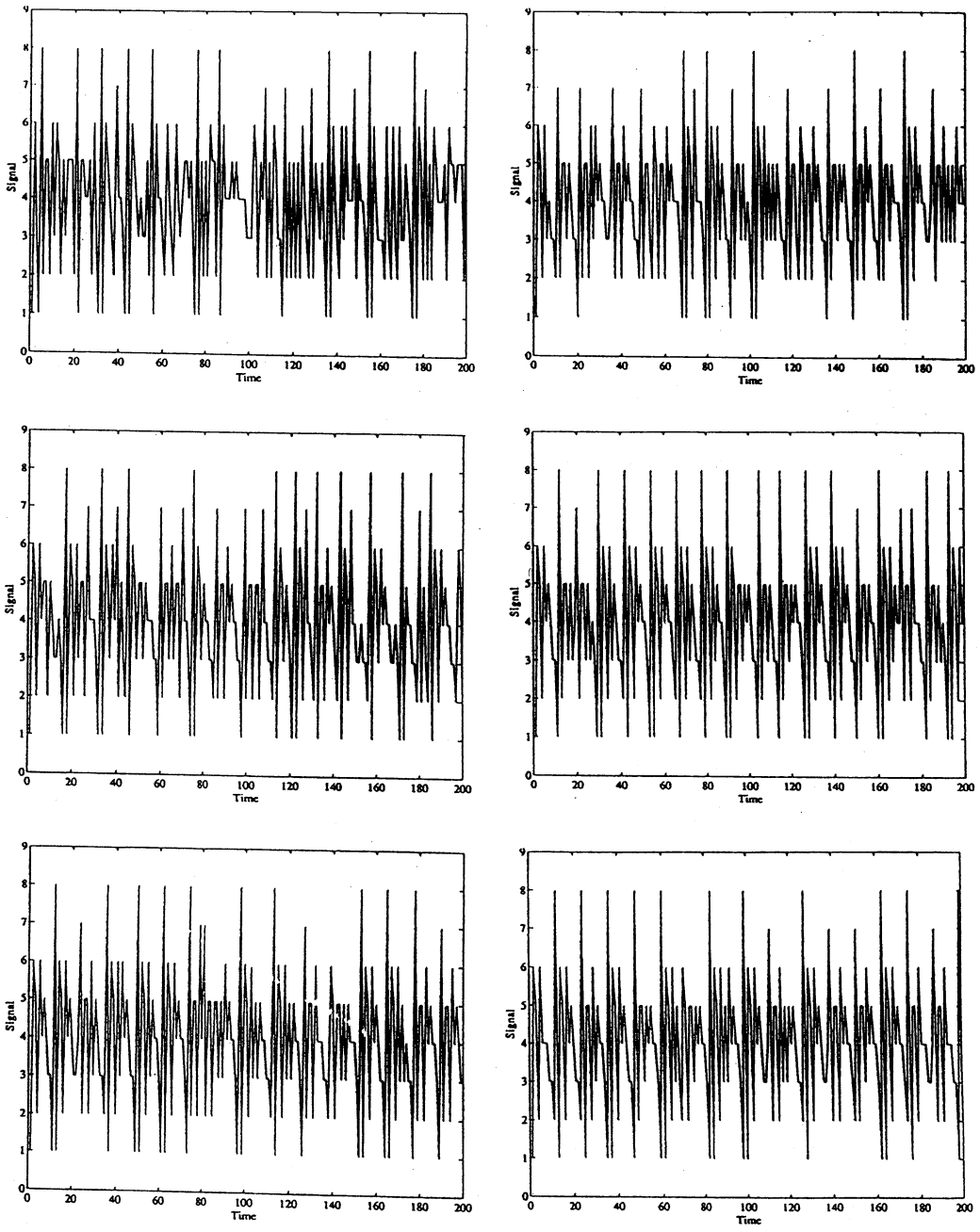


FIG. 2. *HMM's estimated from data in Figure* 1a. *Top to bottom, left-hand side*: 10, 20 *and* 30 *hidden states. Top to bottom, right-hand side*: 40, 50 *and* 60 *hidden states.*

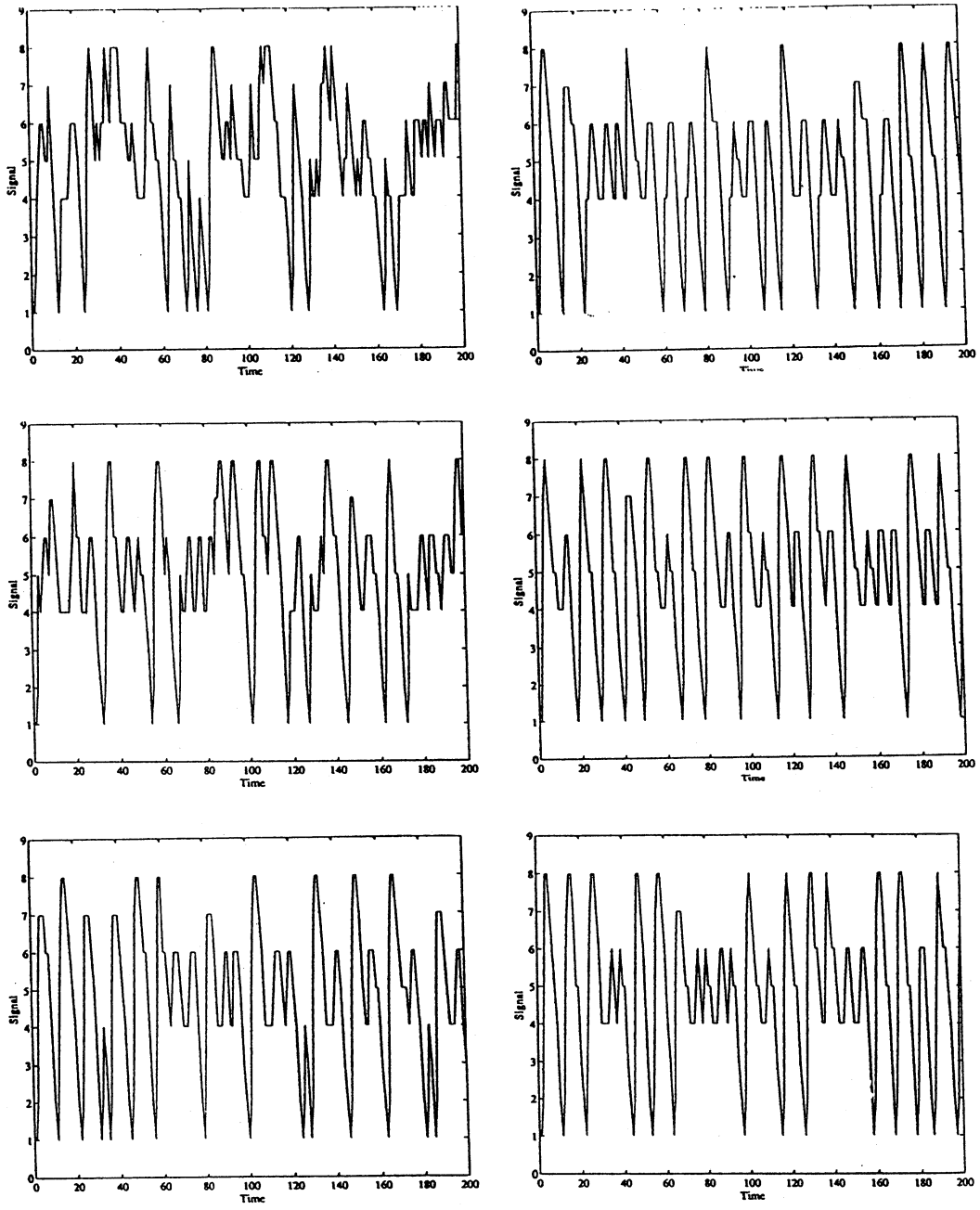
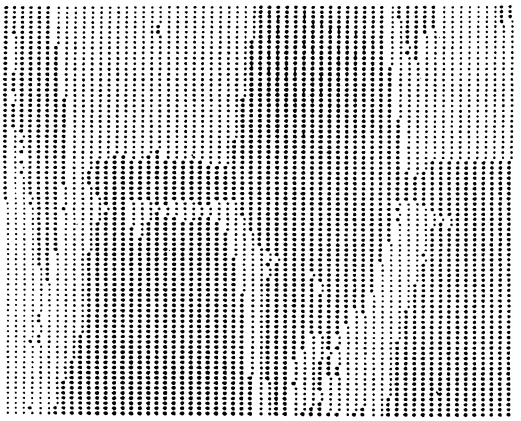
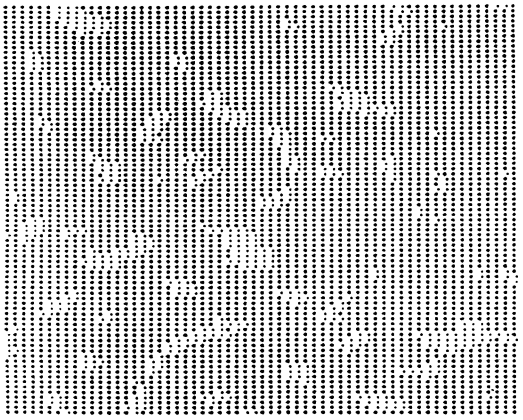


FIG. 3. *HMM's estimated from data in Figure* 1b. *Top to bottom, left-hand side*: 10, 20 *and* 30 *hidden states. Top to bottom, right-hand side*: 40, 50 *and* 60 *hidden states.*



(a)



(b)

FIG. 4. (a) *Thresholded image of straw.* (b) *Thresholded image of paper.*

Following our approach to the speech data, we viewed these images as samples from stationary (spatial) processes, and attempted to fit these processes with N-state hidden Markov models. Specifically, we employed the hiding function $f(x) = x \bmod 2$ and a four-nearest-neighbor Gibbs representation for the hidden process, X_t, $t \in S = \{(i, j)\colon 1 \le i \le 80,\ 1 \le j \le 60\}$. In both experiments, N was fixed at 10, so that $X_t \in \{0, 1, \ldots, 9\}$.

For each texture we fit two matrices $\alpha^h = \{\alpha^h_{kl}\}$ and $\alpha^v = \{\alpha^v_{kl}\}$, where $0 \le k,\ l \le 9$ and h stands for "horizontal" and v for "vertical." These matrices represent the Gibbs potential for X, as follows:

$$\Pi\left(X_{i,j} = k | X_{i-1,j} = l_1, X_{i+1,j} = l_2, X_{i,j-1} = l_3, X_{i,j+1} = l_4\right)$$
$$\propto \exp - \left\{\alpha^v_{l_1 k} + \alpha^v_{k l_2} + \alpha^h_{l_3 k} + \alpha^h_{k l_4}\right\},$$

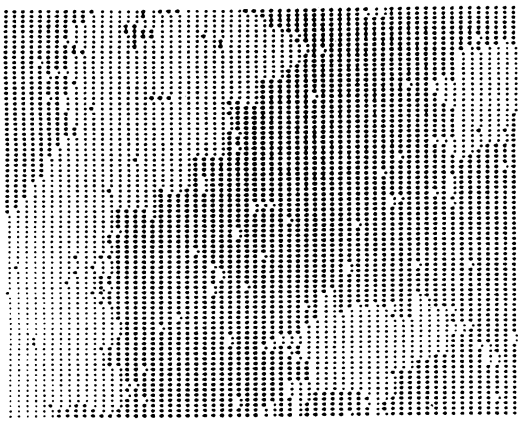
except that terms are dropped when they reference outside of the 80×60 array ("free boundary conditions"). Given a sample $z = \{z_t\}_{t \in S}$, the partial derivative with respect to α_{kl}^h $(0 \leq k, l \leq 9)$ of the log-likelihood of z, under the model $\{f(X_t)\}_{t \in S}$, is

$$E\left[N_{kl}^h\right] - E\left[N_{kl}^h \mid f(X_t) = z_t, t \in S\right], \tag{6}$$

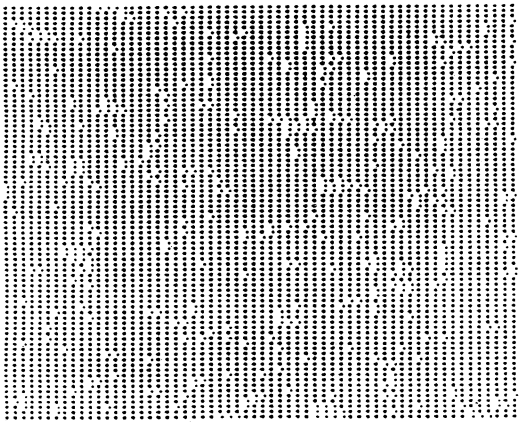
where

$$N_{kl}^h = \#\left\{(i,j)\colon 1 \leq i \leq 80, 1 \leq j \leq 59, X_{i,j} = k, X_{i,j+1} = 1\right\}$$

(a "sufficient statistic"). An analogous expression governs partial derivatives with respect to the components of α^v. One way to estimate the matrices α^h and α^v is via a discrete gradient ascent: compute (6) at the "current" parameter values, take a small step in the direction of the gradient, recom-



(a)



(b)

FIG. 5. (a) *HMM estimated from data in Figure* 4a. (b) *HMM estimated from data in Figure* 4b.

pute (6) and so on. Unfortunately, the computation of (6) is notoriously difficult. We resorted to Monte Carlo methods (cf. Metropolis, Rosenbluth, Rosenbluth, Teller and Teller [39] and Besag and Green [8]), repeatedly using the Gibbs sampler to estimate both expectations.

The approach is unsatisfactory. It is slow and it is difficult to judge convergence, both within an iteration (computation of the expectations) and overall (when to stop?). There have been many suggestions for improving the efficiency of the calculations; see, for example, Younes [45] and Qian and Titterington [42]. We experimented with a variety of alternatives, without much success. In the end we settled on the approach outlined above, which we view as decidedly brute force and last resort.

Having estimated potential functions (α^h and α^v) for both the (binarized) straw and paper textures, we drew samples from the corresponding Gibbs distributions—again, via the Gibbs sampler. The results, viewed through the hiding function f, are shown in Figure 5.

As with the problem of synthesis in speech, texture synthesis is made intriguing by the availability of unlimited amounts of data. Despite this favorable circumstance, there are as of yet no fully satisfactory solutions, especially if one wants to render samples at arbitrary angles and resolution. We have offered a solution, *in principle*: Nearest-neighbor HMM's are dense and can be estimated. Evidently, however, the approach is a long way from being practical. In any case, others have already made good progress: We cite [15], [26], [19], [22], [33] and [21], for some state-of-the-art work on texture estimation and synthesis.

REFERENCES

[1] BAHL, L. R., JELINEK, F. and MERCER, R. L. (1983). A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **5** 179–190.

[2] BALL, F. and SANSOM, M. (1988). Aggregated Markov processes, incorporating time interval omission. *Adv. in Appl. Probab.* **20** 546–572.

[3] BALL, F. G. and RICE, J. A. (1992). Stochastic models for ion channels: introduction and bibliography. *Math. Biosci.* **112** 189–206.

[4] 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 for ecology. *Bull. Amer. Math. Soc.* **73** 360–363.

[5] BAUM, L. E. and PETRIE, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Statist.* **37** 1554–1563.

[6] BAUM, L. E., PETRIE, T., SOULES, G. and WEISS, N. (1970). A maximization technique in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.* **41** 164–171.

[7] BERBEE, H. C. P. and BRADLEY, R. C. (1984). A limitation of Markov representation for stationary processes. *Stochastic Process. Appl.* **18** 33–45.

[8] BESAG, J. and GREEN, P. J. (1993). Spatial statistics and Bayesian computation (with discussion). *J. Roy. Statist. Soc. Ser. B* **55** 25–37.

[9] BICKEL, P. J. and RITOV, Y. (1993). Inference in hidden Markov models I: local asymptotic normality in the stationary case. Technical Report 383, Dept. Statistics, Univ. California, Berkeley.

[10] BILLINGSLEY, P. (1964). *Ergodic Theory and Information.* Wiley, New York.

[11] BLACKWELL, D. and KOOPMANS, L. (1957). On the identifiability problem for functions of finite Markov chains. *Ann. Math. Statist.* **28** 1011–1015.
[12] BRADLEY, R. C. (1993). An addendum to "A limitation of Markov representation for stationary processes." *Stochastic Process. Appl.* **47** 159–166.
[13] BROCKETT, R. W. (1979). Stochastic realization theory and Planck's law for black body radiation. *Ricerche Automat.* **10** 344–362.
[14] BRODATZ, P. (1966). *Texture: A Photographic Album for Artists and Designers.* Dover, New York.
[15] CHELLAPPA, R. and KASHYAP, R. L. (1985). Texture synthesis using 2-D noncausal autoregressive models. *IEEE Trans. Acoust. Speech Signal Process.* **33** 194–203.
[16] CHURCHILL, G. A. (1989). Stochastic models in heterogeneous DNA sequences. *Bull. Math. Biol.* **51** 79–94.
[17] COMETS, F. and GIDAS, B. (1992). Parameter estimation for Gibbs distributions from partially observed data. *Ann. Appl. Probab.* **2** 142–170.
[18] DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39** 1–38.
[19] DERIN, H. and ELLIOTT, H. (1987). Modeling and segmentation of noisy and textured images using Gibbs random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **9** 39–55.
[20] DHARMADHIKARI, S. W. (1963). Sufficient conditions for a stationary process to be a function of a finite Markov chain. *Ann. Math. Statist.* **34** 1033–1041.
[21] ELFADEL, I. M. and PICARD, R. W. (1994). Gibbs random fields, co-occurrences, and texture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **16** 24–37.
[22] FRANCOS, J. M., MEIRI, A. Z. and PORAT, B. (1992). A unified texture model based on a 2-D Wold like decomposition. Technical report, Dept. Electrical Engineering, Technion–Israel Institute of Technology, Haifa, Israel.
[23] FREDKIN, D. R. and RICE, J. A. (1987). Correlation functions of a function of a finite-state Markov process with application to channel kinetics. *Math. Biosci.* **87** 161–172.
[24] Fredkin, F. R. and RICE, J. A. (1992). Maximum likelihood estimation and identification directly from single-channel recordings. *Proc. Roy. Soc. London Ser. B* **249** 125–132.
[25] FRIGESSI, A. and PICCIONI, M. (1990). Parameter estimation for 2 dimensional Ising fields corrupted by noise. *Stochastic Process. Appl.* **34** 297–311.
[26] GAGALOWICZ, A. and MA, S. D. (1985). Sequential synthesis of natural textures. *Computer Vision, Graphics, and Image Processing* **30** 289–315.
[27] GEMAN, S., KEHAGIAS, A. and KÜNSCH, H. (1993). Consistent estimation of stationary processes and stationary random fields. Technical report, Div. Applied Mathematics, Brown Univ.
[28] GEORGII, H.-O. (1988). *Gibbs Measures and Phase Transitions.* de Gruyter, New York.
[29] GILBERT, E. J. (1959). On the identifiability problem for functions of finite Markov chains. *Ann. Math. Statist.* **30** 688–697.
[30] GRENANDER, U. (1978). *Abstract Inference.* Wiley, New York.
[31] ITÔ, H., AMARI, S.-I. and KOBAYASHI, K. (1992). Identifiability of hidden Markov information sources and their minimum degrees of freedom. *IEEE Trans. Inform. Theory* **38** 324–333.
[32] JI, C. (1990). Sieve estimators for pair-interaction potentials and local characteristics in Gibbs random fields. Technical Report 2037, Institute of Statistics, Univ. North Carolina.
[33] JI, C. and SEYMOUR, L. (1992). On the selection of Markov random field texture models. Technical report, Dept. Statistics, Univ. North Carolina.
[34] KALLIANPUR, G. (1980). *Stochastic Filtering Theory.* Springer, New York.
[35] KEHAGIAS, A. (1992). Approximation of stochastic processes by hidden Markov models. Ph.D. dissertation, Div. Applied Mathematics, Brown Univ.
[36] KIENKER, P. (1989). Equivalence of aggregated Markov models of ion-channel gating. *Proc. Roy. Soc. London Ser. B* **236** 269–309.