LA-UR 97-992

TITLE: **MAXIMUM LIKELIHOOD CONTINUITY MAPPING FOR FRAUD DETECTION**

AUTHOR(S): John Hogden

SUBMITTED TO: External Distribution - Hard Copy

Los Alamos

Los Alamos National Laboratory
Los Alamos New Mexico 87545

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, make any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Maximum likelihood continuity mapping for fraud detection

John Hogden
MS B265
Los Alamos National Lab
Los Alamos, NM 87545
(email: hogden@lanl.gov)

We describe a novel time-series analysis technique called maximum likelihood continuity mapping (MALCOM), and focus on one application of MALCOM: detecting fraud in medical insurance claims. Given a training data set composed of typical sequences, MALCOM creates a stochastic model of sequence generation, called a continuity map (CM). A CM maximizes the probability of sequences in the training set given the model constraints. CM's can be used to estimate the likelihood of sequences not found in the training set, enabling anomaly detection and sequence prediction -- important aspects of data mining. Since MALCOM can be used on sequences of categorical data (e.g. sequences of words) as well as real valued data, MALCOM is also a potential replacement for database search tools such as N-gram analysis. In a recent experiment, MALCOM was used to evaluate the likelihood of patient medical histories, where "medical history" is used to mean the sequence of medical procedures performed on a patient. Physicians whose patients had anomalous medical histories (according to MALCOM) were evaluated for fraud by an independent agency. Of the small sample (12 physicians) that has been evaluated, 92% have been determined fraudulent or abusive. Despite the small sample, these results are encouraging.

keywords: anomaly, fraud, time-series, continuity map

I. INTRODUCTION

The main goal of this paper is to present a novel technique for analyzing time series data. The technique has already been used to solve a difficult problem in speech processing (Hogden, 1996b; Hogden, 1997) and is currently being studied as a way to detect fraud in medical insurance claims. Although verifying that a medical insurance claim is fraudulent is a difficult and time-consuming task requiring the aid of medical professionals (and sometimes involving legal proceedings), some evidence that the technique can be used for fraud detection is now available. This evidence, although incomplete, will be presented following the discussion of the theory underlying the time-series analysis. To protect the interests of the agency funding the fraud detection work and the individuals whose medical claims have been evaluated, some details of the data and the fraud detection procedure have been omitted. However, these omissions should not adversely affect the reader's ability to obtain a general understanding of the technique being presented, or it's potential.

Methods for estimating the likelihood of data sequences have obvious applications to detecting anomalous data sequences. Consider the problem of detecting fraud using large databases of medical histories, in which each patient's medical history is composed of a sequence of medical procedures performed on the patient. Given a large enough collection of medical histories, such as are available at large medical insurance companies or the various medical social services, sufficient statistical information should be present to infer whether the care that is being delivered and the charges for the care are normal or unusual. If a patient's medical history is sufficiently anomalous, the patient's medical history or the physician's record should be further reviewed for evidence of fraud.

To accurately estimate the likelihood of a medical procedure being performed on a particular patient, we must take into account the other procedures performed on the patient, e.g. although a kidney transplant may be very unlikely, the transplant should not be considered so unlikely if the patient has already undergone a series of diagnostics tests typically used to detect kidney disease. Furthermore, since doctors are likely to perform simple, noninvasive tests before performing more complex and/or invasive tests, seeing the same set of procedures performed in a different order may radically alter a perception of how likely a medical history is.

One approach to using context to determine the probability of a sequence is to use N-grams, as has been done in language modeling for speech recognition (Lee, 1989; Maltese & Mancini, 1992) and for searching large databases of text documents (Damashek, 1992; Damashek, 1995). For example, in order to determine the probability of the symbol sequence [B, G, Q, D] an algorithm might use stored knowledge about the probability of G following B, Q following G, and D following Q. These probabilities are called bigram probabilities because they use stored information about the probability of two-procedure sequences. If we knew the probability of longer sequences of procedures (sequences of N procedures are called N-grams) estimates of the probabilities of long sequences should improve. While it is theoretically possible to calculate the probabilities of all N-procedure sequences, the number of probabilities that have to be estimated increases as S^N, where S is the number of distinct symbols (note that S is not necessarily the length of a sequence because symbols can be repeated). Therefore, as a practical matter, it is usually impossible to obtain enough data to accurately estimate all the necessary parameters.

The time-series analysis technique discussed below, Maximum Likelihood Continuity Mapping (MALCOM), can also be used for estimating the likelihood of a data sequence. MALCOM uses training data composed of sequences of symbols (words, letters, medical procedure codes, etc.) to create a model of sequence generation. After training, the model can be used to determine the probability of a sequence of symbols. As is discussed below, the number of parameters that need to be estimated to create the model increases linearly with the number of symbols. This means that less data will typically be needed for training MALCOM than for training N-grams. So for data sets for which MALCOM is a good model, MALCOM should be preferred to N-gram analysis.

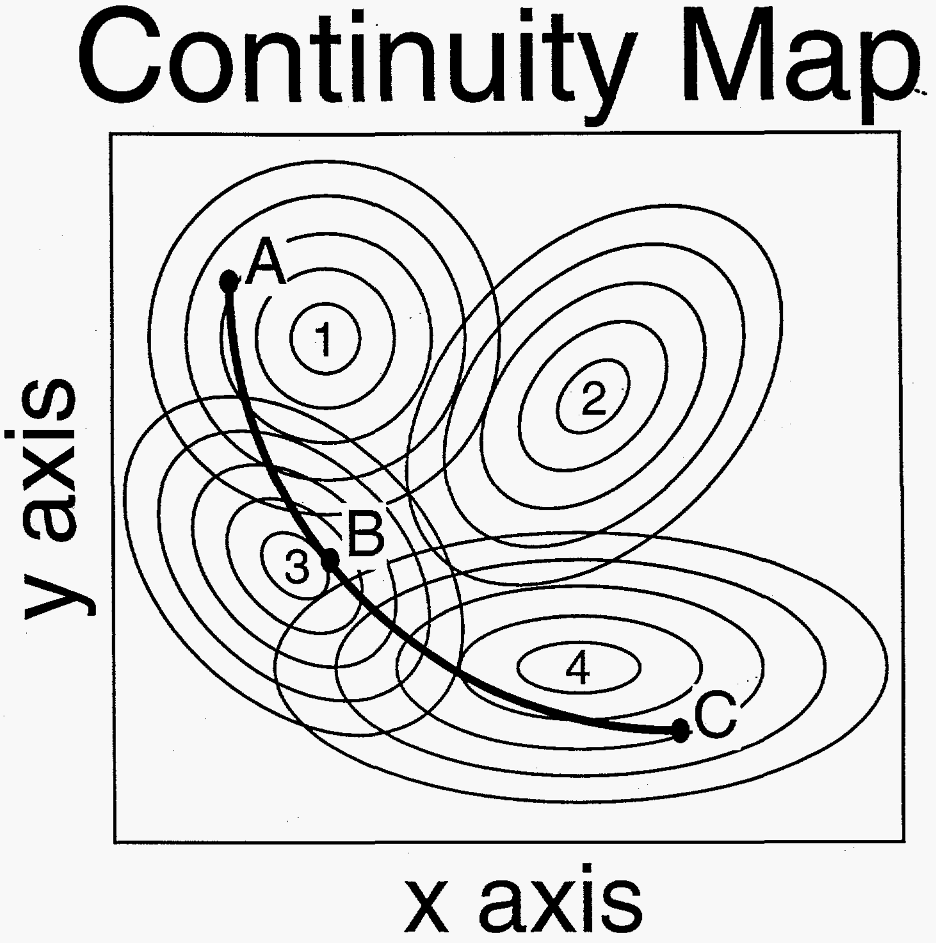

Figure 1

In the MALCOM model of sequence generation (Hogden, Scovel & White, 1997), sequences are produced as a point moves smoothly through an abstract space called a continuity map (CM). Figure 1 shows a hypothetical 2-dimensional CM that will be used to explain MALCOM. Higher dimensional CM's can also be made, but the 2-dimensional map is sufficient as an example. The CM in Figure 1 is characteristic of a CM used to determine the probabilities of sequences composed of symbols in the set {1, 2, 3, 4}, such as the sequence [1, 4, 4, 3, 2]. In Figure 1, the set of concentric ellipses around the number "2" are used to represent equiprobability curves of a probability density function (PDF). The PDF gives the probability that the symbol "2" will be produced from within any region of the CM. Similarly, the ellipses surrounding the numbers 1, 3, and 4, represent equiprobability curves of PDFs giving the probability of producing the symbols "1", "3", and "4", respectively, from any region in the CM. For ease of exposition, call the smallest ellipse centered around the number i, curve L_{i1}, call the next largest curve L_{i2}, etc. In the following discussion it will be assumed that the height of the PDF (on the z axis -- not shown in the figure) of L_{ij} is identical to the height of L_{kj}, i.e. the curve closest to "1" connects points with the same probability density as the points connected by the curve closest to "2", etc. It will also be assumed that the height of each PDF monotonically decreases as we move radially out from the center of the PDF.

From Figure 1, it can be seen that the probability of producing symbol "1" from the point marked "A" is higher than the probability of producing "2", "3", or "4" from position "A", but that there is some probability of producing "2", "3" or "4" from position A. In general, every position in the CM will have some likelihood of producing each of the symbols, and each sequence of n positions in the CM has some likelihood of producing any sequence of n symbols.

To estimate the probability of a sequence of symbols using MALCOM, we find the path through the CM, i.e. a sequence of positions in the CM, that maximizes the probability of the symbol sequence, and use the probability of producing the symbol sequence from the path as the estimate of the probability of the sequence. Assuming that all of the PDFs have the same number of parameters, the number of parameters needed to describe the CM (the positions and shapes of all the PDFs) is the number of symbols times the number of parameters needed to describe the PDF associated with any one symbol. Therefore, the number of parameters to estimate increases linearly with the number of symbols.

In the MALCOM model, if all paths through the CM were equally likely, then all sequences which differed only in the order of the symbols would be equally probable. Since it is important to be able to represent the fact that symbol sequences differing in order may have different probabilities, MALCOM constrains the possible paths through the CM. As discussed in section II, this will allow MALCOM to assign different probabilities to sequences with different orderings of the same symbols. As the smooth curve connecting the points "A", "B", and "C" suggests, MALCOM as currently embodied requires that paths through the CM are smooth, where by "smooth" we mean (speaking loosely) that the Fourier transform of a path has no energy above some cut-off frequency, f_c. Since all signals of less than infinite extent have high frequency components, exactly what is meant by this constraint is discussed further in section III. The smoothness constraint can also be thought of in terms of probability theory. To do so, we incorporate an a priori PDF over all continuity map trajectories, and have the PDF be uniform and non-zero over all trajectories that have no energy above f_c, but 0 for paths with Fourier components above f_c. This smoothness constraint could be replaced by other constraints, e.g. that the probability of a path goes down as the frequencies increase, or the paths must all lie on a circle, etc.

Note that no meaning is attributed to the CM axes in Figure 1 -- the axes are simply labeled "x axis" and "y axis". However, in some cases, the CM trajectories inferred by MALCOM are very interpretable. For example, in a study where continuity maps were constructed using digitized recordings of speech as the training data, the trajectories recovered by MALCOM were highly correlated with tongue positions (Hogden, 1996b). Thus, MALCOM was able to solve the problem of finding a nonlinear function relating speech sounds to tongue positions using training data that did not contain any tongue position measurements. Apparently, when the data sequences are the result of the smooth movement of real objects, MALCOM can sometimes recover the positions of the objects underlying the data production. The importance of this result for speech recognition, speaker recognition, speech coding, and speech synthesis is discussed elsewhere (Hogden, 1996a).

Below we describe the theory that allows us to determine the path through the CM that maximizes the likelihood of the data, as well as the technique for estimating the parameters of the PDFs in the CM. Since the algorithm for creating the CM uses the technique for finding the path that maximizes the probability of the data, the following description will first discuss how the best path is found given a CM, and then discuss how to make a CM.

II. FINDING CONTINUITY MAP TRAJECTORIES THAT MAXIMIZE THE PROBABILITY OF THE OBSERVED DATA

The probability of a symbol sequence given a continuity map trajectory will be derived by first finding the probability of a single symbol given a single continuity map position, and then by combining probabilities over all the symbols in a sequence. Next, a technique for finding the continuity map trajectory that maximizes the conditional probability of a sequence of symbols will be described. In section III, the problem is constrained to find the *smooth* continuity map trajectory (as opposed to any arbitrary continuity map trajectory) that maximizes the probability of the symbol sequence.

The following definitions are used:

n = the number of symbols in a given sequence,
d= the number of dimensions in the continuity map,
$c(t)$ = the t^{th} symbol in the sequence,
$\mathbf{c} = [c(1), c(2), \ldots c(n)]$ = a sequence of symbols,
$x_i(t)$ = the continuity map position on axis i at time t,
$\mathbf{x}(t) = [x_1(t), x_2(t), \ldots x_d(t)]$ = a vector giving the continuity map position at time t, and
$\mathbf{X} = [\mathbf{x}(1), \mathbf{x}(2), \ldots \mathbf{x}(n)]$ = a sequence of continuity map positions.

Further definitions are needed to specify the mapping from continuity map positions to symbols. Let:

$P(c_i)$ = the probability of observing symbol c_i given no information about context,
$P(\mathbf{x}|c_i, \varphi)$ = the probability that continuity map position $\mathbf{x}$ was used to produce symbol c_i,

where φ = a set of model parameters, e.g., φ could include the mean and covariance matrix of a Gaussian probability density function used to model the distribution of $\mathbf{x}$ given c.

Note that the form of the $P(\mathbf{x}|c_i,\varphi)$ distributions is not specified. $P(\mathbf{x}|c_i,\varphi)$ should be chosen to allow the distribution of the continuity map positions to closely match the various possible distributions needed to model different kinds of sequences. For some types of data, it may even be necessary to specify $P(\mathbf{x}|c_i,\varphi)$ as a mixture of Gaussians, or some other multimodal distribution.

With these definitions, the probability of observing symbol c_j given that the current continuity map position is $\mathbf{x}$, is expressed as:

$$P(c_j|\mathbf{x},\varphi) = \frac{P(c_j,\mathbf{x}|\varphi)}{P(\mathbf{x}|\varphi)} = \frac{P(c_j,\mathbf{x}|\varphi)}{\sum_i P(c_i,\mathbf{x}|\varphi)} = \frac{P(\mathbf{x}|c_j,\varphi)P(c_j)}{\sum_i P(\mathbf{x}|c_i,\varphi)P(c_i)} \qquad \text{Eq. 1}$$

The probability of the symbol sequence can be determined by assuming conditional independence, i.e.

$$P[\mathbf{c}|\mathbf{X},\varphi] = \prod_{t=0}^{n} P[c(t)|\mathbf{x}(t),\varphi] \qquad \text{Eq. 2}$$

Note that the probability of observing a symbol is not assumed to be independent of the preceding and subsequent symbols; it is only assumed to be conditionally independent. So if $\mathbf{x}(t)$ is dependent on $\mathbf{x}(t')$ then $c(t)$ is dependent on $c(t')$. If we appropriately constrain continuity map trajectories, MALCOM can assign different probabilities to different orderings of the same symbols.

It is possible to find the continuity map trajectory that maximizes the conditional probability of a sequence of codes, i.e. find the **X** that maximizes $P(\mathbf{c}|\mathbf{X},\varphi)$. This is done by first finding the gradient of $LogP[\mathbf{c}\,|\,\mathbf{X},\varphi]$, and then using a standard gradient maximization algorithm. The gradient can be shown to be:

$$\nabla LogP[\mathbf{c}|\mathbf{X},\varphi]=\frac{\nabla P[\mathbf{x}(t')|c(t'),\varphi]}{P[\mathbf{x}(t')|c(t'),\varphi]}-\frac{\sum_i P[c_i]\nabla P[\mathbf{x}(t')|c_i,\varphi]}{\sum_i P[\mathbf{x}(t')|c_i,\varphi]P[c_i]} \qquad \text{Eq. 3}$$

III. CONSTRAINING THE TRAJECTORIES

The preceding analysis is incomplete because it ignores constraints on the possible continuity map paths. As already mentioned, in this fraud detection study, the constraint is that the trajectories through the CM are composed of only low-frequency components. Realizing that a discrete Fourier transform can be though of as a matrix multiplication, this constraint is equivalent to requiring that the path lie on a hyperplane composed of the axes defined by low frequency sine and cosine waves. Thus, when $\nabla\mathrm{Log}(\mathbf{c}|\mathbf{X},\varphi)$ is perpendicular to the constraining hyperplane, i.e. has no components below the cut-off frequency, so that $\mathrm{Log}(\mathbf{c}|\mathbf{X},\varphi)$ can not increase without **X** traveling off the hyperplane, then a constrained local maximum has been reached (Marsden & Tromba, 1981). To restate this point, the smooth path that maximizes the likelihood of the observed data is the path for which $\nabla\mathrm{Log}(\mathbf{c}|\mathbf{X},\varphi)$ has no components on the hyperplane. In order to apply this constraint using a gradient maximization algorithm, we need to ensure that the initial guess at a path is smooth, and then modify the initial guess using only smooth gradients. In the work described below, we used the conjugate gradient technique to solve the constrained maximization problem.

Solving the constrained optimization problem involves low-pass filtering the continuity map trajectories. By low-pass filtering a multidimensional path we mean low-pass filtering the time series composed of $[x_1(1), x_1(2), \ldots x_1(n)]$ (the path projected onto dimension 1), then low-pass filtering the path composed of $[x_2(1), x_2(2), \ldots x_2(n)]$ (the path projected onto dimension 2), etc. until the path has been low-pass filtered on each dimension. Notice that this process will force the path to be low-pass filtered regardless of any rotation, reflection, translation or scaling of the CM. This result follows because any linear combination of signals having no energy above f_c will have no energy above f_c.

However, the uncertainty principle tells us that signals of finite duration always have some energy at high frequencies, and trajectories through the CM encountered in practice will be of finite duration. This is not a problem in practice because digital filtering theory assumes that the paths are periodic -- after the last element of the time series we restart with the first element of the time series. Thus, the paths are treated as if they are of infinite length.

Treating the paths as if they are periodic causes one other complication. In the hypothetical continuity map shown in Figure 1, the values of x-axis component of the path [A,B,C] increase in value from time 1 to time 3. Therefore, the time series made up of the path positions from Figure 1 projected onto the x-axis, and then repeated periodically, will have large discontinuities. The CM could easily be rotated and translated so that the x-axis component of the path at times 1 and 3 are 0 and the x-axis component of the path at time 2 is some positive value. Thus, by performing simple rotations and translations of the CM we can get time series that either have large discontinuities or are relatively smooth. To avoid problems that would arise from the discontinuities, the trend of the time series should be removed before smoothing the paths, and then added back after the filtering has been performed (Press, Flannery, Teukolsky & Vetterling, 1988 discuss removing the trend before filtering). Similar considerations should be applied to the smoothing of the gradient, which can also be thought of as a multidimensional path.

IV. FINDING THE CONTINUITY MAP PARAMETERS

In the preceding sections, it was assumed that $P(c)$ and $P(\mathbf{x}|c,\varphi)$ are known. However, in general the values of $P(c)$ and $P(\mathbf{x}|c,\varphi)$ will be determined from a data set composed of symbol sequences by iteratively repeating two steps:

1) given a collection of symbol sequences and some initial estimate of mapping from symbols to continuity map positions, use the procedures described in sections II and III to find the paths that maximize the conditional probability of the symbol sequences.
2) given the paths that maximize the probability of the symbol sequences, find the value of φ and the $P(c_i)$ values that will maximize (or at least increase) the conditional probability of the symbol sequences.

Since both of these steps will increase the probability of the data, by iteratively repeating them, the probability of the data is increased until a local (possibly global) maximum is reached. Calculation of φ can be accomplished using standard gradient maximization algorithms using the gradient of the log of the conditional probability with respect to the components of φ, which can be shown to be:

$$\nabla Log P[\mathbf{c}|\mathbf{X},\varphi] = \sum_t \left\{ \frac{\nabla P[\mathbf{x}(t)|c(t),\varphi]}{P[\mathbf{x}(t)|c(t),\varphi]} - \frac{\sum_i \nabla\{P[\mathbf{x}(t)|c_i,\varphi]P[c_i]\}}{\sum_i P[\mathbf{x}(t)|c_i,\varphi]P[c_i]} \right\} \qquad \text{Eq. 4}$$

Gradient techniques are not needed to derive the $P(c)$ values that maximize the conditional probability; the values can be shown to be:

$$P(c_k) = n_k/n \qquad \text{Eq. 5}$$

where n_k is the number of times c_k is observed in the data sample.

IV. A. EXAMPLE

Although many different forms of the $P[\mathbf{x}(t)|c(t),\varphi]$ distributions could be used, in this section we give the gradient equations for the exemplary case where the distribution of continuity map positions that produce symbol c is a multivariate Gaussian characterized by the equation:

$$P[\mathbf{x} \mid c,\varphi] = \frac{1}{(2\pi)^{d/2}|\sigma(c)|^{1/2}} \exp\left\{-\frac{1}{2}[\mathbf{x}-\mu(c)]^t \sigma^{-1}(c)[\mathbf{x}-\mu(c)]\right\} \qquad \text{Eq. 6}$$

where:

$\mu(c)$ is a vector giving the mean of all the continuity map positions used to produce symbol c. For example, $\mu_i(c)$ may give the mean continuity map position on axis i used to create sounds quantized as code c,

$\sigma(c)$ is the covariance matrix of the multivariate Gaussian distribution of continuity map positions that produce symbol c, and

$\mathbf{x}$ is a vector specifying a point in the continuity map.

The gradient with respect to the components of $\mathbf{x}$ is:

$$\nabla LogP[\mathbf{c}|\mathbf{X},\varphi] = -\sigma^{-1}[c(t')]\{\mathbf{x}(t') - \mu[c(t')]\} + \frac{\sum_i P[c_i]P[\mathbf{x}(t')|c_i,\varphi]\sigma^{-1}(c_i)\{\mathbf{x}(t') - \mu(c_i)\}}{\sum_i P[\mathbf{x}(t')|c_i,\varphi]P[c_i]} \qquad \text{Eq. 7}$$

and the gradient with respect to $\mu(c_k)$ is:

$$\nabla LogP[\mathbf{c}|\mathbf{X},\varphi] = \sum_{t\in c(t)=c_k}\sigma^{-1}[c(t)]\{\mathbf{x}(t) - \mu[c(t)]\} - \sum_t \frac{P[c_k]P[\mathbf{x}(t)|c_k,\varphi]\sigma^{-1}(c_k)\{\mathbf{x}(t) - \mu(c_k)\}}{\sum_i P[\mathbf{x}(t)|c_i,\varphi]P[c_i]} \qquad \text{Eq. 8}$$

V. FRAUD DETECTION STUDY

The preliminary study described below suggest that a simplification of MALCOM (MALCOM *1*) is able to learn a model of medical procedure sequences well enough to allow fraud detection. The simplifications in MALCOM *1* significantly decrease training time when trying to learn the CM, but do not completely maximize the probability of the data. Improved performance is expected when the full MALCOM algorithm is applied to this problem.

V.A. MALCOM *1*

Instead of maximizing the conditional probability of the observed data, MALCOM *1* creates a CM by maximizing the probability of the smooth CM paths. Mathematically, this amounts to ignoring the second term in Eqs. 7 & 8. The simplified versions of Eqs. 7 and 8 are, respectively:

$$\nabla LogP[\mathbf{c}|\mathbf{X},\varphi] = -\sigma^{-1}[c(t')]\{\mathbf{x}(t') - \mu[c(t')]\} \qquad \text{Eq. 9}$$

and

$$\nabla LogP[\mathbf{c}|\mathbf{X},\varphi] = \sum_t \sigma^{-1}[c(t)]\{\mathbf{x}(t) - \mu[c(t)]\} \qquad \text{Eq. 10}$$

Notice that Eq. 10 is significantly simpler to maximize than Eq. 8. Eq. 8 requires an iterative maximization algorithm whereas Eq. 10 can be solved analytically. The analytic solution for Eq. 10 is to set

$$\mu(c_i) = \sum_{t \ni c(t)=c_i} \mathbf{x}(t) \Big/ n_i \qquad \text{Eq. 11}$$

P(c) is calculated using Eq. 5. For this study, all the covariance matrices were set to the identity matrix.

One difficulty in using MALCOM *1* is that there is a degenerate solution in which all of the $\mu(c)$'s converge to the same point. This problem was avoided by forcing the variance of the $\mu(c)$'s to be 1.

V.B. A Continuity Map For Medical Claim Sequences

To use MALCOM for detecting fraud in medical claim sequences, a CM was created from a training data set of claim histories for 1,944 patients. While it is not clear whether MALCOM provides an appropriate model for medical sequence generation, we might imagine that positions in the CM created from medical claim sequences can be loosely thought of as describing a patient's health as measured in several ways, e.g. one axis might correspond to the health of the patient's musculoskeletal system, one axis might correspond to the health of the patient's cardiovascular system, etc.; however, the dimensions of the space are determined by the algorithm, not by preconceived notions of how health should be represented. With this interpretation of the CM, the assumption that the paths through the CM are smooth corresponds to the constraint that the health of the patient changes relatively slowly, getting gradually worse, and then, hopefully, getting gradually better. While this is not always the case, future versions should be able to accommodate occasional sudden changes in health, such as breaking a leg.

V.C. Comparing MALCOM To A Context-Free Model

After training, MALCOM was used to rank 28,785 patient medical histories. The rankings were lower if the medical history was less probable, i.e. more likely to be fraudulent. The medical history probabilities were also evaluated using what we will call the *context-free model*, in which context information was not used. In the context-free model, the probability of a sequence was simply the product of the probabilities of the individual symbols, where the probabilities of the individual symbols was calculated using Eq. 5.

Of the 28,785 medical histories in the testing set, 19 were known to be anomalous. There are probably other anomalous medical histories in the data set, but it is impossible to know how many or which ones are anomalous.

When the MALCOM rankings of the 19 known anomalous medical histories were compared to the context-free model rankings, we found that MALCOM ranked the known anomalous medical histories as more anomalous than the context-free model. Approximately 2/3 of the 19 known anomalous medical histories were ranked in the worst 1% by MALCOM, but less than half of the anomalous medical histories were ranked in the worst 1% for the context-free model. To find the same number of previously known anomalous medical histories using the context-fee model, nearly twice as many medical histories would have to be evaluated.

These results do not conclusively show that MALCOM outperforms the context-free model, since they could have been obtained even if the context-free model found more fraudulent medical histories than MALCOM. For example, it could be that the medical histories ranked most anomalous by the context-free model are actually fraudulent, but are cases of fraud that have not yet been detected. Clearly, to adequately compare these techniques, we need to be able to verify whether the most anomalous medical histories are actually fraudulent, something that is not currently practical. Until this kind of verification can be performed, we are limited to concluding that these results suggest that MALCOM may perform better that the context-free model.

V.D. Detecting Fraudulent Physicians

In a related test, MALCOM was used to evaluate the likelihood of medical histories for approximately 2.2 million patients. Approximately 30,000 physicians who treated the patients were then ranked in terms of how anomalous their treatment practices were. Of the 30,000 physicians, 12 of the most anomalous physicians (according to MALCOM) have been evaluated for fraud by an independent fraud detection agency. While the number of evaluated physicians is very small, it should be reiterated that confirming that a physician is fraudulent or abusive is a complex process, sometimes involving legal proceedings, and always expensive and time consuming. Eleven (92%) of the 12 physicians have been determined to be fraudulent or abusive and 7 (58%) of the fraudulent or abusive physicians had not previously been detected by the fraud detection agency. The 95% confidence interval for the percent fraud in the population sampled by using the most anomalous physicians is 55% to 100%, and the 95% confidence interval for the percentage of new fraud cases is approximately 25% to 86%.

V.E. Comparing MALCOM To Other Techniques

Further comparisons between MALCOM and other techniques have also been made. For example, to determine whether the physicians found by MALCOM could be easily found using other techniques, we looked at how many of the most anomalous 5% of the physicians (as ranked by MALCOM) were also ranked in the worst 5% of physicians by 3 other fraud detection techniques that have been studied at Los Alamos National Laboratory. Dividing the number of physicians in common by the number of physicians in the worst 5%, we get a measure of the percent overlap. Interestingly, although the other fraud detection techniques also successfully found fraudulent or abusive providers, the percent overlap between MALCOM and each of the other 3 methods were: 2%, 5%, and 7%. A percent overlap score was also found between MALCOM and a technique that combined MALCOM and 5 other features. The overlap with the combined technique was 18%. Since MALCOM constituted 1/6 (17%) of the features used in the combined analysis, this low percent overlap suggests that MALCOM is finding different kinds of fraud than the other techniques studied.

VI. SUMMARY

The major goal of this paper was to describe how continuity maps can be formed and used as models of sequence generation. We hope that the description of MALCOM will prompt other researchers to study MALCOM's application to anomaly detection and a host of other applications involving time series analysis.

In addition, results of using MALCOM for anomaly/fraud detection have been reported to the extent that they exist, but can only be taken as suggesting that MALCOM should be studied further. While the experimental results are clearly preliminary, it is important to note that they uniformly suggest that MALCOM has potential as a fraud detection algorithm.

BIBLIOGRAPHY

Damashek, M. (1992). What Doc's Up? Language identification, topic identification and information retrieval on the desktop (RBW087 961-1073s): Los Alamos National Laboratory.

Damashek, M. (1995). Gauging similarity with n-grams: Language-independent categorization of text. Science, 267, 843-848.

Hogden, J. (1996a). Improving on hidden markov models: An articulatorily constrained, maximum likelihood approach to speech recognition and speech coding (unclassified LA-UR-96-3945). Los Alamos, NM: Los Alamos National Laboratory.

Hogden, J. (1996b). A maximum likelihood approach to estimating speech articulator positions from speech acoustics. The Journal of the Acoustical Society of America, 100(4(pt. 2)), 2663.

Hogden, J. (1997). Speech processing using maximum likelihood continuity mapping, (pp. 1-35). U.S.A.: Provisional Patent Application for the University of California.

Hogden, J., Scovel, J., & White, J. (1997). Anomaly analysis using maximum likelihood continuity mapping, (pp. 1-20). U.S.A: Provisional Patent Application for the University of California.

Lee, K. F. (1989). Automatic Speech Recognition: The Development of the SPHINX System. Boston: Kluwer Academic Publishers.

Maltese, G., & Mancini, F. (1992). An automatic technique to include grammatical and morphological information in a trigram-based statistical language model. IEEE International Conference on Acoustics, Speech, and Signal Processing, 1, 157-160.

Marsden, J., & Tromba, A. (1981). Vector Calculus. (2 ed.). San Francisco: W. H. Freeman and Company.

Press, W., Flannery, B., Teukolsky, S., & Vetterling, W. (1988). Numerical Recipes in C: The Art of Scientific Computing. Cambridge: Cambridge University Press.