NONPARAMETRIC ESTIMATORS FOR ONLINE SIGNATURE AUTHENTICATION

Alexander T. Ihler, John W. Fisher III, and Alan S. Willsky

$\{$*ihler, willsky*$\}$*@mit.edu*, $\{$*fisher*$\}$*@ai.mit.edu*
Laboratory for Information and Decision Systems / Artificial Intelligence Laboratory
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge MA 02139

ABSTRACT

We present extensions to our previous work in modelling dynamical processes. The approach uses an information theoretic criterion for searching over subspaces of the past observations, combined with a nonparametric density characterizing its relation to one-step-ahead prediction and uncertainty. We use this methodology to model handwriting stroke data, specifically signatures, as a dynamical system and show that it is possible to learn a model capturing their dynamics for use either in synthesizing realistic signatures and in discriminating between signatures and forgeries even though no forgeries have been used in constructing the model. This novel approach yields promising results even for small training sets.

1. INTRODUCTION

Real-world dynamical processes often exhibit characteristics which make them extremely difficult to model with conventional tools. Specifically, nonlinear effects and nongaussian randomness can cause canonical modelling methods to fail or at least to require problem-specific modification (human intervention) to mitigate such difficulties. In general there is a fundamental tradeoff in the capacity of a model and its ease of use. For example, linear methods have great advantages in computation but lack the modelling capacity to consistently handle problems which fall outside the linear-quadratic-gaussian regime. Neural network based approaches lack a complete model of uncertainty, while HMM methods are not well suited to continuous state dynamics. Nonparametric approaches possess the modelling capacity; however, such capacity must be controlled (e.g. via dimensionality reduction).

The approach we present, based on nonparametric modelling of a *subspace* of the data, chosen via an information-theoretic criterion, allows us to exploit the modelling capacity of a nonparametric density estimate while reducing the computational burden. This gives us the flexibility to describe complex, possibly multimodal uncertainty and nonlinear system dynamics while retaining control over the computational complexity. Such a system description has an intrinsic notion of randomness and uncertainty (not restricted to a "noise-like" interpretation) which questions the role of "prediction" as a useful metric, but nevertheless characterize many dynamical systems.

2. DYNAMICAL SYSTEM MODEL DESCRIPTION

We hypothesize our dynamical systems to be of the form depicted in Figure 1. In such systems the state is fully captured by the local past of the process. The conditional distribution $p(x_k|x_{k-1},\ldots,x_{k-N})$ is by assumption stationary. It should be noted that such a description includes both stationary and nonstationary processes (e.g. random walk). Furthermore, we hypothesize that the intrinsic dimensionality of this state is less than N. That is, there exists a (possibly vector-valued) function G such that

$$p(x_k|G(x_{k-1},\ldots,x_{k-N}))$$

is stationary and G is sufficient, i.e. the mutual information satisfies

$$I(x_k,\{x_{k-1},\ldots,x_{k-N}\}) = I(x_k,G(x_{k-1},\ldots,x_{k-N}))$$

This G corresponds to the informative subspace of the delay coordinate space $\{x_k,x_{k-1},\ldots,x_{k-N}\}$. The arbitrariness of G allows an arbitrary data manifold – our assumption is only that its dimension is small

Despite this underlying assumption, it should be noted that as discussed in [1] useful information can be extracted by such a model even when the true process does not quite satisfy the conditions. For example, a process which has a dependence longer than N will cause any lost information to be attributed to randomness in the signal. Note also that although $\hat{G}$ is in some way capturing the relational structure of the data and $\hat{p}$ its uncertainty, the two are intimately related. Note $\hat{G}$ need not equal G exactly; it need only be

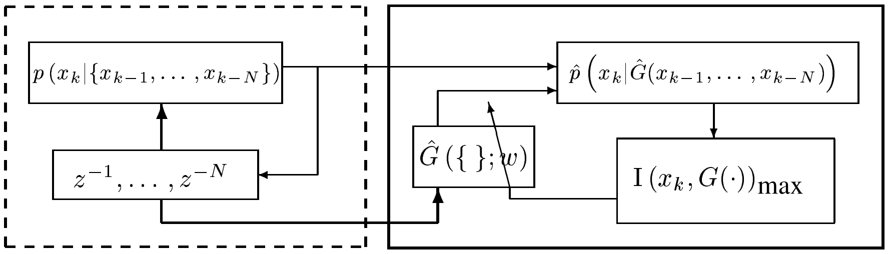


Fig. 1. Dynamical System Model

equivalent to G up to a bijective transformation, which then determines the relationship of p and $\hat{p}$. Thus, representational capacity in the model $\hat{p}$ can reduce the required complexity of $\hat{G}$.

3. TRAINING NONPARAMETRIC DYNAMICAL MODELS

The mutual information between x and any statistic $\hat{G}$ is bounded as follows [2]

$$\begin{aligned}\mathrm{I}(x,\{x_{past}\}) &\geq \mathrm{I}(x,\hat{G}(x_{past}))\\ &= \mathrm{H}(x)+\mathrm{H}(\hat{G}(x_{past}))-\mathrm{H}(x,\hat{G}(x_{past}))\end{aligned}$$

with equality if $\hat{G}$ is sufficient. As described in [1] our training method maximizes $\mathrm{I}(x,\hat{G})$. We construct a density estimate of $p(x,\hat{G}(x_{past}))$ as follows:

$$\hat{p}(x)=\sum_{k=1}^{N}\frac{1}{h}K(\frac{x-X_k}{h}) \tag{1}$$

where $X_k=[x_k,\hat{G}(x_{k-1},\ldots)]$, K is a kernel function (in our case a unit-variance Gaussian) and h is the kernel bandwidth [3]. Once $\hat{G}$ is learned this will become the distribution in our model.

Such a density can be used to estimate the entropy gradient [4, 5]. The method of [5] has several nice properties, such as a computational advantage and a term which discourages saturation which led us to select it for use. We then use this estimate to train our statistics, which we parameterize as single-layer network structures, i.e. the i^{th} statistic is given by

$$\hat{G}_i(x_{k-1},\ldots,x_{k-N})=\sigma\left(\sum_j w_{i,j}x_{k-j}\right) \tag{2}$$

where the $w_{i,j}$ are the network weights and $\sigma(\cdot)$ is the hyperbolic tangent function.

This simple form of $\hat{G}$ does not severely limit its capabilities; as discussed above, the modelling power of $\hat{p}$ compensates for lack of flexibility in $\hat{G}$. In addition, this technique is easily extended to multilayer perceptrons, allowing more complex functional approximation if desired.

4. GENERATIVE MODELS

Having captured the relation of the past to subsequent data points, our model can be used to synthesize sample realizations of the process from any point in time. However, the nonparametric nature of the modelled density raises the issue of how these points should be selected. It may be that this density is not unimodal, in which case the MSE estimate can be arbitrarily unlikely. Another choice, the ML estimate, results in sequences which are not typical [2]. In fact, in such cases the role of prediction is not clear, although sampling is. As a consequence of our complete model of uncertainty, synthesis – the task of generating multiple plausible patterns each of which displays the observed dynamics – is viable. Having modelled uncertainty means we also avoid mere repetition of observed data; this distinction is important in a number of applications such as image or audio tasks where humans are adept at discerning repetitive structure. To generate synthesis paths, we need only sample from the induced conditional distribution of the model. If we have accurately captured the relation and randomness inherent in the process this will preserve its observed structure.

5. DISCRIMINATIVE MODELS FOR PROCESS CLASSIFICATION

As described in [1], such a model also can be used to evaluate the likelihood of a new process sequence of unknown type under the learned dynamical model. When all hypotheses have been modelled, a simple evaluation and likelihood ratio test will allow us to discriminate. However, in the case that we wish to discriminate between a modelled process and a continuum of other possibilities which we are not able to model accurately, the question becomes more difficult. The problem of signature verification considered later is one

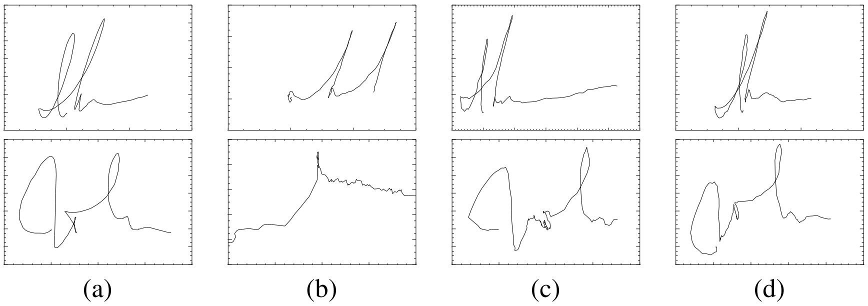


Fig. 2. (a) Example signature; (b) synthesis using only local (dx, dy) information; (c) local (x, y) information; (d) local (dx, dy) augmented by time

such case where it may not be possible to model "alternatives", i.e. forgeries.

However, it is still possible to test discriminatively. It can be shown that in the large-data limit the log-likelihood of the process $\mathcal{X}$ will approach its negative entropy rate $-\mathrm{H}(\mathcal{X})$. We can bound the deviation from this for a given acceptance probability and number of samples, or the required sample size for a given deviation and probability of acceptance. Any other process $\mathcal{Y}$ will approach $-\mathrm{H}(\mathcal{Y}) - \mathrm{D}(\mathcal{Y}\|\mathcal{X})$ where $\mathrm{D}(\cdot)$ is the KL-divergence between the two processes. This quantity may be larger than $-\mathrm{H}(\mathcal{X})$, when $\mathcal{Y}$ represents a "more probable than expected" version of $\mathcal{X}$; but $\mathcal{Y}$ which do not exhibit the same dynamics will generally be unlikely under $\mathcal{X}$'s model and so have a likelihood less than $-\mathrm{H}(\mathcal{X})$.

While we cannot characterize the probability of false-alarm without some model for "all other processes"; we can characterize the probability of rejecting a correct process $\mathcal{X}$ as a function of the number of samples and the acceptance region, or choose a region of acceptance given a fixed number of samples and an acceptable probability of incorrect rejection.

6. HANDWRITING AS A DYNAMICAL SYSTEM

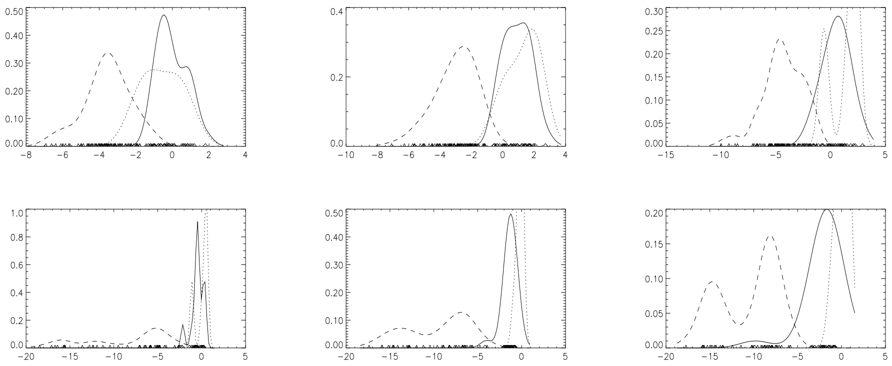
Handwriting represents a highly nonlinear system which exhibits both obvious structure and variability. Signatures are the most extreme example, being so consistent that we regularly use them as verification of identity yet random enough that no two look exactly alike. Even text of arbitrary content contains a considerable amount of information about its writer's identity. Regarding handwriting as a two-dimensional time series, we can then consider the problems of signature synthesis and recognition as problems of modelling this dynamical system.

To acquire our signature examples, we used a CrossPad digitizing tablet, which samples with equal time-spacing and a spatial resolution of 256 pixels/inch. Eight example signatures were taken from each of five subjects and resampled to have the same number of samples (that person's average, between 130 and 200 points); no further warping or feature-matching was performed. Informative statistics were then learned within the recent past, i.e. the previous ten (x, y) pairs using the methodology of [1]. Forgeries were used solely in the likelihood testing stage, not for training, and consisted of so-called "skilled" forgeries, wherein the forger is given access to copies of the true signature, time to practice, and knows that the dynamics of the motion will play a role. For testing, the new signature is resampled to the same length as those of the training set, and the accumulated log-likelihood of the new data conditioned on its statistics of the past is computed.

Inherent in any model is a choice of coordinate systems, and frequently its selection is quite relevant to the difficulty of the task. In this case there are two obvious coordinate systems: relative and differential. This illustrates an essential tradeoff in such problems, namely manually removing information which may or may not be extraneous in order to improve the volume of data available for the density estimate. If the removed information is truly extraneous this improves the estimate; if not it may create bimodalities which could otherwise be differentiated. For the moment we put this issue aside to present some results; we will return to it later.

In Figure 2 we see examples of a true signature and several synthesized sample paths. The first, given only differential information, may possess characteristics of handwriting but fails to capture the process. Essentially, this is due to a lack of context – it has insufficient information to disambiguate position within the word. The second, using relative (x, y) coordinates, possesses context but as we might expect has more consistency near $(0, 0)$ than at later positions which are dependent on long-term factors such as slant or size deviation. Finally, differential coordinates augmented by a time index have good consistency throughout but are not, for example, confined to a straight line.

This illustrates an earlier point. The dynamics of sign-



(x, y) (dx, dy) $(dx, dy) + t$

Fig. 3. Estimated PDF of average log-likelihoods for test signatures (solid), training (dotted), and forgeries (dashed) for "ihler" and "john"

ing are not stationary; it is only through conditioning on statistics which provide this context that we can approximate them as such. If the necessary context is present in the observations, it will be learned; but without it the model cannot hope to capture the true dynamics. Such context is clearly necessary for synthesis, but may be less vital for discrimination. Figure 3 shows a kernel estimate of the distribution of average likelihoods for three cases: signatures from the training set (excluding them from the density first), true signatures from the testing set, and forgery attempts. Notice that despite its small size the training set provides a reasonable estimate of the true likelihood distribution, while forgeries are in general found to be unlikely under the learned models.

It should be noted that this preliminary result was restricted to short, single stroke signatures. It is certainly possible to learn multiple stroke models.

7. CONCLUSIONS

We have shown the flexibility and capacity of our approach to model difficult dynamical systems involving nonlinear and nongaussian dynamics, even a degree of nonstationarity. The system model possesses the ability to synthesize realistic sample paths, and to characterize the likelihood of a new sequence. We discussed the use of these models for synthesis and hypothesis testing, even when no alternatives can be well characterized.

We then take a novel outlook to the handwriting recognition problem. Signature verification is a task in which we wish to differentiate between two hypotheses, but in reality data from only one is available to us. This perhaps makes a dynamical system model and likelihood evaluation uniquely suited for such a test, since we have the assurance that the better our model fits the true dynamics the more difficult a forgery will be. We demonstrate the capability to learn such a dynamical model, even from few examples, and show that it has captured dynamics in two ways – its capability of generating plausible new signatures, and in the clustering of test signatures around the estimated entropy rate. We also show that it is at least reasonably difficult to forge such a dynamic, even with practice.

8. ACKNOWLEDGEMENTS

The authors would like to thank Nick Matsakis, both for the use of his equipment and for all of his help with data aquisition.

9. REFERENCES

[1] J. W. Fisher, A. T. Ihler, and P. Viola, "Learning informative statistics: A nonparametric approach," in *Proceedings, NIPS 1999*.

[2] T. Cover and J. Thomas, *Elements of Information Theory*, John Wiley and Sons, Inc., New York, 1991.

[3] E. Parzen, "On estimation of a probability density function and mode," *Ann. of Math Stats.*, vol. 33, pp. 1065–1076, 1962.

[4] P. Viola, N.N. Schraudolph, and T.J. Sejnowski, "Empirical entropy manipulation for real-world problems," in *Proceedings, NIPS 1995*, pp. 851–7.

[5] J.W. Fisher and J.C. Principe, "A methodology for information theoretic feature extraction," in *Proceedings of the IEEE International Joint Conference on Neural Networks*, 1998.