An Architecture for Statistical Inference of Heterogenous Timeseries Data

Ryan Michael
kerinin@gmail.com

February 2, 2009

1 Introduction

1.1 General Overview

The goal is to create a method of statistical inference capable of processing timeseries data which is both multi-variate and exhibits different behaviors at different times. This type of data is common, and developing a robust method of analysis has applications in many domains. The general approach is to create a series of estimates using subsets of the observed data, and to then combine these estimates in an intelligent manner which captures the relevance of each estimate to the current prediction task. By using a set of 'typical' estimates, we are able to reduce the computational demands of the system, as each estimate is a condensed representation of the data from which it was derived. This approach also allows us to reduce data redundancy by only using distinct estimates.

The most basic operation used in this system is the estimation of probability densities. Based on a set of observations drawn from some random process, we generate an estimate of the underlying probability distribution. This estimate tells us the probability of each point in the input space being observed. Areas of the input space in which a dense set of observations are observed are given high probability, while areas of the input space with few observations are given low probability. This basic operation, in combination with some basic laws of statistics allow us to build the full inference system.

We assume that observations are pulled from multiple independent sources, all of which respond to some underlying phenomena. For instance one set of observations could be from a microphone and another from a light detector. We do not know how the inputs are related or to what extent their behavior changes over time. For example one input could be a steady sine wave and another could be based on the decay of a radioactive isotope. In the former

case previous observations of the source are useful, in the latter they're not. Alternately two inputs could be light detectors in the same room or they could be on different continents; in the former their input would be highly similar, in the latter not as much.

Our general strategy has three phases; generating estimates, correlating estimates, and applying estimates to a given prediction task.

1.2 Generating Estimates

Estimates are generated by picking a time window at random and only dealing with observations which take place in that window. Using these observations we create an estimate of the probability distribution underlying the observations (this distribution would include time as a variable). This process is repeated until we feel we have a reasonable sample of the system as a whole, at which time we can start to correlate the estimates.

1.3 Correlating Estimates

The goal of correlating estimates is to determine the relationships between estimates at different times and from different sources. Estimates are correlated by treating them as variables whose value for a given time window is determined by the extent to which the observed data corresponds to the estimate. For each time window there exist a set of estimates which have some value describing their accuracy at predicting the observed value. We can treat the accuracy measurement of each estimate as a multi-dimensional point, each dimension determined by an estimate's accuracy. Using a set of these points taken at different times, we create a probability distribution estimate. The domain of this probability distribution has the same dimensionality as the number of estimates we have generated. This probability density allows us to predict the probability of an estimate in of one source based on the probability of an estimate in of another source because frequent combinations of accuracy values will have higher probability than other combinations. This 'correlation density' also tells us which estimates are most commonly observed - information we can use in conjunction with estimate similarity to determine which estimates to use and which to discard.

1.4 Making Predictions

Once the estimates have been correlated, we are able to generate predictions. For simplicity, we'll assume that the first two phases occur on as set of 'training' data which is representative of the underlying data, while prediction takes place continuously using new sets of observations which occur in some time window. We make predictions for a given source by combining the existing estimates we

have for that source based on their accuracy at predicting the given observations. Because we have determined the correlation between estimate accuracy of different sources, we can use other sources to refine our confidence in each estimate of the given source; the influence of estimates of the given source which do not correspond to a likely 'point' in the correlation density are suppressed while the influence of estimates which correspond to likely points in the correlation density are enhanced.

2 Problem Setting

We begin with a hidden random variable

$$X = (\Omega, \mathcal{F}, \mathcal{P})$$

Our knowledge of X comes from a set of independant sources which we treat as random variable generated by X:

$$X^n : \Omega \mapsto \Omega^n \in \mathbb{R}^d \times \mathbb{R}_+$$
$$X^n = (\Omega^n, \mathcal{F}^n, \mathcal{P}^n)$$
$$\mathbf{X} = [X^0, ..., X^c]$$

We refer to each of these sources as a channel, and refer to each channel as the i^{th} element of the set $\mathbf{X}$:

$$C^n = [\Omega^n, \mathcal{F}^n, \mathcal{P}^n]$$

For each channel we are given a set of ℓ observations of dimension d, each with a time value:

$$X^n = [(x_0^n, t_0^n), ..., (x_\ell^n, t_\ell^n)] \in \mathbb{R}^d \times \mathbb{R}_+$$

We assume that the probability of X is a time-dependent mixture of some set of distributions whose influence is unknown and changes over time. We can refer to the specific mixture which corresponds to a given time window $\mathcal{T}(t, \theta) = [t_i : t \leq t_i < t + \theta]$ as

$$F(t, \theta) = \sum_i \delta_i \cdot f_i$$

Finally, we assume that a similar mixture of densities can be determined for each channel:

$$F(t, \theta)^n = \sum_i \delta_i^n \cdot f_i^n$$

3 Single Variable w/ Memory

3.1 Estimation

We begin by considering the case where only one channel exists, so for now will omit the superscript and refer to $F^n = (\Omega^n, \mathcal{F}^n, \mathcal{P}^n)$ as $F = (\Omega, \mathcal{F}, \mathcal{P})$. Our task is to determine both a set of underlying distributions $[f_0, ..., f_i]$ and their relative influence over time. We begin by estimating a set of probability distributions Φ over various windows $[(t, \theta) \in \mathcal{T}]$ using the samples which fall into the window using some function ψ_E:

$$\begin{aligned} S_{(t,\theta)} &= [(x_i, \frac{t_i - t}{\theta}) \in X : t \leq t_i < t + \theta] \\ \psi_E(S_{(t,\theta)} : \alpha_E) &\mapsto \phi_n \simeq F(t, \theta) \\ \Phi &= [\phi_0, ..., \phi_p] \end{aligned}$$

The estimation algorithm $\psi_{estimate}$ operates on the set of samples X_n in a given window $\mathcal{T}_n$ and produces an estimate ϕ_n of the probability distribution F of the random variable X localized at the window $\mathcal{T}_n$. Multiple algorithms exist to accomplish such density estimations this, so details will be omitted.

3.2 Correlation

The entropy value defines a new random variable, for which we will compute a density estimate using ψ_C as we have with the input data previously:

$$\begin{aligned} H_u(S_{(t,\theta)}) \mapsto \delta_v &= -\sum_{(x,t) \in S_{(t,\theta)}} \phi_u(x, t) \log \phi_u(x, t) \\ \Delta &= [\delta_0, ..., \delta_q] \\ \psi_C(\Delta, \alpha_C) &\mapsto \phi_\Delta \\ \phi_\Delta(S_{(t,\theta)}, \delta_v) &\simeq Pr(H_u(S_{(t,\theta)}) = \delta_v) \end{aligned}$$

3.3 Prediction

The derivation algorithm produces predictions of X over a time window (t, θ) based on a set of observations which occur in that time window $S_{(t,\theta)}$. Rather than using the estimation algorithm ψ_E to predict these values, however, the derivation algorithm ψ_D predicts the probability distribution using bayesian modification of existing estimates. To accomplish this, we use the entropy value of the conditional probability given $S_{(t,\theta)}$

$$Pr(X = (x,t)|H_n(S_{(t,\theta)} = \delta_u) = \frac{Pr(H_n(S_{(t,\theta)}) = \delta_u|X = (x,t)) \cdot Pr(X = (x,t))}{Pr(H_n(S_{(t,\theta)} = \delta_u)}$$

We can easily determine one of these terms:

$$Pr(X = (x,t)) = \prod_{0 \leq i < q} \phi_i(x,t)$$

Because the entropy is calculated as a sum of the entropy of each observation, we can determine the conditional probability $H_n(S_{(t,\theta)}) = \delta_u|X = (x,t)$ by subtracting the entropy of $H_n((x,t))$ from the value of δ_u, and then using our correlation estimate ϕ_Δ to predict the probability of $\delta_u - H_n((x,t))$. We include the expanded form of $Pr(H_n(S_{(t,\theta)}))$ to clarify the cancellation of terms:

$$\begin{aligned}
H_n(X) \mapsto \delta &= -\sum_{x \in X} \phi(x) \log \phi(x) \\
Pr(H_n(S_{(t,\theta)}) = \delta_u) &= \frac{|y : H_n(y) = \delta_u|}{|y : H_n(y) \neq \delta_u|} \\
Pr(H_n(S_{(t,\theta)}) = \delta_u|X = S_{(t,\theta)}) &= \frac{|y : H_n(y) = \delta_u + H_n((x,t))|}{|y : H_n(y) \neq \delta_u|} \\
\frac{Pr(H_n(S_{(t,\theta)}) = \delta_u|X = S_{(t,\theta)})}{Pr(H_n(S_{(t,\theta)}) = \delta_u)} &= \frac{|y : H_n(y) = \delta_u + H_n((x,t))| \cdot |y : H_n(y) \neq \delta_u|}{|y : H_n(y) = \delta_u| \cdot |y : H_n(y) \neq \delta_u|} \\
&= \frac{Pr(H_n((x,t)) = \delta_u + H_n((x,t)))}{Pr(H_n((x,t)) = \delta_u)}
\end{aligned}$$

The derivation algorithm can easily be extended to multiple windows by using joint entropy probabilities. Because we do not know anything about the conditional relationships between estimates, we must assume they are i.i.d., and that their joint probability is equal to the sum of the marginal probabilities.

$$Pr(A|B \cap C) = \frac{Pr(B \cap C|A) \cdot Pr(A)}{Pr(B \cap C)}$$

$$Pr(X = (x,t) \mid H_n(S_{(t,\theta)}) = \delta_u \ \cap \ H_m(S_{(t,\theta)}) = \delta_v) =$$

$$\frac{Pr(H_n(S_{(t,\theta)}) = \delta_u \ \cap \ H_m(S_{(t,\theta)}) = \delta_v \mid X = (x,t)) \cdot Pr(X = (x,t))}{Pr(H_n(S_{(t,\theta)}) = \delta_u \ \cap \ H_m(S_{(t,\theta)}) = \delta_v)}$$

$$Pr(H_n(S_{(t,\theta)}) = \delta_u \ \cap \ H_m(S_{(t,\theta)}) = \delta_v | X = (x,t)) =$$

$$\prod_{i=[n,m],j=[u,v]} Pr(H_i(S_{(t,\theta)}) = \delta_j + H_i((x,t)))$$

$$= \prod_{i=[n,m],j=[i,j]} \phi_\Delta(S_{(t,\theta)}, \delta_j + H_i((x,t)))$$

$$Pr(H_n(S_{(t,\theta)}) = \delta_u \ \cap \ H_m(S_{(t,\theta)}) = \delta_v) =$$

$$\prod_{i=[n,m],j=[u,v]} Pr(H_i(S_{(t,\theta)}) = \delta_j)$$

$$= \prod_{i=[n,m],j=[u,v]} \phi_\Delta(S_{(t,\theta)}, \delta_j)$$

Which gives us the following as our prediction algorithm:

$$\psi_P((x,t) : S_{(t,\theta)}, \Phi) \mapsto \mathbb{P}$$

$$= \frac{\prod_{0 \le n < q} \phi_\Delta(S_{(t,\theta)}, H_n(S_{(t,\theta)})) \cdot \prod_{0 \le n < q} \phi_n((x,t))}{\prod_{0 \le n < q} \phi_\Delta(S_{(t,\theta)}, H_n(S_{(t,\theta)}) + H_n((x,t)))}$$

$$= \prod_{0 \le n < q} \phi_n((x,t)) \cdot \frac{\phi_\Delta(S_{(t,\theta)}, H_n(S_{(t,\theta)}))}{\phi_\Delta(S_{(t,\theta)}, H_n(S_{(t,\theta)}) + H_n((x,t)))}$$

4 Multiple Variables w/ Memory

Algorithms for determining probability density functions tend to scale exponentially with the dimensionality of the input data. For this reason it would be helpful if the algorithm could operate on independent channels of data and only calculate relationships between channels in cases where the channel's probability distribution is conditional on such relationships.

Extending the existing theory to this situation, return to our original setting of the problem:

$$
\begin{aligned}
X^n : \Omega &\mapsto \Omega^n \in \mathbb{R}^d \times \mathbb{R}_+ \\
X^n &= (\Omega^n, \mathcal{F}^n, \mathcal{P}^n) \\
\mathbf{X} &= [X^0, ..., X^c] \\
C^A &= [\Omega^A, \mathcal{F}^A, \mathcal{P}^A]
\end{aligned}
$$

We will now refer to observations, and estimates using superscript to denote their channel. We previously developed three generalized algorithms, ψ_E, ψ_C, ψ_D; we will now extend each one to handle multiple channels.

4.1 Estimation

The estimation algorithm is unchanged in the context of multiple channels. Each channel generates its independent estimates over independent time windows. The estimation process is intended to give each channel an understanding of its own behavior at a specific time window, and as such the generation of estimates does not rely on previous behaviors of the channel or the behavior of other channels. We therefore re-write the estimation algorithm to reflect the new notation:

$$
\begin{aligned}
S^A_{t,\theta} &= [(x^A_i, \frac{t_i - t}{\theta}) \in X^A : t \leq t_i < t + \theta] \\
\psi_E(S^A_{t,\theta} : \alpha^A_E) &\mapsto \phi^A_n \ \simeq \ F^A(t, \theta) \\
\Phi^A &= [\phi^A_0, ..., \phi^A_p]
\end{aligned}
$$

4.2 Correlation

The correlation algorithm is where the most substantial changes must be made to accomodate multiple channels. The biggest difference from the single-channel approach is that each channel takes place in an independent abstract space. Recall that a critical component of the prediction algorithm is the evaluation of the entropy of a set of observations given an estimate:

$$
H_n(S_{t,\theta}), H_n(X) = -\sum_{x \in X} \phi_n(x) \log \phi_n(x)
$$

If the probability density $\phi(X)$ operates over a different abstract space Ω than the observations $S_{t,\theta}$ are taken from, we cannot calculate the entropy. In other

words, we cannot evaluate the entropy of an estimate from channel C^B using observations of channel C^A.

To address this, we begin with the observation that all channels share the time dimension t as part of their abstract space Ω^A. This allows us to specify a uniform time window (t, θ) for all channels $C^A \in \mathbf{C}$, and to then evaluate the entropy H_n^A of each estimate $\phi_n^A \in \Phi^A$ given the subset of each channel's observations $S_{t,\theta}^A$ for the time window. Given the time window (t, θ), we can treat these entropy measurements as a coherent set and interpret them as a multi-dimensional random vector in much the same way we treated different time windows in the single-channel case.

Using the same algorithm ψ_C as we used previously, we can estimate the probability density ϕ_Δ of this random vector.

$$\Delta^A = [\delta_0^A, ..., \delta_p^A]$$
$$\Delta = [\Delta^i \forall i]$$
$$\psi_C(\Delta, \alpha_C) \mapsto \phi_\Delta$$

4.3 Prediction

We begin by considering the situation in which we wish to make predictions on channel C^A using information from another, C^B. The derivation algorithm therefore produces predictions of X^A over a time window (t, θ) based on a set of observations which occur in that time window $S_{t,\theta}^B$. To do so we select estimates in both channels:

$$C^A \rightarrow \phi_n^A$$
$$C^B \rightarrow \phi_m^B$$

We calculate the entropy of each estimate in its own context:

$$\delta^A = H_n^A(S_{t,\theta}^A)$$
$$\delta^B = H_m^B(S_{t,\theta}^B)$$

Then calculate the probability of each point in X^A based on the joint probability of the entropy values $Pr(\delta^A \cap \delta^B)$.

$$Pr(X^A = (x,t)|H_n^A = \delta^A \cap H_m^B = \delta^B) =$$
$$\frac{Pr(H_n^A = h^A \cap H_m^B = \delta^B|X^A = (x,t)) \cdot Pr(X^A = (x,t))}{Pr(H_n^A = h^A \cap H_m^B = \delta^B)}$$

This time, we can easily determine two of these terms:

$$Pr(X^A = (x,t)) = \prod_{0 \leq i < q} \phi_i^A(x,t)$$
$$Pr(H_n^A = \delta^A \cap H_m^B = \delta^B) = \phi_\Delta\Big(\{A : \delta^A, B : \delta^B\}\Big)$$

Which leaves us with the process of calculating the effect of $X^A = (x,t)$ on the entropy of ϕ_n^A and ϕ_m^B. We adopt the same process used in the single-channel case, replacing δ^A with $\delta^A + H_n^A((x,t))$, and recalculating the joint probability. Since ϕ_m^B is in a different abstract space, we cannot make any useful statements about the dependence of its entropy given a point outside the abstract space being predicted. We can now determine the third of the three terms:

$$Pr(H_n^A = \delta^A \cap H_m^B = \delta^B|X^A = (x,t)) = Pr(H_n^A = \delta^A + H_n^A((x,t)) \cap H_m^B = \delta^B)$$
$$= \phi_\Delta\Big(\{A : \delta^A + H_n^A((x,t)), B : \delta^B\}\Big)$$

To state the problem in a more general setting, let us assume we have a set of channels $\mathbf{C}$ with an associated set of estimates $\mathbf{\Phi}$ and we are attempting to predict the values of $X^0; C^0 \in \mathbf{C}$ based on a set of observations $\mathbf{S}_{t,\theta}$ over all the channels:

$$\mathbf{C} = [C^0, ..., C^c]$$

$$\mathbf{\Phi} = [\Phi^0, ..., \Phi^c]$$
$$\mathbf{S}_{t,\theta} = [S_{t,\theta}^0, ..., S_{t,\theta}^c]$$

$$\psi_D((x^0,t) : \mathbf{S}_{t,\theta}, \mathbf{\Phi}) \mapsto \mathbb{P} = Pr(X^0 = (x^0,t)|\mathbf{S}_{t,\theta})$$

$$= \prod_{0 \leq n < q} \phi_n^0((x^0,t)) \cdot \frac{\phi_\Delta\Big(\Big\{0 : H_n^0(S_{t,\theta}^0)\Big\} \cup \Big[\Big\{c : H_m^c(S_{t,\theta}^c)\Big\} : \forall m, \forall c\Big]\Big)}{\phi_\Delta\Big(\Big\{0 : H_n^o(S_{t,\theta}^0) + H_n^0((x,t))\Big\} \cup \Big[\Big\{c : H_m^c(S_{t,\theta}^c)\Big\} : \forall m, \forall c)\Big]\Big)}$$

5 Optimization

The system as described thus far makes several assumptions for simplicity which would lead to unnecessary computational performance. We now spend some time discussion methods of reducing the computational demands of the system. We will look at each component of the system in turn.

5.1 Estimation Optimization

The generation of estimates has been described as a process of selecting time windows at random from the full set of training data. Selecting windows at random is not necessary to satisfy the i.i.d. requirements of producing an accurate probability estimate, so long as the selection of time windows for which entropy is calculated is random. Let's consider the characteristics of a 'good' set of estimates for prediction.

The most obvious characteristic of a 'good' estimate is that it has a low average entropy, which is to say that it frequently captures the behavior of observed data. If an estimate is not applicable to the observed data, its influence on predictions will be marginal, and the computational time required to calculate its entropy and influence on a prediction will have been wasted.

Another characteristic of a 'good' set of estimates is that they have minimal redundancy. Again, if two estimates are essentially identical, their influence on prediction tasks will be identical, and the computational demands of evaluating their entropy and influence will be wasted.

Unfortunately, these two parameters will likely be mutually exclusive, so some method of balancing them is required.

5.2 Correlation

It is critical that correlation takes place using i.i.d. data in order to produce accurate probability estimates. We had previously assumed that each time window used to generate estimates would be used to correlate the estimates, however it is not necessary that the time windows used to determine the entropy values used to correlate estimates correspond to the time windows used to generate estimates. This allows us to select a number of time windows to use which balances the computational demands of density estimation with the need for accuracy.

5.3 Prediction

For prediction, it is not necessary to actually use each of the estimates in generating predictions, so long as we have accurate estimates of their individual probabilities and the joint probability of any two estimates. Since we must evaluate a number of joint probabilities equal to the square of the number of estimates being considered, selecting a useful subset of the defined estimates can provide substantial reductions in computational demands.

Our objective is to consider only estimates which will be relevant to the prediction task. If we restrict our attention to a single channel, this can be done by selecting a subset of estimates which have the lowest entropy in the context of the prediction task. Extending this approach to a situation with multiple channels becomes slightly more complex, as the influence of a given estimate depends not only on its entropy, but also on the entropy values of predictions in other channels. We cannot restrict our attention to the set of all estimates below a given entropy, as with the single-channel case, because an estimate's influence may be increased based on a high entropy value of an estimate in another channel.

To restrict the estimates which are used in the multi-channel situation, we must establish the extent to which each estimate's probability correlates with each other estimate's. If a given estimate's probability is only marginally dependent on another's, we can omit the other estimate from the set considered.