MUTUAL INFORMATION BETWEEN RANDOM PROCESSES FROM HIGH DIMENSIONAL DATA

Victor Solo

Department of Electrical Engineering and Computer Science
University of Michigan, Ann Arbor
Ann Arbor, MI, 48109
email: vsolo@umich.edu

ABSTRACT

A number of signal estimation problems are arising where a relatively low dimensional state is to be estimated from a high dimensional observation sequence. In previous work we have shown this leads to considerable simplification in the structure of optimal state estimators even in non-linear problems. In these and other state estimation problems there is a growing interest in the computation of mutual information between unobserved state and observed sequence. Here we show that the mutual information computation can be likewise considerably simplified.

1. INTRODUCTION

Where once, but a few disciplines such as astronomy, medical imaging and geophysics were plagued by large data problems, today the problem is ubiquitos. In computer vision, image sequence data generates large dimensional observations of low dimensional phenomena (such as rigid body rotation)[1],[2]. In neuroscience direct multielectrode recordings of spiking neurons from tens and even hundreds of brain cells can now be made all relating to a single phenomenon, (such as the position and velocity of an arm) of low state dimension [3]. And even in econometrics there are now available scores of daily stock price series going back to 1960, all relating to a 'market index' [4].

In previous work [5] we have shown that under reasonable conditions the considerable information available in the high-dimensional observation sequence allows optimal filters to be approximated with much simpler static estimators with almost no loss of statistical efficiency.

Given the growing interest in computation of mutual information in a number of areas, including those mentioned above, we investigate these computational simplifications for mutual information estimation in nonlinear state estimation problems. In contrast to much recent work on the problem of estimating mutual information our results are developed for correlated random processes not just independent sequences.

In sections 2 we review our earlier work on the problem; for lack of space treating only nonlinear state models. In section 3 we discuss computation of mutual information for a nonlinear state estimation problem. Conclusions are offered in section 4.

2. NON-LINEAR FILTERS WITH HIGH DIMENSIONAL POINT PROCESS OBSERVATIONS

In our previous work [5] we derived results for linear state space models estimated with the Kalman filter and then extended the results to the nonlinear case using Laplace asymptotics. Lack of space precludes a full development here so we proceed directly to the more interesting nonlinear case. It proves most efficient to repeat some of the Laplace asymptotics from [5]. This not only provides clarity, and allows for further comments but also facilitates a very quick development of the new mutual information results in the next section. For lack of space and because it is the motivating application for this work, only the case of point process observations is treated. Results for analog observations are attainable with a similar development.

We treat a continuous time non-linear state space model but for simplicity suppose time is discretised into tiny intervals of extent δ so that in each tiny interval > 1 events occur with probability $o(\delta)$. In neural coding practice we can think of δ as being of order 1ms. This is not to be confused then, with more usual binning of order 30ms-100ms [3] which will lead to degradation of the estimators.

The associated discretised non-linear state space model is (with $x_k = x(k\delta)$): **State Equation**

$$x_{k+1} = [x_k + \delta f_c(k\delta, x_k)] + n_k\sqrt{\delta} = f_k(x_k) + n_k\sqrt{\delta} \quad (2.1)$$

where n_k is a Gaussian white noise (independent of past observations and past states) with conditional variance $Q_k(x_k) = Q(k, x_k)$. The observations are a vector of conditional (possibly inhomogeneous) Poisson process observations from p cells and we consider discrete time (binned) observations,

$N_k^{(\delta,r)} = \#$ events in $k\delta, (k\delta+\delta), r = 1, \cdots, p$ which we can write as: **Observation Equation**

$$N_k^{(\delta,r)} = \lambda^r(k\delta, x_k)\delta + m_k^r = \lambda_k^{(r)}(x_k)\delta + m_k^r, r = 1, \cdots, p$$

c.f.[6]. We can think of m_k^r as a white noise of conditional variance $\lambda_k^{(r)}(x_k)\delta$. Also denote $N_k^\delta = (N_k^{(\delta,1)}, \cdots, N_k^{(\delta,p)})$ as well as

$$N_{1,k}^{(r)} = (N_1^{(\delta,r)}, \cdots, N_k^{(\delta,r)}) \Rightarrow N_{1,k} = (N_{1,k}^{(1)}, \cdots, N_{1,k}^{(p)})$$

We wish initially to approximate the filtered and predicted state estmators $\hat{x}_{k|k}, \hat{x}_{k|k-1}$ and their corresponding error covariances $P_{k|k}, P_{k|k-1}$. We deal with $\hat{x}_{k|k}, P_{k|k}$ and quote results for the other two whose details will be given elsewhere.

We need to approximate the conditional density $p(x_{k+1}|N_{1,k+1})$. We have

$$p(x_{k+1}|N_{1,k+1}) = \frac{\overline{\rho}(x_{k+1}, N_{1,k+1})}{P(N_{1,k+1})}$$

where $\overline{\rho}(\cdot)$ is a joint density and $P(\cdot)$ is a marginal density. Using the observation equation we can write (with $p(\cdot|\cdot)$ a conditional density)

$$\overline{\rho}(x_{k+1}, N_{1,k+1}) = p(N_{k+1}^\delta|x_{k+1})\rho(x_{k+1}, N_{1,k})$$

Simple conditional Bernoulli calculations give then (where $P_*(\cdot)$ is a reference density for a unit rate Poisson model)

$$\begin{aligned}
&\overline{\rho}(x_{k+1}, N_{1,k+1}) = e^{-U_{k+1}(x_{k+1})}\rho(x_{k+1}, N_{1,k}) \\
&\times P_*(N_{k+1}^\delta) \\
&\quad P_*(N_{k+1}^\delta) = \Pi_1^p P_*(N_{k+1}^{(\delta,r)}) \\
&\quad P_*(N_{k+1}^{(\delta,r)}) = \Pi_1^{k+1}[\delta^{N_i^{\delta,r}} e^{-\delta}/N_i^{\delta,r}!] \\
&U_{k+1}(x) = -\Sigma_1^p[N_{k+1}^{(\delta,r)} log\lambda_{k+1}^{(r)}(x) - (\lambda_{k+1}^r(x) - 1)\delta]
\end{aligned}$$

The conditional mean estimator is then

$$\begin{aligned}
\hat{x}_{k+1|k+1} &= E(x_{k+1}|N_{1,k+1}) \\
&= \frac{\int x_{k+1} e^{-U_{k+1}(x_{k+1})}\rho(x_{k+1}, N_{1,k})dx_{k+1}}{\int e^{-U_{k+1}(x_{k+1})}\rho(x_{k+1}, N_{1,k})dx_{k+1}}
\end{aligned}$$

We approximate these integrals using Laplace asymptotics [7] based on the fact that $U_{k+1}(x)$ grows without bound as $p \to \infty$ and the integral will be dominated by behavior near its minimum. For this to work we make a 'high-dimensional' assumption:

HD. $\Sigma_1^p \lambda_{k+1}^r(x) \to \infty, as, p \to \infty$.

Note then that, as $p \to \infty$,

$var(\Sigma_1^p(N_{k+1}^{(\delta,r)} - \lambda_{k+1}^r(x)\delta) = \Sigma_1^p \lambda_{k+1}^r(x)\delta \to \infty$

This means $U_{k+1}(x) \to \infty$ in probability as $p \to \infty$. So we can apply Laplace asymptotics which leads us to introduce the static estimator

$$x_{k+1}^* = arg.min.U_{k+1}(x) \tag{2.2}$$

which obeys: $U_{k+1}^{'}\big|_{x=x_{k+1}^*} = 0$, where

$$U_{k+1}^{'} = -\Sigma_1^p(N_{k+1}^{(\delta,r)} - \lambda_{k+1}^r(x)\delta)\frac{dln\lambda_{k+1}^r(x)}{dx}$$

We now approximate $U_{k+1}(x)$ in the conditional mean integrals by a second order Taylor series about x_{k+1}^*. The integrals become Gaussian integrals and evaluating them yields

$$\begin{aligned}
\hat{x}_{k+1|k+1} &\approx \frac{x_{k+1}^*\psi(x_{k+1}^*)}{\psi(x_{k+1}^*)} = x_{k+1}^* \\
\psi(x) &= e^{-U_{k+1}(x)}\rho(x, Y_{1,k})|U_{k+1}^{''}(x)|^{-\frac{1}{2}}
\end{aligned}$$

We similarly find: $P_{k+1|k+1} \approx (U_{k+1}^{''}(x_{k+1}^*))^{-1}$.

Note that x_k^* is a maximum likelihood estimator based on an inhomogeneous Poisson model. The bracketed term is then a sample Fisher information.

And we then obtain the remarkable result (given first in [5]),

Result 1a. Under HD, the optimal mean square non-linear filter is well approximated by the static non-linear regression estimator (2.2).

In a similar way it can be shown ([5]),

Result 1b. Under HD, the optimal mean square one step ahead predictor is well approximated by the static predictor which achieves, asymptotically (as $p \to \infty$) the same state error variance that would be obtained were the state observed,

$$\hat{x}_{k+1|k} \approx f_k(x_{k+1}^*), P_{k+1|k} \approx Q_k(x_{k+1}^*)$$

To sum up again, with high dimensional non-redundant data the mean square optimal nonlinear filter is no better than an instantaneous population coding (static) maximum likelihood estimator. Further, we now see that to predict future state values using current data, we just advance the (static) maximum likelihood estimator according to the nonlinear dynamics. And we then obtain the same performance we would obtain had we observed the state (and used the prediction $f_k(x_k)$)! Finally we must point to the influence of δ. The smaller is δ the smaller is $U_{k+1}(x)$. So as δ increases the Laplace asymptotics becomes more accurate.

Computation.

Here we offer comments additional to those in [5]. The maximum likelihood estimator must be found iteratively and a Newton algorithm is a natural choice. A good start value is available from the previous timestep. If we then try to shorten things by identifying iteration with time we get an update

$$\overline{x}_{k+1} = \overline{x}_k - (U_{k+1}^{''}(\overline{x}_k))^{-1}U_{k+1}^{'}(\overline{x}_k)$$

We can further replace $U_{k+1}^{''}(\overline{x}_k)$ by its sample scoring approximation (to ensure postive definiteness) $\Gamma_k(\overline{x}_k)$

$$\Gamma_k(x) = \delta\Sigma_1^p \frac{\partial\lambda_{k+1}^{(r)}(x)}{\partial x}\frac{\partial\lambda_{k+1}^{(r)}(x)}{\partial x^T}\frac{1}{\lambda_k^{(r)}(x)}$$

we call the resulting algorithm the pseudo-scoring algorithm. On the other hand we can consider a standard approximate filter such as the extended Kalman filter (EKF) [8]. This is given by a time-varying version of the Kalman filter equations listed above with e.g. $F_k = I + \delta \frac{\partial f_c(k\delta, \hat{x}_k)}{\partial \hat{x}_k}$. If we drop the $O(\delta)$ term here we can show the EKF becomes

$\hat{x}_{k+1} = \hat{x}_k - P_k U_{k+1}^{'}(\hat{x}_k)$
$P_k = (P_{k|k-1}^{-1} + \Gamma(\hat{x}_k))^{-1}$
$P_{k|k-1} = P_{k-1} + Q_k\delta$

Now following the argument above we could replace this with a static estimator i.e. replace P_k by $(\Gamma_k(\hat{x}_k))^{-1}$. We thus recover the pseudo-scoring algorithm!

3. MUTUAL INFORMATION

Mutual information and entropy computation has become of great interest in the neural information processing area [3] although as [9],[10] discusses not always with a solid justification. Our concerns here are in direct estimation of mutual information for its own sake and capacity plays no role. We now consider calculation of the mutual information between the spike train and the state trajectory. The only previous attempt at such a computation is due to [11] who use Gaussian approximations which lack a full justification. Also [12] have discussed computation of bounds. Elsewhere in the neural information processing literature the current approach to mutual information computation (for single spike trains) is by brute force calculation based on the defintion [3],[13],[14],[15],[16],[17]. This requires large amounts of data including particularly many trials (or replicates).

However there is another approach, which is to implement formulae for mutual information expressed in terms of underlying stochastic intensity functions. For a single spike train a formula (for entropy) was first given by [18] and for the problem of interest here, such a formula has been developed by [19],[20].

In recent work of the author [21] the results of [19],[20] have been rederived in a much simpler way and also new formulae developed in more complex situations. But the effect of high-dimensional data is not addressed at all in [21].

However in order to help motivate the results we state a new result (not derived in [21]).

Result 2: Mutual Information for Partially Observed Nonlinear State Space Systems.
Given an observation interval $(0, T)$ denote $T = n\delta$, $X_{1,n} = (x_1, \cdots, x_n)^T$ and use $H(\cdot|\cdot)$ for conditional entropy, then

$$\mathcal{I}(N_{1,n}, X_{1,n}) = \Sigma_1^n H(N_k^\delta | N_{1,k-1}) - \Sigma_1^n H(N_k^\delta | x_k) \quad (3.1)$$

Proof: from the definition, $\mathcal{I}(N_{1,n}, X_{1,n}) = H(N_{1,n}) - H(N_{1,n}|X_{1,n})$. Using the chain rule [22] gives

$H(N_{1,n}) = \Sigma_1^n H(N_k^\delta | N_{1,k-1})$

which gives the first term of the result. And so we turn to the second term. The conditional chain rule gives

$$\begin{aligned} H(N_{1,n}|X_{1,n}) &= H(N_n^\delta, N_{1,n-1}|X_{1,n}) \\ = H(N_{1,n-1}|X_{1,n}) &+ H(N_n^\delta | N_{1,n-1}, X_{1,n}) \end{aligned}$$

Now from the observation equation, the conditional distribution of N_k^δ given past states and past observations depends only on the most recent state and so further the conditional distribution of $N_{1,n-1}$ given $X_{1,n-1}$ is fully determined. Thus we find by direct calculation from the last expression

$$\begin{aligned} H(N_{1,n}|X_{1,n}) &= H(N_{1,n-1}|X_{1,n-1}, x_n) + H(N_n^\delta | x_n) \\ &= H(N_{1,n-1}|X_{1,n-1}) + H(N_n^\delta | x_n) \end{aligned}$$

and summing up gives the result. This result seems to be new although a related result has been given by [23]. We did not use any point process properties in the derivation so Result 2 applies to analog observations as well.

With some extra work (essentially described in [21]) we obtain the discretised version of the result of Bremaud,

$$\mathcal{I}(N_{1,n}, X_{1,n}) = \quad (3.2)$$
$$\delta \Sigma_1^n \Sigma_1^p [E(\lambda_k^{(r)}(x_k) ln \lambda_k^{(r)}(x_k)) - E(\hat{\lambda}_k^{(r)} ln \hat{\lambda}_k^{(r)})]$$
$$\text{where } \hat{\lambda}_k^{(r)} = \hat{\lambda}_{(k\delta)}^{(r)} = E(\lambda_k^{(r)}(x_k)|N_{1,k-1}) \quad (3.3)$$

The two terms correspond to those in (3.1). Now we use Laplace asymptotics to approximate the $\hat{\lambda}_k^{(r)}$ terms. Now using the Markov property of the state model

$\hat{\lambda}_{k+1}^{(r)} = \int \lambda_{k+1}^{(r)}(x_{k+1}) p(x_{k+1}|N_{1,k}) dx_{k+1}$
$= \int \lambda_{k+1}^{(r)}(x_{k+1}) (\int p(x_{k+1}|x_k) p(x_k|N_{1,k}) dx_k) dx_{k+1}$

Now applying Laplace asymptotics to the inner integral as before yields

$\hat{\lambda}_{k+1}^{(r)} \approx \int \lambda_{k+1}^{(r)}(x_{k+1}) p(x_{k+1}|x_k^*) dx_{k+1}$

But now in view of (2.1) for small δ, $p(x_{k+1}|x_{k+1}^*)$ behaves like a Dirac-delta concentrated on $x_{k+1} = f_k(x_k^*)$ and using this yields $\hat{\lambda}_{k+1}^{(r)} \approx \lambda_{k+1}^{(r)}(f_k(x_k^*))$. Thus we obtain:

Result 3: Approximate Mutual Information. Under HD

$$\mathcal{I}(N_{1,n}, X_{1,n}) \approx \delta \Sigma_1^n \Sigma_1^p E(A_k^{(r)} - B_k^{(r)}) \quad (3.4)$$
$$A_k^{(r)} = \lambda_k^{(r)}(x_k) ln \lambda_k^{(r)}(x_k) \quad (3.5)$$
$$B_k^{(r)} = \lambda_{k+1}^{(r)}(f_{k-1}(x_{k-1}^*)) ln \lambda_{k+1}^{(r)}(f_{k-1}(x_{k-1}^*))$$

This is much simpler to calculate than (3.2) which requires (via (3.3)) generation of the full conditional density of the state given the observations. Simple Monte-Carlo averaging can be used to get the expectations in (3.4). But for the second term, which is data dependent, this will still be demanding.

But a much simpler idea occurs, namely to estimate $\mathcal{I}_n = \mathcal{I}(N_{1,n}, X_{1,n})$ by the sum $\hat{\mathcal{I}}_n = \delta \Sigma_1^n \Sigma_1^p \mathcal{I}_k^{(r)}$, $\mathcal{I}_k^{(r)} = (E(A_k^{(r)})) - B_k^{(r)})$. To see why this might work suppose $\lambda_k(x), f_k(x)$ are not time-varying i.e. are only functions of x. Then under reasonable assumptions, $E(A_k^{(r)}) - B_k^{(r)}$

will be stationary in k. Now $\hat{\mathcal{I}}_n$ has mean $\mathcal{I}_n$ which is of order $\delta np = Tp$ while $var(\hat{\mathcal{I}}_n) = \delta^2 \Sigma\Sigma cov(\mathcal{I}_k^{(r)}, \mathcal{I}_{k'}^{(r')})$ will be of order $\delta^2 pn = \frac{p}{n}T^2$ provided the correlations are not too high. So the estimator precision $= \frac{\text{mean}}{\text{standard error}}$ will be of order $= \frac{Tp}{\frac{\sqrt{p}}{\sqrt{n}}T} = \sqrt{np}$ i.e. large. Finally $E(A_k^{(r)})$ can be easily estimated by Monte-Carlo averaging since it does not depend on the observed data. This would then allow estimation of mutual information off a single realization.

4. SUMMARY

In this paper we have discussed state and information estimation from high-dimensional data. We had previously shown that high-dimensional observations yields significant simplification with optimal state estimators collapsing to simple static estimators. In a nonlinear model with point process observations we have here interpreted these static estimators as instantaneous maximum likelihood population coding schemes and have also further elaborated on their computation. Further we have shown that an associated model based computation of mutual information between observed spike train and unobserved state can be approximated in a very simple way. These results throw interesting light then on the computational capabilities of large neuronal assemblies.

5. REFERENCES

[1] M. Tekalp, *Digital Video Processing*, Prentice-Hall, Englewood Cliffs, N.J., 1995.

[2] R.M. Haralick and L.G. Shapiro, *Computer and Robot Vision, Vol II*, Addison-Wesley, New York, 1993.

[3] F. Rieke, D. Warland, R. de Ruyter van Stevenink, and W. Bialek, *Spikes: Exploring the Neural Code*, MIT Press, Boston, 1997.

[4] J Bai, "Inferential theory for factor models of large dimension," *Econometrica*, vol. 71, pp. 135–171, 2003.

[5] V Solo, "State estimation from high dimensional data," in *Proc IEEE ICASSP 2004, Montreal Canada , May 2004*. IEEE, 2004, pp. –.

[6] V Solo, "'Unobserved' Monte Carlo method for identification of partially observed nonlinear state space systems, Part II: Counting process observations," in *Proc 39th IEEE Conference on Decision and Control, Sydney , Australia*. IEEE, 2000, pp. 3331–3336.

[7] N. Bleistein and R A Handelsman, *Asymptotic Expansions of Integrals*, Dover, New York, 1986.

[8] T Kailath, A H Sayeed, and B Hassibi, *Linear Estimation*, Prentice Hall, New York, 2000.

[9] DH Johnson, "Dialogue concerning neural coding and information theory," Tech. Rep. -, Department of Electrical & Computer Engineering, Rice University, 2003.

[10] DH Johnson, "Four top reasons mutual information does not quantify neural information processing," Tech. Rep. -, Dept ECE, Rice University, 2004.

[11] R Barbieri, LM Frank, DP Nguyen, MC Quirk, V Solo, MA Wilson, and EN Brown, "Dynamic analyses of information encoding in neural ensembles," *Neural Computation*, vol. 16, pp. 277–307, 2004.

[12] CJ Rozell and DH Johnson, "Examining methods for estimating mutual information in spiking neural systems," Tech. Rep. -, Department of Electrical & Computer Engineering Rice University, 2004.

[13] Dayan P and Abbott LF, *Theoretical Neuroscience*, MIT Press, Cambridge MA, 2001.

[14] SP Strong, R Koberle, R de Ruyter van Steveninck, and W Bialek, "Entropy and information in neural spike trains," *Physical Review Letters*, vol. 80, pp. 197–200, 1998.

[15] A Borst and FE Theunissen, "Information theory and neural coding," *Nature Neuroscience*, vol. 2, pp. 947–957, 1999.

[16] Liam Paninski, "Estimation of entropy and mutual information," *Neural Computation*, vol. 15, pp. 1191–1253, 2003.

[17] JD Victor, "Binless strategies for estimation of information from neural data," *PHYSICAL REVIEW E*, vol. 66, pp. 051903, 2002.

[18] J A McFadden, "The entropy of a point process," *J Soc Indust Appl Math*, vol. 13, pp. 988–994, 1965.

[19] P Bremaud, "On the information carried by a stochastic point process," *Revue du Cethededc*, vol. 43, pp. 45–70, 1975.

[20] P Bremaud, *Point Processes and Queues*, Springer Verlag, Heidelberg, 1981.

[21] V Solo, "System identification with analog and counting process observations II: Mutual information.," in *Proc IEEE Conf on Decision and Control 2005*. IEEE, 2005, p. to appear.

[22] T Cover and J Thomas, *Elements of Information Theory*, J Wiley, New York, 1991.

[23] AJ Goldsmith and PP Varaiya, "Capacity, mutual information, and coding for finite-state markov channels," *IEEE Trans on Information Theory*, vol. 42, pp. 868–886, 1996.