STATE ESTIMATION FROM HIGH-DIMENSIONAL DATA

Victor Solo

School of Electrical Engineering
University of New South Wales
Sydney, NSW 2052, AUSTRALIA
email: vsolo@syscon.ee.unsw.edu.au

ABSTRACT

It is implicit in traditional discussions of linear or nonlinear state estimation filters that there is no relation specified between the dimension of the state and the observation vector dimension. If anything though, the state would often be thought to have higher dimension. But increasingly in practice problems are arising where the reverse is the case. In this paper we show that state estimation filters, such as the Kalman filter undergo a remarkable simplification in structure and computation when the observation dimension is much larger than the state dimension. Both linear and nonlinear cases (including point processes) are discussed.

1. INTRODUCTION

In traditional discussions of state space estimation e.g. [1], [2] no relation is specified between the state and observation dimensions. Although it would often be implicit that the state dimension is larger. However with the tremendous recent growth in sensor diversity and capacity, communication channel throughputs and computational processing speeds, significant applications are arising where the reverse is the case. In econometrics hundreds of time series may be collected relating to a macroeconomic phenomenon of low dimension [3]; in neuroscience (which itself is undergoing tremendous growth in generation of high dimensional data) multielectrode recordings of spiking neurons from tens and even hundreds of cells are recorded all relating to a single phenomenon of low state dimension [4]; and in computer vision image sequence data similarly generates large dimensional observations of low dimensional phenomenon (e.g. rigid body rotation)[5],[6].

In all these areas, state estimation of one kind or another is of interest but it seems to have escaped notice that the structure of optimal filters in such cases undergoes considerable simplification. In this work we show that the Kalman filter undergoes remarkable structural and computational simplification (section 2). In section 3 we extend this to nonlinear analog systems and in section 4 to analog systems with point process observations. Conclusions are offered in section 5.

Notation. In the sequel if A, B are positive definite matrices, by $A \leq B$ we mean $B - A$ is positive semidefinite. Also $\lambda_{max}(A), \lambda_{min}(A)$ denotes respectively the largest, smallest eigenvalues of postive definite A. Also $Y_{1,k}$ denotes $(y_1^T, \cdots, y_k^T)^T$. And n =dim(x),p=dim(y).

This work was partly supported by the Australian Research Council

2. THE HIGH-DIMENSIONAL KALMAN FILTER

We consider a standard linear *time invariant* stochastic state space model in discrete time, with,

State Equation: $x_{k+1} = Fx_k + w_k$

Observation Equation: $y_k = Hx_k + v_k$

where (w_k, v_k) is a Gaussian white noise sequence with covariance $\left(\begin{smallmatrix} Q & 0 \\ 0 & R \end{smallmatrix}\right)$. We assume:

A1. R is positive definite.

For our development we need formulations for both the predicted state $\hat{x}_{k|k-1}$, the filtered state $\hat{x}_{k|k}$ and their associated error covariances $P_{k|k-1}, P_{k|k}$ respectively. We recall then the following well known Kalman filter update expressions.

State Update

$$\begin{aligned} \hat{x}_{k|k} &= \hat{x}_{k|k-1} + P_{k|k-1}H^T R_{ek}^{-1} e_k \qquad (2.1) \\ \hat{x}_{k+1|k} &= F\hat{x}_{k|k} = F\hat{x}_{k|k-1} + K_k e_k \end{aligned}$$

Kalman Gain: $K_k = FP_{k|k-1}H^T R_{ek}^{-1}$

Prediction Error: $e_k = y_k - H\hat{x}_{k|k-1}$

Variance:$R_{ek} = var(e_k) = R + HP_{k|k-1}H^T$

State Error Variance Update

$$\begin{aligned} P_{k+1|k} &= FP_{k|k}F^T + Q \qquad (2.2) \\ P_{k|k} &= P_{k|k-1} - P_{k|k-1}H^T R_{ek}^{-1} HP_{k|k-1} \end{aligned}$$

and associated Information forms

$$\begin{aligned} P_{k|k}^{-1} &= P_{k|k-1}^{-1} + H^T R^{-1} H \qquad (2.3) \\ K_k &= FP_{k|k}H^T R^{-1} \end{aligned}$$

Less familiar perhaps is the filtering error

$$\nu_k = y_k - H\hat{x}_{k|k}, var(\nu_k) = RR_{ek}^{-1}R$$

It is elementary to show from (2.1) that: $R^{-1}\nu_k = R_{ek}^{-1}e_k$ which explains the variance formula.

High Dimensional Observations

We now suppose $n = dim(x) < p = dim(y)$ and that:

A2. Rank(H)=n.

Our discussion begins with the crucial observation that under A1,A2: $W = H^T R^{-1} H$ has rank =n.

We now suppose:

A3. $\lambda_{min}(W) \to \infty$ as $p \to \infty$.

This encapsulates our assumption of high dimensional observations in an operational way. If $R = \sigma^2 I$ this means $\lambda_{min}(H^T H) \to \infty$. We might e.g. find in a typical application that $H^T H/p \to$ some finite full rank M. In any case we note an important special case:

A3s. $\lambda_{min}(H^T H) \to \infty$ as $p \to \infty$; $sup_p \lambda_{max}(R) < \infty$.

The second part of this condition ensures that adding highly correlated data to 'beef up' p does not help.

Continuing we note that W^{-1} exists and so from the information filter formula (2.3)

$$P_{k|k}^{-1} \geq W \Rightarrow P_{k|k} \leq W^{-1} \tag{2.4}$$

This means $P_{k|k}$ is a bounded (matrix) sequence and so must have at least one limit point (matrix) say P_f, which may not be unique (there may of course be a continuum of limit points). But then every such limit point obeys, $P_f \leq W^{-1}$ and so every limit point converges on 0;

$P_f \to 0$ as $p \to \infty$.

Putting (2.4) into (2.2) gives $P_{k+1|k} \leq FW^{-1}F^T + Q$. So $P_{k+1|k}$ is a bounded (matrix) sequence and must have a (possibly non-unique) limit point (matrix) P_p. Also for any such limit point (matrix): $P_{k+1|k} \geq Q \Rightarrow P_p \geq Q$

$$\Rightarrow 0 \leq P_p - Q \leq FW^{-1}F^T \to 0$$

as $W \to \infty$ i.e. every limit point (matrix) converges on Q: $P_p \to Q$. This is a startling result for the following reason.

If we return to the state space model we see that Q is the innovation variance for the state. If we observed the state directly the best prediction error we could achieve would have variance Q. Our result shows we are achieving this instead with high dimensional noisy data. We can summarise the discussion so far in:

Result 1A. Under A1,A2,A3 we have:

<u>Prediction Asymptotics</u>: Any limit point (matrix) of $P_{k+1|k}$ say P_p obeys: $P_p \to Q$ as $W \to \infty$.

<u>Filtering Asymptotics</u>: Any limit point (matrix) of $P_{k|k}$ say P_f obeys: $P_f \to 0$ as $W \to \infty$.

But this is only half the story.

Static State Estimators

We now show how the Kalman filter can be approximated by a *static* estimator. We introduce the filtered *regression estimator* $x^*_{k|k}$ and its associated forecast $x^*_{k|k-1}$

$$\begin{aligned} x^*_{k|k} &= W^{-1}H^T R^{-1} y_k \\ x^*_{k|k-1} &= F\hat{x}_{k-1|k-1} \end{aligned}$$

and compare their performance to that of the Kalman filter.

The filtered state estimation error is

$$\begin{aligned} x_k - x^*_{k|k} &= -W^{-1}H^T R^{-1}(y_k - Hx_k) \\ &= -W^{-1}H^T R^{-1} v_k \end{aligned}$$

which has variance$P^*_{k|k} = W^{-1}$ and this is (as expected) larger than $P_{k|k} \leq W^{-1}$. However let us consider the relative increase in variance in replacing $\hat{x}_{k|k}$ by $x^*_{k|k}$.

$$\begin{aligned} P_{k|k}^{-1}(P^*_{k|k} - P_{k|k}) &= (P_{k|k-1}^{-1} + W)W^{-1} - I \\ &= (WP_{k|k-1})^{-1} \end{aligned}$$

and this has a (possibly non-unique) limit point $(WP_f)^{-1}$ and under A3 this $\to 0$ as $p \to \infty$, since $P_f \to Q$. Thus the relative increase is small.

Note that $x^*_{k|k}, \hat{x}_{k|k}$ differ by a white noise

$$\begin{aligned} x^*_{k|k} - \hat{x}_{k|k} &= W^{-1}H^T R^{-1}(y_k - H\hat{x}_{k|k}) \\ &= W^{-1}H^T R^{-1}\nu_k \\ &= W^{-1}H^T R_{ek}^{-1} e_k \end{aligned}$$

with consequent variance

$$var(x^*_{k|k} - \hat{x}_{k|k}) = W^{-1}H^T R_{ek}^{-1} H W^{-1} \tag{2.5}$$

We now look at the prediction error.

$$\begin{aligned} P^*_{k|k-1} &= E(x_k - x^*_{k|k-1})(x_k - x^*_{k|k-1})^T \\ &= E(Fx_{k-1} + w_k - Fx^*_{k-1|k-1}) \\ &\times (Fx_{k-1} + w_k - Fx^*_{k-1|k-1})^T \\ &= FP^*_{k-1|k-1}F^T + Q \\ &= FW^{-1}F^T + Q \\ &= P^*_f, \text{ say} \end{aligned}$$

We immediately deduce, under A3: $P^*_f \to Q$, as $p \to \infty$

Again we look at the relative difference in prediction variances. By orthogonality of the optimal estimator

$$\begin{aligned} P^*_{k+1|k} &= P_{k+1|k} + E(u_k u_k^T); u_k = x^*_{k+1|k} - \hat{x}_{k+1|k} \\ &= P_{k+1|k} + FE(x^*_{k|k} - \hat{x}_{k|k})(x^*_{k|k} - \hat{x}_{k|k})^T F^T \\ &= P_{k+1|k} + FW^{-1}H^T R_{ek}^{-1} H W^{-1} F^T, \text{ by (2.5)} \end{aligned}$$

Thus the norm of the relative difference is

$$\begin{aligned} \| \cdot \| &= \| P_{k+1|k}^{-1}(P^*_{k+1|k} - P_{k+1|k}) \| \\ &= \| P_{k+1|k} F W^{-1} H^T R_{ek}^{-1} H W^{-1} F^T \| \end{aligned}$$

However: $R_{ek} \geq R \Rightarrow R^{-1} \geq R_{ek}^{-1} \Rightarrow H^T R_{ek}^{-1} H \leq W$

$$\Rightarrow \| \cdot \| \leq \| Q^{-1} F W^{-1} F^T \|$$

which, under A3, $\rightarrow 0$ as $p \rightarrow \infty$. To sum up:

Result 1B. Under A1,A2,A3: The static regression state estimator is asymptotically as good as the Kalman filter

$$P^*_{k|k-1} = P^*_f = FW^{-1}F^T + Q \rightarrow Q, \text{ as } W \rightarrow \infty$$

and the relative difference $\rightarrow 0$ as $p \rightarrow \infty$i.e.

$$sup_k \| P^{-1}_{k|k-1}(P^*_{k|k-1} - P_{k|k-1}) \| \rightarrow 0$$

3. NON-LINEAR FILTERS WITH HIGH DIMENSIONAL ANALOG OBSERVATIONS

A suitable version of the above results can be developed for the non-linear case.

For simplicity we consider a discrete time nonlinear model, (The continuous time case will be discussed elsewhere).

$$\begin{aligned} x_{k+1} &= f(k, x_k) + \sigma(k, x_k) n_k \\ &= f_k(x_k) + \sigma_k(x_k) n_k \\ y_k &= h(k, x_k) + v_k = h_k(x_k) + v_k \end{aligned}$$

(n_k, v_k) is a Gaussian white noise (independent of past y_k) with conditional variance $\left(\begin{smallmatrix} Q_k & 0 \\ 0 & R_k \end{smallmatrix}\right) = \left(\begin{smallmatrix} Q(k,x_k) & 0 \\ 0 & R(k,x_k) \end{smallmatrix}\right)$. We wish 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 then to approximate the conditional density $p(x_{k+1}|Y_{1,k+1})$. We have

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

where $\overline{\rho}(\cdot,\cdot)$ denotes a joint density and $P(\cdot)$ a marginal density. In view of the observation equation we can write

$$\overline{\rho}(x_{k+1}, Y_{1,k+1}) = p(y_{k+1}|x_{k+1})\rho(x_{k+1}, Y_{1,k})$$

and some simple Gaussian algebra gives

$$\begin{aligned} \overline{\rho}(x_{k+1}, Y_{1,k+1}) &= p_*(y_{k+1}) e^{U_{k+1}(x_{k+1})} \rho(x_{k+1}, Y_{1,k}) \\ p_*(y_{k+1}) &= e^{\frac{1}{2} y^T_{k+1} R^{-1}_{k+1} y_{k+1}} \\ U_{k+1}(x) &= y^T_{k+1} R^{-1}_{k+1} h_{k+1}(x) \\ &- \frac{1}{2} h_{k+1}(x)^T R^{-1}_{k+1} h_{k+1}(x) \end{aligned}$$

The conditional mean estimator is then

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

We will approximate these integrals using Laplace asymptotics [7] assuming

N1. $h^T_{k+1}(x) R^{-1}_{k+1}(x) h_{k+1}(x) \rightarrow \infty$ as $p \rightarrow \infty$.

Note that N1 ensures that as $p \rightarrow \infty$,

$$\begin{aligned} & var(y^T_{k+1} R^{-1}_{k+1}(x) h_{k+1}(x)) \\ = \; & h^T_{k+1}(x) R^{-1}_{k+1}(x) h_{k+1}(x) \rightarrow \infty \end{aligned}$$

Applying then Laplace asymptotics we introduce

$$x^*_{k+1} = arg.min.U_{k+1}(x)$$

which satisfies: $U'_{k+1}(x^*_{k+1}) = \frac{dU_{k+1}}{dx}\big|_{x=x^*_{k+1}} = 0.$

and we may expand $U_{k+1}(x)$ is a Taylor series about x^*_{k+1}. This Taylor series is plugged into the integrals above yielding Gaussian integrals which are easily evaluated giving

$$\begin{aligned} \hat{x}_{k+1|k+1} &\approx \frac{x^*_{k+1} N(x^*_{k+1})}{N(x^*_{k+1})} = x^*_{k+1} \\ N(x) &= e^{U_{k+1}(x)} \rho(x, Y_{1,k}) |U''_{k+1}(x)|^{\frac{1}{2}} \\ U''_{k+1}(x) &= \frac{d^2 U}{dx dx^T} \end{aligned}$$

Thus we get a remarkable extension of the linear result.

Result 2.Under the high-dimensional observation assumption N1 the filtered estimate is well approximated by the static non-linear regression estimator x^*_{k+1}.

In a similar way we find easily: $P_{k+1|k+1} \approx U''_{k+1}(x^*_{k+1})$

To compute the non-linear regression estimator a natural candidate is the Newton Raphson algorithm (with i the iteration index)

$$x^{(i+1)} = x^{(i)} - [U''_{k+1}(x^{(i)})]^{-1} U'_{k+1}(x^{(i)})$$

In a similar way it can be shown that

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

This last result again is remarkable in showing that the prediction error achieves (approximately) the true state innovations variance.

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

We now turn to develop similar results in a counting process context. We suppose a continuous non-linear state space model (of which the model above can be treated as a discretisation) but now a vector of conditional (possibly inhomogeneous) Poisson process observations

$$dN^r_t = \lambda^r(t, x(t))dt + dM^r_t, r = 1, \cdots, p$$

and conditional on the whole state trajectory, $N_t = (N^r_t, r = 1, \cdots, p)$ is a (vector) inhomogeneous Poisson process with rate function $\lambda_t = (\lambda^r(t, x(t)), r = 1, \cdots, p)$.

$M_t = (M_t^r, r = 1, \cdots, p)$ is then a vector martingale. We consider discrete time (binned) observations (wherein δ is the sampling interval) with $x_k = x(k\delta)$,

$$\begin{aligned} y_k^r &= N_k^{(\delta,r)} = \lambda^r(k\delta, x_k)\delta + m_k^r \\ &= \#\text{ events in } k\delta, (k\delta + \delta) \end{aligned}$$

c.f.[8]. We can think of m_k^r as a white noise of conditional variance $\lambda^r(k, x_k)\delta$. Also denote

$$\begin{aligned} N_k^\delta &= (N_k^{(\delta,1)}, \cdots, N_k^{(\delta,p)}) \\ N_{1,k}^{(r)} &= (N_1^{(\delta,r)}, \cdots, N_k^{(\delta,r)}) \\ N_{1,k} &= (N_{1,k}^{(1)}, \cdots, N_{1,k}^{(p)}) \end{aligned}$$

We are interested in approximating $\hat{x}_{k|k}, P_{k|k}$ etc. as before.

As before write the conditional density $p(x_{k+1}|N_{1,k+1})$

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

and again using the observation equation

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

Simple inhomogeneous Poisson calculations c.f. [8], give

$$\overline{\rho}(x_{k+1}, N_{1,k+1}) = P_*(N_{k+1}^\delta)e^{U_{k+1}(x_{k+1})}\rho(x_{k+1}, N_{1,k})$$

$$P_*(N_{k+1}^\delta) = \Pi_1^p P_*(N_{k+1}^{(\delta,r)})$$

$$P_*(N_{k+1}^{(\delta,r)}) = \Pi_1^{k+1}[\delta^{N_i^{\delta,r}} e^{-\delta}/N_i^{\delta,r}!]$$

$$U_{k+1}(x) = \Sigma_1^p[N_{k+1}^{(\delta,r)} log\lambda_{k+1}^{(r)}(x) - (\lambda_{k+1}^r(x) - 1)\delta]$$

The conditional mean estimator is then

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

As before, estimate this with Laplace asymptotics assuming **C1**. $\Sigma_1^p \lambda_{k+1}^r(x) \to \infty, as, p \to \infty$.

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

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

So introduce: $x_{k+1}^* = arg.minU_{k+1}(x)$

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

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

Again expanding $U_{k+1}(x)$ in a Taylor series,inserting in the integrals above and evaluating yields $\hat{x}_{k+1|k+1} \approx x_{k+1}^*$.

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

And again we then recover the remarkable result.

Result 3. Under the high-dimensional observation assumption C1, the optimal mean square non-linear filter is well approximated by a static non-linear regression estimator.

The nonlinear results for the predictor given above similarly follow here.

5. CONCLUSIONS

In this paper we have discussed state estimation from high-dimensional data i.e. when state dimension $<<$ observation dimension. Under some reasonable regularity conditions (that exclude addition of degenerate observations) we have shown that optimal dynamic filters and predictors exhibit a remarkable simplification. In particular the high-dimensional data provides such strong state information that optimal static regression estimators are just as good. And knowledge of the state dynamics provides little value for efficient state estimation. Of course predictions based on the static estimator still need knowledge of system matrices. Also dynamics will remain inportant for system identification. We have developed versions of this result in both linear and non-linear cases and with both analog and digital observations. These results promise considerable computational gains for many contemporary state estimation applications.

6. REFERENCES

[1] A.E. Bryson and Y.C. Ho, *Applied Optimal Control*, Blaisdell, Cambridge, MA, 1969.

[2] B D O Anderson and J Moore, *Optimal Filtering*, Prentice Hall, New Jersey, 1979.

[3] J Stock and M Watson, "Diffusion indices," Tech. Rep., NBER, 1999.

[4] F. Rieke, D. Warland, R. de Ruyter van Stvenink, and W. Bialek, *Spikes: Exploring the neural code*, MIT Press, Boston, 1997.

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

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

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

[8] V Solo, "'Unobserved' Monte Carlo method for system identification of partially observed nonlinear state space systems, Part II: Counting process observations," in *Proc 39th IEEE CDC, Sydney Australia*. IEEE, 2000.