Sufficient Statistics

Suppose we are given samples

$$X_1, \ldots, X_N \sim P_\theta(x)$$

Is there a sufficient statistic

- i.e. we can compress to $T(X_1, \ldots, X_N)$ w/o losing any information?

Factorization Theorem [Neyman, ...]
A statistic is sufficient iff

$$P_\theta(X_1, \ldots, X_N) = u(X_1, \ldots X_N)\, v(T(X_1, \ldots X_N), \theta)$$

Graphically this means

$$\boxed{\theta} \rightarrow \boxed{T(X_1, \ldots X_N) = t} \rightarrow \boxed{X_1 \ldots X_N}$$

i.e. $X_1 \ldots X_N \perp\!\!\!\perp \theta \mid T(X_1 \ldots X_N) = t$

There is a canonical way to satisfy this condition

Def: An exponential family has the form

$$P_\theta(x) = \frac{h(x)\, e^{\langle \theta, T(x) \rangle}}{Z(\theta)}$$

e.g. for the Ising model, we can take

$$T(x) = [\text{vec}(xx^T), x]$$

$$\theta = [\text{vec}(A), h]$$

Corollary: For an exponential family, any estimator

$$\hat{\theta}(X_1 \ldots X_N)$$

can be turned into another one $\tilde{\theta}(T(X_1 \ldots X_N) = t)$

Problem: Looking inside

$\tilde{\theta}$:

$T(X_1 \ldots X_N) = t \rightarrow X_1 \ldots X_N \rightarrow \hat{\theta}$

But sampling can be hard

Thm [Montanari] [Bresler et al] There are graphical models that can be efficiently learned, but not if you reduce to sufficient statistics

Open: Are there computational-vs-statistical tradeoffs for learning exponential families?

Linear Dynamical Systems

Canonical model for time series

$$x_{t+1} = Ax_t + B\underbrace{u_t}_{\text{observed}} + w_t$$

x_t: hidden state; u_t: control; w_t: process noise

$$\underbrace{y_t}_{\text{observed}} = Cx_t + Du_t + z_t$$

y_t: output; z_t: observation noise

e.g. Could represent discretization of an ODE

Many questions

① Inference: If A, B, C, D are known, how can we estimate the trajectory

Proposition [Kalman] There is a statistically optimal and computationally efficient estimator, if w_t, z_t, x_0 are Gaussian, based on least squares

Kalman filter was used in the moon landing, won national medal of science

② learning: Given one long trajectory how can we estimate A, B, C, D?

def: The Markov parameters, (up to order s) are

$$[D, CB, CAB, \ldots CA^sB]$$

Intuition: These govern the input-output (time-lagged) relationships for the LDS

① How do we estimate the Markov parameters?

② And how do we use them to estimate A, B, C, D?

Observation: Can only learn A, B, C up to similarity

$$\hat{A} \leftarrow TAT^{-1} \quad \hat{B} \leftarrow TB \quad \hat{C} \leftarrow CT^{-1}$$

In particular, this transformation preserves the Markov parameters and yields equivalent LDS

In 1966, Ho and Kalman gave an algorithm

Markov parameters $\longrightarrow$ estimates for A, B, C, D

We'll see it's a noncommutative version of the matrix pencil method

def: The observability matrix is

$$O_s = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{s-1} \end{bmatrix}$$

and the controllability matrix is

$$\mathcal{O}_s = [B, AB, \dots A^{s-1}B]$$

Step #1: Form the Hankel matrix

$$H = \begin{bmatrix} X_1 & X_2 & X_3 & \cdots & X_{s+1} \\ X_2 & X_3 & & & \\ X_3 & & & & \\ & & & & X_{2s} \end{bmatrix}$$

H^- (first s rows) | H^+ (last s rows)

$$= \begin{bmatrix} D & CB & CAB & \cdots \\ CB & CAB & & \\ CAB & & & \\ & & & \end{bmatrix}$$

where $[X_0, X_1, \dots X_{2s}]$ are Markov params

Note: This is what we did in superresolution too

low frequency measurements → entries of Hankel matrix

As before, H has a hidden factorization we are looking for given by

$$H = \mathcal{O}_{s+1} \mathcal{C}_{s+1}$$

Step #2: Let H^- = first s blocks of columns of H

Compute $U \Sigma V^T = H^-$ and let $\hat{\mathcal{O}} = U\Sigma^{1/2}$, $\hat{\mathcal{C}} = \Sigma^{1/2} V^T$

(H^-: rank n)

Now $\mathcal{O}_{s+1}$ and $\hat{\mathcal{O}}$ can be quite different but

Lemma: $O_{s+1} = \hat{O}T$ for some invertible transformation T, provided that O_{s+1} and $\hat{O}$ are full col/row rank resp.

Proof: since O_{s+1} and $\hat{O}$ have full col/row rank resp, we have that $H^- = O_{s+1}\hat{Q}$ and

$$\text{colspan}(H^-) = \text{colspan}(O_{s+1})$$

And by properties of the SVD

$$\text{colspan}(\hat{O}) = \text{colspan}(H^-)$$

And since O_{s+1} and $\hat{O}$ have n cols and the same colspan, there exists an invertible T s.t.

$$O_{s+1} = \hat{O}T \text{ as desired.}$$

Similarly, under same conditions we have

$$\hat{Q} = S\hat{Q}$$

(S: invertible)

But since $O_{s+1}\hat{Q} = H^- = \hat{O}\hat{Q}$ we must have that $S = T^{-1}$

Step #3: Let H^+ = last s blocks of cols of H

Then $H^+ = O_{s+1} A \hat{Q}$

$$= \hat{O}TAT^{-1}\hat{Q}$$

So if we compute $\hat{A} = \hat{O}^+ H^+ \hat{Q}^+ = TAT^{-1}$

rows of
Moreover the first block of $\hat{O}$ is

$$\hat{C} = CT$$

and the first block of cols of $\hat{\mathcal{O}}$ is

$$\hat{B} = T^{-1}B$$

This works in the noiseless case. But when is it stable?

Another parallel to super resolution:

Thm [Oymak, Ozay] The Ho-Kalman algorithm is stable if the observability and controllability matrices are well-conditioned

These conditions are natural

(1) If $\text{rank}(\mathcal{O}_s) < n$ there is some direction in the hidden state space that you never see, even after s steps

(2) If $\text{rank}(\mathcal{C}_s) < n$, there is some direction in the hidden state space that the control never touches

Can show a sample complexity lower bound, for estimating A, B, C up to similarity when either $\mathcal{O}$ or $\mathcal{C}$ are ill conditioned.

Now how do you estimate the Markov parameters?

Many approaches based on regression, but strong assumptions like

(1) assumptions about characteristic poly. of A, or phases of its roots

e.g. $\|p_\lambda\|_1 \leq C$

$\uparrow$

ℓ_1-norm of coefficients

(2) strict stability, i.e. $\rho(A) < 1$

spectral radius

but then no long-range correlations, converges to stationarity, cannot discretize ODEs

(3) restrictions or exponential dependence on the dimension

Thm [Bakshi, Liu, Moitra, Yau] There is a polynomial time algorithm for estimating the Markov parameters, just assuming O and Q are well-conditioned from a single trajectory

In fact, the approach is the method of moments, extends to hypercontractive noise

The starting point is

If the control/noises are ⊥ and mean zero then

Observation: $\mathbb{E}[y_{t+j} u_t^T] = \begin{cases} D & \text{if } j = 0 \\ CA^{j-1}B & \text{else} \end{cases}$

Proof: Expand the recurrence, e.g. for $j = 1$

$$y_{t+1} = Cx_{t+1} + Du_{t+1} + z_{t+1}$$

$$= CAx_t + CBu_t + Cw_t$$

$$+ Du_{t+1} + z_{t+1}$$ ☐

So why aren't we done?

We could try to estimate $CA^{j-1}B$ as

$$\frac{1}{T} \sum_{t=1}^{T} y_{t+j} u_t^T$$

e.g. Gaussian random walk

but this estimator does not have bounded variance because of dependencies across time steps

Main Idea: Can we form a new time series

$$\hat{y}_t \triangleq y_t - \sum_{j=1}^{n} c_j y_{t-j}$$

such that

① $\mathbb{E}[\hat{y}_{t+j} u_t^T]$ is unchanged but

② Its variance remains bounded

For example, if we take c_j's = coefficients of characteristic poly

Then the Cayley-Hamilton theorem tells us

$$A^n - \sum_{j=1}^{n} c_j A^{n-j} = 0$$

and we can expand (assuming $D, w_t, z_t = 0$ for simplicity)

$$\hat{y}_t = y_t - \sum_{j=1}^{n} c_j y_{t-j}$$

$$= \sum_{i=1}^{n} \left(CA^{i-1}B - \sum_{j=1}^{i-1} c_j CA^{i-j-1}B \right) u_{t-i}$$

$$+ \sum_{i=n+1}^{t} \left(CA^{i-n-1} \left(A^n - \sum_{j=1}^{n} c_j A^{n-j} \right) B \right) u_{t-i}$$

$\left(A^n - \sum_{j=1}^{n} c_j A^{n-j} \right) \to 0$

Thus the variance of $\hat{y}_{t+j} u_t^T$ is bounded, independently of t

The expressions become more complicated, but still well behaved in general, and can make sure the expectation is unchanged by instead setting

$$\hat{y}_{t+k} = y_{t+k} - \underbrace{\sum_{j=1}^{s} \alpha_j y_{t-j}}$$

u_t doesn't show up here

Main ideas:

① use assumptions about observability/controllability to reason about existence of bounded α's that make the estimator good (avoid characteristic polynomial)

② write a convex program to search for α_j's
i.e. bound α's

③ argue that any bad α_j's that do not stabilize the estimator whp are infeasible by anticoncentration

Epilogue: Augment Ho-Kalman with tensor methods to learn mixtures of LDS's from short trajectories