

Sufficient Statistics

Suppose we are given samples

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

Is there a sufficient statistic

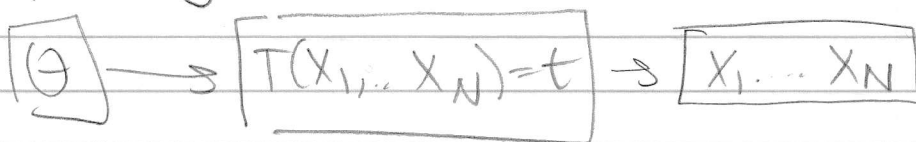
- i.e. we can compress to $T(X_1, \dots, X_N)$
w/o losing any information?

Factorization Theorem [Neyman, ...]

A statistic is sufficient iff

$$P_\theta(x_1, \dots, x_N) = u(x_1, \dots, x_N) v(T(x_1, \dots, x_N), \theta)$$

Graphically this means



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

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

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

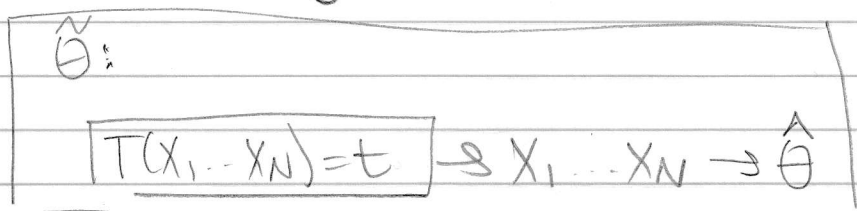
$$\Theta = [\text{vec}(A), h]$$

Corollary: For an exponential family, any estimator

$$\hat{\Theta}(x_1, \dots, x_N)$$

can be turned into another one $\hat{\Theta}(T(x_1, \dots, x_N) = t)$

Problem: Looking inside



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

$$x_{t+1} = \underset{\substack{\uparrow \\ \text{hidden state}}}{A} x_t + \underset{\substack{\uparrow \\ \text{control}}}{B} \overset{\substack{\text{observed}}}{u_t} + \underset{\substack{\uparrow \\ \text{process noise}}}{w_t}$$

$$\overset{\substack{\text{observed}}}{y_t} = \underset{\substack{\uparrow \\ \text{output}}}{C} x_t + \underset{\substack{\uparrow \\ \text{observation noise}}}{D} u_t + z_t$$

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

$$[D, CB, CAB, \dots, CA^s B]$$

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

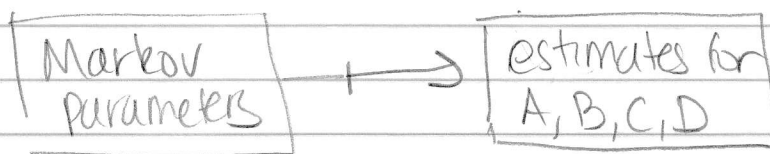
- ① How do we estimate the Markov parameters?
- ② And how do we use them to estimate A, B, C, D ?

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

$$\hat{A} \leftarrow T A T^{-1} \quad \hat{B} \leftarrow T B \quad \hat{C} \leftarrow C T^{-1}$$

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

In 1966, Ho and Kalman gave an algorithm



We'll see it's a non commutative 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

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

Step #1: Form the Hankel matrix

$$H = \begin{matrix} H^T & & & & \\ \vdots & X_1 & X_2 & X_3 & \dots & X_{s+1} \\ \vdots & X_2 & X_3 & & & \\ \vdots & X_3 & & & & \\ \vdots & & & & & X_{2s} \\ & D & CB & CAB & \dots & \end{matrix} H^T$$

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

Note: This is what we did in superresolution too

low frequency measurements \rightarrow entries of Hankel matrix

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

$$H = \underset{s+1}{\mathbb{O}} \underset{s+1}{\mathbb{P}}$$

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

$$\text{Compute } U \Sigma V^T = H^- \text{ and let } \hat{\mathbb{O}} = U \Sigma^{1/2} \\ \hat{\mathbb{P}} = \Sigma^{1/2} V^T$$

\uparrow
rank n

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

lemma: $O_{sH} = \hat{O}T$ for some invertible transformation T ,
provided that O_{sH} and \hat{O} are full coln/row rank resp.

Proof: since O_{sH} and \hat{O} have full coln/row rank resp,
we have that $H^- = O_{sH} \hat{O}$ and

$$\text{colspan}(H^-) = \text{colspan}(O_{sH})$$

And by properties of the SVD

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

And since O_{sH} and \hat{O} have n colns and the
same colspan, there exists an invertible T s.t.

$$O_{sH} = \hat{O}T \text{ as desired. } \square$$

Similarly, under same conditions we have

$$\text{then } H^+ = \hat{O}^+ \hat{O}$$

invertible

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

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

$$\text{Then } H^+ = O_{sH} A \hat{O}^+$$

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

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

moreover the first block of ^{rows of} \hat{O} is

$$\hat{C} = CT$$

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

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

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

Another parallel to superresolution:

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

These conditions are natural

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

② If $\text{rank}(Q_s) < n$, 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 O or Φ 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

$$\text{eg. } \|P_\lambda\|_1 \leq C$$

↑
 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, Motra, 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 \perp
and mean zero then

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

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

$$\begin{aligned} y_{t+1} &= Cx_{t+1} + Du_{t+1} + z_{t+1} \\ &= CAx_t + CBu_t + Cw_t \\ &\quad + Du_{t+1} + z_{t+1} \quad \square \end{aligned}$$

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

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

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

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

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

$$\hat{y}_{t+k} = y_{t+k} - \underbrace{\sum_{j=1}^n \alpha_j y_{t-j}}_{u_t \text{ doesn't show up here}}$$

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_j '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