Stat 991: Multivariate Analysis, Dimensionality Reduction, and Spectral Methods **Lecture: 19**

Spectral Methods for Learning Vector State-Space Models

Instructor: Sham Kakade

1 Linear Algebra Review

Let $M \in \mathbb{R}^{d\times d'}$.

- If M has rank k, then:
  - M has k linearly independent rows and k linearly independent columns.
  - the range of the linear map $f(x) = Mx$ is of dimension k (same holds for the map $f(x) = M^\top x$).
  - $k \le \min(d, d')$
- M has full rank if M $= \min(d, d')$. We say M has full row rank if rank$(M) = d$ (in which case, it must be that $d \le d'$). We say M has full column rank if rank$(M) = d'$ (in which case, it must be that $d' \le d$).
- If $M = UDV^\top$ is the "thin" SVD of M, then the diagonal matrix D is $k \times k$ with all non-zero entries in the diagonal; U is $d \times k$ matrix of rank k with orthonormal columns; V is $d' \times k$ matrix of rank k with orthonormal columns.
- For a matrix $B \in \mathbb{R}^{k\times d}$ where B has rank k, then rank$(AB) =$ rank(A).
- Let $M = UDV^\top$ be the "thin" SVD of M. Then the pseudo-inverse is defined as $M^+ = VD^{-1}U^\top$.
- Let M be a $k \times d$ matrix of rank k (so $k \le d$), $CC^+ =$ I. However, $C^+C \ne$ I (unless $k = d$ in which case C is invertible and $C^+ = C^{-1}$.)
- If $A = BC$ where C is a $k \times d$ matrix of rank k, then $AC^+ = B$.

2 Vector State-Space Models

At each timestep t, there is a vector valued hidden state $h_t \in \mathbb{R}^k$ and observation $x_t \in \mathbb{R}^d$. We assume that $d \ge k$.

Typically, state space models have the natural conditional independence structure (where conditioned on a hidden state, we have that the past, future, and current observation are all independent). Here, we consider the weaker setting of only pairwise independencies. In particular, assume that:

1. The next hidden state h_{t+1}, conditioned on the current state h_t, is not correlated with the previous $h_{t'}$ and $x_{t'}$ for $t' \le t$ (clearly, with appropriate conditional independence assumptions this would be satisfied. however, this condition is weaker).
2. The observation x_t, conditioned on the state h_t, is not correlated with any other $h_{t'}$ and $x_{t'}$. (Again, with the natural independence assumption, this is satisfied).
3. In expectation, the next state h_{t+1} and (current) observation x_t are linearly related to the current state h_t:

$$\mathbb{E}[h_{t+1}|h_t] = Th_t \quad \text{and} \quad \mathbb{E}[x_t|h_t] = Oh_t \tag{1}$$

   where $T \in \mathbb{R}^{k\times k}$ and $O \in \mathbb{R}^{d\times k}$.

2.1 Kalman filters

Linear state space models (e.g. those with additive Gaussian noise and appropriate independent assumptions) fall into this setting. Here Kalman filters provide the Bayes optimal predictions (in terms of square loss).

2.2 Hidden Markov Models

HMMs also fall into this setting, if we represent both h_t and x_t as binary vectors (with only entry being 1). However, observe that the noise process η_t, where $x_t = Oh_t + \eta_t$, is heteroskedastic, i.e. the noise η_t depends on h_t (this must be the case since x_t is binary vector).

3 Learning

The goal of learning is to estimate a model, using only samples, which accurately predicts the joint probability of long sequences $x_1, x_2, \ldots$ or (accurately predicts the conditional probability of future events given past events). As our vector space model is not completely specified, this is not possible to do exactly (with any amount of data). Though we address this later (for linear Gaussian noise models and HMMs).

However, we may hope to recover both T and O using samples. Clearly, this also is not possible as the hidden state can be transformed linearly (e.g. written in a different basis), and this would alter both T and O, yet give the same observable probabilities. Instead, the best we can hope for is estimating T and O up to a linear transformation.

The key to this lecture is showing that this is possible (under a non-degenerate assumptions) with essentially a closed form solution — which depends on a certain SVD/CCA.

3.1 Estimation

Now let us show that we can recover both T and O for this class of models, up to a linear transformation (which is all that we could hope for in general).

Assumption 1 (Full Rank)**.** *Assume T, O and $\mathbb{E}[h_t h_t^\top]$ (for all t) are full rank. In other words, these matrices are all rank k.*

Now let us examine some properties with regards to the CCA of the cross correlation matrix.

Lemma 3.1 (CCA properties)**.** *Say τ is some arbitrary timestep Let the "thin" SVD of the cross correlation matrix at some timestep τ be $E[x_{\tau+1}x_\tau^\top] = UDV^\top$ (where all zeros have been removed form D appropriately). Then*

1. *D is a $\mathbb{R}^{k\times k}$ matrix.*
2. *The range of U equals the range of O.*
3. *$U^\top O$ is invertible.*
4. *Define $M = U^\top O$. Consider the transformed hidden state variables:*

$$\widetilde{h}_t = Mh_t \quad \textit{and} \quad \widetilde{T} = MTM^{-1}$$

Then $h_t = M^{-1}\widetilde{h}_t$ (since M is invertible) and

$$\mathbb{E}[\widetilde{h}_{t+1}|\widetilde{h}_t] = \widetilde{T}\widetilde{h}_t \quad \textit{and} \quad \mathbb{E}[x_t|\widetilde{h}_t] = U\widetilde{h}_t$$

In other words, this provides another representation of our time series.

Proof. First, let observe that:

$$\begin{aligned}
UDV^\top &= \mathbb{E}[x_{\tau+1}x_\tau^\top] \\
&= \mathbb{E}[\mathbb{E}[x_{\tau+1}x_\tau^\top|h_\tau]] \\
&= \mathbb{E}[\mathbb{E}[x_{\tau+1}|h_\tau]\mathbb{E}[x_\tau^\top|h_\tau]] \\
&= \mathbb{E}[OTh_\tau(Oh_\tau)^\top] \\
&= OT\mathbb{E}[h_\tau h_\tau^\top]O^\top
\end{aligned}$$

Hence, by assumption, the righthand side must have rank k. Thus, D must have rank k, which completes the first claim. Also, as both U and O are rank k, then the above shows they have the same range, which proves the second claim.

Since $U^\top U = \mathrm{I} \in \mathbb{R}^{k\times k}$ (though $UU^\top$ is not necessarily I as $UU^\top$ is $d \times d$ while being only rank k).

$$DV^\top = (U^\top O)T\mathbb{E}[h_\tau h_\tau^\top]O^\top$$

Thus $U^\top O$ must have rank k since $DV^\top$ is rank k. Hence, it is invertible as it is a $k \times k$ matrix, which proves the third claim.

Due to the invertibility of M we have that $\mathbb{E}[\widetilde{h}_{t+1}|\widetilde{h}_t] = \widetilde{T}\widetilde{h}_t$. Now note that $UU^\top$ is a projection onto the range of U (as U has orthogonal columns). As U and O have the same range, we have $UU^\top O = O$. Hence,

$$\mathbb{E}[x_t|\widetilde{h}_t] = OM^{-1}\widetilde{h}_t = UU^\top OM^{-1}\widetilde{h}_t = UMM^{-1}\widetilde{h}_t = U\widetilde{h}_t$$

which proves the final claim. □

Theorem 3.2. *Under the previous assumptions and conditions, we have that:*

$$\widetilde{T} = (U^\top\mathbb{E}[x_{t+2}x_t^\top])(U^\top\mathbb{E}[x_{t+1}x_t^\top])^+. \tag{2}$$

where A^+ denotes the pseudo-inverse of A.

Proof. First, as have shown above:

$$\mathbb{E}[x_{t+1}x_t^\top] = OT\mathbb{E}[h_t h_t^\top]O^\top$$

Now observe that $\mathbb{E}[x_{t+1}|h_t] = OTh_t$. Now,

$$\begin{aligned}
\mathbb{E}[x_{t+2}x_t^\top] &= \mathbb{E}[\mathbb{E}[x_{t+2}x_t^\top|h_t]] \\
&= \mathbb{E}[\mathbb{E}[x_{t+2}|h_t]\mathbb{E}[x_t^\top|h_t]] \\
&= OT^2\mathbb{E}[h_t h_t^\top]O^\top \\
&= OT(U^\top O)^{-1}(U^\top O)T\mathbb{E}[h_t h_t^\top]O^\top \\
&= OT(U^\top O)^{-1}U^\top\mathbb{E}[x_{t+1}x_t^\top].
\end{aligned}$$

where the last step uses our expression for $\mathbb{E}[x_{t+1}x_t^\top]$. Hence,

$$U^\top\mathbb{E}[x_{t+2}x_t^\top] = \widetilde{T}(U^\top\mathbb{E}[x_{t+1}x_t^\top]).$$

Also, as we have shown,

$$U^\top\mathbb{E}[x_{t+1}x_t^\top] = U^\top OT\mathbb{E}[h_t h_t^\top]O^\top$$

which implies $U^\top\mathbb{E}[x_{t+1}x_t^\top]$ must be rank k. The claim now follows (see the last property in the linear algebra review above). □

4 Acknowledgements

Daniel Hsu provided many of the observations in this lecture, which was adapted from his notes.