Learning Multiple Systems And Error Rate Improvements

Dalton Jones

March 23, 2022

Suppose we have a set of D linear systems with measured input $u^{(i)}$ and output $y^{(i)}$. In particular, we have for some state space realization of each system

$$x_t^{(i)} = A_i x_{t-1}^{(i)} + B_i u_{t-1}^{(i)} + v_{t-1}^{(i)}$$
$$y_t^{(i)} = C_i x_{t-1}^{(i)} + w_t^{(i)}$$

where v, w are some iid noise. We can rewrite the input output relationship using markov parameters as

$$y_t^{(i)} = \sum_{j=1}^{t} C_i A_i^j B_i u_{t-j}^{(i)} + \sum_{j=1}^{t} C_i A_i^j v_{t-j}^{(i)} + w_t^{(i)}$$
$$y_t^{(i)} = \sum_{j=1}^{t} G_j^{(i)} u_{t-j}^{(i)} + \sum_{j=1}^{t} F_j^{(i)} v_{t-j}^{(i)} + w_t^{(i)}.$$

Previous results show that we can use least squares to estimate the first T terms of the markov parameters of G^i with error decaying as $\sim \sqrt{T}/\sqrt{N}$ where N is the length of the timeseries used in each case.

However, suppose that there exists some low dimensional structure in the systems. In other words, suppose that the $G^{(i)}$ are spanned by a set of "elementary" markov parameter vectors $G_1^*, ...G_r^*$. In other words if we stack the markov parameters as row vectors into a matrix

$$G = \begin{pmatrix} \langle G^{(1)} \rangle \\ \langle G^{(2)} \rangle \\ \vdots \\ \langle G^{(D)} \rangle \end{pmatrix}$$

the singular value decompostion of $G = U\Sigma V^\top$ where V is a matrix with rows given by G_i^* (the elementary system markov parameters).

Suppose that I know $U\Sigma$ and I consider the first r rows of G, denoted as $G[1:r]$. From the knowledge of $U\Sigma$ I know that I can take a weighted sum of the rows of $G[1:r]$ to obtain G_1^*. In particular, we have that

$G_1^* = \sum_{i=1}^{r} \alpha_i G^{(i)}$. In which case, due to the linearity of the systems, we have

$$y_1^* = \sum_{i=1}^{r} \alpha_i y^{(i)} = G_1^* * \sum_{i=1}^{r} u^{(i)} + \text{noise}.$$

Hence, instead of learning D systems separately, each at a rate of $1/\sqrt{N}$, we can learn r subsystems by combining the data above. Note that if we partition the datasets equally into r groups of size D/r, then within each of those groups we need r data sets to play the trick above, we should have D/r^2 synthetic "runs" of the system G_1^*. The upshot of this is now that we go from an error rate of $1/\sqrt{N}$ to an error rate of $r/\sqrt{DN}$.

Now what happens if we don't know the linear combination coefficients α_i to form the synthetic system inputs and outputs of G_1^*? We should be able to learn them by looking at the SVD of $\hat{G} = G + E$ where E is a matrix with frobenius norm bounded by $||E|| \leq \sqrt{D/N}$ with high probability.

Additionally, note that the coefficients α_i used to synthesize the inputs and outputs of G_1^* can be calculated as

$$\begin{aligned}\alpha &= G_1^* G[1:r]^\dagger \\ &= G_1^* V \Sigma^{-1} U \\ &= e_1 \Sigma^{-1} U[1:r]^T \\ &= e_1 (1/\sigma_1) U[1]\end{aligned}$$

Where e_1 is the standard basis vector and $U[1]$ is the first row of $U[1:r]$.

Suppose we can estimate $\hat{\sigma_1} = \sigma_1 + \delta_\sigma$ and $\hat{U[1]} = U[1] + \delta_U$ via singular value decomposition. Where with high probability $|\delta_\sigma| \leq \Delta_\sigma$ and $||\delta_U||_2 \leq \Delta_U$ (QUESTION: What is the error rate here. Can we show this decays at a particular rate?)

Finally, suppose that $\Delta_\sigma/\sigma_1 \leq c < 1$, then using the taylor series expansion of $1/(\sigma_1 + \delta_\sigma)$ we can bound the error

$$\left|\frac{1}{\sigma_1} - \frac{1}{\sigma_1 + \delta_\sigma}\right| \leq \frac{\Delta_\sigma}{1 + (\delta_\sigma/\sigma_1)} \leq \frac{\Delta_\sigma}{1-c}$$

Then we have that

$$\begin{aligned}||\alpha - \hat{\alpha}||_2 &= ||\sigma_1^{-1} U[1] - \hat{\sigma_1}^{-1} \hat{U[1]}|| \\ &\leq \Delta_U \frac{\Delta_\sigma}{1-c} + \Delta_U \sigma_1^{-1} + \Delta_\sigma.\end{aligned}$$

Suppose that we call the vector $e_\alpha = \alpha - \hat{\alpha}$. Then if we use $\hat{\alpha}$ to synthesize the inputs and outputs of the system G_1^*, the solution of the least squares estimator becomes

$$\hat{G}_1^* = G_1^* + \delta_G + \sum_{j=1}^{r} e_\alpha[j] G^{(j)} + \text{error from noise}$$

where δ_G is the error in estimating the singular vector G_1^* from the SVD of $\hat{G}$ and the error from noise scales as $1/\sqrt{N}$. If we could bound $||\delta_G||_2 \leq \Delta_G$ and $||G^(j)||_2 \leq G_{max}$ for any j we have that the overall error rate becomes

$$\Delta_G + G_{max}\left(\Delta_U \frac{\Delta_\sigma}{1-c} + \Delta_U \sigma_1^{-1} + \Delta_\sigma\right) + \mathcal{O}(\sqrt{1/ND}).$$

Hence, if we can show that $\Delta_\sigma, \Delta_U, \Delta_G \sim \mathcal{O}(\sqrt{1/ND})$, we can show that this estimation scheme combining multiple systems actually improves the error rate, even when the relationship between the systems is unknown.