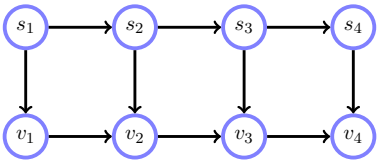

Figure 24.8: A first order switching AR model. In terms of inference, conditioned on $v_{1:T}$, this is a HMM.

where † denotes the pseudo inverse, see `LDSsubspace.m`. Estimates for the covariance matrices can also be obtained from the residual errors in fitting the block Hankel matrix and extended observability matrix. Whilst this derivation formally holds only for the noise free case one can nevertheless apply this in the case of non-zero noise and hope to gain an estimate for $\mathbf{A}$ and $\mathbf{B}$ that is correct in the mean. In addition to forming a solution in its own right, the subspace method forms a potentially useful way to initialise the EM algorithm.

24.5.4 Structured LDSs

Many physical equations are local both in time and space. For example in weather models the atmosphere is partitioned into cells $h_i(t)$ each containing the pressure at that location. The equations describing how the pressure updates only depend on the pressure at the current cell and small number of neighbouring cells at the previous time $t-1$. If we use a linear model and measure some aspects of the cells at each time, then the weather is describable by a LDS with a highly structured sparse transition matrix $\mathbf{A}$. In practice, the weather models are non-linear but local linear approximations are often employed[270]. A similar situation arises in brain imaging in which voxels (local cubes of activity) depend only on their neighbours from the previous timestep[111].

Another application of structured LDSs is in temporal independent component analysis. This is defined as the discovery of a set of independent latent dynamical processes, from which the data is a projected observation. If each independent dynamical process can itself be described by a LDS, this gives rise to a structured LDS with a block diagonal transition matrix $\mathbf{A}$. Such models can be used to extract independent components under prior knowledge of the likely underlying frequencies in each of the temporal compoments[62].

24.5.5 Bayesian LDSs

The extension to placing priors on the transition and emission parameters of the LDS leads in general to computational difficulties in computing the likelihood. For example, for a prior on $\mathbf{A}$, the likelihood is $p(\mathbf{v}_{1:T}) = \int_{\mathbf{A}} p(\mathbf{v}_{1:T}|\mathbf{A})p(\mathbf{A})$ which is difficult to evaluate since the dependence of the likelihood on the matrix $\mathbf{A}$ is a complicated function. Approximate treatments of this case are beyond the scope of this book, although we briefly note that sampling methods[57, 105] are popular in this context, in addition to deterministic variational approximations[27, 23, 62].

24.6 Switching Auto-Regressive Models

Whilst the linear dynamical models considered so far in this chapter are powerful, they nevertheless have some inherent restrictions. For example, they cannot model abrupt changes in the observation. Here we describe an extension of the AR model. We consider a set of S different AR models, each with associated coefficients $\mathbf{a}(s)$, $s = 1, \ldots, S$, and allow the model to select one of these AR models at each time. For a @@ time-series of scalar values $v_{1:T}$ an L^{th} order switching AR model can be written as

$$v_t = \hat{\mathbf{v}}_{t-1}^{\mathsf{T}} \mathbf{a}(s_t) + \eta_t, \qquad \eta_t \sim \mathcal{N}\left(\eta_t | 0, \sigma^2(s_t)\right) \tag{24.6.1}$$

where we now have a set of AR coefficients $\theta = \left\{\mathbf{a}(s), \sigma^2(s), s \in \{1, \ldots, S\}\right\}$. The discrete switch variables themselves have a Markov transition $p(s_{1:T}) = \prod_t p(s_t|s_{t-1})$ so that the full model is, see fig(24.7),

$$p(v_{1:T}, s_{1:T}|\theta) = \prod_t p(v_t|v_{t-1}, \ldots, v_{t-L}, s_t, |\theta)p(s_t|s_{t-1}) \tag{24.6.2}$$

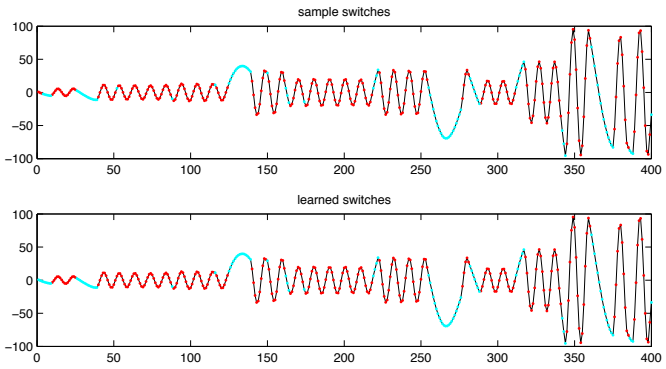

Figure 24.9: Learning a Switching AR model. The upper plot shows the train data. The colour indicates which of the two AR models is active at that time. Whilst this information is plotted here, this is assumed unknown to the learning algorithm, as are the coefficients $\mathbf{a}(s)$. We assume that the order $L = 2$ and number of switches $S = 2$ however is known. In the bottom plot we show the time series again after training in which we colour the points according to the most likely smoothed AR model at each timestep. See `demoSARlearn.m`.

24.6.1 Inference

Given an observed sequence $v_{1:T}$ and parameters θ inference is straightforward since this is a form of HMM. To make this more apparent we may write

$$p(v_{1:T}, s_{1:T}) = \prod_t \hat{p}(v_t|s_t)p(s_t|s_{t-1}) \tag{24.6.3}$$

where

$$\hat{p}(v_t|s_t) \equiv p(v_t|v_{t-1}, \ldots, v_{t-L}, s_t) = \mathcal{N}\left(v_t|\hat{\mathbf{v}}_{t-1}^\mathsf{T}\mathbf{a}(s_t), \sigma^2(s_t)\right) \tag{24.6.4}$$

Note that the emission distribution $\hat{p}(v_t|s_t)$ is time-dependent. The filtering recursion is then

$$\alpha(s_t) = \sum_{s_{t-1}} \hat{p}(v_t|s_t)p(s_t|s_{t-1})\alpha(s_{t-1}) \tag{24.6.5}$$

Smoothing can be achieved using the standard recursions, modified to use the time-dependent emissions, see `demoSARinference.m`.

With high frequency data it is unlikely that a change in the switch variable is reasonable at each time t. A simple constraint to account for this is to use a modified transition

$$\hat{p}(s_t|s_{t-1}) = \begin{cases} p(s_t|s_{t-1}) & \mathrm{mod}\,(t, T_{skip}) = 0 \\ \delta\,(s_t - s_{t-1}) & \text{otherwise} \end{cases} \tag{24.6.6}$$

24.6.2 Maximum likelihood learning using EM

To fit the set of AR coefficients and innovation variances, $\mathbf{a}(s), \sigma^2(s), s = 1, \ldots, S$, using maximum likelihood training for a set of data $v_{1:T}$, we may make use of the EM algorithm.

M-step

Up to negligible constants, the energy is given by

$$E = \sum_t \langle \log p(v_t|\hat{\mathbf{v}}_{t-1}, \mathbf{a}(s_t)) \rangle_{p^{old}(s_t|v_{1:T})} + \sum_t \langle \log p(s_t|s_{t-1}) \rangle_{p^{old}(s_t, s_{t-1})} \tag{24.6.7}$$

which we need to maximise with respect to the parameters θ. Using the definition of the emission and isolating the dependency on $\mathbf{a}$, we have

$$-2E = \sum_t \left\langle \frac{1}{\sigma^2(s_t)} \left(v_t - \hat{\mathbf{v}}_{t-1}^\mathsf{T}\mathbf{a}(s_t)\right)^2 + \log \sigma^2(s_t) \right\rangle_{p^{old}(s_t|v_{1:T})} + \text{const.} \tag{24.6.8}$$