1.4 Parameter estimation *

We can compute ML or MAP estimates of the parameters of an LDS using EM. The method is conceptually quite similar to the Baum-Welch algorithm for HMMs, except we use Kalman smoothing instead of forwards-backwards in the E step, and use different calculations in the M step. We give the detals below. For simplicity, we ignore the inputs $\mathbf{u}_{1:T}$ and the matrices $\mathbf{B}$ and $\mathbf{D}$.

Let ℓ_c represent the expected complete data log-likelihood, which, for a single sequence, is given by

$$\ell_c(\boldsymbol{\theta}) = \log p(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}) = -\sum_{t=1}^{T} \left(\frac{1}{2}(\mathbf{y}_t - \mathbf{C}\mathbf{z}_t)^T \mathbf{R}^{-1}(\mathbf{y}_t - \mathbf{C}\mathbf{z}_t) \right) - \frac{T}{2}\log|\mathbf{R}| \tag{1.45}$$

$$-\sum_{t=1}^{T} \left(\frac{1}{2}(\mathbf{z}_t - \mathbf{A}\mathbf{z}_{t-1})^T \mathbf{Q}^{-1}(\mathbf{z}_t - \mathbf{A}\mathbf{z}_{t-1}) \right) - \frac{T}{2}\log|\mathbf{Q}| \tag{1.46}$$

$$-\frac{1}{2}(\mathbf{z}_1 - \mathbf{m}_1)^T \boldsymbol{\Sigma}_1^{-1}(\mathbf{z}_1 - \mathbf{m}_1) - \frac{1}{2}\log|\boldsymbol{\Sigma}_1| + \text{const} \tag{1.47}$$

It will help to rewrite this using the trace trick:

$$\ell_c(\boldsymbol{\theta}) = -\sum_{t=1}^{T} \left(\frac{1}{2}\text{tr}(\mathbf{R}^{-1}(\mathbf{y}_t\mathbf{y}_t^T + \mathbf{C}\mathbf{z}_t\mathbf{z}_t^T\mathbf{C}^T - \mathbf{C}\mathbf{z}_t\mathbf{y}_t^T - \mathbf{y}_t\mathbf{z}_t^T\mathbf{C}^T)) \right) + \frac{T}{2}\log|\mathbf{R}^{-1}| \tag{1.48}$$

$$-\sum_{t=1}^{T} \left(\frac{1}{2}\text{tr}(\mathbf{Q}^{-1}(\mathbf{z}_t\mathbf{z}_t^T + \mathbf{A}\mathbf{z}_{t-1}\mathbf{z}_{t-1}^T\mathbf{A}^T - \mathbf{A}\mathbf{z}_{t-1}\mathbf{z}_t^T - \mathbf{z}_t\mathbf{z}_{t-1}\mathbf{A}^T)) \right) + \frac{T}{2}\log|\mathbf{Q}^{-1}| \tag{1.49}$$

$$-\frac{1}{2}\text{tr}(\boldsymbol{\Sigma}_1^{-1}(\mathbf{z}_1\mathbf{z}_1^T + \mathbf{m}_1\mathbf{m}_1^T - \mathbf{m}_1\mathbf{z}_1^T - \mathbf{z}_1\mathbf{m}_1^T)) + \frac{1}{2}\log|\boldsymbol{\Sigma}_1^{-1}| + \text{const} \tag{1.50}$$

We need to take expectations of this, so we define the following notation:

$$\boldsymbol{\mu}_t = \mathbb{E}[\mathbf{z}_t|\mathbf{y}_{1:T}] = \boldsymbol{\mu}_{t|T} \tag{1.51}$$

$$\mathbf{P}_t = \mathbb{E}\left[\mathbf{z}_t\mathbf{z}_t^T|\mathbf{y}_{1:T}\right] = \mathbf{V}_{t|T} + \boldsymbol{\mu}_{t|T}\boldsymbol{\mu}_{t|T}^T \tag{1.52}$$

$$\mathbf{P}_{t,t-1} = \mathbb{E}\left[\mathbf{z}_t\mathbf{z}_{t-1}^T|\mathbf{y}_{1:T}\right] = \mathbf{V}_{t,t-1|T} + \boldsymbol{\mu}_{t|T}\boldsymbol{\mu}_{t-1|T}^T \tag{1.53}$$

We can compute these two-sliced smoothed marginals as follows. If we define $\gamma_t = \alpha_t\beta_t$, where $\gamma_t(z) = p(\mathbf{z}_t|\mathbf{y}_{1:T})$ is the smoothed posterior, $\alpha_t(z) = p(\mathbf{z}_t|\mathbf{y}_{1:t})$ is the filtered posterior, and $\beta_t(z) \propto p(\mathbf{y}_{t+1}|\mathbf{z}_t)$ as the conditional likelihood, we can write, by analogy to Equation **??**,

$$p(\mathbf{z}_{t-1}, \mathbf{z}_t|\mathbf{y}_{1:T}) \propto \alpha(\mathbf{z}_{t-1})p(\mathbf{y}_t|\mathbf{z}_t)p(\mathbf{z}_t|\mathbf{z}_{t-1})\beta(\mathbf{z}_t) \tag{1.54}$$

One can show (Exercise 13.31 of [**?**]) that the mean of this is $\mathbb{E}[\mathbf{z}_{t-1}, \mathbf{z}_t|\mathbf{y}_{1:T}] = (\boldsymbol{\mu}_{t-1|T}, \boldsymbol{\mu}_{t|T})$, and the covariance is

$$\mathbf{V}_{t,t-1|T} = \text{cov}[\mathbf{z}_t, \mathbf{z}_{t-1}|\mathbf{y}_{1:T}] = \mathbf{J}_{t-1}\mathbf{V}_{t|T} \tag{1.55}$$

With these quantities in hand, it is easy to derive the expected complete data log likelihood

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \sum_{i=1}^{N} \mathbb{E}_{p(\mathbf{z}|\mathbf{y}_{1:T,i}, \boldsymbol{\theta}^{old})}[\log p(\mathbf{y}_{1:T,i}, \mathbf{z}_{1:T,i}|\boldsymbol{\theta})] \tag{1.56}$$

1.4.1 ML estimation

For notational simplicity, we derive the gradient for a single sequence, although we state the M step for multiple sequences. In the multiple sequence case, we define $T = \sum_{i=1}^{N} T_i$ as the total length of all the sequences. These results are based on [GH96b]; we leave their detailed derivation as an exercise.

- Output matrix

$$\frac{\partial \mathcal{Q}}{\partial \mathbf{C}} = \sum_{t=1}^{T} \mathbf{R}^{-1}\mathbf{y}_t\boldsymbol{\mu}_t^T - \sum_{t=1}^{T} \mathbf{R}^{-1}\mathbf{C}\mathbf{P}_t = 0 \tag{1.57}$$

$$\hat{\mathbf{C}} = (\sum_{i=1}^{N}\sum_{t=1}^{T_i} \mathbf{y}_{ti}\boldsymbol{\mu}_{ti}^T)(\sum_{i=1}^{N}\sum_{t=1}^{T_i} \mathbf{P}_{t,i})^{-1} \tag{1.58}$$

- Output noise covariance

$$\frac{\partial Q}{\partial \mathbf{R}^{-1}} = \frac{T}{2}\mathbf{R} - \frac{1}{2}\sum_{t=1}^{T}\left(\mathbf{y}_t\mathbf{y}_t^T - \mathbf{C}\boldsymbol{\mu}_t\mathbf{y}_t^T - \mathbf{y}_t\boldsymbol{\mu}_t^T\mathbf{C}^T + \mathbf{C}\mathbf{P}_t\mathbf{C}^T\right) = 0 \tag{1.59}$$

$$\hat{\mathbf{R}} = \frac{1}{T}\sum_{i=1}^{N}\sum_{t=1}^{T_i}\left[\mathbf{y}_{ti}\mathbf{y}_{ti}^T - \hat{\mathbf{C}}\boldsymbol{\mu}_{ti}\mathbf{y}_{ti}^T - \mathbf{y}_{ti}\boldsymbol{\mu}_{ti}^T(\hat{\mathbf{C}})^T + \hat{\mathbf{C}}\mathbf{P}_{ti}\hat{\mathbf{C}}^T\right] \tag{1.60}$$

$$= \frac{1}{T}\left[(\sum_{i=1}^{N}\sum_{t=1}^{T_i}\mathbf{y}_{ti}\mathbf{y}_{ti}^T) - \hat{\mathbf{C}}\left(\sum_{i=1}^{N}\sum_{t=1}^{T_i}\boldsymbol{\mu}_{ti}\mathbf{y}_{ti}^T\right)\right] \tag{1.61}$$

- State dynamics matrix

$$\frac{\partial Q}{\partial \mathbf{A}} = -\sum_{t=2}^{T}\mathbf{Q}^{-1}\mathbf{P}_{t,t-1} + \sum_{t=2}^{T}\mathbf{Q}^{-1}\mathbf{A}\mathbf{P}_{t-1} = 0 \tag{1.62}$$

$$\hat{\mathbf{A}} = \left(\sum_{i=1}^{N}\sum_{t=2}^{T_i}\mathbf{P}_{t,t-1,i}\right)\left(\sum_{i=1}^{N}\sum_{t=2}^{T_i}\mathbf{P}_{t-1,i}\right)^{-1} \tag{1.63}$$

- State noise covariance

$$\frac{\partial Q}{\partial \mathbf{Q}^{-1}} = \frac{T-1}{2}\mathbf{Q} - \frac{1}{2}\sum_{t=2}^{T}(\mathbf{P}_t - \mathbf{A}\mathbf{P}_{t-1,t} - \mathbf{P}_{t,t-1}\mathbf{A}^T + \mathbf{A}\mathbf{P}_{t-1}\mathbf{A}) = 0 \tag{1.64}$$

$$\hat{\mathbf{Q}} = \frac{1}{T-N}\sum_{i=1}^{N}\sum_{t=2}^{T_i}\left[\mathbf{P}_{t,i} - \hat{\mathbf{A}}\mathbf{P}_{t-1,t,i} - \mathbf{P}_{t,t-1,i}\hat{\mathbf{A}}^T + \hat{\mathbf{A}}\mathbf{P}_{t-1,i}\hat{\mathbf{A}}^T\right] \tag{1.65}$$

$$= \frac{1}{T-N}\left[\sum_{i=1}^{N}\sum_{t=2}^{T_i}\mathbf{P}_{t,i} - \hat{\mathbf{A}}\left(\sum_{i=1}^{N}\sum_{t=2}^{T}\mathbf{P}_{t-1,t,i}\right)\right] \tag{1.66}$$

- Initial mean

$$\hat{\mathbf{m}}_1 = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\mu}_{i,1} \tag{1.67}$$

- Initial covariance

$$\hat{\boldsymbol{\Sigma}}_1 = \frac{1}{N}\sum_{i=1}^{N}\left[\mathbb{E}\left[\mathbf{z}_{1i}\mathbf{z}_{1i}^T\right] + \hat{\mathbf{m}}_1\hat{\mathbf{m}}_1^T - \hat{\mathbf{m}}_1\mathbb{E}\left[\mathbf{z}_{1i}\right]^T - \mathbb{E}\left[\mathbf{z}_{1i}\right]\hat{\mathbf{m}}_1^T\right] \tag{1.68}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left[\mathbf{V}_{i,1} + (\hat{\mathbf{m}}_1 - \boldsymbol{\mu}_{i,1})(\hat{\mathbf{m}}_1 - \boldsymbol{\mu}_{i,1})^T\right] \tag{1.69}$$

Often we require $\boldsymbol{\Sigma}_1$ to be diagonal, to ensure a stable estimate.

Note that we can set $\mathbf{Q} = \mathbf{I}$ without loss of generality, since an arbitrary noise covariance can be modeled by appropriately modifying $\mathbf{A}$. Also, by analogy with factor analysis, we can require $\mathbf{R}$ to be diagonal. Again, this is without loss of generality. Doing this reduces the number of free parameters and improves numerical stability.

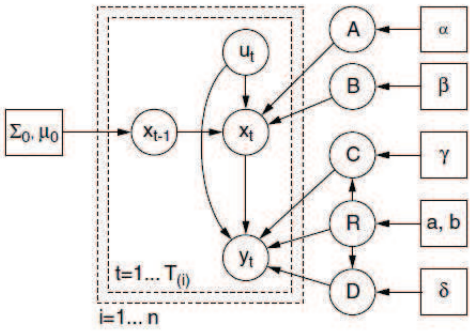

Figure 1.5: Representation of the priors we are using for a linear dynamical system. Source: Figure 5.3 of [**?**].

1.4.1.1 Offset term

It is often useful to add an offset term $\mathbf{b}$ to the output variables, that is, the observation model becomes

$$p(\mathbf{y}_t|\mathbf{z}_t, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{y}_t|\mathbf{b} + \mathbf{C}\mathbf{z}_t, \mathbf{R}) \tag{1.70}$$

We cannot use the technique adopted in Section **??**, where we centered the response and the input, since the input is not observed. So instead we define $\tilde{\mathbf{z}}_t = [1; \mathbf{z}_t]$, and $\tilde{\mathbf{C}} = [\mathbf{b},\ \mathbf{C}]$, so $\tilde{\mathbf{C}}\tilde{\mathbf{z}}_t = \mathbf{b} + \mathbf{C}\mathbf{z}_t$. (Hence $\tilde{\mathbf{C}}$ is a $D \times (L+1)$ matrix, where D is the size of $\mathbf{y}_t$ and L is the size of $\mathbf{z}_t$.) Also, let us define $\tilde{\boldsymbol{\mu}}_t = [1; \boldsymbol{\mu}_t]$, and

$$\tilde{\mathbf{P}}_t = \mathbb{E}\left[\begin{pmatrix}1\\ \mathbf{z}_t\end{pmatrix}\begin{pmatrix}1 & \mathbf{z}_t\end{pmatrix}\right] = \mathbb{E}\left[\begin{pmatrix}1 & \mathbf{z}_t^T\\ \mathbf{z}_t & \mathbf{z}_t\mathbf{z}_t^T\end{pmatrix}\right] = \begin{pmatrix}1 & \boldsymbol{\mu}_t^T\\ \boldsymbol{\mu}_t & \mathbf{P}_t\end{pmatrix} \tag{1.71}$$

Then the M step becomes

$$\hat{\tilde{\mathbf{C}}} = (\sum_{i=1}^{N}\sum_{t=1}^{T_i}\mathbf{y}_{ti}\tilde{\boldsymbol{\mu}}_{ti}^T)(\sum_{i=1}^{N}\sum_{t=1}^{T_i}\tilde{\mathbf{P}}_{t,i})^{-1} \tag{1.72}$$

$$\hat{\mathbf{R}} = \frac{1}{T}\left[(\sum_{i=1}^{N}\sum_{t=1}^{T_i}\mathbf{y}_{ti}\mathbf{y}_{ti}^T) - \hat{\tilde{\mathbf{C}}}\left(\sum_{i=1}^{N}\sum_{t=1}^{T_i}\tilde{\boldsymbol{\mu}}_{ti}\mathbf{y}_{ti}^T\right)\right] \tag{1.73}$$

1.4.2 MAP estimation

Direct application of the above equations often results in numerical instabilities. One way to regularize the problem is to use MAP estimation. We use the same priors and notation as [**?**, ch5], who present a variational Bayesian treatment of this model. We simplify their results (to avoid notational overload) by just giving point estimates of the parameters, and by omitting consideration of the input sequence $\mathbf{u}_{1:T}$. and the $\mathbf{B}$ and $\mathbf{D}$ matrices.

We pick conjugate Gaussian or conditionally Gaussian priors for each row of the various weight matrices, and Gamma priors for the elements precision matrix $\mathbf{R}$, More precisely, we have the following, where $\mathbf{A}_{j,:}$ refers to row j of $\mathbf{A}$ treated as a column vector:

$$p(\mathbf{A}) = \prod_j \mathcal{N}(\mathbf{A}_{j,:}|\mathbf{0}, \text{diag}(\boldsymbol{\alpha})^{-1}) \tag{1.74}$$

$$p(\mathbf{C}|\mathbf{R}) = \prod_j \mathcal{N}(\mathbf{C}_{j,:}|\mathbf{0}, R_{jj}\text{diag}(\boldsymbol{\gamma})^{-1}) \tag{1.75}$$

$$p(\text{diag}(\mathbf{R})) = \prod_j \text{Ga}(R_{jj}^{-1}|a, b) \tag{1.76}$$

(We do not condition $\mathbf{A}$ on $\mathbf{Q}$ because we assume $\mathbf{Q} = \mathbf{I}$.) We do not put a prior on the initial distribution $\mathbf{m}_1$, $\mathbf{\Sigma}_1$. Our assumptions are illustrated in Figure 1.5. To derive the modified M step, we just need to maximize $\mathcal{Q}'(\boldsymbol{\theta}) = \mathcal{Q}(\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})$. We explain how to do this below.

- State dynamics matrix

$$\log p(\mathbf{A}) = -\frac{1}{2}\sum_j \mathbf{A}_{j,:}^T \text{diag}(\boldsymbol{\alpha})\mathbf{A}_{j,:} \tag{1.77}$$

$$\frac{\partial \mathcal{Q}'}{\partial \mathbf{A}} = -\sum_{t=2}^{T}\mathbf{P}_{t,t-1} + \sum_{t=2}^{T}\mathbf{A}\mathbf{P}_{t-1} - \text{diag}(\boldsymbol{\alpha}).*\mathbf{A} = 0 \tag{1.78}$$

$$\hat{\mathbf{A}} = \left(\sum_{i=1}^{N}\sum_{t=2}^{T_i}\mathbf{P}_{t,t-1,i}\right)\left(\text{diag}(\boldsymbol{\alpha}) + \sum_{i=1}^{N}\sum_{t=2}^{T_i}\mathbf{P}_{t-1,i}\right)^{-1} \tag{1.79}$$

- Output matrix

$$\log p(\mathbf{C}|\mathbf{R}) = -\frac{1}{2}\sum_j \mathbf{C}_{j,:}^T(\text{diag}(\boldsymbol{\gamma}).*\mathbf{R}^{-1})\mathbf{C}_{j,:} \tag{1.80}$$

$$\frac{\partial \mathcal{Q}'}{\partial \mathbf{C}} = +\sum_{t=1}^{T}\mathbf{R}^{-1}\mathbf{y}_t\boldsymbol{\mu}_t^T - \sum_{t=1}^{T}\mathbf{R}^{-1}\mathbf{C}\mathbf{P}_t - \mathbf{R}^{-1}.*\text{diag}(\boldsymbol{\gamma})\mathbf{C} = 0 \tag{1.81}$$

$$\hat{\tilde{\mathbf{C}}} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T_i}\mathbf{y}_{ti}\tilde{\boldsymbol{\mu}}_{ti}^T\right)\left(\text{diag}(\tilde{\boldsymbol{\gamma}}) + \sum_{i=1}^{N}\sum_{t=1}^{Ti}\tilde{\mathbf{P}}_{t,i}\right)^{-1} \tag{1.82}$$

where $\tilde{\boldsymbol{\gamma}}$ is $D+1$ in length. (Often we set the first component of $\tilde{\boldsymbol{\gamma}}$ to 0, so that $\mathbf{b}$ will not be regularized.) From $\tilde{\mathbf{C}}$ we can extract $\mathbf{b}$ and $\mathbf{C}$.

- Observation covariance. The log prior has the form

$$\log p(\mathbf{R}) = \sum_k (a-1)\log R_{kk}^{-1} - R_{kk}^{-1}b \tag{1.83}$$

But we also have a dependence on $\mathbf{R}$ from the $p(\mathbf{C}|\mathbf{R})$ term. So the objective function (for a single sequence) has the form

$$\mathcal{Q}'(\mathbf{R}) = \frac{T}{2}\log|\mathbf{R}^{-1}| - \frac{1}{2}\sum_{t=1}^{T}\text{tr}\left(\mathbf{R}^{-1}(\mathbf{y}_t\mathbf{y}_t^T - \hat{\tilde{\mathbf{C}}}\sum_{t=1}^{T}\tilde{\boldsymbol{\mu}}_t\mathbf{y}_t^T)\right) \tag{1.84}$$

$$+\sum_k (a-1)\log R_{kk}^{-1} - bR_{kk}^{-1} - \frac{1}{2}\sum_j\sum_k \tilde{C}_{jk}^2\gamma_k R_{jj}^{-1} \tag{1.85}$$

Taking derivatives wrt R_{jj}^{-1} we have

$$\frac{\partial \mathcal{Q}'}{\partial R_{jj}^{-1}} = \frac{T}{2}R_{jj} - \frac{1}{2}G_{jj} + (a-1)R_{jj} - b - \frac{1}{2}\sum_k \tilde{C}_{jk}^2\gamma_k \tag{1.86}$$

where

$$\mathbf{G} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T_i}\mathbf{y}_{ti}\mathbf{y}_{ti}^T\right) - \hat{\tilde{\mathbf{C}}}\left(\sum_{i=1}^{N}\sum_{t=1}^{T_i}\tilde{\boldsymbol{\mu}}_{ti}\mathbf{y}_{ti}^T\right) \tag{1.87}$$

Hence the MAP estimate is

$$R_{jj} = \frac{G_{jj} + 2b + \sum_k \tilde{C}_{jk}^2\gamma_k}{2(a-1)+T} \tag{1.88}$$