DS-GA 3001.001 Special Topics in Data Science: Modeling Time Series
Handout: Expectation maximization for LDS

1 LSD model specification

The latent variables $\mathbf{z}_i$ have first order Markov structure with linear dynamics and i.i.d. gaussian noise:

$$\mathbf{z}_i = \mathbf{A}\mathbf{z}_{i-1} + \mathbf{w}_i \tag{1}$$

where $\mathbf{z}_0 \sim \mathcal{N}(\mu_0, \Sigma_0)$ and $\mathbf{w}_i \sim \mathcal{N}(0, \mathbf{Q})$.

Observations are obtained by another linear projection with i.i.d. gaussian noise:

$$\mathbf{x}_i = \mathbf{C}\mathbf{z}_i + \mathbf{v}_i \tag{2}$$

with $\mathbf{v}_i \sim \mathcal{N}(0, \mathbf{R})$.

2 Inference in LDS summary

During filtering (forward pass) information flows rightwards from index 1 to t with the beliefs about current state $\mathbf{z}_i$ updated as a function of past state $\mathbf{z}_{i-1}$ and current observation $\mathbf{x}_i$:

$$\begin{aligned}
\mu_{i|i} &= \mu_{i|i-1} + \mathbf{K}_i \left(x_i - \mathbf{C}\mu_{i|i-1}\right) && (3)\\
\Sigma_{i|i} &= \Sigma_{i|i-1} - \mathbf{K}_i \mathbf{C} \Sigma_{i|i-1} && (4)\\
\mathbf{K}_i &= \Sigma_{i|i-1}\mathbf{C}^{\mathrm{t}} \left(\mathbf{C}\Sigma_{i|i-1}\mathbf{C}^{\mathrm{t}} + R\right)^{-1} && (5)\\
\mu_{i|i-1} &= \mathbf{A}\mu_{i-1|i-1} && (6)\\
\Sigma_{i|i-1} &= \mathbf{A}\Sigma_{i-1|i-1}\mathbf{A}^{\mathrm{t}} + \mathbf{Q} && (7)
\end{aligned}$$

In contrast, during smoothing (backward pass) the beliefs about the current state $\mathbf{z}_i$ are updated based on how well it predicts the next state $\mathbf{z}_{i+1}$ relative to its already computed posterior marginal:

$$\begin{aligned}
\mu_{i|t} &= \mu_{i|i} + \mathbf{F}_i \left(\mu_{i+1|t} - \mu_{i+1|i}\right) && (8)\\
\Sigma_{i|t} &= \Sigma_{i|i} + \mathbf{F}_i \left(\Sigma_{i+1|t} - \Sigma_{i+1|i}\right) \mathbf{F}_i^{\mathrm{t}} && (9)\\
\mathbf{F}_i &= \Sigma_{i|i}\mathbf{A}^{\mathrm{t}}\Sigma_{i+1|i}^{-1} && (10)
\end{aligned}$$

3 Parameter learning

The *expectation maximization* algorithm provides a general way for maximum likelihood estimation of parameters in models with latent variables. Here I'm using the Bishop version of the EM derivation, which differs slightly in the way it sets up the lower bound on the likelihood (but see Roweis and Ghahramani 1999 for the interpretation of EM as coordinate ascent on the free energy, as discussed in the lecture). We use $\mathbf{x}$ to denote all observed variables and $\mathbf{z}$ to denote all latents. To compute the ML estimates, we need to find the parameters that maximize the likelihood function

$$\mathrm{P}\left(\mathbf{x}|\theta\right) = \int_Z \mathrm{P}\left(\mathbf{x}, \mathbf{z}|\theta\right) \mathbf{dz} \tag{11}$$

where we have marginalized out the unknown latents.

In general, this quantity is hard to optimize w.r.t. θ (and possibly hard to compute to begin with), whereas the complete-data likelihood $\mathrm{P}\left(\mathbf{x}, \mathbf{z}|\theta\right)$ is more tractable. We can introduce a distribution $q(\mathbf{z})$,

and use it to rewrite the likelihood in a more convenient form:

$$\begin{aligned}
\log \mathrm{P}\left(\mathbf{x}|\theta\right) &= \log \mathrm{P}\left(\mathbf{x}|\theta\right) \int_Z q(\mathbf{z}) \mathbf{dz} = \int_Z q(\mathbf{z}) \log \mathrm{P}\left(\mathbf{x}|\theta\right) \mathbf{dz} && (12)\\
&= \int_Z q(\mathbf{z}) \left(\log \mathrm{P}\left(\mathbf{x}, \mathbf{z}|\theta\right) - \log \mathrm{P}\left(\mathbf{z}|\mathbf{x}, \theta\right)\right) \mathbf{dz} && (13)\\
&= \int_Z q(\mathbf{z}) \left(\log \mathrm{P}\left(\mathbf{x}, \mathbf{z}|\theta\right) - \log q(\mathbf{z}) + \log q(\mathbf{z}) - \log \mathrm{P}\left(\mathbf{z}|\mathbf{x}, \theta\right)\right) \mathbf{dz} && (14)\\
&= \int_Z q(\mathbf{z}) \log \frac{\mathrm{P}\left(\mathbf{x}, \mathbf{z}|\theta\right)}{q(\mathbf{z})} \mathbf{dz} - \int_Z q(\mathbf{z}) \log \frac{\mathrm{P}\left(\mathbf{z}|\mathbf{x}, \theta\right)}{q(\mathbf{z})} \mathbf{dz} && (15)\\
&= \mathcal{L}(q, \theta) + \mathrm{KL}\left(q(\mathbf{z}) || \mathrm{P}\left(\mathbf{z}|\mathbf{x}, \theta\right)\right) && (16)
\end{aligned}$$

where KL denotes the Kullback-Leibler divergence. Here we have used first the fact that $q(\mathbf{z})$ is a distribution so it integrates to 1, and then the chain rule in log form $\log \mathrm{P}\left(\mathbf{x}, \mathbf{z}|\theta\right) = \log \mathrm{P}\left(\mathbf{z}|\mathbf{x}, \theta\right) + \log \mathrm{P}\left(\mathbf{x}|\theta\right)$. Since $\mathrm{KL}(p||q) \geq 0$ for any pairs of distributions p, q, the functional $\mathcal{L}(q, \theta)$ is a lower bound for $\log \mathrm{P}\left(\mathbf{x}|\theta\right)$. Moreover, the bound becomes tight when the KL is zero, i.e. when $q(\mathbf{z}) = \mathrm{P}\left(\mathbf{z}|\mathbf{x}, \theta\right)$.

The optimization procedure proceeds in 2 steps: E- and M-. In the E-step, we are starting from a current value for parameters θ^{old}, and optimize $\mathcal{L}(q, \theta)$ with respect to q, which, as discussed above, will yield $q(\mathbf{z}) = \mathrm{P}\left(\mathbf{z}|\mathbf{x}, \theta^{\mathrm{old}}\right)$. In the M-step, q is kept fixed and $\mathcal{L}(q, \theta)$ is optimized w.r.t. θ, which will further increase the lower bound (unless we have already reached an optimum). Once we've fixed q to its optimum, the expression for the $\mathcal{L}(q, \theta)$ becomes:

$$\begin{aligned}
\mathcal{L}(q, \theta) &= \int_Z \mathrm{P}\left(\mathbf{z}|\mathbf{x}, \theta^{\mathrm{old}}\right) \log \frac{\mathrm{P}\left(\mathbf{x}, \mathbf{z}|\theta\right)}{\mathrm{P}\left(\mathbf{z}|\mathbf{x}, \theta^{\mathrm{old}}\right)} \mathbf{dz} && (17)\\
&= \int_Z \mathrm{P}\left(\mathbf{z}|\mathbf{x}, \theta^{\mathrm{old}}\right) \log \mathrm{P}\left(\mathbf{x}, \mathbf{z}|\theta\right) - \int_Z \mathrm{P}\left(\mathbf{z}|\mathbf{x}, \theta^{\mathrm{old}}\right) \log \mathrm{P}\left(\mathbf{z}|\mathbf{x}, \theta^{\mathrm{old}}\right) && (18)\\
&= \mathcal{Q}(\theta, \theta^{\mathrm{old}}) + \mathrm{const} && (19)
\end{aligned}$$

where the second term is the entropy of q, which is independent of θ. Hence in the M-step the quantity to optimize is the complete data (log)likelihood, with the parameter to optimize θ appearing only in the log (which means easy to optimize as long as the joint is the exponential family).

How do we apply this ideas to LDS? First, the E-step we will set $q(\mathbf{z}_*) = \mathrm{P}\left(\mathbf{z}_*|\mathbf{x}_*, \theta\right)$, where now the index * is used to mark the fact that we are looking at the whole data sequence. The marginals of the posterior distribution are the quantities computed by the Kalman smoother (we are always conditioning on the full observed data sequence during learning). We'll see in a bit that we actually need to also compute joint marginals at this point (details of why and how below).

In the M-step, we optimize $\mathcal{Q}(\theta, \theta^{\mathrm{old}}) = \mathbb{E}_{\mathbf{z}|\theta^{\mathrm{old}}}[\log \mathrm{P}\left(\mathbf{x}, \mathbf{z}|\theta\right)]$ with respect to the individual parameters. Given the dependency structure of the LDS generative model, the log of the joint separates into terms corresponding to the initial conditions, latent dynamics and observations, such that we can optimize their corresponding parameters separately:

$$\begin{aligned}
\mathcal{Q}(\theta, \theta^{\mathrm{old}}) &= \mathbb{E}_{\mathbf{z}|\theta^{\mathrm{old}}} \left[-\frac{1}{2} \log |\Sigma_0| - \frac{1}{2} \left(\mathbf{z}_0 - \mu_0\right)^{\mathrm{t}} \Sigma_0^{-1} \left(\mathbf{z}_0 - \mu_0\right)\right]\\
&+ \mathbb{E}_{\mathbf{z}|\theta^{\mathrm{old}}} \left[-\frac{t}{2} \log |\mathbf{Q}| - \frac{1}{2} \sum_i \left(\mathbf{z}_{i+1} - \mathbf{A}\mathbf{z}_i\right)^{\mathrm{t}} \mathbf{Q}^{-1} \left(\mathbf{z}_{i+1} - \mathbf{A}\mathbf{z}_i\right)\right]\\
&+ \mathbb{E}_{\mathbf{z}|\theta^{\mathrm{old}}} \left[-\frac{t}{2} \log |\mathbf{R}| - \frac{1}{2} \sum_i \left(\mathbf{x}_i - \mathbf{C}\mathbf{z}_i\right)^{\mathrm{t}} \mathbf{R}^{-1} \left(\mathbf{x}_i - \mathbf{C}\mathbf{z}_i\right)\right]\\
&+ \mathrm{const}
\end{aligned}$$

First, to determine the parameters of the initial state we have to optimize:

$$\mathcal{Q}(\theta, \theta^{\mathrm{old}}) = \mathbb{E}_{\mathbf{z}|\theta^{\mathrm{old}}} \left[-\frac{1}{2} \log |\Sigma_0| - \frac{1}{2} \left(\mathbf{z}_0 - \mu_0\right)^{\mathrm{t}} \Sigma_0^{-1} \left(\mathbf{z}_0 - \mu_0\right)\right] + \mathrm{const}$$

where all terms that do not depend on μ_0 and Σ_0 have been absorbed in the constant. If we note that the expression is a log multivariate gaussian in $\mathbf{z}_1$ and using the known ML parameter estimates for a multivariate gaussian, this yields:[1]

$$\mu_0^{\text{new}} = \mathbb{E}_{\mathbf{z}|\theta^{\text{old}}}[\mathbf{z}_1] = \mu_{0|t} \tag{20}$$

$$\Sigma_0^{\text{new}} = \mathbb{E}_{\mathbf{z}|\theta^{\text{old}}}\left[\mathbf{z}_1\mathbf{z}_1^{\text{t}}\right] - \mathbb{E}_{\mathbf{z}|\theta^{\text{old}}}[\mathbf{z}_1]\,\mathbb{E}_{\mathbf{z}|\theta^{\text{old}}}\left[\mathbf{z}_1^{\text{t}}\right] = \Sigma_{0|t} \tag{21}$$

To optimize the parameters $\mathbf{C}$ and $\mathbf{R}$ describing the observation model we need to optimize:

$$\mathcal{Q}(\theta, \theta^{\text{old}}) = \mathbb{E}_{\mathbf{z}|\theta^{\text{old}}}\left[-\frac{t}{2}\log|\mathbf{R}| - \frac{1}{2}\sum_i (\mathbf{x}_i - \mathbf{C}\mathbf{z}_i)^{\text{t}}\,\mathbf{R}^{-1}\,(\mathbf{x}_i - \mathbf{C}\mathbf{z}_i)\right]$$

If we treat $\mathbf{z}_i$ as observed, this would correspond to simple multivariate linear regression, after which we will have to take again expectations w.r.t. q (see appendix):[2]

$$\mathbf{C}^{\text{new}} = \left(\sum_i \mathbf{x}_i\mathbb{E}[\mathbf{z}_i^{\text{t}}]\right)\left(\sum_i \mathbb{E}[\mathbf{z}_i\mathbf{z}_i^{\text{t}}]\right)^{-1} \tag{22}$$

$$\mathbf{R}^{\text{new}} = \frac{1}{t}\sum_i \left\{\mathbf{x}_i\mathbf{x}_i^{\text{t}} - \mathbf{C}^{\text{new}}\mathbb{E}[\mathbf{z}_i]\mathbf{x}_i^{\text{t}} - \mathbf{x}_i\mathbb{E}[\mathbf{z}_i^{\text{t}}]\mathbf{C}^{\text{new t}} + \mathbf{C}^{\text{new}}\mathbb{E}[\mathbf{z}_i\mathbf{z}_i^{\text{t}}]\mathbf{C}^{\text{new t}}\right\} \tag{23}$$

where $\mathbb{E}[\mathbf{z}_i] = \mu_{i|t}$, $\mathbb{E}[\mathbf{z}_i\mathbf{z}_i^{\text{t}}] = \Sigma_{i|t} + \mu_{i|t}\mu_{i|t}^{\text{t}}$, as inferred by the Kalman smoother. Sanity check for dimensions: $\mathbf{C}$ has d_x rows and d_z columns; $\mathbf{x}\mathbf{x}^{\text{t}}$ is $d_x \times d_x$ dimensional, as is $\mathbf{R}$; $\mathbf{z}\mathbf{x}^{\text{t}}$ has d_z rows and d_x columns, which when multiplied with $\mathbf{C}$ on the left side gives a matrix of size $d_x \times d_x$; third term is just the transpose of the same, last term has the same dimensionality as $\mathbf{C}\mathbf{C}^{\text{t}}$ which is $d_x \times d_x$.

Lastly, we can optimize the parameters $\mathbf{A}$ and $\mathbf{Q}$ describing the latent space dynamics:

$$\mathcal{Q}(\theta, \theta^{\text{old}}) = \mathbb{E}_{\mathbf{z}|\theta^{\text{old}}}\left[-\frac{t}{2}\log|\mathbf{Q}| - \frac{1}{2}\sum_i (\mathbf{z}_{i+1} - \mathbf{A}\mathbf{z}_i)^{\text{t}}\,\mathbf{Q}^{-1}\,(\mathbf{z}_{i+1} - \mathbf{A}\mathbf{z}_i)\right]$$

This yields (e.g. taking the form of the expression above and replacing x_i with z_{i+1} in expectation):

$$\mathbf{A}^{\text{new}} = \left(\sum_i \mathbb{E}[\mathbf{z}_{i+1}\mathbf{z}_i^{\text{t}}]\right)\left(\sum_i \mathbb{E}[\mathbf{z}_i\mathbf{z}_i^{\text{t}}]\right)^{-1} \tag{24}$$

$$\mathbf{Q}^{\text{new}} = \frac{1}{t}\sum_i \left\{\mathbb{E}[\mathbf{z}_{i+1}\mathbf{z}_{i+1}^{\text{t}}] - \mathbf{A}^{\text{new}}\mathbb{E}[\mathbf{z}_i\mathbf{z}_{i+1}^{\text{t}}] - \mathbb{E}[\mathbf{z}_{i+1}\mathbf{z}_i^{\text{t}}]\mathbf{A}^{\text{new,t}} + \mathbf{A}^{\text{new}}\mathbb{E}[\mathbf{z}_i\mathbf{z}_i^{\text{t}}]\mathbf{A}^{\text{new t}}\right\} \tag{25}$$

The last thing to note is that beyond the Kalman soother computed moments $\mathbb{E}[\mathbf{z}_i]$, and $\mathbb{E}[\mathbf{z}_i\mathbf{z}_i^{\text{t}}]$, here we additionally need the joint posterior covariance $\mathbb{E}[\mathbf{z}_{i+1}\mathbf{z}_i^{\text{t}}] = \Sigma_{i+1|t}\mathbf{F}_i^{\text{t}} + \mu_{i+1|t}\mu_{i|t}^{\text{t}}$.

To prove this last expression, we need to explicitly write out the joint posterior for $\mathbf{z}_i$ and $\mathbf{z}_{i+1}$:

$$\mathrm{P}(\mathbf{z}_i, \mathbf{z}_{i+1}|\mathbf{x}_{1:t}) = \frac{\mathrm{P}(\mathbf{x}_{1:t}|\mathbf{z}_i, \mathbf{z}_{i+1})\,\mathrm{P}(\mathbf{z}_i, \mathbf{z}_{i+1})}{\mathrm{P}(\mathbf{x}_{1:t})} \tag{26}$$

$$= \frac{\mathrm{P}(\mathbf{x}_{1:i}|\mathbf{z}_i)\,\mathrm{P}(\mathbf{x}_{i+1}|\mathbf{z}_{i+1})\,\mathrm{P}(\mathbf{x}_{i+2:t}|\mathbf{z}_{i+1})\,\mathrm{P}(\mathbf{z}_{i+1}|\mathbf{z}_i)\,\mathrm{P}(\mathbf{z}_i)}{\mathrm{P}(\mathbf{x}_{1:t})} \tag{27}$$

$$= \frac{\mathrm{P}(\mathbf{x}_{1:i}, \mathbf{z}_i)\,\mathrm{P}(\mathbf{x}_{i+1}|\mathbf{z}_{i+1})\,\mathrm{P}(\mathbf{x}_{i+2:t}|\mathbf{z}_{i+1})\,\mathrm{P}(\mathbf{z}_{i+1}|\mathbf{z}_i)}{\mathrm{P}(\mathbf{x}_{1:i})\,\mathrm{P}(\mathbf{x}_{i+1}|\mathbf{x}_{1:i})\,\mathrm{P}(\mathbf{x}_{i+2:t}|\mathbf{x}_{1:i+1})} \tag{28}$$

$$= \mathrm{P}(\mathbf{z}_i|\mathbf{x}_{1:i})\,\frac{\mathrm{P}(\mathbf{z}_{i+1}|\mathbf{z}_i)\,\mathrm{P}(\mathbf{x}_{i+1}|\mathbf{z}_{1+i})}{\mathrm{P}(\mathbf{x}_{i+1}|\mathbf{x}_{1:i})}\,\frac{\mathrm{P}(\mathbf{z}_{i+1}|\mathbf{x}_{1:t})}{\mathrm{P}(\mathbf{z}_{i+1}|\mathbf{x}_{1:i+1})} \tag{29}$$

where we have used Bayes' rule, then the conditional independence relationships in LDS, the chain rule for joint $\mathrm{P}(\mathbf{x}_{1:t})$, and then reorganized some of the factors. All the individual components are multivariate gaussians. In particular, $\mathrm{P}(\mathbf{x}_{i+1}|\mathbf{x}_{1:i}) = \mathcal{N}(\mathbf{x}_{i+1}; \mathbf{C}\mathbf{A}\mu_{i|i}, \mathbf{C}\Sigma_{i|i}\mathbf{C}^{\text{t}} + \mathbf{R})$. Hence the result is also multivariate gaussian. We simply need to write out the exponent and massage it in the appropriate multivariate gaussian canonical form (tedious but doable, see also Bishop for more details).

[1] One way to think of this is that we first pretend that latents $\mathbf{z}$ are given, derive the ML estimates as if that were true, then at the very last step take the expectation under q.

[2] To simplify notation we drop the explicit distribution indexing in the expectation here.

Appendix: maximum likelihood for multivariate regression

Assume that observation pairs $\left(\mathbf{x}^{(n)}, \mathbf{y}^{(n)}\right),\ n = 1 : N$ are generated independently according to the process:[3]

$$\mathbf{y} = \mathbf{W}\mathbf{x} + \epsilon \tag{30}$$

where the term ϵ is zero mean gaussian noise, $\epsilon \sim \mathcal{N}(\mathbf{0}, \Sigma)$, and $\theta = \{\mathbf{W}, \Sigma\}$ is shorthand for the model parameters.

The corresponding likelihood function is:[4]

$$\mathrm{P}\left(\mathbf{y}_*|\mathbf{x}_*, \theta\right) = \prod_n \mathrm{P}\left(\mathbf{y}_n|\mathbf{x}_n\right) = \prod_n \mathcal{N}\left(\mathbf{y}_n; \mathbf{W}\mathbf{x}_n, \Sigma\right) \tag{31}$$

Maximum likelihood learning aims to find a setting for θ that maximizes the log likelihood:

$$\mathcal{L}(\theta) = \log \mathrm{P}\left(\mathbf{y}_*|\mathbf{x}_*, \theta\right) = -\frac{N}{2}\log|\Sigma| - \frac{1}{2}\sum_n \left(\mathbf{y}^{(n)} - \mathbf{W}\mathbf{x}^{(n)}\right)^{\mathrm{t}} \Sigma^{-1} \left(\mathbf{y}^{(n)} - \mathbf{W}\mathbf{x}^{(n)}\right) + \text{const} \tag{32}$$

where we have plugged in the definition of a multivariate gaussian.

First, let's consider the maximization with respect to $\mathbf{W}$. Taking the derivative of $\mathcal{L}$ w.r.t. $\mathbf{W}$ and setting the result to 0 yields:

$$\sum_n \mathbf{y}^{(n)}\mathbf{x}^{(n)\,\mathrm{t}} = \mathbf{W}\sum_n \mathbf{x}^{(n)}\mathbf{x}^{(n)\,\mathrm{t}} \tag{33}$$

Rearranging the terms we get the final maximum likelihood estimate for $\mathbf{W}$:

$$\hat{\mathbf{W}} = \left(\sum_n \mathbf{y}^{(n)}\mathbf{x}^{(n)\,\mathrm{t}}\right)\left(\sum_n \mathbf{x}^{(n)}\mathbf{x}^{(n)\,\mathrm{t}}\right)^{-1} \tag{34}$$

Sanity check for dimensionality: $\mathbf{y}\mathbf{x}^{\mathrm{t}}$ has d_y rows and d_x columns, just as $\mathbf{W}$; $\mathbf{x}\mathbf{x}^{\mathrm{t}}$ is $d_x \times d_x$ dimensional.

We also need find the ML estimate for the noise covariance, which takes the form:

$$\hat{\Sigma} \;=\; \frac{1}{N}\sum_n \left(\mathbf{y}^{(n)} - \mathbf{W}\mathbf{x}^{(n)}\right)\left(\mathbf{y}^{(n)} - \mathbf{W}\mathbf{x}^{(n)}\right)^{\mathrm{t}} \tag{35}$$

This is essentially the empirical covariance of the residuals.

[3] $\mathbf{x}$ and $\mathbf{y}$ are column vectors of size d_x and d_y, respectively.

[4] To simplify notation we shorthand all $\mathbf{x}^{(n)}$ data points as $\mathbf{x}_*$ and similarly all $\mathbf{y}^{(n)}$ data points as $\mathbf{y}_*$.