4G10 BMI: notes for lecture 2
linear dynamical systems

Yashar Ahmadian

October 16, 2023

The models we discussed for dimensionality in Lecture 2 ignored temporal inter-dependencies that exist in the neural data. This is certainly a problematic assumption from the point of view of generative or causal modeling, considering that neural activity results from the *dynamics* of recurrently connected neural circuits; but it is also problematic as it leads to sub-optimal inferences that can significantly hinder the proper operation of a BMI. To see this, consider a (hypothetical) population of neurons in your motor cortex that represents the momentary speed of your hand along a given direction. As we tend to perform smooth movements with hand speed changing continuously in time, the latent state $\boldsymbol{z}_t$ that we infer at time t should be constrained not only by the neural activity $\boldsymbol{x}_t$ observed at time t, but also by other temporally adjacent neural observations (e.g. $\boldsymbol{x}_{t-1}$ and $\boldsymbol{x}_{t+1}$), because of their statistical dependence (in the language of information theory, not only $\boldsymbol{x}_t$ but also neural activity at other times contain information about z_t). In other words, latent state inference ought to involve some form of 'filtering' or 'smoothing' of the observations.

The GPFA model discussed in Lecture 3 does take into account temporal inter-dependencies. However, the specific form of the Gaussian process prior on $\boldsymbol{z}_{1:T}$ in that specific model: (a) does not account for possible cross-correlations/interactions between different components of $\boldsymbol{z}$, and (b) hides the dynamical interactions responsible for creating the temporal dependencies. The model class we discuss below explicitly incorporates, in the generative model, the assumption of temporal continuity and smoothness in the underlying latent state trajectories, and moreover does so by explicit formulation in terms of dynamical systems. Bayesian inference on such a generative model then automatically uses "surrounding observations" to improve the inference of the latent states.

1 Gaussian linear dynamical systems

We start by describing the generative model and then, in the next section, discuss inference based on that model. As in the models in previous lectures, the generative model is specified by its two components: the prior on the latent trajectories, and the observation model relating the latent variables to the observed ones.

The generative prior on the latent trajectory. A simple form of prior over latent states that captures temporal continuity is the prior induced by linear stochastic dynamics of the form:

$$\boldsymbol{z}_{t+1} = A\boldsymbol{z}_t + B\boldsymbol{\epsilon}_{t+1}, \qquad \boldsymbol{\epsilon}_{t+1} \sim \mathcal{N}(0, I_M), \tag{1}$$

with the initial condtion distribution

$$\boldsymbol{z}_1 \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0). \tag{2}$$

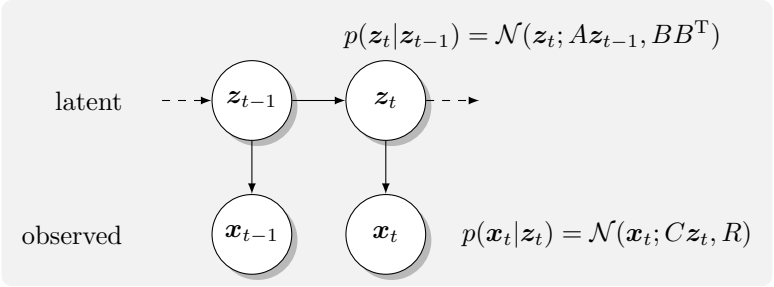


Figure 1: Graphical model associated with the LDS model.

Here A is the so-called "state matrix" (which describes the interaction between different components of $\boldsymbol{z}_t$), B is an "input matrix", and $\boldsymbol{\epsilon}_{t+1}$ is Gaussian white[1] noise. In other words, $p(\boldsymbol{z}_{t+1}|\boldsymbol{z}_t) = \mathcal{N}(\boldsymbol{z}_{t+1}; A\boldsymbol{z}_t, BB^{\mathrm{T}})$. Note that under the above model, $\boldsymbol{z}_{1:T}$ form a Markov chain.[2]

The system described above is mathematically a stochastic linear dynamical system; however, in the context of generative models, the term **"linear dynamical system" (LDS)** refers to the full generative model obtained when we combine the above prior with the FA-like Gaussian observation model, that we will now introduce.

The generative observation model. Similar to GPFA, we use a linear Gaussian model with independent observation noise, in which conditioned on $\boldsymbol{z}_t$, $\boldsymbol{x}_t$ is independent of $\boldsymbol{z}$ at other times. The dependence of $\boldsymbol{x}_t$ on $\boldsymbol{z}_t$ is given by

$$\boldsymbol{x}_t|\boldsymbol{z}_t \sim \mathcal{N}(C\boldsymbol{z}_t, R) \tag{3}$$

where R is a diagonal matrix.

The full generative model is described graphically in Figure 1. Note that this model thus constitutes a hidden Markov model (HMM).[3]

1.1 Inference

In LDSs, two different inference problems arise. The first is the "filtering" problem of inferring the latent state $\boldsymbol{z}_t$ based only on the observations *up until time t*. The second is called "smoothing", and consists of inferring $\boldsymbol{z}_t$ based on *all* observations, including future ones. While filtering is naturally more relevant to online BMI decoding applications, smoothing is an important building block for learning the parameters of the generative model.

Kalman filtering The filtering distribution $p(\boldsymbol{z}_t|\boldsymbol{x}_{0:t})$ is also known as the Kalman filter, which you will have come across in other 3F/4F modules. To derive the Kalman filter, we proceed by induction: we assume

[1] Both spatially and temporally; *i.e.*, the noise vectors at different times are independent.

[2] Furthermore, since Gaussian variables remain Gaussian under linear transformations, it can be checked that the joint distribution of the sequence $\boldsymbol{z}_{1:T}$ is multivariate Gaussian; with zero mean and a $TM \times TM$ covariance matrix. This observation can also be extended to continuous time Gaussian LDS. Thus, mathematically speaking, the Gaussian LDS described above, or their continuous time generalizations, constitute a subclass of Gaussian processes, but of a form that does not fall in the class of GP's employed in GPFA. For various reasons, LDS are not typically formulated in the language used for discussing GP's, namely in terms of a covariance kernel.

[3] But prior exposition to or experience with HMM's is not required – these notes are self-contained.

that $p(\boldsymbol{z}_t|\boldsymbol{x}_{0:t})$ is known and equal to $\mathcal{N}(\boldsymbol{z}_t; \boldsymbol{\mu}_t, \Sigma_t)$, and go on showing that $p(\boldsymbol{z}_{t+1}|\boldsymbol{x}_{0:t+1})$ is also Gaussian with moments $\boldsymbol{\mu}_{t+1}$ and Σ_{t+1} which we can obtain as a function of $\boldsymbol{\mu}_t$, Σ_t and $\boldsymbol{x}_{t+1}$:

$$p(\boldsymbol{z}_{t+1}|\boldsymbol{x}_{0:t+1}) = p(\boldsymbol{z}_{t+1}|\boldsymbol{x}_{t+1}, \boldsymbol{x}_{0:t}) \propto p(\boldsymbol{x}_{t+1}|\boldsymbol{z}_{t+1}, \boldsymbol{x}_{0:t})\, p(\boldsymbol{z}_{t+1}|\boldsymbol{x}_{0:t}) \quad \text{(Bayes' theorem)} \tag{4}$$

filtering distrib

$$= p(\boldsymbol{x}_{t+1}|\boldsymbol{z}_{t+1})\, p(\boldsymbol{z}_{t+1}|\boldsymbol{x}_{0:t}) \quad \text{(conditional independence of observations)} \tag{5}$$

$$= p(\boldsymbol{x}_{t+1}|\boldsymbol{z}_{t+1}) \int p(\boldsymbol{z}_{t+1}|\boldsymbol{z}_t, \boldsymbol{x}_{0:t})\, p(\boldsymbol{z}_t|\boldsymbol{x}_{0:t})\, d\boldsymbol{z}_t \quad \text{(sum and product rules)} \tag{6}$$

$$= p(\boldsymbol{x}_{t+1}|\boldsymbol{z}_{t+1}) \int p(\boldsymbol{z}_{t+1}|\boldsymbol{z}_t)\, p(\boldsymbol{z}_t|\boldsymbol{x}_{0:t})\, d\boldsymbol{z}_t \quad \text{(conditional independence: Markov property)} \tag{7}$$

Note that the above is valid for any HMM. For our Gaussian LDS model, substituting the Gaussian distributions for the observation and transition distributions, and using the hypothesis of our induction, we obtain

$$p(\boldsymbol{z}_{t+1}|\boldsymbol{x}_{0:t+1}) \propto \mathcal{N}(\boldsymbol{x}_{t+1}; C\boldsymbol{z}_{t+1}, R) \int \mathcal{N}(\boldsymbol{z}_{t+1}; A\boldsymbol{z}_t, BB^{\mathrm{T}})\, \mathcal{N}(\boldsymbol{z}_t; \boldsymbol{\mu}_t, \Sigma_t)\, d\boldsymbol{z}_t. \tag{8}$$

Luckily, we do not have to explicitly evaluate the integral in the last equation: first note that (as in our earlier calculation of the marginal likelihood in pPCA) this integral is nothing but an expression for the marginal density of $\boldsymbol{z}_{t+1} = A\boldsymbol{z}_t + B\boldsymbol{\epsilon}_{t+1}$, with $\boldsymbol{z}_t$ distributed according to $\boldsymbol{z}_t \sim \mathcal{N}(\boldsymbol{\mu}_t, \Sigma_t)$ (by the assumption of our induction) and $\boldsymbol{\epsilon}_{t+1} \sim \mathcal{N}(\mathbf{0}, I_M)$. As $\boldsymbol{z}_{t+1}$ is a linear combination of two Gaussian random variables, it is itself Gaussian (*i.e.*, its marginal distribution is Gaussian). In particular, it suffices to evaluate the mean and covariance matrix of $\boldsymbol{z}_{t+1}$ in order to determine its marginal (Gaussian) distribution. Since $\boldsymbol{\epsilon}_{t+1}$ has zero mean, the mean of $\boldsymbol{z}_{t+1}$ is $A\boldsymbol{\mu}_t$, and since $\boldsymbol{\epsilon}_{t+1}$ is independent of $\boldsymbol{z}_t$, the covariance matrix of $\boldsymbol{z}_{t+1}$ is the sum of the covariance matrices of the two terms $A\boldsymbol{z}_t$ and $B\boldsymbol{\epsilon}_{t+1}$, which is $A\Sigma_t A^{\mathrm{T}} + BB^{\mathrm{T}}$. Thus, we can rewrite Equation (8) as

$$p(\boldsymbol{z}_{t+1}|\boldsymbol{x}_{0:t+1}) \propto \mathcal{N}(\boldsymbol{x}_{t+1}; C\boldsymbol{z}_{t+1}, R)\, \mathcal{N}(\boldsymbol{z}_{t+1}; A\boldsymbol{\mu}_t, A\Sigma_t A^{\mathrm{T}} + BB^{\mathrm{T}}). \tag{9}$$

It is easy to see that the right hand side is again a Gaussian density over $\boldsymbol{z}_{t+1}$ (because both factors are exponentials of quadratic expressions in z_{t+1}). The mean, $\boldsymbol{\mu}_{t+1}$, and covariance, Σ_{t+1}, can then be obtained using the same type of calculation as those that led to the expression for mean and covariance of the posterior distribution in pPCA (lecture 2). The result is

$$\Sigma_{t+1}^{-1} = C^{\mathrm{T}} R^{-1} C + \left(A\Sigma_t A^{\mathrm{T}} + BB^{\mathrm{T}}\right)^{-1} \tag{10}$$

$$\boldsymbol{\mu}_{t+1} = \Sigma_{t+1}\left[(A\Sigma_t A^{\mathrm{T}} + BB^{\mathrm{T}})^{-1} A\boldsymbol{\mu}_t + C^{\mathrm{T}} R^{-1} \boldsymbol{x}_{t+1}\right] \tag{11}$$

Given that $p(\boldsymbol{z}_0)$ is Gaussian, inductive reasoning allows us to conclude that the filtering distribution is indeed Gaussian at all times. Finally, Equation (10) can be simplified using the Woodbury identity $(A+UBU^{\mathrm{T}})^{-1} = A^{-1} - A^{-1}U(B^{-1} + U^{\mathrm{T}}A^{-1}U)^{-1}U^{\mathrm{T}}A^{-1}$, leading to a more interpretable form:

$$\boldsymbol{\mu}_{t+1} = \underbrace{A\boldsymbol{\mu}_t}_{\text{prediction}} + \underbrace{\underbrace{\Sigma_{t+1}C^{\mathrm{T}}R^{-1}}_{\text{Kalman gain}}\underbrace{(\boldsymbol{x}_{t+1} - CA\boldsymbol{\mu}_t)}_{\text{prediction error}}}_{\text{correction}} \tag{12}$$

Without the "correction" term μ_t would evolve according to a deterministic (noise-free) version of the latent linear dynamics. This is indeed how the marginal means of $\boldsymbol{z}_t$ evolve under the *prior* distribution on the latent trajectory. However, *a posteriori* (*i.e.*, after having observed $\boldsymbol{x}_{1:T}$) this prior prediction has to be corrected depending on the new observation $\boldsymbol{x}_{t+1}$, and according to Equation (12), the correction is proportional to the prediction error $\boldsymbol{x}_{t+1} - CA\boldsymbol{\mu}_t$.

$$\Sigma_{t+1} = \left(\Sigma^{-1} + C^{\mathrm{T}}R^{-1}C\right)^{-1} = \Sigma - \Sigma C^{\mathrm{T}}\left(R + C\Sigma C^{\mathrm{T}}\right)^{-1}C\Sigma$$

$$\Sigma_{t+1} = \Sigma - K_t C\Sigma = (I - K_t C)\Sigma$$

$$\mu_{t+1} = (I - K_t C)\Sigma\left[\Sigma^{-1}A\mu_t + C^{\mathrm{T}}R^{-1}x_{t+1}\right] = (I - K_t C)A\mu_t + (I - K_t C)\Sigma C^{\mathrm{T}}R^{-1}x_{t+1}$$

$$= \left[I - \underbrace{\Sigma C^{\mathrm{T}}\left(R + C\Sigma C^{\mathrm{T}}\right)^{-1}}_{K_t}C\right]\Sigma$$

Kalman smoothing Given an entire time series of T observations, the smoothing (marginal) distribution $p(\boldsymbol{z}_t|\boldsymbol{x}_{0:T})$ can be computed using an efficient algorithm known as Kalman smoothing (aka the Rauch-Tung-Striebel algorithm), which consists of a forward pass of Kalman filtering, as above, followed by a backwards pass. Indeed, we first note that the smoothing distribution at time step T is also the Gaussian $p(\boldsymbol{z}_T|\boldsymbol{x}_{0:T})$ obtained via filtering. From there, we can work backwards in much the same way as we worked forwards in the filtering pass. Suppose $p(\boldsymbol{z}_{t+1}|\boldsymbol{x}_{0:T})$ is known to be a Gaussian with mean $\tilde{\boldsymbol{\mu}}_{t+1}$ and covariance $\tilde{\Sigma}_{t+1}$. To obtain $\tilde{\boldsymbol{\mu}}_t$ and $\tilde{\Sigma}_t$ (and convince ourselves that $p(\boldsymbol{z}_t|\boldsymbol{x}_{0:T})$ is indeed Gaussian), we marginalise out $\boldsymbol{z}_{t+1}$:

$$p(\boldsymbol{z}_t|\boldsymbol{x}_{0:T}) = \int d\boldsymbol{z}_{t+1}\, \underbrace{p(\boldsymbol{z}_{t+1}|\boldsymbol{x}_{0:T})}_{\mathcal{N}(\tilde{\boldsymbol{\mu}}_{t+1},\tilde{\Sigma}_{t+1})}\, p(\boldsymbol{z}_t|\boldsymbol{z}_{t+1},\boldsymbol{x}_{0:T}) \tag{13}$$

Moreover, we can use the conditional independence assumptions of the LDS model to split:

$$\begin{aligned}
p(\boldsymbol{z}_t|\boldsymbol{z}_{t+1},\boldsymbol{x}_{0:T}) &\propto p(\boldsymbol{z}_t,\boldsymbol{z}_{t+1},\boldsymbol{x}_{0:T}) && \text{(proportional as a function of } \boldsymbol{z}_t) && (14)\\
&= p(\boldsymbol{z}_t|\boldsymbol{x}_{0:t})\, p(\boldsymbol{z}_{t+1}|\boldsymbol{z}_t)\, p(\boldsymbol{x}_{t+1:T}|\boldsymbol{z}_{t+1}) && \text{(conditional independence)} && (15)\\
&\propto p(\boldsymbol{z}_t|\boldsymbol{x}_{0:t})\, p(\boldsymbol{z}_{t+1}|\boldsymbol{z}_t) && \text{(just dropping the last term, constant in } \boldsymbol{z}_t) && (16)\\
&= \mathcal{N}(\boldsymbol{z}_t;\boldsymbol{\mu}_t,\Sigma_t)\,\mathcal{N}(\boldsymbol{z}_{t+1};A\boldsymbol{z}_t,BB^{\mathrm{T}}). && && (17)
\end{aligned}$$

The product of the two normal densities in the last line is a (unnormalized) Gaussian density in $\boldsymbol{z}_t$ with mean[4] $\boldsymbol{\mu}_t - \Sigma_t A^{\mathrm{T}}(A\Sigma_t A^{\mathrm{T}} + BB^{\mathrm{T}})^{-1}A(\boldsymbol{z}_{t+1} - \boldsymbol{\mu}_t)$ and covariance $(\Sigma_t^{-1} + A^{\mathrm{T}}(BB^{\mathrm{T}})^{-1}A)^{-1}$. Defining $P_t = A\Sigma_t A^{\mathrm{T}} + BB^{\mathrm{T}}$, we can now proceed with the marginalization in Equation (13), yielding

$$\tilde{\boldsymbol{\mu}}_t = \boldsymbol{\mu}_t - \Sigma_t A^{\mathrm{T}} P^{-1} A(\tilde{\boldsymbol{\mu}}_{t+1} - \boldsymbol{\mu}_t) \tag{18}$$

$$\tilde{\Sigma}_t = \left(\Sigma_t^{-1} + A^{\mathrm{T}}(BB^{\mathrm{T}})^{-1}A\right)^{-1} + \Sigma_t A^{\mathrm{T}} P^{-1} A \tilde{\Sigma}_{t+1} A^{\mathrm{T}} P^{-1} A \Sigma_t \tag{19}$$

This backward update step can again be massaged into a more interpretable form:

$$G_t \equiv \Sigma_t A^{\mathrm{T}}(A\Sigma_t A^{\mathrm{T}} + BB^{\mathrm{T}})^{-1} \tag{20}$$

$$\tilde{\boldsymbol{\mu}}_t = \boldsymbol{\mu}_t + G_t(\tilde{\boldsymbol{\mu}}_{t+1} - A\boldsymbol{\mu}_t) \tag{21}$$

$$\tilde{\Sigma}_t = \Sigma_t + G_t(\tilde{\Sigma}_{t+1} - P_t)G_t^{\mathrm{T}} \tag{22}$$

1.2 Parameter estimation

Here we will give a brief and incomplete summary of parameter estimation in LDS. Similar to pPCA and FA, the parameter estimation in LDS can be done by maximum likelihood, i.e. by maximizing $p(\{\boldsymbol{x}_{1:T}^{(k)}\}_{k=1}^K|\boldsymbol{\theta})$, or equivalently the log-likelihood

$$\log p(\{\boldsymbol{x}_{1:T}^{(k)}\}_{k=1}^K|\boldsymbol{\theta}) = \sum_{k=1}^K \log p(\boldsymbol{x}_{1:T}^{(k)}|\boldsymbol{\theta}) \tag{23}$$

Here $\{\boldsymbol{x}_{1:T}^{(k)}\}_{k=1}^K$ denotes a dataset of observed trajectories in K trials, and $\boldsymbol{\theta}$ is the vector of parameters of the LDS model, namely a concatentation of A, B, C, $\mathrm{diag}(R)$, $\boldsymbol{\mu}_0$, and Σ_0. In writing the equation above, we have assumed that the observed neural acitivity in different trials are statistically independent, and therefore their log-probabilities sum over trials.

[4] The Woodbury identity is again used here.

The difficulty is that $p(\boldsymbol{x}_{1:T}^{(k)}|\boldsymbol{\theta})$ (which is actually a Gaussian distribution) has a complicated expression in terms of the parameters (more generally, *i.e.*, when we deviate from assumptions of linearity or Gaussianity, even calculating the density $p(\boldsymbol{x}_{1:T}^{(k)}|\boldsymbol{\theta})$, aka the marginal likelihood, is often intractible as it requires integration over the high-dimensional latent variable trajectories).

One classic approach for obtaining the maximum likelihood parameter estimates is the Expectation-Maximization (EM) algorithm. More modern gradient based (or second order) optimization methods are also popular and can be more efficient. Finally, one can go beyond parameter estimation and calculate an approximation to the Bayesian posterior distribution over the parameters, which also encodes their posterior uncertainty. A popular method for doing that is variational inference.

We will not cover these methods in this lecture. Here, we will only briefly sketch aspects of the EM algorithm for LDS; more details can be found in Section 13.3.2 of the Pattern Recognition and Machine Learning book by C.M. Bishop.

The expectation maximizimation (EM) algorithm. The EM algorithm is an iterative algorithm that alternates between the inference of the latent variables (E step) given the current estimate of the parameters, and updating of parameter estimates (M step) using the posterior distribution of the latent variables obtained in the previous E step.

In general, the E step consists of calculating the joint posterior distribution of the latent variables $p(\boldsymbol{z}_{1:T}|\boldsymbol{x}_{1:T})$. However, due to the HMM structure and Gaussianity of the posterior in our LDS, and also because of the structure of the sufficient statistics for the model parameters, all that we need from this posterior are the single-time-point posterior expectations $\mathbb{E}[\boldsymbol{z}_t]$, and posterior expectations of the pairwise products of latents over the same and consecutive time points, namely $\mathbb{E}[\boldsymbol{z}_t\boldsymbol{z}_t^{\mathrm{T}}]$ and $\mathbb{E}[\boldsymbol{z}_{t+1}\boldsymbol{z}_t^{\mathrm{T}}]$; here and below we denote posterior averages (over $\boldsymbol{z}$) by $\mathbb{E}[\cdot]$, keeping the conditioning on $\boldsymbol{x}_{1:T}$ implicit. The former are the smoothing posterior expectations $\tilde{\boldsymbol{\mu}}$ that, as described in the previous section, are calculated using the forward-backward algorithm. It turns out that a small addition to the forward-backward algorithm also yields the needed pairwise expectations (see Eq. 13.104 of the Bishop book).

In the M-step, we optimise the parameters given the above posterior expectations of the latents (which in turn were calculated based on the optimised parameters in the previous M-step). More precisely, the parameters are optimized to maximise the posterior expectation of the *joint* log-probability density, namely by solving the optimization problem[5]

$$\max_{\boldsymbol{\theta}} \sum_{k=1}^{K} \mathbb{E}[\log p(\boldsymbol{z}_{1:T}^{(k)}, \boldsymbol{x}_{1:T}^{(k)}|\boldsymbol{\theta})]. \tag{24}$$

Here, we will only review the M-step update of A (constituting a subset of model parameters). In particular, this will clarify why we (only in this case) need the pairwise posterior expectations. Based on the HMM structure (see Figure 1), let us first write out the components of the joint probabilities (for trial k) more explicitly as

$$\log p(\boldsymbol{z}_{1:T}^{(k)}, \boldsymbol{x}_{1:T}^{(k)}|\boldsymbol{\theta}) = \log p(\boldsymbol{z}_1^{(k)}|\boldsymbol{\theta}) + \sum_{t=2}^{T} \log p(\boldsymbol{z}_t^{(k)}|\boldsymbol{z}_{t-1}^{(k)}, \boldsymbol{\theta}) + \sum_{t=1}^{T} \log p(\boldsymbol{x}_t^{(k)}|\boldsymbol{z}_t^{(k)}, \boldsymbol{\theta}), \tag{25}$$

$$= \log p(\boldsymbol{z}_1^{(k)}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) + \sum_{t=2}^{T} \log p(\boldsymbol{z}_t^{(k)}|\boldsymbol{z}_{t-1}^{(k)}, A, B) + \sum_{t=1}^{T} \log p(\boldsymbol{x}_t^{(k)}|\boldsymbol{z}_t^{(k)}, C, R). \tag{26}$$

[5]Note how this is different from the direct maximization of model likelihood or log-likelihood, namely $\log p(\boldsymbol{x}_{1:T}|\boldsymbol{\theta})$; however, the fundamental theorem underlying and justifying the EM shows that the EM iterations converge at a local maximum of the model ("marginal") likelihood $\log p(\boldsymbol{x}_{1:T}|\boldsymbol{\theta})$.

Since only the second sum above depends on A, we can focus on that. The summands of this sum are given by (as shown in previous section – see *e.g.*, Equation (8))

$$\log \mathcal{N}(\boldsymbol{z}_t^{(k)}; A\boldsymbol{z}_{t-1}^{(k)}, BB^{\mathrm{T}}) = -\frac{1}{2}(\boldsymbol{z}_t^{(k)} - A\boldsymbol{z}_{t-1}^{(k)})^{\mathrm{T}}(BB^{\mathrm{T}})^{-1}(\boldsymbol{z}_t^{(k)} - A\boldsymbol{z}_{t-1}^{(k)}) + \text{const.} \tag{27}$$

where const. means independent of A. We see that this is quadratic in the $\boldsymbol{z}$'s and moreover only involves products of $\boldsymbol{z}$'s over the same or consecutive time steps. In particular, we see that the posterior expectation of this term, which we need to maximise in A, only involves the pairwise posterior expectations $\mathbb{E}[\boldsymbol{z}_t\boldsymbol{z}_t^{\mathrm{T}}]$ and $\mathbb{E}[\boldsymbol{z}_t\boldsymbol{z}_{t-1}^{\mathrm{T}}]$. Taking gradients with respect to A yields[6]

$$\sum_{k=1}^{K}\sum_{t=2}^{T}(BB^{\mathrm{T}})^{-1}\mathbb{E}\left[(\boldsymbol{z}_t^{(k)} - A\boldsymbol{z}_{t-1}^{(k)})\boldsymbol{z}_{t-1}^{(k)\mathrm{T}}\right] \tag{28}$$

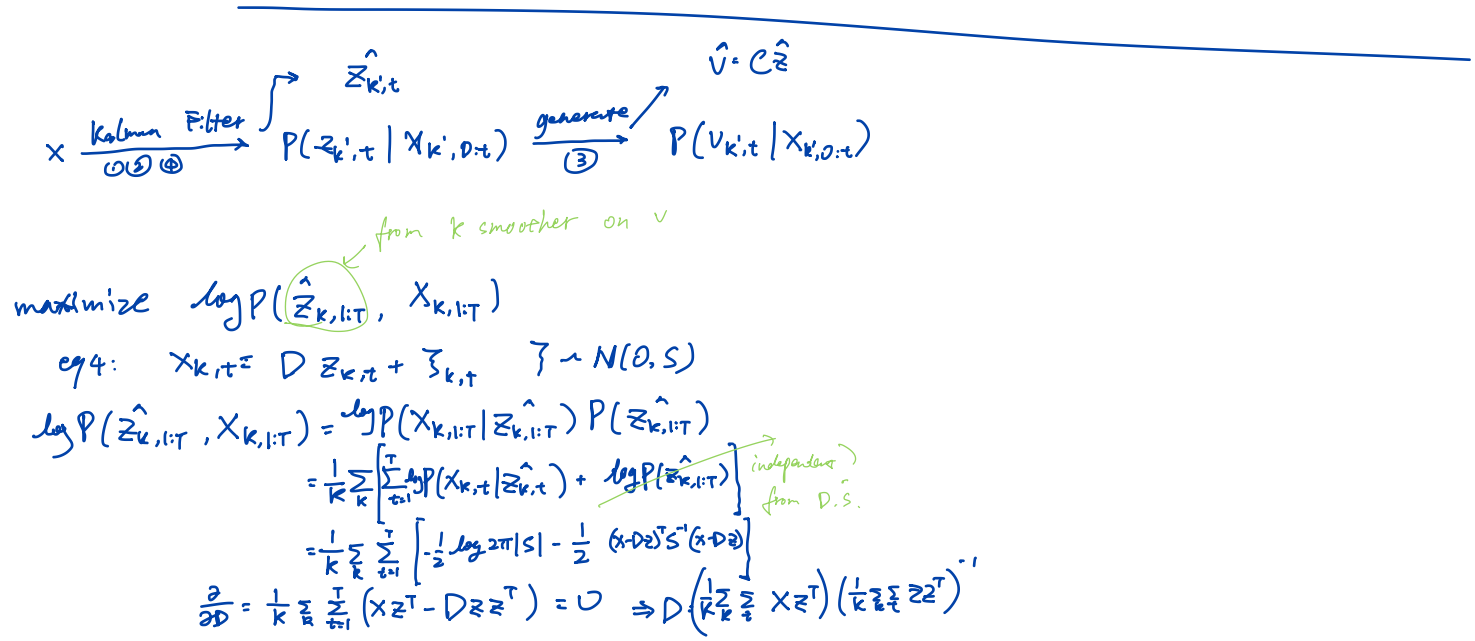
Setting this equal to zero (the necessary condition for local maximum), after multiplying by BB^{T} (and dividing by $K(T-1)$), yields the update rule

$$A^{\text{new}} = \overline{\mathbb{E}}[\boldsymbol{z}_t\boldsymbol{z}_{t-1}^{\mathrm{T}}]\,\overline{\mathbb{E}}[\boldsymbol{z}_{t-1}\boldsymbol{z}_{t-1}^{\mathrm{T}}]^{-1} \tag{29}$$

where $\overline{\mathbb{E}}[\cdot]$ denotes expectation over the posterior as well as averaging over trials and time steps $t = 2$ to T (and we have hidden the trial superscripts). Finally, note how this update generalizes the least squares solution derived in the lecture slides, in the case where the $\boldsymbol{z}$'s are fully observed, namely

$$A = \left(\overline{\boldsymbol{z}_t\boldsymbol{z}_{t-1}^{\mathrm{T}}}\right)\left(\overline{\boldsymbol{z}_{t-1}\boldsymbol{z}_{t-1}^{\mathrm{T}}}\right)^{-1} \tag{30}$$

where the bar denotes trial and time averaging; thus when the $\boldsymbol{z}$'s are latent, the formula is the same except trial/time-averaging has to be supplemented by a posterior average, because the $\boldsymbol{z}$'s were not directly observed.



[6] To derive this, it is more straightforward to write Equation (27) explicitly in component form, as sums over matrix/vector components, and take a partial derivative w.r.t. A_{ij} and then at the end re-express the result in matrix/vector form.

$\frac{\partial}{\partial S} = \Sigma\Sigma - \frac{1}{2}\cdot$