Sheet 3

Time Series Models

By Elias and Michaela

June 2025

1 Task 1. Kullback-Leibler divergence of two multivariate normal distributions

1.1 Derive the analytical expression for the KL divergence given the distributions above.

The KL divergence is defined by

$$KL(\mathcal{P}||\mathcal{Q}) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx$$

Inserting the Gaussian densities, we find

$$KL(\mathcal{P}\|\mathcal{Q}) = \int_{-\infty}^{\infty} \frac{1}{2} p(x) \left(\log \left(\frac{|\Sigma_B|}{|\Sigma_A|} \right) - (x-\mu_A)^T \Sigma_A^{-1} (x-\mu_A) + (x-\mu_B)^T \Sigma_B^{-1} (x-\mu_B) \right) dx$$

By using logarithm rules and that the $(2\pi)^{-n/2}$ term disappears. Since $p(x)$ is a probability density we have

$$KL(\mathcal{P}\|\mathcal{Q}) = \frac{1}{2} \log \left(\frac{|\Sigma_B|}{|\Sigma_A|} \right) - \int_{-\infty}^{\infty} \frac{1}{2} p(x) \left((x-\mu_A)^T \Sigma_A^{-1} (x-\mu_A) + (x-\mu_B)^T \Sigma_B^{-1} (x-\mu_B) \right) dx$$

Now we use that for a multivariate random variable X with $E(X) = \mu$ and $\mathrm{Var}(X) = \Sigma$ it holds $E(X^T A X) = Tr(A\Sigma) + \mu^T A \mu$.

$$\begin{aligned}
KL(\mathcal{P}\|\mathcal{Q}) &= \frac{1}{2} \log \left(\frac{|\Sigma_B|}{|\Sigma_A|} \right) - \frac{1}{2} \left(E_P \left((X-\mu_A)^T \Sigma_A^{-1} (X-\mu_A) \right) + E_P \left((X-\mu_B)^T \Sigma_B^{-1} (X-\mu_B) \right) \right) \\
&= \frac{1}{2} \log \left(\frac{|\Sigma_B|}{|\Sigma_A|} \right) - \frac{1}{2} (Tr(\Sigma_A^{-1}\Sigma_A) - (\mu_A-\mu_A)^T \Sigma_A^{-1} (\mu_A-\mu_A) \\
&+ Tr(\Sigma_B^{-1}\Sigma_A) - (\mu_A-\mu_B)^T \Sigma_B^{-1} (\mu_A-\mu_B)) \\
&= \frac{1}{2} \left(\log \left(\frac{|\Sigma_B|}{|\Sigma_A|} \right) - n + Tr(\Sigma_B^{-1}\Sigma_A) - (\mu_A-\mu_B)^T \Sigma_B^{-1} (\mu_A-\mu_B) \right)
\end{aligned}$$

Where we have used that $E_P(x-\mu_A) = (\mu_A - \mu_A)$, $E_P(x-\mu_B) = (\mu_A - \mu_B)$, $\mathrm{Var}_P(X-\mu_A) = \Sigma_A$, and $\mathrm{Var}_P(X-\mu_B) = \Sigma_A$

1.2 How does the KL divergence simplify if we assume the covariances to be diagonal.

In the previous exercise we found

$$KL(\mathcal{P}\|\mathcal{Q}) = \frac{1}{2} \left(\log \left(\frac{|\Sigma_B|}{|\Sigma_A|} \right) - n + Tr(\Sigma_B^{-1}\Sigma_A) - (\mu_A-\mu_B)^T \Sigma_B^{-1} (\mu_A-\mu_B) \right)$$

Assuming the variance matrices to be diagonal, it follows that $|\Sigma_B| = \sum_{i=1} \sigma_{Bi}$, $|\Sigma_A| = \sum_{i=1} \sigma_{Ai}$. The inverse of a diagonal is easy to compute, it is also a diagonal matrix where each diagonal element is the reciprocal of the diagonal element of the original matrix, thus $Tr(\Sigma_B^{-1}\Sigma_A) = \sum_{i=1}^{T} \frac{\sigma_{Ai}}{\sigma_{Bi}}$. Furthermore $(\mu_A-\mu_B)^T \Sigma_B^{-1} (\mu_A-\mu_B) = \sum_{i=1}^{n} (\mu_A-\mu_B)^2 \sigma_{Bi}$. The KL divergence then simplifies to

$$KL(\mathcal{P}\|\mathcal{Q}) = \frac{1}{2} \left(\log \left(\frac{\sum_{i=1}^{T} \sigma_{Bi}}{\sum_{i=1}^{T} \sigma_{Ai}} \right) - n + \sum_{i=1}^{T} \frac{\sigma_{Ai}}{\sigma_{Bi}} - \sum_{i=1}^{T} (\mu_A-\mu_B)^2 \sigma_{Bi} \right)$$

2 Task 2. M-Step in a linear Gaussian state space model

2.1 Derive the expressions for all remaining parameters.

In class we, with use of Jensens inequality, found the following lower bound of the log likelihood

$$\log p_\theta(x) \geq E_Z[\log(p_\theta(x, Z)] + H_q(q(z|x)) = ELBO(Z|\theta)$$

In the M-step of the EM algorithm we find the parameters of the model that maximize the ELBO considering the density $q(z|x)$ as fixed. Since the second term is fixed, the task is

$$\arg\max_\theta ELBO(Z|\theta) = \arg\max_\theta E_Z[\log(p_\theta(x, Z))]$$

Remembering the model assumptions regarding conditional independencies and the Markov property we find

$$\begin{aligned}
E_Z[\log(p_\theta(x, Z))] &= E_Z\left[\log\left(p(z_1)\prod_{t=1}^{T} p(x_t|z_t)\prod_{t=2}^{T} p(z_t|z_{t-1})\right)\right] \\
&= E_Z\left[\log p(z_1) + \sum_{t=1}^{T}\log p(x_t|z_t) + \sum_{t=2}^{T}\log p(z_t|z_{t-1})\right] \\
&= -\frac{1}{2}E_Z[(z_1-\mu_0)^T\Sigma^{-1}(z_1-\mu_0) + T\log|\Sigma| + T\log|\Gamma| \\
&+ \sum_{t=1}^{T}(x_t - Bz_t)^T\Gamma^{-1}(x_t - Bz_t) + \sum_{t=2}^{T}(z_t - Az_{t-1})^T\Sigma^{-1}(z_t - Az_{t-1})]
\end{aligned}$$

Where we have used the assumptions made regarding the distribution of the random variables. Since the expression is convex in the parameters (hopefully), we can find the optimum by taking the derivative and setting it to 0. Taking the derivative and interchanging the integral and the derivative results in

$$\begin{aligned}
\frac{\partial}{\partial B}E_Z[\log(p_\theta(x, Z))] &= E_Z\left[\frac{\partial}{\partial B}\log(p_\theta(x, Z))\right] \\
&= -\frac{1}{2}E_Z\left[\frac{\partial}{\partial B}\sum_{t=1}^{T}(x_t - Bz_t)^T\Gamma^{-1}(x_t - Bz_t)\right] \\
&= -\frac{1}{2}E_Z\left[\sum_{t=1}^{T}\frac{\partial}{\partial B}\left(x_t^T\Gamma^{-1}x_t - x_t^T\Gamma^{-1}Bz_t - z_t^TB^T\Gamma^{-1}x_t + z_t^TB^T\Gamma^{-1}Bz_t\right)\right] \\
&= -\frac{1}{2}E_Z\left[\sum_{t=1}^{T} -\Gamma^{-1}x_tz_t^T - \Gamma^{-1}x_tz_t^T + \left(\Gamma^{-1}\right)^T Bz_tz_t^T + \Gamma^{-1}Bz_tz_t^T\right]
\end{aligned}$$

Where the derivatives are taken with help from the matrix cookbook page 11. The above expression further simplifies due to symmetry of the variance matrix

$$\frac{\partial}{\partial B}E_Z[\log(p_\theta(x, Z))] = E_Z\left[\sum_{t=1}^{T}\Gamma^{-1}x_tz_t^T - \Gamma^{-1}Bz_tz_t^T\right]$$

Setting the expression equal to 0 and rearranging we find

$$\begin{aligned}
\Gamma^{-1}\sum_{t=1}^{T}x_tE_Z\left[z_t^T\right] = \Gamma^{-1}B\sum_{t=1}^{T}E_Z\left[z_tz_t^T\right] \Leftrightarrow \\
B = \frac{\sum_{t=1}^{T}x_tE_Z\left[z_t^T\right]}{\sum_{t=1}^{T}E_Z\left[z_tz_t^T\right]}
\end{aligned}$$

Which would be the M-step expression for the B matrix.

Now turning to maximizing the ELBO with respect to Σ, we consider only the terms of $E_Z[\log(p_\theta(x,Z))]$ that concern Σ and denote them by Q_Σ. as

$$\begin{aligned}
Q_\Sigma &= -\frac{1}{2}E_Z\left[(z_1-\mu_0)^T\Sigma^{-1}(z_1-\mu_0) + T\log|\Sigma| + \sum_{t=2}^{T}(z_t - Az_{t-1})^T\Sigma^{-1}(z_t-Az_{t-1})\right] \\
&= -\frac{1}{2}E_Z\left[Tr\left((z_1-\mu_0)^T\Sigma^{-1}(z_1-\mu_0)\right) + T\log|\Sigma| + \sum_{t=2}^{T}Tr\left((z_t - Az_{t-1})^T\Sigma^{-1}(z_t-Az_{t-1})\right)\right] \\
&= -\frac{1}{2}E_Z\left[Tr\left(\Sigma^{-1}(z_1-\mu_0)(z_1-\mu_0)^T\right) + T\log|\Sigma| + \sum_{t=2}^{T}Tr\left(\Sigma^{-1}(z_t - Az_{t-1})(z_t-Az_{t-1})^T\right)\right] \\
&= -\frac{1}{2}E_Z\left[Tr\left(\Sigma^{-1}\left((z_1-\mu_0)(z_1-\mu_0)^T + \sum_{t=2}^{T}(z_t - Az_{t-1})(z_t-Az_{t-1})^T\right)\right) + T\log|\Sigma|\right]
\end{aligned}$$

Where we have used that the trace of a scalar is just the scalar the self in the first equality. In the second equality we have used the circular property of trace. And at last in the third, the linearity of trace. Let $S = (z_1-\mu_0)(z_1-\mu_0)^T + \sum_{t=2}^{T}(z_t - Az_{t-1})(z_t-Az_{t-1})^T$, then

$$Q_\Sigma = -\frac{1}{2}E_Z\left[Tr\left(\Sigma^{-1}S\right) + T\log|\Sigma|\right]$$

Taking the derivative of $E_Z[\log(p_\theta(x,Z))]$ with respect to Σ we find

$$\begin{aligned}
\frac{\partial}{\partial\Sigma}E_Z\left[\log(p_\theta(x,Z))\right] = \frac{\partial}{\partial\Sigma}Q_\Sigma &= -\frac{1}{2}E_Z\left[\frac{\partial}{\partial\Sigma}Tr\left(\Sigma^{-1}S\right) + \frac{\partial}{\partial\Sigma}T\log|\Sigma|\right] \\
&= -\frac{1}{2}E_Z\left[-\Sigma^{-1}S\Sigma^{-1} + T\Sigma^{-1}\right]
\end{aligned}$$

Using matrix calculus rules. Setting the derivative to 0 we obtain

$$\begin{aligned}
&\frac{\partial}{\partial\Sigma}E_Z\left[\log(p_\theta(x,Z))\right] = 0 \Leftrightarrow \\
&-\frac{1}{2}E_Z\left[-\Sigma^{-1}S\Sigma^{-1} + T\Sigma^{-1}\right] = 0 \Leftrightarrow \\
&-\Sigma^{-1}E_Z\left[S\right] = TI \Leftrightarrow \\
&\Sigma = -\frac{1}{T}\left(E_Z\left[(z_1-\mu_0)(z_1-\mu_0)^T\right] + \sum_{t=2}^{T}E_Z\left[(z_t-Az_{t-1})(z_t-Az_{t-1})^T\right]\right) \Leftrightarrow \\
&\Sigma = -\frac{1}{T}\Big(E_Z(z_1z_1^T) - E_Z[z_1]\mu_0^T - \mu_0E_Z[z_1^T] + \mu_0\mu_0^T \\
&\quad + \sum_{t=2}^{T}E_Z(z_tz_t^T) - E_Z(z_tz_t^T)A^T - AE_Z(z_tz_t^T) + AE_Z(z_tz_t^T)A^T\Big)
\end{aligned}$$

Inserting the update expression for A we obtain the M-step expression for the Σ matrix.

Now turning to maximizing the ELBO with respect to Γ, we consider only the terms of $E_Z[\log(p_\theta(x,Z))]$

that concern Γ and denote them by Q_Γ. as

$$\begin{aligned}
Q_\Gamma &= -\frac{1}{2}E_Z\left[T\log|\Gamma| + \sum_{t=1}^{T}(x_t - Bz_t)^T\Gamma^{-1}(x_t - Bz_t)\right] \\
&= -\frac{1}{2}E_Z\left[T\log|\Gamma| \sum_{t=2}^{T} Tr\left((x_t - Bz_t)^T\Gamma^{-1}(x_t - Bz_t)\right)\right] \\
&= -\frac{1}{2}E_Z\left[T\log|\Sigma| + \sum_{t=2}^{T} Tr\left(\Gamma^{-1}(x_t - Bz_t)(x_t - Bz_t)^T\right)\right] \\
&= -\frac{1}{2}E_Z\left[Tr\left(\Gamma^{-1}\sum_{t=2}^{T}(x_t - Bz_t)(x_t - Bz_t)^T\right) + T\log|\Gamma|\right]
\end{aligned}$$

Taking the derivative of $E_Z[\log(p_\theta(x, Z))]$ with respect to Γ we find

$$\begin{aligned}
\frac{\partial}{\partial\Gamma}E_Z\left[\log(p_\theta(x, Z))\right] &= \frac{\partial}{\partial\Gamma}Q_\Gamma \\
&= \frac{\partial}{\partial\Gamma}Tr\left(\Gamma^{-1}\sum_{t=2}^{T}(x_t - Bz_t)(x_t - Bz_t)^T\right) + T\log|\Gamma|] \\
&= -\frac{1}{2}E_Z\left[\Gamma^{-1}\sum_{t=2}^{T}(x_t - Bz_t)(x_t - Bz_t)^T\Gamma^{-1} + T\Gamma^{-1}\right]
\end{aligned}$$

Setting the derivative to zero to find the maximum, we get

$$\begin{aligned}
0 &= -\frac{1}{2}E_Z\left[\Gamma^{-1}\sum_{t=2}^{T}(x_t - Bz_t)(x_t - Bz_t)^T\Gamma^{-1} + T\Gamma^{-1}\right] \Leftrightarrow \\
0 &= -E_Z\left[\Gamma^{-1}\sum_{t=2}^{T}(x_t - Bz_t)(x_t - Bz_t)^T + TI\right] \Leftrightarrow \\
\Gamma &= -\frac{1}{T}E_Z\left[\sum_{t=2}^{T}(x_t - Bz_t)(x_t - Bz_t)^T\right] \Leftrightarrow \\
\Gamma &= -\frac{1}{T}\Big(\sum_{t=2}^{T}x_t x_t^T - x_t E_Z[z_t^T]B^T - BE_Z(z_t)x_t^T + BE_Z(z_t z_t^T)B^T\Big)
\end{aligned}$$

for the parameter Γ.