Kalman Filters and Bayesian Mixture Models

To be completed individually or in groups of two people. **Please be sure that matriculation numbers are clearly included at the top of your submission.** Submissions can be handwritten or in LaTeX formatting, but hard-to-read handwritten submissions will not be graded. Please submit via Ilias. Submissions should be a single PDF document (note that Jupyter notebooks can and should also be downloaded as PDFs, and not submitted as .ipynb files). Each question will be graded "pass" (full points) or "fail" (no points). We award 0.5 bonus points for the exam for each theory and practical question solved. You must complete 50% of all exercises to enter the final exam.

1. **EXAMple Question — Kalman filter**

   Let's say we have a Gauss-Markov model with latent state x_t at time step t, and transition distribution: $x_t|x_{t-1} \sim \mathcal{N}(Ax_{t-1}, Q)$. We further know that our previous hidden state x_{t-1}, given our observations Y up till then, is normally distributed with $p(x_{t-1}|Y_{0:t-1}) = \mathcal{N}(x_{t-1}; m_{t-1}, P_{t-1})$

   Calculate the mean and covariance m_t^- and P_t^- of the prediction distribution for x_t based on all previous observations; of $p(x_t|Y_{0:t-1})$. Express m_t^- and P_t^- as functions of the parameters specified above.

   Hint: you can make use of the identities in Lecture 17 (updated) or other known identities for Gaussians.

2. **Theory Question — Bayesian Mixture models**

   In the last exercise sheet, we derived and implemented the Expectation Maximization (EM) algorithm to find the Maximum Likelihood Estimate (MLE) for a Gaussian Mixture Model (GMM). The point estimates of all relevant parameters were usually pretty good, but when EM gets stuck in a local minimum, the result can be bad (you may have encountered this in the practical exercise).

   EM is not a fully Bayesian method; we only obtain a point estimate of the model parameters θ. In this exercise, we consider a Bayesian Gaussian Mixture Model (BGMM) and obtain a posterior over the parameters and latents.

   For simplicity, we will fix the cluster covariance matrices to $\Sigma_k = \sigma I$ for all clusters. As in the last exercise, data is generated as follows:

   - draw a discrete cluster identity $\boldsymbol{z}_i \in \{0,1\}^k, \sum_j z_{ij} = 1$ ("one-hot") with

   $$p(\boldsymbol{z}_i \mid \pi) = \prod_{j=1}^{k} \pi_j^{z_{ij}}$$

   - draw the datum $x_i \in \mathbb{R}^d$ from one of k Gaussian distributions, selected by $\boldsymbol{z}_i$ with probability

   $$p(\boldsymbol{x}_i \mid \boldsymbol{z}_i, \boldsymbol{\mu}) = \prod_{j=1}^{k} \mathcal{N}(\boldsymbol{x}_i; \mu_j, \sigma I)^{z_{ij}}$$

   Yet additionally, we introduce a prior distribution on $\boldsymbol{\mu}$ and π.

   - draw π_j from a Dirichlet distribution, which guarantees that $\sum_{j=1}^{k} \pi_j = 1$ i.e.

   $$p(\pi) = Dir(\pi|\alpha) = \frac{1}{B(\alpha)} \prod_{j=1}^{k} \pi_j^{\alpha_j - 1}$$

- draw each μ_j from a standard normal distribution independently i.e.

$$p(\boldsymbol{\mu}) = \prod_{j=1}^{k} \mathcal{N}(\mu_j; 0, I)$$

This leads to the following generative model

$$p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\mu}, \pi) = \prod_{i=1}^{n} p(x_i|z_i, \boldsymbol{\mu})p(z_i|\pi)p(\pi)p(\boldsymbol{\mu})$$

a) Exactly computing the posterior distribution $p(\boldsymbol{Z}, \boldsymbol{\mu}, \pi|\boldsymbol{X})$ is hard. In fact, the posterior distribution will be highly multimodal (i.e. has multiple modes). Briefly argue why this is the case.

b) To tackle situations like these, you learned MCMC methods for approximate Bayesian inference. We will derive a Gibbs sampling algorithm for this model. Recall that for this, we require all full conditional distributions, i.e. $p(\pi|\boldsymbol{\mu}, \boldsymbol{X}, \boldsymbol{Z})$, $p(\boldsymbol{\mu}|\boldsymbol{X}, \boldsymbol{Z}, \pi)$ and $p(\boldsymbol{Z}|\boldsymbol{X}, \pi, \boldsymbol{\mu})$. Obtain expressions for these distributions.
Hints: Recall the notion of conditional independence and conjugacy. You can use all identities on the wikipedia article on conjugate priors.

c) Write down the Gibbs sampling algorithm in pseudo-code (you don't have to solve b for this).

3. **Practical Question** In this week's practical question, we will gain an intuition for the Kalman filter and take a trip to space: we want to infer the location of so called "AstroCat" and unfortunately only have access to noisy observations of the location. We will follow the material provided by Neuromatch - an online summer school, that all of you who are interested in (Probabilistic) Machine Learning for Neuroscience should certainly check out. You will find the required information in notebook `Exercise_10.ipynb`, which can be solved without solving the questions above. Note that the notation they used is slightly different from the one we use for this course $y \to m$ for *m*easurements and $x \to s$ for the latent *s*tates.

No 12-2pm tutorial on Monday, July 11th Please note that on Monday, July 11th, we will only have the later tutorial from 2pm to 4pm. Everyone who usually attends the first one from 12 pm to 2 pm, please come either later that day or on Thursday. Please reach out to us if that is not possible for you and we will try to find an alternative solution.