Supplementary Material

This is the supplementary material for ‘State space methods for efficient inference in Student-t process regression’ by Solin and Särkkä published in Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS). The references in this document point to the bibliography in the article.

1.1 Proof of Lemma 2.2

Proof. Let $\gamma \sim IG(\alpha, \beta)$ be inverse gamma distributed with parameters $\alpha$ and $\beta$ and $y \mid \gamma \sim N(\mu, \gamma K)$. The scale mixture form of the probability density function can be written as

$$p(y) = \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} \gamma^{-\alpha-1} \exp\left(-\frac{\beta}{\gamma}\right) \frac{1}{\Gamma(\frac{\nu}{2})} \frac{1}{|K|^\frac{n}{2}} \exp\left(-\frac{\Delta^2}{\gamma}\right) d\gamma$$ (12)

$$= \frac{1}{\Gamma(\alpha)} \frac{1}{(2\beta\pi)^\frac{n}{2}} |K|^\frac{1}{2} \int_0^\infty \xi^{\alpha+\frac{n}{2}-1} \exp\left(-\xi\left(1 + \frac{\Delta^2}{2\beta}\right)\right) d\xi$$ (13)

$$= \frac{\Gamma(\alpha + \frac{n}{2})}{\Gamma(\alpha)} \frac{1}{(2\beta\pi)^\frac{n}{2}} |K|^{\frac{1}{2}} \left(1 + \frac{\Delta^2}{2\beta}\right)^{-\left(\alpha+\frac{n}{2}\right)}$$ (14)

where $\Delta^2 = (y - \mu)^T K^{-1} (y - \mu)$. We now recognize this as the Student-t density in Definition 2.1 by parametrizing $\alpha = \frac{\nu}{2}$ and $\beta = \frac{\nu}{2}$. Thus $y \sim MVT(\mu, K, \nu)$. Note the redundancy in $\gamma \sim IG(\frac{\nu}{2}, \rho^\nu\frac{\nu-2}{2})$ and $y \mid \gamma \sim N(\mu, \frac{\nu}{\rho}K)$ for $\rho > 0$. Without loss of generality, we choose $\rho = 1$. \qed

1.2 Marginal likelihood for the naive TP

We write down the negative log marginal likelihood (energy) function and its derivatives with respect to the degrees of freedom $\nu$ and the covariance hyperparameters $\theta = (\sigma_n^2, \theta_1, \theta_2, \ldots)$. The negative log marginal likelihood, $\mathcal{L} = -\log p(y \mid \nu, \theta)$, is given by

$$\mathcal{L} = \frac{n}{2} \log((\nu-2)\pi) + \frac{1}{2} \log(|K_\theta|) + \log\left(\Gamma\left(\frac{\nu+n}{2}\right)\right)$$

$$+ \log\left(\Gamma\left(\frac{n}{2}\right)\right) + \frac{\nu+n}{2} \log\left(1 + \frac{\beta}{\nu-2}\right),$$ (15)

where $\beta = y^T K_\theta^{-1} y$. The derivatives can now be given as

$$\frac{\partial}{\partial \nu} \mathcal{L} = \frac{1}{2} \frac{n}{\nu-2} - \frac{1}{2} \psi\left(\frac{\nu+n}{2}\right) + \frac{1}{2} \psi\left(\frac{\nu}{2}\right)$$

$$+ \frac{1}{2} \psi\left(1 + \frac{\beta}{\nu-2}\right) - \frac{1}{2} \frac{\nu+n}{\nu-2} (\nu-2 + \beta),$$ (16)

$$\frac{\partial}{\partial \theta_i} \mathcal{L} = \frac{1}{2} \text{Tr}\left(K_\theta^{-1} \frac{\partial K_\theta}{\partial \theta_i}\right) + \frac{1}{2} \frac{\nu+n}{\nu-2 + \beta} y^T K_\theta^{-1} \frac{\partial K_\theta}{\partial \theta_i} K_\theta^{-1} y,$$ (17)

where $\psi(\cdot)$ is the digamma function.
1.3 Marginal likelihood for the state space TP

The negative log marginal likelihood can be evaluated recursively starting from $\mathcal{L}_0 = 0$:

$$
\mathcal{L}_k = \mathcal{L}_{k-1} + \frac{1}{2} \log((\nu - 2)\pi) + \frac{1}{2} \log(\|S_k\|) + \log \Gamma\left(\frac{\nu_{k-1}}{2}\right) - \log \Gamma\left(\frac{\nu_k}{2}\right) - \frac{1}{2} \log \left(\frac{\nu_{k-1} - 2}{\nu - 2}\right) - \frac{\nu_k}{2} \log \left(1 + \frac{v_k^T S_k^{-1} v_k}{\nu_{k-1} - 2}\right),
$$

where $v_k$ and $S_k$ are the innovation mean and covariance evaluated by the filter update step, and $\nu_k = \nu_{k-1} + n_k$. Formally differentiating $\mathcal{L}_k$ gives a recursion algorithm for evaluating the gradient along with the filtering steps:

$$
\frac{\partial \mathcal{L}_k(\theta)}{\partial \theta_i} = \frac{\partial \mathcal{L}_{k-1}(\theta)}{\partial \theta_i} + \frac{1}{2} \text{Tr}\left(S_k^{-1}(\theta) \frac{\partial S_k(\theta)}{\partial \theta_i}\right)
$$

$$
+ \frac{\nu_k}{\nu_{k-1} - 2 + v_k^T S_k^{-1} v_k} \left(v_k^T(\theta) S_k^{-1}(\theta) \frac{\partial v_k(\theta)}{\partial \theta_i} - \frac{1}{2} v_k^T(\theta) S_k^{-1}(\theta) \frac{\partial S_k(\theta)}{\partial \theta_i} S_k^{-1}(\theta) v_k(\theta)\right). 
$$

The formal differentiation of the function also includes differentiating the filter prediction and update steps. This leads to the following rather lengthy recursion formulas, which include a lot of small matrix operations. On the filter prediction step we compute:

$$
\frac{\partial m_{k|k-1}(\theta)}{\partial \theta_i} = \frac{\partial A_{k-1}(\theta)}{\partial \theta_i} m_{k-1|k-1}(\theta) + A_{k-1}(\theta) \frac{\partial m_{k-1|k-1}(\theta)}{\partial \theta_i},
$$

$$
\frac{\partial P_{k|k-1}(\theta)}{\partial \theta_i} = \frac{\partial A_{k-1}(\theta)}{\partial \theta_i} P_{k-1|k-1}(\theta) A_{k-1}(\theta)^T + A_{k-1}(\theta) \frac{\partial P_{k-1|k-1}(\theta)}{\partial \theta_i} A_{k-1}(\theta)^T
$$

$$
+ A_{k-1}(\theta) P_{k-1|k-1}(\theta) \frac{\partial A_{k-1}(\theta)}{\partial \theta_i} + A_{k-1}(\theta) \frac{\partial Q_{k-1}(\theta)}{\partial \theta_i} + \frac{\partial \gamma_{k-1}(\theta)}{\partial \theta_i} Q_{k-1}(\theta),
$$

and on the filter update step we compute:

$$
\frac{\partial v_k(\theta)}{\partial \theta_i} = -H \frac{\partial m_{k|k-1}(\theta)}{\partial \theta_i},
$$

$$
\frac{\partial S_k(\theta)}{\partial \theta_i} = H \frac{\partial P_{k|k-1}(\theta)}{\partial \theta_i} H^T,
$$

$$
\frac{\partial K_k(\theta)}{\partial \theta_i} = \frac{\partial P_{k|k-1}(\theta)}{\partial \theta_i} H^T S_k^{-1}(\theta) - P_{k|k-1}(\theta) H^T S_k^{-1}(\theta) \frac{\partial S_k(\theta)}{\partial \theta_i} S_k^{-1}(\theta),
$$

$$
\frac{\partial m_{k|k}(\theta)}{\partial \theta_i} = \frac{\partial m_{k|k-1}(\theta)}{\partial \theta_i} + \frac{\partial K_k(\theta)}{\partial \theta_i} v_k(\theta) + K_k(\theta) \frac{\partial v_k(\theta)}{\partial \theta_i},
$$

$$
\frac{\partial P_{k|k}(\theta)}{\partial \theta_i} = \gamma_k(\theta) \left(\frac{\partial P_{k|k-1}(\theta)}{\partial \theta_i} - \frac{\partial K_k(\theta)}{\partial \theta_i} S_k(\theta) K_k^T(\theta) K_k(\theta)\right)
$$

$$
- K_k(\theta) \frac{\partial S_k(\theta)}{\partial \theta_i} K_k^T(\theta) - K_k(\theta) S_k(\theta) \frac{\partial K_k(\theta)}{\partial \theta_i}
$$

$$
\frac{\partial \gamma_k(\theta)}{\partial \theta_i} = \frac{\gamma_{k-1}(\theta)}{\nu_k - 2 + v_k^T S_k^{-1} v_k} \nu_k - 2
$$

$$
+ \frac{\gamma_{k-1}(\theta)}{\nu_k - 2} \left(2 v_k^T(\theta) S_k^{-1}(\theta) \frac{\partial v_k(\theta)}{\partial \theta_i} - v_k^T(\theta) S_k^{-1}(\theta) \frac{\partial S_k(\theta)}{\partial \theta_i} S_k^{-1}(\theta) v_k(\theta)\right).
$$

Note that, the derivative $\frac{\partial \mathcal{L}_k}{\partial \theta_i}$ can be evaluated as given in Equation (16), if the $\beta = \beta_n$ is evaluated along the filtering recursion such that $\beta_k = \beta_{k-1} + \gamma_{k-1} v_k^T S_k^{-1} v_k$ and starting from $\beta_0 = 0$. For maximum a posteriori estimation, the recursion should be started from the initial condition $\frac{\partial \mathcal{L}_0(\theta)}{\partial \theta_i} = -\frac{\partial \log p(\theta)}{\partial \theta_i}$. For a similar formulation for the Gaussian filter, see [6] and the references therein.