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 18$^{\text{th}}$ 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 $\gamma \sim \mathrm{IG}(\alpha, \beta)$ be inverse gamma distributed with parameters α and β and $\mathbf{y} \mid \gamma \sim \mathrm{N}(\boldsymbol{\mu}, \gamma\mathbf{K})$. The scale mixture form of the probability density function can be written as

$$
\begin{aligned}
p(\mathbf{y}) &= \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} \gamma^{-\alpha-1} \exp\left(-\frac{\beta}{\gamma}\right) \frac{1}{(2\pi)^{\frac{n}{2}}} \frac{1}{|\gamma\mathbf{K}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\frac{\Delta^2}{\gamma}\right) \mathrm{d}\gamma && (12)\\
&= \frac{1}{\Gamma(\alpha)} \frac{1}{(2\beta\pi)^{\frac{n}{2}}} \frac{1}{|\mathbf{K}|^{\frac{1}{2}}} \int_0^\infty \xi^{\alpha+\frac{n}{2}-1} \exp\left(-\xi\left(1+\frac{\Delta^2}{2\beta}\right)\right) \mathrm{d}\xi && (13)\\
&= \frac{\Gamma(\alpha+\frac{n}{2})}{\Gamma(\alpha)} \frac{1}{(2\beta\pi)^{\frac{n}{2}}} \frac{1}{|\mathbf{K}|^{\frac{1}{2}}} \left(1+\frac{\Delta^2}{2\beta}\right)^{-(\alpha+\frac{n}{2})}, && (14)
\end{aligned}
$$

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

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 ν and the covariance hyperparameters $\boldsymbol{\theta} = (\sigma_\mathrm{n}^2, \theta_1, \theta_2, \ldots)$. The negative log marginal likelihood, $\mathcal{L} = -\log p(\mathbf{y} \mid \nu, \boldsymbol{\theta})$, is given by

$$
\begin{aligned}
\mathcal{L} = \frac{n}{2}\log((\nu-2)\pi) + \frac{1}{2}\log(|\mathbf{K}_{\boldsymbol{\theta}}|) - \log\left(\Gamma\left(\frac{\nu+n}{2}\right)\right) \\
+ \log\left(\Gamma\left(\frac{\nu}{2}\right)\right) + \frac{\nu+n}{2}\log\left(1+\frac{\beta}{\nu-2}\right), \quad (15)
\end{aligned}
$$

where $\beta = \mathbf{y}^\mathsf{T}\mathbf{K}_{\boldsymbol{\theta}}^{-1}\mathbf{y}$. The derivatives can now be given as

$$
\begin{aligned}
\frac{\partial}{\partial\nu}\mathcal{L} &= \frac{1}{2}\frac{n}{\nu-2} - \frac{1}{2}\psi\left(\frac{\nu+n}{2}\right) + \frac{1}{2}\psi\left(\frac{\nu}{2}\right) \\
&\quad + \frac{1}{2}\log\left(1+\frac{\beta}{\nu-2}\right) - \frac{1}{2}\frac{(\nu+n)\beta}{(\nu-2)(\nu-2+\beta)}, && (16)\\
\frac{\partial}{\partial\theta_i}\mathcal{L} &= \frac{1}{2}\mathrm{Tr}\left(\mathbf{K}_{\boldsymbol{\theta}}^{-1}\frac{\partial\mathbf{K}_{\boldsymbol{\theta}}}{\partial\theta_i}\right) + \frac{1}{2}\frac{\nu+n}{\nu-2+\beta}\mathbf{y}^\mathsf{T}\mathbf{K}_{\boldsymbol{\theta}}^{-1}\frac{\partial\mathbf{K}_{\boldsymbol{\theta}}}{\partial\theta_i}\mathbf{K}_{\boldsymbol{\theta}}^{-1}\mathbf{y}, && (17)
\end{aligned}
$$

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 $\mathcal{L}_0 = 0$:

$$\mathcal{L}_k = \mathcal{L}_{k-1} + \frac{1}{2}\log((\nu-2)\pi) + \frac{1}{2}\log(|\mathbf{S}_k|) + \log\Gamma\left(\frac{\nu_{k-1}}{2}\right) - \log\Gamma\left(\frac{\nu_k}{2}\right) + \frac{1}{2}\log\left(\frac{\nu_{k-1}-2}{\nu-2}\right) + \frac{\nu_k}{2}\log\left(1+\frac{\mathbf{v}_k^\mathsf{T}\mathbf{S}_k^{-1}\mathbf{v}_k}{\nu_{k-1}-2}\right), \quad (18)$$

where $\mathbf{v}_k$ and $\mathbf{S}_k$ 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:

$$\frac{\partial\mathcal{L}_k(\boldsymbol{\theta})}{\partial\theta_i} = \frac{\partial\mathcal{L}_{k-1}(\boldsymbol{\theta})}{\partial\theta_i} + \frac{1}{2}\mathrm{Tr}\left(\mathbf{S}_k^{-1}(\boldsymbol{\theta})\,\frac{\partial\mathbf{S}_k(\boldsymbol{\theta})}{\partial\theta_i}\right) + \frac{\nu_k}{\nu_{k-1}-2+\mathbf{v}_k^\mathsf{T}\mathbf{S}_k^{-1}\mathbf{v}_k}\left(\mathbf{v}_k^\mathsf{T}(\boldsymbol{\theta})\,\mathbf{S}_k^{-1}(\boldsymbol{\theta})\,\frac{\partial\mathbf{v}_k(\boldsymbol{\theta})}{\partial\theta_i} - \frac{1}{2}\mathbf{v}_k^\mathsf{T}(\boldsymbol{\theta})\,\mathbf{S}_k^{-1}(\boldsymbol{\theta})\,\frac{\partial\mathbf{S}_k(\boldsymbol{\theta})}{\partial\theta_i}\,\mathbf{S}_k^{-1}(\boldsymbol{\theta})\,\mathbf{v}_k(\boldsymbol{\theta})\right). \quad (19)$$

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:

$$\frac{\partial\mathbf{m}_{k|k-1}(\boldsymbol{\theta})}{\partial\theta_i} = \frac{\partial\mathbf{A}_{k-1}(\boldsymbol{\theta})}{\partial\theta_i}\,\mathbf{m}_{k-1|k-1}(\boldsymbol{\theta}) + \mathbf{A}_{k-1}(\boldsymbol{\theta})\,\frac{\partial\mathbf{m}_{k-1|k-1}(\boldsymbol{\theta})}{\partial\theta_i}, \quad (20)$$

$$\frac{\partial\mathbf{P}_{k|k-1}(\boldsymbol{\theta})}{\partial\theta_i} = \frac{\partial\mathbf{A}_{k-1}(\boldsymbol{\theta})}{\partial\theta_i}\,\mathbf{P}_{k-1|k-1}(\boldsymbol{\theta})\,\mathbf{A}_{k-1}^\mathsf{T}(\boldsymbol{\theta}) + \mathbf{A}_{k-1}(\boldsymbol{\theta})\,\frac{\partial\mathbf{P}_{k-1|k-1}(\boldsymbol{\theta})}{\partial\theta_i}\,\mathbf{A}_{k-1}^\mathsf{T}(\boldsymbol{\theta}) + \mathbf{A}_{k-1}(\boldsymbol{\theta})\,\mathbf{P}_{k-1|k-1}(\boldsymbol{\theta})\,\frac{\partial\mathbf{A}_{k-1}^\mathsf{T}(\boldsymbol{\theta})}{\partial\theta_i} + \gamma_{k-1}(\boldsymbol{\theta})\frac{\partial\mathbf{Q}_{k-1}(\boldsymbol{\theta})}{\partial\theta_i} + \frac{\partial\gamma_{k-1}(\boldsymbol{\theta})}{\partial\theta_i}\mathbf{Q}_{k-1}(\boldsymbol{\theta}), \quad (21)$$

and on the filter update step we compute:

$$\frac{\partial\mathbf{v}_k(\boldsymbol{\theta})}{\partial\theta_i} = -\mathbf{H}\,\frac{\partial\mathbf{m}_{k|k-1}(\boldsymbol{\theta})}{\partial\theta_i}, \quad (22)$$

$$\frac{\partial\mathbf{S}_k(\boldsymbol{\theta})}{\partial\theta_i} = \mathbf{H}\frac{\partial\mathbf{P}_{k|k-1}(\boldsymbol{\theta})}{\partial\theta_i}\,\mathbf{H}^\mathsf{T}, \quad (23)$$

$$\frac{\partial\mathbf{K}_k(\boldsymbol{\theta})}{\partial\theta_i} = \frac{\partial\mathbf{P}_{k|k-1}(\boldsymbol{\theta})}{\partial\theta_i}\,\mathbf{H}^\mathsf{T}\,\mathbf{S}_k^{-1}(\boldsymbol{\theta}) - \mathbf{P}_{k|k-1}(\boldsymbol{\theta})\,\mathbf{H}^\mathsf{T}\,\mathbf{S}_k^{-1}(\boldsymbol{\theta})\,\frac{\partial\mathbf{S}_k(\boldsymbol{\theta})}{\partial\theta_i}\,\mathbf{S}_k^{-1}(\boldsymbol{\theta}), \quad (24)$$

$$\frac{\partial\mathbf{m}_{k|k}(\boldsymbol{\theta})}{\partial\theta_i} = \frac{\partial\mathbf{m}_{k|k-1}(\boldsymbol{\theta})}{\partial\theta_i} + \frac{\partial\mathbf{K}_k(\boldsymbol{\theta})}{\partial\theta_i}\,\mathbf{v}_k(\boldsymbol{\theta}) + \mathbf{K}_k(\boldsymbol{\theta})\,\frac{\partial\mathbf{v}_k(\boldsymbol{\theta})}{\partial\theta_i}, \quad (25)$$

$$\frac{\partial\mathbf{P}_{k|k}(\boldsymbol{\theta})}{\partial\theta_i} = \frac{\gamma_k(\boldsymbol{\theta})}{\gamma_{k-1}(\boldsymbol{\theta})}\left(\frac{\partial\mathbf{P}_{k|k-1}(\boldsymbol{\theta})}{\partial\theta_i} - \frac{\partial\mathbf{K}_k(\boldsymbol{\theta})}{\partial\theta_i}\,\mathbf{S}_k(\boldsymbol{\theta})\,\mathbf{K}_k^\mathsf{T}(\boldsymbol{\theta}) - \mathbf{K}_k(\boldsymbol{\theta})\,\frac{\partial\mathbf{S}_k(\boldsymbol{\theta})}{\partial\theta_i}\,\mathbf{K}_k^\mathsf{T}(\boldsymbol{\theta}) - \mathbf{K}_k(\boldsymbol{\theta})\,\mathbf{S}_k(\boldsymbol{\theta})\,\frac{\partial\mathbf{K}_k^\mathsf{T}(\boldsymbol{\theta})}{\partial\theta_i}\right) + \frac{1}{\gamma_{k-1}(\boldsymbol{\theta})}\left(\frac{\partial\gamma_k(\boldsymbol{\theta})}{\partial\theta_i} - \frac{\gamma_k(\boldsymbol{\theta})}{\gamma_{k-1}(\boldsymbol{\theta})}\,\frac{\partial\gamma_{k-1}(\boldsymbol{\theta})}{\partial\theta_i}\right)\left(\mathbf{P}_{k|k-1} - \mathbf{K}_k\mathbf{S}_k\mathbf{K}_k^\mathsf{T}\right), \quad (26)$$

$$\frac{\partial\gamma_k(\boldsymbol{\theta})}{\partial\theta_i} = \frac{\partial\gamma_{k-1}(\boldsymbol{\theta})}{\partial\theta_i}\,\frac{\nu_{k-1}-2+\mathbf{v}_k^\mathsf{T}\mathbf{S}_k^{-1}\mathbf{v}_k}{\nu_k-2} + \frac{\gamma_{k-1}(\boldsymbol{\theta})}{\nu_k-2}\left(2\,\mathbf{v}_k^\mathsf{T}(\boldsymbol{\theta})\,\mathbf{S}_k^{-1}(\boldsymbol{\theta})\,\frac{\partial\mathbf{v}_k(\boldsymbol{\theta})}{\partial\theta_i} - \mathbf{v}_k^\mathsf{T}(\boldsymbol{\theta})\,\mathbf{S}_k^{-1}(\boldsymbol{\theta})\,\frac{\partial\mathbf{S}_k(\boldsymbol{\theta})}{\partial\theta_i}\,\mathbf{S}_k^{-1}(\boldsymbol{\theta})\,\mathbf{v}_k(\boldsymbol{\theta})\right). \quad (27)$$

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