Supplementary Material: The Variational Coupled Gaussian Process Dynamical Model

Dmytro Velychko, Benjamin Knopp, Dominik Endres*

March 31, 2017

1 Partial optimization of variational distribution for simplified ELBO

While optimizing the full variational posterior in augmented Gaussian Processes models the following type of term appears often in the ELBO equation:

$$\begin{aligned}\mathcal{R}(q(\boldsymbol{u}), q(\boldsymbol{v})) &= \int q(\boldsymbol{u})(f(q(\boldsymbol{v}), \boldsymbol{u}) + \log \frac{p(\boldsymbol{u})}{q(\boldsymbol{u})})d\boldsymbol{u} \\ &= \int q(\boldsymbol{u}) f(q(\boldsymbol{v}), \boldsymbol{u})d\boldsymbol{u} + \int q(\boldsymbol{u}) \log p(\boldsymbol{u})d\boldsymbol{u} - \int q(\boldsymbol{u}) \log q(\boldsymbol{u})d\boldsymbol{u}\end{aligned} \tag{S1}$$

To simplify the optimization of such terms, we would like to carry out the optimization with respect to the density $q(\boldsymbol{u})$ analytically, so as to remove the dependency on $q(\boldsymbol{u})$. To this end, we calculate for the optimal variational $q^*(\boldsymbol{u})$ in the above equation. This approach was suggested in [1], however, it is not well described there. Here we give an extended derivation. A necessary condition for maximality is a vanishing functional derivative under the constraint that the density $q(\boldsymbol{u})$ is normalized to one:

$$\int q(\boldsymbol{u})d\boldsymbol{u} - 1 = 0 \tag{S2}$$

which is fulfilled at the stationary points of the Lagrangian

$$\mathcal{X}(q(\boldsymbol{u}), q(\boldsymbol{v})) = \mathcal{R}(q(\boldsymbol{u}), q(\boldsymbol{v})) + \lambda \left(\int q(\boldsymbol{u})d\boldsymbol{u} - 1 \right) \tag{S3}$$

*Theoretical Neuroscience Group, Dept. Psychology, University of Marburg, Gutenbergstr. 18, 35032 Marburg, Germany
{dmytro.velychko,benjamin.knopp,dominik.endres}@uni-marburg.de

where λ is chosen so that (S2) holds. Taking the derivative of $\mathcal{X}(q(\boldsymbol{u}), q(\boldsymbol{v}))$ and setting it to zero yields

$$\frac{\delta \mathcal{X}(q(\boldsymbol{u}), q(\boldsymbol{v}))}{\delta q(\boldsymbol{u})} = f(q(\boldsymbol{v}), \boldsymbol{u}) + \log p(\boldsymbol{u}) - \log q(\boldsymbol{u}) + 1 + \lambda = 0 \tag{S4}$$

and therefore, denoting $Z = \exp(-\lambda - 1)$

$$q^*(\boldsymbol{u}) = \exp(f(q(\boldsymbol{v}), \boldsymbol{u}) + \log p(\boldsymbol{u}) + 1 + \lambda) \tag{S5}$$

$$q^*(\boldsymbol{u}) = \frac{1}{Z} p(\boldsymbol{u}) \exp(f(q(\boldsymbol{v}), \boldsymbol{u})) \tag{S6}$$

$$Z = \exp(-\lambda - 1) = \int p(\boldsymbol{u}) \exp(f(q(\boldsymbol{v}), \boldsymbol{u})) d\boldsymbol{u} \tag{S7}$$

Substituting the optimal $q^*(\boldsymbol{u})$ into the original term, we get:

$$\begin{aligned}
\mathcal{R}(q(\boldsymbol{v})) &= \int \frac{1}{Z} p(\boldsymbol{u}) \exp(f(q(\boldsymbol{v}), \boldsymbol{u}))(f(q(\boldsymbol{v}), \boldsymbol{u}) + \log \frac{p(\boldsymbol{u})}{\frac{1}{Z} p(\boldsymbol{u}) \exp(f(q(\boldsymbol{v}), \boldsymbol{u}))}) d\boldsymbol{u} \\
&= \int \frac{1}{Z} p(\boldsymbol{u}) \exp(f(q(\boldsymbol{v}), \boldsymbol{u})) \left(\log \frac{p(\boldsymbol{u}) \exp(f(q(\boldsymbol{v}), \boldsymbol{u}))}{\frac{1}{Z} p(\boldsymbol{u}) \exp(f(q(\boldsymbol{v}), \boldsymbol{u}))} \right) d\boldsymbol{u} \\
&= \log(Z) \frac{1}{Z} \int p(\boldsymbol{u}) \exp(f(q(\boldsymbol{v}), \boldsymbol{u})) d\boldsymbol{u} \\
&= \log \int p(\boldsymbol{u}) \exp(f(q(\boldsymbol{v}), \boldsymbol{u})) d\boldsymbol{u}
\end{aligned} \tag{S8}$$

This is the simplified version of (S1), which depends only on $q(\boldsymbol{v})$ because $q^(\boldsymbol{u}) = q^*(\boldsymbol{u})$ has been determined by optimization.

2 vCGPDM ELBO derivation

Here we give the extended derivation of the ELBO from the main paper. Let's assume we deal with M parts. The model reads:

$$\boldsymbol{X} = \{\boldsymbol{X}^0 \dots \boldsymbol{X}^M\} \tag{S9}$$

$$\boldsymbol{X}^i = \{\boldsymbol{x}_0^i \dots \boldsymbol{x}_T^i\}; \boldsymbol{x}_t^i \in \mathbb{R}^{Q^i} \tag{S10}$$

$$\boldsymbol{Y}^i = \{\boldsymbol{y}_0^i \dots \boldsymbol{y}_T^i\}; \boldsymbol{y}_t^i \in \mathbb{R}^{D^i} \tag{S11}$$

$$f^{j,i}(\boldsymbol{X}_{-t}^j) \sim \mathcal{GP}(0, k_f^j(\boldsymbol{x}_{-t}^j, \boldsymbol{x}_{-t}^{j\prime})) \tag{S12}$$

$$p(\boldsymbol{x}_t^i) \sim \prod_{j=1}^{M} \mathcal{N}(\boldsymbol{x}_t^i | f^{j,i}(\boldsymbol{x}_{-t}^j), \boldsymbol{I}\alpha^{j,i}); \alpha > 0 \tag{S13}$$

$$g_d^i(\boldsymbol{X}^i) \sim \mathcal{GP}(0, k_g(\boldsymbol{x}_t^i, \boldsymbol{x}_t^{i\prime})) \tag{S14}$$

$$\boldsymbol{g}_d^i = g_d^i(\boldsymbol{X}^i) \tag{S15}$$

$$\boldsymbol{Y}_{:,d}^i \sim \mathcal{N}(\boldsymbol{g}_d^i, \boldsymbol{I}\beta^i); \beta > 0 \tag{S16}$$

$$p(\boldsymbol{x}_0^i) = \mathcal{N}(\boldsymbol{x}_0^i | 0, \boldsymbol{I}) \tag{S17}$$

Here and following upper indexes are part-related, lower indexes indicate dimensions.

We have $M \times M$ latent dynamics mappings, which are combined into M mappings with the Product of Experts - multiplying and renormalizing the distribution from separate partial predictions. Each of the $M \times M$ mappings is augmented with some inducing inputs and outputs: $f^{j,i} : \boldsymbol{z}^{j,i} \to \boldsymbol{u}^{j,i}$. The full augmented joint distribution of the model reads:

$$
\begin{aligned}
p(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{f}, \boldsymbol{v}, \boldsymbol{g}, \boldsymbol{y}) &= p(\boldsymbol{y}|\boldsymbol{g}) p(\boldsymbol{g}|\boldsymbol{x}, \boldsymbol{v}) p(\boldsymbol{v}) p(\boldsymbol{x}, \boldsymbol{f}|\boldsymbol{u}) p(\boldsymbol{u}) \\
&= \left[\prod_{i=1}^{M} p(\boldsymbol{y}^i|\boldsymbol{g}^i) p(\boldsymbol{g}^i|\boldsymbol{x}^i, \boldsymbol{v}^i) p(\boldsymbol{v}^i) \right] p(\boldsymbol{x}, \boldsymbol{f}|\boldsymbol{u}) p(\boldsymbol{u}) \\
&= \left[\prod_{i=1}^{M} \left[\prod_{d=1}^{D_i} p(\boldsymbol{y}_d^i|\boldsymbol{g}_d^i) p(\boldsymbol{g}_d^i|\boldsymbol{x}^i, \boldsymbol{v}^i) \right] p(\boldsymbol{v}^i) \right] \\
&\times \left[\prod_{t=1}^{T} \left[\prod_{i=1}^{M} p(\boldsymbol{x}_t^i | \{\boldsymbol{f}_t^{:,i}, \alpha^{:,i}\}) \right] \left[\prod_{i=1}^{M} \prod_{j=1}^{M} p(\boldsymbol{f}_t^{j,i} | \boldsymbol{f}_{1:t-1}^{j,i}, \boldsymbol{x}_{0:t-1}^{j}, \boldsymbol{u}^{j,i}) \right] \right] \\
&\times \left[\prod_{i=1}^{M} \prod_{j=1}^{M} p(\boldsymbol{u}^{j,i}) \right] \left[\prod_{i=1}^{M} p(\boldsymbol{x}_0^i) \right]
\end{aligned} \tag{S18}
$$

The full proposal variational posterior is:

$$
\begin{aligned}
q(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{f}, \boldsymbol{v}, \boldsymbol{g}) &= p(\boldsymbol{g}|\boldsymbol{x}, \boldsymbol{v}) q(\boldsymbol{v}) p(\boldsymbol{f}|\boldsymbol{x}, \boldsymbol{u}) q(\boldsymbol{x}) q(\boldsymbol{u}) \\
&= p(\boldsymbol{g}|\boldsymbol{x}, \boldsymbol{v}) q(\boldsymbol{v}) \left[\prod_{t=1}^{T} \prod_{i=1}^{M} \prod_{j=1}^{M} p(\boldsymbol{f}_t^{j,i} | \boldsymbol{f}_{1:t-1}^{j,i}, \boldsymbol{x}_{0:t-1}^{j}, \boldsymbol{u}^{j,i}) \right] q(\boldsymbol{x}) q(\boldsymbol{u})
\end{aligned} \tag{S19}
$$

The ELBO is:

$$\mathcal{L}(\boldsymbol{\theta}) = \int_{\boldsymbol{x},\boldsymbol{v},\boldsymbol{g}} q(\boldsymbol{x},\boldsymbol{u},\boldsymbol{f},\boldsymbol{v},\boldsymbol{g}) \log\left(\frac{p(\boldsymbol{x},\boldsymbol{u},\boldsymbol{f},\boldsymbol{v},\boldsymbol{g},\boldsymbol{y})}{q(\boldsymbol{x},\boldsymbol{u},\boldsymbol{f},\boldsymbol{v},\boldsymbol{g})}\right) \tag{S20}$$

$$= \sum_{i=1}^{M}\sum_{d=1}^{D} \int_{\boldsymbol{x}^i,\boldsymbol{v}^i,\boldsymbol{g}_d^i} p(\boldsymbol{g}_d^i|\boldsymbol{x}^i,\boldsymbol{v}^i) q(\boldsymbol{x}^i) q(\boldsymbol{v}^i) \log\frac{p(\boldsymbol{y}_d^i|\boldsymbol{g}_d^i)}{q(\boldsymbol{v}^i)} \tag{S21}$$

$$+ \int q(\boldsymbol{u}) \left[\sum_{t=1}^{T} \int q(\boldsymbol{x}_t) q(\boldsymbol{x}_{-t}) \left(\int \left[\prod_{i=1}^{M}\prod_{j=1}^{M} p(\boldsymbol{f}_t^{j,i}|\boldsymbol{f}_{1:t-1}^{j,i}, \boldsymbol{x}_{0:t-1}^{j}, \boldsymbol{u}^{j,i})\right]\right.\right.$$

$$\left.\left.\times \log\prod_{i=1}^{M} p(\boldsymbol{x}_t^i|\{\boldsymbol{f}_t^{:,i}, \alpha^{:,i}\}) d\boldsymbol{f}_t\right) d\boldsymbol{x}_t d\boldsymbol{x}_{-t}\right]$$

$$+ q(\boldsymbol{u}) \log\frac{p(\boldsymbol{u})}{q(\boldsymbol{u})} d\boldsymbol{u}$$

$$+ \int q(\boldsymbol{x}_0) \log p(\boldsymbol{x}_0) d\boldsymbol{x}_0 + H(q(\boldsymbol{x})) \tag{S22}$$

The (S21) part is the GPLVM ELBO and is given in [2]. Next, we consider only the ELBO component which is relevant for the dynamics and apply the sufficient statistics assumption: knowing $\boldsymbol{x}_{-t}^j$ and $\boldsymbol{u}^{j,i}$ is sufficient for the $\boldsymbol{f}_t^{j,i}$ distribution. The innermost integral is:

$$\begin{aligned}
\mathcal{A} &= \int \left[\prod_{i=1}^{M}\prod_{j=1}^{M} p(\boldsymbol{f}_t^{j,i}|\boldsymbol{x}_{-t}^j, \boldsymbol{u}^{j,i})\right] \log\prod_{i=1}^{M} p(\boldsymbol{x}_t^i|\{\boldsymbol{f}_t^{:,i}, \alpha^{:,i}\}) d\boldsymbol{f}_t \\
&= \sum_{i=1}^{M} \int \left[\prod_{j=1}^{M} p(\boldsymbol{f}_t^{j,i}|\boldsymbol{x}_{-t}^j, \boldsymbol{u}^{j,i})\right] \log p(\boldsymbol{x}_t^i|\{\boldsymbol{f}_t^{:,i}, \alpha^{:,i}\}) d\boldsymbol{f}_t^i \\
&= \sum_{i=1}^{M} \int \left[\prod_{j=1}^{M} \mathcal{N}(\boldsymbol{f}_t^{j,i}|\boldsymbol{\mu}_{\boldsymbol{f}_t^{j,i}}, \boldsymbol{S}_{\boldsymbol{f}_t^{j,i}})\right] \log\mathcal{N}(\boldsymbol{x}_t^i|\alpha_i \sum_{j=1}^{M} \alpha_{j,i}^{-1} \boldsymbol{f}_t^{j,i}, \boldsymbol{I}\alpha_i) d\boldsymbol{f}_t^i \\
&= \sum_{i=1}^{M} \left[-\frac{1}{2}\operatorname{tr}\left[\alpha_i \sum_{j=1}^{M} (\alpha_{j,i}^{-1})^2 \boldsymbol{S}_{\boldsymbol{f}_t^{j,i}}\right] + \log\mathcal{N}(\boldsymbol{x}_t^i|\alpha_i \sum_{j=1}^{M} \alpha_{j,i}^{-1} \boldsymbol{\mu}_{\boldsymbol{f}_t^{j,i}}, \boldsymbol{I}\alpha_i)\right]
\end{aligned} \tag{S23}$$

$$\boldsymbol{\mu}_{\boldsymbol{f}_t^{j,i}} = \boldsymbol{K}_{\boldsymbol{x}_{-t}^j, \boldsymbol{z}^{j,i}}^{j,i} \left(\boldsymbol{K}_{\boldsymbol{z}^{j,i}, \boldsymbol{z}^{j,i}}^{j,i}\right)^{-1} \boldsymbol{u}^{j,i} \tag{S24}$$

$$\boldsymbol{S}_{\boldsymbol{f}_t^{j,i}} = \boldsymbol{K}_{\boldsymbol{x}_{-t}^j, \boldsymbol{x}_{-t}^j}^{j,i} - \boldsymbol{K}_{\boldsymbol{x}_{-t}^j, \boldsymbol{z}^{j,i}}^{j,i} \left(\boldsymbol{K}_{\boldsymbol{z}^{j,i}, \boldsymbol{z}^{j,i}}^{j,i}\right)^{-1} \boldsymbol{K}_{\boldsymbol{x}_{-t}^j, \boldsymbol{z}^{j,i}}^{j,i} \tag{S25}$$

$$\alpha_i = (\sum_{j=1}^{M} \alpha_{j,i}^{-1})^{-1} \tag{S26}$$

Now let's take the integral over $\boldsymbol{x}$:

$$\begin{aligned}
\mathcal{B} &= \int q(\boldsymbol{x}_t)q(\boldsymbol{x}_{-t})\mathcal{A}d\boldsymbol{x}_t d\boldsymbol{x}_{-t} \\
&= \sum_{i=1}^{M}\Bigg(-\frac{1}{2}\alpha_i\sum_{j=1}^{M}(\alpha_{j,i}^{-1})^2\,\mathrm{tr}\left[\Psi_0^{j,i}(\boldsymbol{x}_{-t}^j) - \left(\boldsymbol{K}_{\boldsymbol{z}^{j,i},\boldsymbol{z}^{j,i}}^{j,i}\right)^{-1}\Psi_2^{j,i}(\boldsymbol{x}_{-t}^j)\right] \\
&- \log Z(\boldsymbol{I}\alpha_i) - \frac{1}{2}\alpha_i^{-1}\left[\mathrm{tr}(\boldsymbol{S}_{\boldsymbol{x}_t^i}) + \boldsymbol{\mu}_{\boldsymbol{x}_t^i}^T\boldsymbol{\mu}_{\boldsymbol{x}_t^i}\right] \\
&+ \boldsymbol{\mu}_{\boldsymbol{x}_t^i}^T\Big(\sum_{j=1}^{M}\alpha_{j,i}^{-1}\Psi_1^{j,i}(\boldsymbol{x}_{-t}^j)\left(\boldsymbol{K}_{\boldsymbol{z}^{j,i},\boldsymbol{z}^{j,i}}^{j,i}\right)^{-1}\boldsymbol{u}^{j,i}\Big) \\
&- \frac{1}{2}\alpha_i\sum_{j=1}^{M}\sum_{k=1}^{M}\alpha_{j,i}^{-1}\alpha_{k,i}^{-1}\boldsymbol{u}^{j,iT}\left(\boldsymbol{K}_{\boldsymbol{z}^{j,i},\boldsymbol{z}^{j,i}}^{j,i}\right)^{-1}\Psi_2^{j,k,i}(\boldsymbol{x}_{-t}^j,\boldsymbol{x}_{-t}^k)\left(\boldsymbol{K}_{\boldsymbol{z}^{k,i},\boldsymbol{z}^{k,i}}^{k,i}\right)^{-1}\boldsymbol{u}^{k,i}\Bigg)
\end{aligned} \tag{S27}$$

The sum over time points:

$$\begin{aligned}
\mathcal{C} &= \sum_{t=1}^{T}\mathcal{B} \\
&= \sum_{i=1}^{M}\Bigg(\sum_{t=1}^{T}\Bigg[-\frac{1}{2}\alpha_i\sum_{j=1}^{M}(\alpha_{j,i}^{-1})^2\,\mathrm{tr}\left[\Psi_0^{j,i}(\boldsymbol{x}_{-t}^j) - \left(\boldsymbol{K}_{\boldsymbol{z}^{j,i},\boldsymbol{z}^{j,i}}^{j,i}\right)^{-1}\Psi_2^{j,i}(\boldsymbol{x}_{-t}^j)\right] \\
&- \log Z(\boldsymbol{I}\alpha_i) - \frac{1}{2}\alpha_i^{-1}\left[\mathrm{tr}(\boldsymbol{S}_{\boldsymbol{x}_t^i}) + \boldsymbol{\mu}_{\boldsymbol{x}_t^i}^T\boldsymbol{\mu}_{\boldsymbol{x}_t^i}\right]\Bigg] \\
&+ \sum_{j=1}^{M}\alpha_{j,i}^{-1}\left[\sum_{t=1}^{T}\boldsymbol{\mu}_{\boldsymbol{x}_t^i}^T\Psi_1^{j,i}(\boldsymbol{x}_{-t}^j)\right]\left(\boldsymbol{K}_{\boldsymbol{z}^{j,i},\boldsymbol{z}^{j,i}}^{j,i}\right)^{-1}\boldsymbol{u}^{j,i} \\
&- \frac{1}{2}\alpha_i\sum_{j=1}^{M}\sum_{k=1}^{M}\alpha_{j,i}^{-1}\alpha_{k,i}^{-1}\boldsymbol{u}^{j,iT}\left(\boldsymbol{K}_{\boldsymbol{z}^{j,i},\boldsymbol{z}^{j,i}}^{j,i}\right)^{-1}\left[\sum_{t=1}^{T}\Psi_2^{j,k,i}(\boldsymbol{x}_{-t}^j,\boldsymbol{x}_{-t}^k)\right]\left(\boldsymbol{K}_{\boldsymbol{z}^{k,i},\boldsymbol{z}^{k,i}}^{j,i}\right)^{-1}\boldsymbol{u}^{k,i}\Bigg)
\end{aligned} \tag{S28}$$

For every $i \in 1 \ldots M$ we may stack up the $\boldsymbol{u}^{:,i}$ into $\boldsymbol{u}^i$ and construct a large block matrices $\mathcal{F}^i$ and stacked matrices $\mathcal{G}^i$ with elements

$$\mathcal{F}_{j,k}^i = \alpha_i\alpha_{j,i}^{-1}\alpha_{k,i}^{-1}\left(\boldsymbol{K}_{\boldsymbol{z}^{j,i},\boldsymbol{z}^{j,i}}^{j,i}\right)^{-1}\left[\sum_{t=1}^{T}\Psi_2^{j,k,i}(\boldsymbol{x}_{-t}^j,\boldsymbol{x}_{-t}^k)\right]\left(\boldsymbol{K}_{\boldsymbol{z}^{k,i},\boldsymbol{z}^{k,i}}^{k,i}\right)^{-1} \tag{S29}$$

$$\mathcal{G}_j^i = \left(\alpha_{j,i}^{-1}\left[\sum_{t=1}^{T}\boldsymbol{\mu}_{\boldsymbol{x}_t^i}^T\Psi_1^{j,i}(\boldsymbol{x}_{-t}^j)\right]\left(\boldsymbol{K}_{\boldsymbol{z}^{j,i},\boldsymbol{z}^{j,i}}^{j,i}\right)^{-1}\right)^T \tag{S30}$$

For $j \neq k$: $\Psi_2^{j,k,i}(\boldsymbol{x}_{-t}^j, \boldsymbol{x}_{-t}^k) = \Psi_1^{j,i}(\boldsymbol{x}_{-t}^j)\Psi_1^{k,i}(\boldsymbol{x}_{-t}^k)$. Otherwise $\Psi_2^{j,j,i}(\boldsymbol{x}_{-t}^j, \boldsymbol{x}_{-t}^j) = \Psi_2^{j,i}(\boldsymbol{x}_{-t}^j)$.

The sum over time points must be expressed as a quadratic form w.r.t. the augmenting outputs $\boldsymbol{u}$:

$$\begin{aligned}
\mathcal{C} &= \sum_{i=1}^{M}\left[-\frac{1}{2}\boldsymbol{u}^{iT}\mathcal{F}^i\boldsymbol{u}^i + \boldsymbol{u}^{iT}\mathcal{G}^i + \mathcal{H}^i\right] \\
&= \sum_{i=1}^{M}\left[-\frac{1}{2}(\boldsymbol{u}^i - \mathcal{F}^{i-1}\mathcal{G}^i)^T\mathcal{F}^i(\boldsymbol{u}^i - \mathcal{F}^{i-1}\mathcal{G}^i) + \frac{1}{2}\mathcal{G}^{iT}\mathcal{F}^{i-1}\mathcal{G}^i + \mathcal{H}^i\right] \quad \text{(S31)} \\
\mathcal{C} &= \sum_{i=1}^{M}\mathcal{C}^i \quad \text{(S32)} \\
\mathcal{C}^i &= -\frac{1}{2}(\boldsymbol{u}^i - \mathcal{F}^{i-1}\mathcal{G}^i)^T\mathcal{F}^i(\boldsymbol{u}^i - \mathcal{F}^{i-1}\mathcal{G}^i) + \frac{1}{2}\mathcal{G}^{iT}\mathcal{F}^{i-1}\mathcal{G}^i + \mathcal{H}^i \quad \text{(S33)} \\
\mathcal{H}^i &= \sum_{t=1}^{T}\left[-\frac{1}{2}\alpha_i\sum_{j=1}^{M}(\alpha_{j,i}^{-1})^2\,\mathrm{tr}\left[\Psi_0^{j,i}(\boldsymbol{x}_{-t}^j) - \left(\boldsymbol{K}_{\boldsymbol{z}^{j,i},\boldsymbol{z}^{j,i}}\right)^{-1}\Psi_2^{j,i}(\boldsymbol{x}_{-t}^j)\right]\right. \\
&\quad \left. - \log Z(\boldsymbol{I}\alpha_i) - \frac{1}{2}\alpha_i^{-1}\left[\mathrm{tr}(\boldsymbol{S}_{\boldsymbol{x}_t^i}) + \boldsymbol{\mu}_{\boldsymbol{x}_t^i}^T\boldsymbol{\mu}_{\boldsymbol{x}_t^i}\right]\right] \quad \text{(S34)}
\end{aligned}$$

Finally, we may write the dynamics ELBO, also accounting for the optimal variational $q(\boldsymbol{u})$ (see eq. S8) and using $p(\boldsymbol{u}^i) = \prod_{j=1}^{M} p(\boldsymbol{u}^{ji}) = \prod_{j=1}^{M}\mathcal{N}(\boldsymbol{u}^{ji}|0, \boldsymbol{K}_{\boldsymbol{z}^{j,i},\boldsymbol{z}^{j,i}}) = \mathcal{N}(\boldsymbol{u}^i|0, \boldsymbol{K}_{\boldsymbol{z}^{:,i},\boldsymbol{z}^{:,i}})$ where $\boldsymbol{K}_{\boldsymbol{z}^{:,i},\boldsymbol{z}^{:,i}}$ is a block-diagonal covariance matrix:

$$\begin{aligned}
\mathcal{L}_{dyn}(\boldsymbol{\theta}) &\geq \log\int p(\boldsymbol{u})exp(\mathcal{C})d\boldsymbol{u} + H(q(\boldsymbol{x})) \\
&= \log\prod_{i=1}^{M}\int p(\boldsymbol{u}^i)exp(\mathcal{C}^i)d\boldsymbol{u}^i + H(q(\boldsymbol{x})) \\
&= \sum_{i=1}^{M}\left[\log\int p(\boldsymbol{u}^i)exp(-\frac{1}{2}(\boldsymbol{u}^i - \mathcal{F}^{i-1}\mathcal{G}^i)^T\mathcal{F}^i(\boldsymbol{u}^i - \mathcal{F}^{i-1}\mathcal{G}^i))d\boldsymbol{u}^i + \frac{1}{2}\mathcal{G}^{iT}\mathcal{F}^{i-1}\mathcal{G}^i + \mathcal{H}^i\right] + H(q(\boldsymbol{x})) \\
&= \sum_{i=1}^{M}\left[-\log Z(\mathcal{F}^{i-1} + \boldsymbol{K}_{\boldsymbol{z}^{:,i},\boldsymbol{z}^{:,i}}) - \frac{1}{2}\mathcal{G}^{iT}\mathcal{F}^{i-1}(\mathcal{F}^{i-1} + \boldsymbol{K}_{\boldsymbol{z}^{:,i},\boldsymbol{z}^{:,i}})^{-1}\mathcal{F}^{i-1}\mathcal{G}^i + \log Z(\mathcal{F}^{i-1})\right] \\
&\quad + \sum_{i=1}^{M}\left[\frac{1}{2}\mathcal{G}^{iT}\mathcal{F}^{i-1}\mathcal{G}^i + \mathcal{H}^i\right] + H(q(\boldsymbol{x})) \quad \text{(S35)}
\end{aligned}$$

This is even lower bound on the ELBO due to the sufficient statistics assumption for the $\boldsymbol{f}_t^{j,i}$ distribution. It is easy to notice that the optimal proposal $q(\boldsymbol{u}) = \prod_{i=1}^{M} q(\boldsymbol{u}^i)$ - factorized.

We optimize the full ELBO w.r.t. the parameters of $q(\boldsymbol{x})$, augmenting inputs $\boldsymbol{z}$, kernel parameters, and couplings $\boldsymbol{\alpha}$. After the optimization the optimal $q(\boldsymbol{u})$

is computed as:

$$
\begin{aligned}
q(\boldsymbol{u}^i) &= \frac{1}{Z} p(\boldsymbol{u}^i) exp(\mathcal{C}^i) \\
&= \frac{1}{Z} \mathcal{N}(\boldsymbol{u}^i | 0, \boldsymbol{K}_{\boldsymbol{z}^{:,i}, \boldsymbol{z}^{:,i}}) exp(-\frac{1}{2}(\boldsymbol{u}^i - \mathcal{F}^{i-1}\mathcal{G}^i)^T \mathcal{F}^i (\boldsymbol{u}^i - \mathcal{F}^{i-1}\mathcal{G}^i)) \\
&= \mathcal{N}(\boldsymbol{u}^i | (\boldsymbol{K}^{-1}_{\boldsymbol{z}^{:,i}, \boldsymbol{z}^{:,i}} + \mathcal{F}^i)^{-1}\mathcal{G}^i, (\boldsymbol{K}^{-1}_{\boldsymbol{z}^{:,i}, \boldsymbol{z}^{:,i}} + \mathcal{F}^i)^{-1})
\end{aligned} \tag{S36}
$$

3 ARD RBF kernel Ψ statistics. Full covariance variational parameters case.

Consider the following form of the approximate variational posterior distribution of $\boldsymbol{X}$:

$$
q(X) = \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_n, S_n) \tag{S37}
$$

Here we derive the Ψ statistics for the variational lower bound for the case when the $\{S_n\}_{n=1\ldots N}$ are *full covariance matrices* and the ARD SE kernel is defined as:

$$
k(\boldsymbol{x}, \boldsymbol{x}') = \sigma_f^2 exp\left(-\frac{1}{2}\sum_{q=1}^{Q} \frac{(\boldsymbol{x}_q - \boldsymbol{x}'_q)^2}{\lambda_q}\right) \tag{S38}
$$

In matrix notation:

$$
\lambda = diag(\lambda_1 \ldots \lambda_Q) \tag{S39}
$$

$$
k(\boldsymbol{x}, \boldsymbol{x}') = \sigma_f^2 exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}')^T \lambda^{-1} (\boldsymbol{x} - \boldsymbol{x}')\right) \tag{S40}
$$

where λ_q are the ARD factors.

The Ψ_0 statistic is easy to calculate and it does not depend on the covariance matrix:

$$
\begin{aligned}
\Psi_0^n &= \int k(\boldsymbol{x}_n, \boldsymbol{x}_n) \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_n, S_n) d\boldsymbol{x}_n \\
&= \int \sigma_f^2 \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_n, S_n) d\boldsymbol{x}_n \\
&= \sigma_f^2 \int \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_n, S_n) d\boldsymbol{x}_n \\
&= \sigma_f^2
\end{aligned} \tag{S41}
$$

$$
\Psi_0 = \sum_{n=1}^{N} \Psi_0^n = N\sigma_f^2 \tag{S42}
$$

The Ψ_1 statistic:

$$
\begin{aligned}
(\Psi_1)_{nm} &= \int k(\boldsymbol{x}_n, \boldsymbol{z}_m)\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_n, S_n)d\boldsymbol{x}_n \\
&= \int \sigma_f^2 exp\left(-\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{z}_m)^T\lambda^{-1}(\boldsymbol{x}_n - \boldsymbol{z}_m)\right)\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_n, S_n)d\boldsymbol{x}_n \quad \text{(S43)}
\end{aligned}
$$

Next, complete the ARD SE kernel to a scaled Gaussian distribution:

$$
\begin{aligned}
(\Psi_1)_{nm} &= \sigma_f^2 \int \frac{Z(\lambda)}{Z(\lambda)}\sigma_f^2 exp\left(-\frac{1}{2}(\boldsymbol{x}_r - \boldsymbol{x}_m)^T\lambda^{-1}(\boldsymbol{x}_r - \boldsymbol{x}_m)\right)\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_n, S_n)d\boldsymbol{x}_n \\
&= \sigma_f^2 Z(\lambda)\int \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{z}_m, \lambda)\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_n, S_n)d\boldsymbol{x}_n \quad \text{(S44)}
\end{aligned}
$$

Noticing the product of two Gaussians, which is an unnormalized Gaussian, the integral boils down to the scaling coefficient, which is a Gaussian value:

$$
\begin{aligned}
(\Psi_1)_{nm} &= \sigma_f^2 Z(\lambda)\mathcal{N}(\boldsymbol{z}_m|\boldsymbol{\mu}_n, \lambda + S_n) \\
&= \sigma_f^2 (2\pi)^{Q/2}\sqrt{|\lambda|}\frac{1}{(2\pi)^{Q/2}\sqrt{|\lambda + S_n|}} exp\left(-\frac{1}{2}(\boldsymbol{z}_m - \boldsymbol{\mu}_n)^T(\lambda + S_n)^{-1}(\boldsymbol{z}_m - \boldsymbol{\mu}_n)\right) \\
&= \sigma_f^2 \frac{\sqrt{|\lambda|}}{\sqrt{|\lambda + S_n|}} exp\left(-\frac{1}{2}(\boldsymbol{z}_m - \boldsymbol{\mu}_n)^T(\lambda + S_n)^{-1}(\boldsymbol{z}_m - \boldsymbol{\mu}_n)\right) \\
&= \sigma_f^2 \sqrt{\frac{\prod_{q=1}^{Q}\lambda_q}{|\lambda + S_n|}} exp\left(-\frac{1}{2}(\boldsymbol{z}_m - \boldsymbol{\mu}_n)^T(\lambda + S_n)^{-1}(\boldsymbol{z}_m - \boldsymbol{\mu}_n)\right) \quad \text{(S45)} \\
& \quad \text{(S46)}
\end{aligned}
$$

The Ψ_2 statistic integral can be solved in the same manner:

$$
\begin{aligned}
(\Psi_2)_{mm'} &= \int k(\boldsymbol{x}_n, \boldsymbol{z}_m)k(\boldsymbol{z}_{m'}, \boldsymbol{x}_n)\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_n, S_n)d\boldsymbol{x}_n \\
&= (\sigma_f^2 Z(\lambda))^2 \int \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{z}_m, \lambda)\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{z}_{m'}, \lambda)\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_n, S_n)d\boldsymbol{x}_n \quad \text{(S47)}
\end{aligned}
$$

Here we have to multiply the Gaussians twice:

$$
\begin{aligned}
(\Psi_2)_{mm'} &= (\sigma_f^2 Z(\lambda))^2 \int \mathcal{N}(\boldsymbol{z}_{m'}|\boldsymbol{z}_m, 2\lambda) \\
&\times \mathcal{N}(\boldsymbol{x}_n|(\lambda^{-1}+\lambda^{-1})^{-1}(\lambda^{-1}\boldsymbol{z}_m + \lambda^{-1}\boldsymbol{z}_{m'}), (\lambda^{-1}+\lambda^{-1})^{-1})\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_n, S_n)d\boldsymbol{x}_n \\
&= (\sigma_f^2 Z(\lambda))^2 \int \mathcal{N}(\boldsymbol{z}_{m'}|\boldsymbol{z}_m, 2\lambda)\mathcal{N}(\boldsymbol{x}_n|\frac{\boldsymbol{z}_m+\boldsymbol{z}_{m'}}{2}, \frac{\lambda}{2})\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_n, S_n)d\boldsymbol{x}_n \\
&= (\sigma_f^2 Z(\lambda))^2 \mathcal{N}(\boldsymbol{z}_{m'}|\boldsymbol{z}_m, 2\lambda) \int \mathcal{N}(\boldsymbol{x}_n|\frac{\boldsymbol{z}_m+\boldsymbol{z}_{m'}}{2}, \frac{\lambda}{2})\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_n, S_n)d\boldsymbol{x}_n \\
&= (\sigma_f^2 Z(\lambda))^2 \mathcal{N}(\boldsymbol{z}_{m'}|\boldsymbol{z}_m, 2\lambda) \int \mathcal{N}(\boldsymbol{\mu}_n|\frac{\boldsymbol{z}_m+\boldsymbol{z}_{m'}}{2}, \frac{\lambda}{2}+S_n)\mathcal{N}(\boldsymbol{x}_n|mean, precision)d\boldsymbol{x}_n \\
&= (\sigma_f^2 Z(\lambda))^2 \mathcal{N}(\boldsymbol{z}_{m'}|\boldsymbol{z}_m, 2\lambda)\mathcal{N}(\boldsymbol{\mu}_n|\frac{\boldsymbol{z}_m+\boldsymbol{z}_{m'}}{2}, \frac{\lambda}{2}+S_n) \int \mathcal{N}(\boldsymbol{x}_n|mean, precision)d\boldsymbol{x}_n \\
&= (\sigma_f^2 Z(\lambda))^2 \mathcal{N}(\boldsymbol{z}_{m'}|\boldsymbol{z}_m, 2\lambda)\mathcal{N}(\boldsymbol{\mu}_n|\frac{\boldsymbol{z}_m+\boldsymbol{z}_{m'}}{2}, \frac{\lambda}{2}+S_n) \\
&= \sigma_f^4 (2\pi)^Q (\prod_{q=1}^{Q} \lambda_q)\mathcal{N}(\boldsymbol{z}_{m'}|\boldsymbol{z}_m, 2\lambda)\mathcal{N}(\boldsymbol{\mu}_n|\frac{\boldsymbol{z}_m+\boldsymbol{z}_{m'}}{2}, \frac{\lambda}{2}+S_n) \qquad \text{(S48)}
\end{aligned}
$$

For the case of the diagonal covariance S_n the Ψ_3 statistic looks simpler [2] as it does not require inversion of the $\lambda + S_n$ and $\frac{\lambda}{2} + S_n$ matrices.

4 Relationship between ELBO and MSE

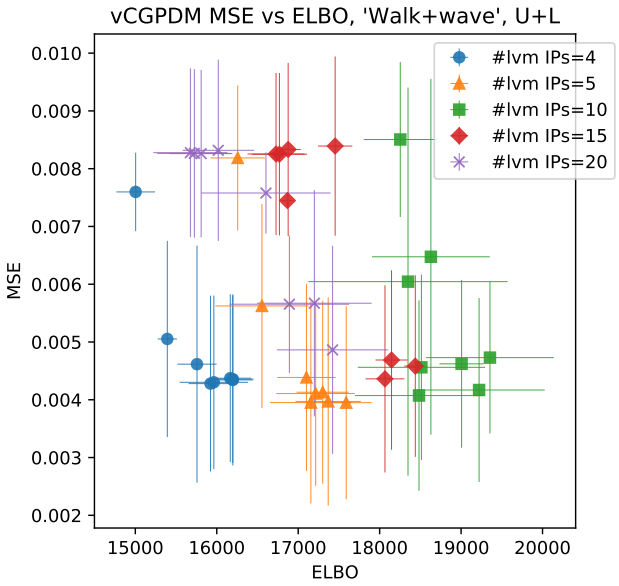

Figure 1: Mean-squared kinematics error, MSE, vs. evidence lower bound, ELBO, for different number of LVM IPs, indicated by symbols. The negative correlation between ELBO and MSE for a given number of LVM IPs is clearly visible. Furthermore, note that the highest ELBO corresponds to an MSE that is very close to the optimal (lowest) one, i.e. ELBO can be used for model selection.

5 Learning time of vCGPDM and MAP CGPDM

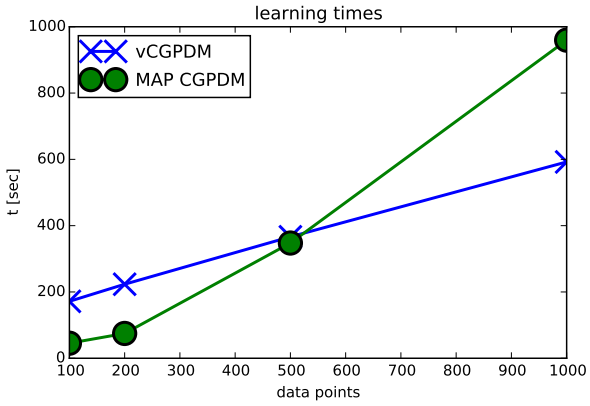

Figure 2: Learning times, including Theano compile times, for a three-part vCGPDM with 10 IPs for all parts and corresponding MAP CGPDM. The linear learning time scaling of the vCGPDM is evident, whereas the MAP CGPDM shows cubic scaling. Thus, the vCGPDM can be used on large data sets which is infeasible for the MAP CGPDM.

References

[1] Michalis K. Titsias. Variational learning of inducing variables in sparse gaussian processes. In David A. Van Dyk and Max Welling, editors, *AISTATS*, volume 5 of *JMLR Proceedings*, pages 567–574. JMLR.org, 2009.

[2] Michalis K. Titsias and Neil D. Lawrence. Bayesian gaussian process latent variable model. In *Proc. 13th AISTATS*, pages 844–851, 2010.