Appendix: Variational Gaussian Process Dynamical Systems

Andreas C. Damianou[1], **Michalis K. Titsias**[2], **Neil D. Lawrence**[1]

[1]*Dept. of Computer Science & Sheffield Institute for Translational Neuroscience, University of Sheffield, UK*
[2]*School of Computer Science, University of Manchester, UK*

A Derivation of the variational bound

We wish to approximate the marginal likelihood:

$$p(Y|\mathbf{t}) = \int p(Y, F, X|\mathbf{t})\mathrm{d}X\mathrm{d}F, \tag{25}$$

by computing a lower bound:

$$\mathcal{F}_v(q, \boldsymbol{\theta}) = \int q(\Theta) \log \frac{p(Y, F, X|\mathbf{t})}{q(\Theta)}\mathrm{d}X\mathrm{d}F. \tag{26}$$

This can be achieved by first augmenting the joint probability density of our model with inducing inputs $\tilde{X}$ along with their corresponding function values U:

$$p(Y, F, U, X, \tilde{X}|\mathbf{t}) = \prod_{d=1}^{D} p(\mathbf{y}_d|\mathbf{f}_d)p(\mathbf{f}_d|\mathbf{u}_d, X)p(\mathbf{u}_d|\tilde{X})p(X|\mathbf{t}) \tag{27}$$

where $p(\mathbf{u}_d|\tilde{X}) = \prod_{d=1}^{D}\mathcal{N}\left(\mathbf{u}_d|\mathbf{0}, K_{MM}\right)$. For simplicity, $\tilde{X}$ is dropped from our expressions for the rest of this supplementary material. Note that after including the inducing points, $p(\mathbf{f}_d|\mathbf{u}_d, X)$ remains analytically tractable and it turns out to be [9]):

$$p(\mathbf{f}_d|\mathbf{u}_d, X) = \mathcal{N}\left(\mathbf{f}_d|K_{NM}K_{MM}^{-1}\mathbf{u}_d, K_{NN} - K_{NM}K_{MM}^{-1}K_{MN}\right). \tag{28}$$

For tractability, we now define a variational density $q(\Theta)$:

$$q(\Theta) = q(F, U, X) = q(F|U, X)q(U)q(X) = \prod_{d=1}^{D} p(\mathbf{f}_d|\mathbf{u}_d, X)q(\mathbf{u}_d)q(X), \tag{29}$$

where $q(X) = \prod_{q=1}^{Q}\mathcal{N}\left(\mathbf{x}_q|\boldsymbol{\mu}_q, S_q\right)$. Now, we return to (26) and replace the joint distribution with its augmented version (27) and the variational distribution with its factorised version (29):

$$\begin{aligned}
\mathcal{F}_v(q, \boldsymbol{\theta}) &= \int q(\Theta) \log \frac{p(Y, F, U, X|\mathbf{t})}{q(F, U, X)}\mathrm{d}X\mathrm{d}F, \\
&= \int \prod_{d=1}^{D} p(\mathbf{f}_d|\mathbf{u}_d, X)q(\mathbf{u}_d)q(X) \log \frac{\prod_{d=1}^{D} p(\mathbf{y}_d|\mathbf{f}_d)\cancel{p(\mathbf{f}_d|\mathbf{u}_d, X)}p(\mathbf{u}_d|\tilde{X})p(X|\mathbf{t})}{\prod_{d=1}^{D}\cancel{p(\mathbf{f}_d|\mathbf{u}_d, X)}q(\mathbf{u}_d)q(X)}\mathrm{d}X\mathrm{d}F \\
&= \int \prod_{d=1}^{D} p(\mathbf{f}_d|\mathbf{u}_d, X)q(\mathbf{u}_d)q(X) \log \frac{\prod_{d=1}^{D} p(\mathbf{y}_d|\mathbf{f}_d)p(\mathbf{u}_d|\tilde{X})}{\prod_{d=1}^{D} q(\mathbf{u}_d)q(X)}\mathrm{d}X\mathrm{d}F \\
&\quad - \int \prod_{d=1}^{D} q(X) \log \frac{q(X)}{p(X|\mathbf{t})}\mathrm{d}X \\
&= \hat{\mathcal{F}}_v - \mathrm{KL}(q \parallel p),
\end{aligned} \tag{30}$$

with:

$$\hat{\mathcal{F}}_v = \sum_{d=1}^{D} \left(\int q(\mathbf{u}_d) q(X) \left\langle \log p(\mathbf{y}_d|\mathbf{f}_d) \right\rangle_{p(\mathbf{f}_d|\mathbf{u}_d,X)} d\mathbf{u}_d \, dX + \log \left\langle \frac{p(\mathbf{u}_d)}{q(\mathbf{u}_d)} \right\rangle_{q(\mathbf{u}_d)} \right) = \sum_{d=1}^{D} \hat{\mathcal{F}}_d \tag{31}$$

Both terms in (30) are analytically tractable, with the first having the same analytical solution as the one derived in [8]. Further, from (31) and similarly to [8], the optimal setting for $q(\mathbf{u}_d)$ is:

$$q(\mathbf{u}_d) = e^{\left\langle \log \mathcal{N}\left(\mathbf{y}_d|\mathbf{a}_d,\beta^{-1}I_d\right)\right\rangle_{q(X)}} p(\mathbf{u}_d), \tag{32}$$

where $\mathbf{a}_d$ is the mean of (28). Therefore, $q(\mathbf{u}_d)$ is a Gaussian distribution since $p(\mathbf{u}_d) = \mathcal{N}(\mathbf{u}_d|\mathbf{0}, K_{MM})$.

After a few more calculations, we find that the complete form of the Jensen's lower bound turns out to be:

$$\begin{aligned}
\mathcal{F}_v(q,\boldsymbol{\theta}) &= \sum_{d=1}^{D} \hat{\mathcal{F}}_d(q,\boldsymbol{\theta}) - \mathrm{KL}(q \parallel p) \\
&= \sum_{d=1}^{D} \log \left(\frac{(\beta)^{\frac{N}{2}} |K_{MM}|^{\frac{1}{2}}}{(2\pi)^{\frac{N}{2}} |\beta\Psi_2 + K_{MM}|^{\frac{1}{2}}} e^{-\frac{1}{2}\mathbf{y}_d^T W \mathbf{y}_d} \right) - \frac{\beta\psi_0}{2} + \frac{\beta}{2}\mathrm{Tr}\left(K_{MM}^{-1}\Psi_2\right) \\
&\quad - \frac{Q}{2}\log|K_t| - \frac{1}{2}\sum_{q=1}^{Q}\left[\mathrm{Tr}\left({K_t}^{-1}S_q\right) + \mathrm{Tr}\left({K_t}^{-1}\boldsymbol{\mu}_q\boldsymbol{\mu}_q^T\right)\right] + \frac{1}{2}\sum_{q=1}^{Q}\log|S_q| + const
\end{aligned} \tag{33}$$

where the last line corresponds to the KL term. Also:

$$\Psi_0 = \mathrm{Tr}(\langle K_{NN}\rangle_{q(X)}) \ , \quad \Psi_1 = \langle K_{NM}\rangle_{q(X)} \ , \quad \Psi_2 = \langle K_{MN}K_{NM}\rangle_{q(X)} \tag{34}$$

The Ψ quantities can be computed analytically as in [8].

B Predictions

B.1 Predictions given only the test time points

To approximate the predictive density, we will need to introduce the underlying latent function values $F_* \in \mathbb{R}^{N_* \times D}$ (the noisy-free version of Y_*) and the latent variables $X_* \in \mathbb{R}^{N_* \times Q}$. We write the predictive density as

$$p(Y_*|Y) = \int p(Y_*, F_*, X_*|Y)\mathrm{d}F_*\mathrm{d}X_* = \int p(Y_*|F_*)p(F_*|X_*,Y)p(X_*|Y)\mathrm{d}F_*\mathrm{d}X_*. \tag{35}$$

The term $p(F_*|X_*,Y)$ is approximated according to

$$q(F_*|X_*) = \int \prod_{d\in D} p(\mathbf{f}_{*,d}|\mathbf{u}_d, X_*)q(\mathbf{u}_d)\mathrm{d}\mathbf{u}_d = \prod_{d\in D} q(\mathbf{f}_{*,d}|X_*), \tag{36}$$

where $q(\mathbf{f}_{*,d}|X_*)$ is a Gaussian that can be computed analytically , since $q(\mathbf{u}_d)$ is also a Gaussian as shown in (32). The term $p(X_*|Y)$ in eq. (35) is approximated by a Gaussian variational distribution $q(X_*)$,

$$p(X_*|Y) \approx \int p(X_*|X)q(X)\mathrm{d}X = \langle p(X_*|X)\rangle_{q(X)} = q(X_*) = \prod_{q=1}^{Q} q(\mathbf{x}_{*,q}), \tag{37}$$

where $p(X_{*,q}|X)$ can be found from the conditional GP prior (see [9]). We can then write

$$\mathbf{x}_{*,q} = \boldsymbol{\alpha}\mathbf{x}_q + \boldsymbol{\epsilon}, \tag{38}$$

where $\boldsymbol{\alpha} = K_{*N}K_t^{-1}$ and $\boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}, K_{**} - K_{*NK_t^{-1}K_{N*}}\right)$. Also, $K_t = k_x(\mathbf{t},\mathbf{t})$, $K_{*N} = k_x(\mathbf{t}_*,\mathbf{t})$ and $K_{**} = k_x(\mathbf{t}_*\mathbf{t}_*)$. Given the above, we know a priori that (37) is a Gaussian and by taking expectations over

$q(X)$ in the r.h.s. of (38) we find the mean and covariance of $q(X_*)$. Substituting for the equivalent forms of $\boldsymbol{\mu}_q$ and S_q from section 2.2 we obtain the final solution

$$\mu_{x_{*,q}} = \mathbf{k}_{*N}\bar{\mu}_q \tag{39}$$

$$\text{var}(x_{*,q}) = k_{**} - \mathbf{k}_{*N}(K_t + \Lambda_q^{-1})^{-1}\mathbf{k}_{N*}. \tag{40}$$

(35) can then be written as:

$$p(Y_*|Y) = \int p(Y_*|F_*)q(F_*|X_*)q(X_*)\mathrm{d}F_*\mathrm{d}X_* = \int p(Y_*|F_*)\left\langle q(F_*|X_*)\right\rangle_{q(X_*)}\mathrm{d}F_* \tag{41}$$

Although the expectation appearing in the above integral is not a Gaussian, its moments can be found analytically [9, 13],

$$\mathbb{E}(F_*) = B^\top\Psi_1^* \tag{42}$$

$$\text{Cov}(F_*) = B^\top\left(\Psi_2^* - \Psi_1^*(\Psi_1^*)\top\right)B + \Psi_0^* I - \text{Tr}\left[\left(K_{MM}^{-1} - \left(K_{MM} + \beta\Psi_2\right)^{-1}\right)\Psi_2^*\right]I, \tag{43}$$

where $B = \beta\left(K_{MM} + \beta\Psi_2\right)^{-1}\Psi_1^\top Y$, $\Psi_0^* = \langle k_f(X_*, X_*)\rangle$, $\Psi_1^* = \langle K_{M*}\rangle$ and $\Psi_2^* = \langle K_{M*}K_{*M}\rangle$. All expectations are taken w.r.t. $q(X_*)$ and can be calculated analytically, while K_{M*} denotes the cross-covariance matrix between the training inducing inputs $\tilde{X}$ and X_*. Finally, since Y_* is just a noisy version of F_*, the mean and covariance of (41) is just computed as: $\mathbb{E}(Y_*) = \mathbb{E}(F_*)$ and $\text{Cov}(Y_*) = \text{Cov}(F_*) + \beta^{-1}I_{N_*}$.

B.2 Predictions given the test time points and partially observed outputs

The expression for the predictive density $p(Y_*^m|Y_*^p, Y)$ follows exactly as in section B.1 but we need to compute probabilities for Y_*^m instead of Y_* and Y is replaced with (Y, Y_*^p) in all conditioning sets. Similarly, F is replaced with F^m. Now $q(X_*)$ cannot be found analytically as in section B.1; instead, it is optimised so that Y_*^p are taken into account. This is done by maximising the variational lower bound on the marginal likelihood:

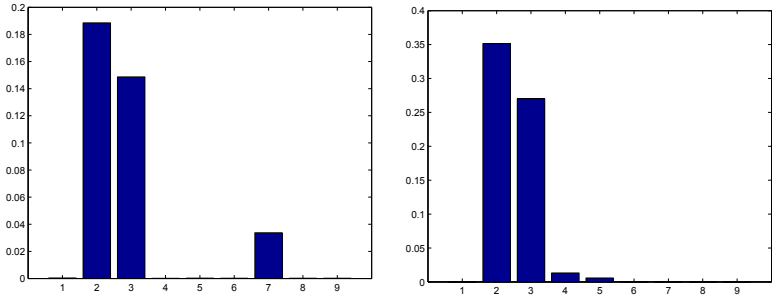
$$\begin{aligned} p(Y_*^p, Y) &= \int p(Y_*^p, Y|X_*, X)p(X_*, X)\mathrm{d}X_*\mathrm{d}X \\ &= \int p(Y^m|X)p(Y_*^p, Y^p|X_*, X)p(X_*, X)\mathrm{d}X_*\mathrm{d}X, \end{aligned}$$

Notice that here, unlike the main paper, we work with the likelihood after marginalising F, for simplicity. Assuming a variational distribution $q(X_*, X)$ and using Jensen's inequality we obtain the lower bound

$$\begin{aligned} &\int q(X_*, X)\log\frac{p(Y^m|X)p(Y_*^p, Y^p|X_*, X)p(X_*, X)}{q(X_*, X)}\mathrm{d}X_*\mathrm{d}X \\ &= \int q(X)\log p(Y^m|X)\mathrm{d}X + \int q(X_*, X)\log p(Y_*^p, Y^p|X_*, X)\mathrm{d}X_*\mathrm{d}X \\ &- \text{KL}[q(X_*, X)||p(X_*, X)] \end{aligned} \tag{44}$$

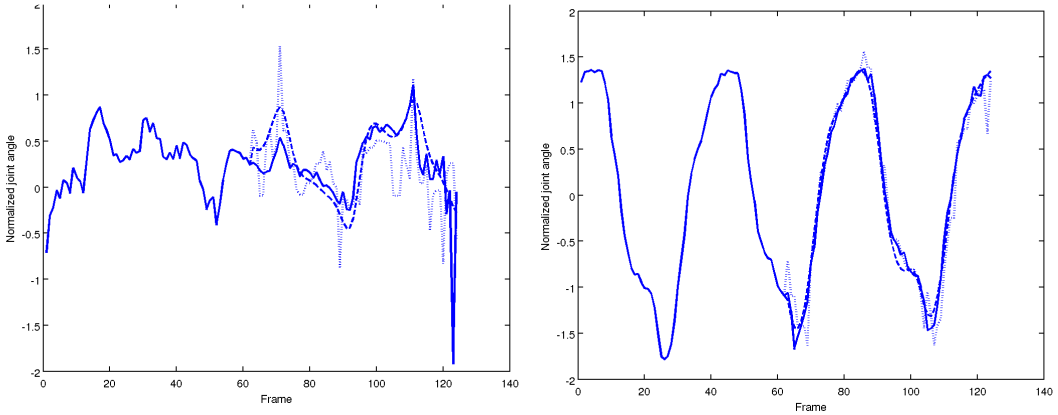
This quantity can now be maximized in the same manner as for the bound of the training phase. Unfortunately, this means that the variational parameters that are already optimised from the training procedure cannot be used here because X and X_* are coupled in $q(X_*, X)$. A much faster but less accurate method would be to decouple the test from the training latent variables by imposing the factorisation $q(X_*, X) = q(X)q(X_*)$. Then, equation (44) would break into terms containing X, X_* or both. The ones containing only X could then be treated as constants.

C Additional results from the experiments



(a) (b)

Figure 3: The values of the scales of the ARD kernel after training on the motion capture dataset using the RBF (fig: (a)) and the Matérn (fig: (b)) covariance function to model the dynamics for VGPDS. The scales that have zero value "switch off" the corresponding dimension of the latent space. The latent space is, therefore, 3-D for (a) and 4-D for (b). Note that the scales were initialized with very similar values.



(a) (b)

Figure 4: The prediction for two of the test angles for the body (fig: 4(a)) and for the legs part (fig: 4(a)). Continuous line is the original test data, dotted line is nearest neighbour in scaled space, dashed line is VGPDS (using the RBF covariance function for the body reconstruction and the Matérn for the legs).

(a) (b) (c) (d)

Figure 5: Some more examples for the reconstruction achieved for the 'dog' dataset. 40% of the test image's pixels (figures (a) and (c)) were presented to the model, which was able to successfully reconstruct them, as can be seen in (b) and (d).