1 Derivation of the variational bound

We wish to approximate the marginal likelihood:

$$p(Y|\mathbf{t}) = \int p(Y, F, X|\mathbf{t}) \mathrm{d}X \mathrm{d}F, \tag{1}$$

by computing a lower bound:

$$\mathcal{F}_v(q, \boldsymbol{\theta}) = \int q(\Theta) \log \frac{p(Y, F, X|\mathbf{t})}{q(\Theta)} \mathrm{d}X \mathrm{d}F, \tag{2}$$

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:

$$p(Y, F, U, X, \tilde{X}|\mathbf{t}) = \prod_{d=1}^{D} p(\mathbf{y}_d|\mathbf{f}_d) p(\mathbf{f}_d|\mathbf{u}_d, X) p(\mathbf{u}_d|\tilde{X}) p(X|\mathbf{t}) \tag{3}$$

where $p(\mathbf{u}_d|\tilde{X}) = \prod_{d=1}^{D} \mathcal{N}(\mathbf{u}_d|\mathbf{0}, K_{MM})$. 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 [**?**]):

$$p(\mathbf{f}_d|\mathbf{u}_d, X) = \mathcal{N}\left(\mathbf{f}_d|K_{NM}K_{MM}^{-1}\mathbf{u}_d, K_{NN} - K_{NM}K_{MM}^{-1}K_{MN}\right). \tag{4}$$

We are now able to define a variational distribution $q(\Theta)$ which factorises as: For tractability we now define a variational density, $q(\Theta)$:

$$q(\Theta) = q(F, U, X) = q(F|U, X)q(U)q(X) = \prod_{d=1}^{D} p(\mathbf{f}_d|\mathbf{u}_d, X) q(\mathbf{u}_d) q(X), \tag{5}$$

where $q(X) = \prod_{q=1}^{Q} \mathcal{N}(\mathbf{x}_q|\boldsymbol{\mu}_q, S_q)$. Now, we return to (2) and replace the joint distribution with its augmented version (3) and the variational distribution with its factorised version (5):

$$\begin{aligned}
\mathcal{F}_v(q, \boldsymbol{\theta}) &= \int q(\Theta) \log \frac{p(Y, F, U, X|\mathbf{t})}{q(F, U, X)} \mathrm{d}X \mathrm{d}F, \\
&= \int \prod_{d=1}^{D} p(\mathbf{f}_d|\mathbf{u}_d, X) q(\mathbf{u}_d) q(X) \log \frac{\prod_{d=1}^{D} p(\mathbf{y}_d|\mathbf{f}_d) \cancel{p(\mathbf{f}_d|\mathbf{u}_d, X)} p(\mathbf{u}_d|\tilde{X}) p(X|\mathbf{t})}{\prod_{d=1}^{D} \cancel{p(\mathbf{f}_d|\mathbf{u}_d, X)} q(\mathbf{u}_d) q(X))} \mathrm{d}X \mathrm{d}F \\
&= \int \prod_{d=1}^{D} p(\mathbf{f}_d|\mathbf{u}_d, X) q(\mathbf{u}_d) q(X) \log \frac{\prod_{d=1}^{D} p(\mathbf{y}_d|\mathbf{f}_d) p(\mathbf{u}_d|\tilde{X})}{\prod_{d=1}^{D} q(\mathbf{u}_d) q(X))} \mathrm{d}X \mathrm{d}F, \\
&\quad - \int \prod_{d=1}^{D} q(X) \log \frac{q(X)}{p(X|\mathbf{t})} \mathrm{d}X \\
&= \hat{\mathcal{F}}_v - \mathrm{KL}(q \parallel p),
\end{aligned} \tag{6}$$

with $\hat{\mathcal{F}}_v = \int q(X) \log p(Y|F) p(F|X) \, \mathrm{d}X \mathrm{d}F = \sum_{d=1}^{D} \hat{\mathcal{F}}_d$. Both terms in (6) are analytically tractable, with the first having the same analytical solution as the one derived in [**?**]. Further calculations in the the $\hat{\mathcal{F}}_v$ term reveal that the optimal setting for $q(\mathbf{u}_d)$ is also a Gaussian.

The complete form of the jensen's lower bound turns out to be:

$$\begin{aligned}
\mathcal{F}_v(q, \boldsymbol{\theta}) &= \sum_{d=1}^{D} \hat{\mathcal{F}}_d(q, \boldsymbol{\theta}) - \mathrm{KL}(q \parallel p) \\
&= \sum_{d=1}^{D} \log \left(\frac{(\beta)^{\frac{N}{2}} |K_{MM}|^{\frac{1}{2}}}{(2\pi)^{\frac{N}{2}} |\beta\Psi_2 + K_{MM}|^{\frac{1}{2}}} e^{-\frac{1}{2}\mathbf{y}_d^T W \mathbf{y}_d} \right) - \frac{\beta\psi_0}{2} + \frac{\beta}{2} \mathrm{Tr}\left(K_{MM}^{-1}\Psi_2\right) \\
&\quad - \frac{Q}{2} \log |K_t| - \frac{1}{2} \sum_{q=1}^{Q} \left[\mathrm{Tr}\left(K_t{}^{-1} S_q\right) + \mathrm{Tr}\left(K_t{}^{-1} \boldsymbol{\mu}_q \boldsymbol{\mu}_q^T\right) \right] + \frac{1}{2} \sum_{q=1}^{Q} \log |S_q| + const
\end{aligned} \tag{7}$$

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

$$\Psi_0 = \text{Tr}(\langle K_{NN} \rangle_{q(X)}) \ , \ \ \Psi_1 = \langle K_{NM} \rangle_{q(X)} \ , \ \ \Psi_2 = \langle K_{MN} K_{NM} \rangle_{q(X)} \tag{8}$$

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

2 Derivatives of the variational bound

Before giving the expressions for the derivatives of the variational bound (6), it should be reminded that the variational parameters μ_q and S_q (for all qs) have been reparametrised as $S_q = \left(K_t^{-1} + diag(\boldsymbol{\lambda}_q)\right)^{-1}$ and $\boldsymbol{\mu}_q = K_t \bar{\boldsymbol{\mu}}_q$, where the function $diag(\cdot)$ transforms a vector into a square diagonal matrix and vice versa. Given the above, the set of the parameters to be optimised is $(\boldsymbol{\theta}_f, \boldsymbol{\theta}_x, \{\bar{\boldsymbol{\mu}}_q, \boldsymbol{\lambda}_q\}_{q=1}^Q, \tilde{X}$. The gradient w.r.t the inducing points $\tilde{X}$, however, has exactly the same form as for $\boldsymbol{\theta}_f$ and, therefore, is not presented here. Also notice that from now on we will often use the term "variational parameters" to refer to the new quantities $\bar{\boldsymbol{\mu}}_q$ and $\boldsymbol{\lambda}_q$.

Some more notation:

1. λ_q is a scalar, an element of the vector $\boldsymbol{\lambda}_q$ which, in turn, is the main diagonal of the diagonal matrix Λ_q.
2. $S_{ij} \triangleq S_{q;ij}$ the element of S_q found in the i-th row and j-th column.
3. $\mathbf{s}_q \triangleq \{S_{q;ii}\}_{i=1}^N$, i.e. it is a vector with the diagonal of S_q.

2.1 Derivatives w.r.t the variational parameters

$$\frac{\vartheta \mathcal{F}_v}{\vartheta \bar{\boldsymbol{\mu}}_q} = K_t \left(\frac{\vartheta \hat{\mathcal{F}}}{\vartheta \boldsymbol{\mu}_q} - \bar{\boldsymbol{\mu}}_q \right) \text{ and } \frac{\vartheta \mathcal{F}_v}{\vartheta \boldsymbol{\lambda}_q} = -(S_q \circ S_q) \left(\frac{\vartheta \hat{\mathcal{F}}}{\vartheta \mathbf{s}_q} + \frac{1}{2} \boldsymbol{\lambda}_q \right). \tag{9}$$

where:

$$\begin{aligned} \frac{\hat{\mathcal{F}}(q, \boldsymbol{\theta})}{\vartheta \mu_q} = & -\frac{\beta D}{2} \frac{\vartheta \Psi_0}{\vartheta \mu_q} + \beta \text{Tr} \left(\frac{\vartheta \Psi_1^T}{\vartheta \mu_q} Y Y^T \Psi_1 A^{-1} \right) \\ & + \frac{\beta}{2} \text{Tr} \left[\frac{\vartheta \Psi_2}{\vartheta \mu_q} \left(D K_{MM}^{-1} - \beta^{-1} D A^{-1} - A^{-1} \Psi_1^T Y Y^T \Psi_1 A^{-1} \right) \right] \end{aligned} \tag{10}$$

$$\begin{aligned} \frac{\vartheta \hat{\mathcal{F}}(q, \boldsymbol{\theta})}{\vartheta S_{q;i,j}} = & -\frac{\beta D}{2} \frac{\vartheta \Psi_0}{\vartheta S_{q;i,j}} + \beta \text{Tr} \left(\frac{\vartheta \Psi_1^T}{\vartheta S_{q;i,j}} Y Y^T \Psi_1 A^{-1} \right) \\ & + \frac{\beta}{2} \text{Tr} \left[\frac{\vartheta \Psi_2}{\vartheta S_{q;i,j}} \left(D K_{MM}^{-1} - \beta^{-1} D A^{-1} - A^{-1} \Psi_1^T Y Y^T \Psi_1 A^{-1} \right) \right] \end{aligned} \tag{11}$$

with $A = \beta^{-1} K_{MM} + \Psi_2$.

2.2 Derivatives w.r.t $\boldsymbol{\theta} = (\boldsymbol{\theta}_f, \boldsymbol{\theta}_x)$ and β

Given that the KL term involves only the temporal prior, its gradient w.r.t the parameters $\boldsymbol{\theta}_f$ is zero. Therefore:

$$\frac{\vartheta \mathcal{F}_v}{\vartheta \theta_f} = \frac{\vartheta \hat{\mathcal{F}}}{\vartheta \theta_f} \tag{12}$$

with:

$$\frac{\vartheta\hat{\mathcal{F}}}{\vartheta\theta_f} = \text{const} - \frac{\beta D}{2}\frac{\vartheta\Psi_0}{\vartheta\theta_f} + \beta\text{Tr}\left(\frac{\vartheta\Psi_1^T}{\vartheta\theta_f}YY^T\Psi_1 A^{-1}\right)$$
$$+ \frac{1}{2}\text{Tr}\left[\frac{\vartheta K_{MM}}{\vartheta\theta_f}\left(DK_{MM}^{-1} - \beta^{-1}DA^{-1} - A^{-1}\Psi_1^T YY^T\Psi_1 A^{-1} - \beta DK_{MM}^{-1}\Psi_2 K_{MM}^{-1}\right)\right]$$
$$+ \frac{\beta}{2}\text{Tr}\left[\frac{\vartheta\Psi_2}{\vartheta\theta_f}\quad\left(DK_{MM}^{-1} - \beta^{-1}DA^{-1} - A^{-1}\Psi_1^T YY^T\Psi_1 A^{-1}\right)\right] \tag{13}$$

The expression above is identical for the derivatives w.r.t the inducing points. For the gradients w.r.t the β term, we have a similar expression:

$$\frac{\vartheta\hat{\mathcal{F}}}{\vartheta\beta} = \frac{1}{2}\Big[D\left(\text{Tr}(K_{MM}^{-1}\Psi_2) + (N-M)\beta^{-1} - \Psi_0\right) - \text{Tr}(YY^\top) + \text{Tr}(A^{-1}\Psi_1^\top YY^\top\Psi_1)$$
$$+\beta^{-2}D\text{Tr}(K_{MM}A^{-1}) + \beta^{-1}\text{Tr}\left(K_{MM}^{-1}A^{-1}\Psi_1^\top YY^\top\Psi_1 A^{-1}\right)\Big] \tag{14}$$

In contrast to the above, the term $\hat{\mathcal{F}}_v$ does involve parameters $\boldsymbol{\theta}_x$, because it involves the variational parameters that are now reparametrized with K_t, which in turn depends on $\boldsymbol{\theta}_x$. To demonstrate that, we will forget for a moment the reparametrization of S_q and we will express the bound as $F(\boldsymbol{\theta}_x, \mu_q(\boldsymbol{\theta}_x))$ (where $\mu_q(\boldsymbol{\theta}_x) = K_t\bar{\boldsymbol{\mu}}_q$) so as to show explicitly the dependency on the variational mean which is now a function of $\boldsymbol{\theta}_x$. Our calculations must now take into account the term $\left(\frac{\vartheta\hat{\mathcal{F}}(\boldsymbol{\mu}_q)}{\vartheta\boldsymbol{\mu}_q}\right)^\top\frac{\vartheta\mu_q(\boldsymbol{\theta}_x)}{\vartheta\boldsymbol{\theta}_x}$ that is what we "miss" when we consider $\mu_q(\boldsymbol{\theta}_x) = \boldsymbol{\mu}_q$:

$$\frac{\vartheta\mathcal{F}_v(\boldsymbol{\theta}_x, \mu_q(\boldsymbol{\theta}_x))}{\vartheta\theta_x} = \frac{\vartheta\mathcal{F}_v(\boldsymbol{\theta}_x, \boldsymbol{\mu}_q)}{\vartheta\theta_x} + \left(\frac{\vartheta\hat{\mathcal{F}}(\boldsymbol{\mu}_q)}{\vartheta\boldsymbol{\mu}_q}\right)^\top\frac{\vartheta\mu_q(\boldsymbol{\theta}_x)}{\vartheta\theta_x}$$
$$= \cancel{\frac{\vartheta\hat{\mathcal{F}}(\boldsymbol{\mu}_q)}{\vartheta\theta_x}} + \frac{\vartheta(-\text{KL})(\boldsymbol{\theta}_x, \boldsymbol{\mu}_q(\boldsymbol{\theta}_x))}{\vartheta\theta_x} + \left(\frac{\vartheta\hat{\mathcal{F}}(\boldsymbol{\mu}_q)}{\vartheta\boldsymbol{\mu}_q}\right)^\top\frac{\vartheta\mu_q(\boldsymbol{\theta}_x)}{\vartheta\theta_x} \tag{15}$$

We do the same for S_q and then we can take the resulting equations and replace $\boldsymbol{\mu}_q$ and S_q with their equals so as to take the final expression which only contains $\bar{\boldsymbol{\mu}}_q$ and $\boldsymbol{\lambda}_q$:

$$\frac{\vartheta\mathcal{F}_v(\boldsymbol{\theta}_x, \mu_q(\boldsymbol{\theta}_x), S_q(\boldsymbol{\theta}_x))}{\vartheta\theta_x} = \text{Tr}\Big[\Big[-\frac{1}{2}\left(\hat{B}_q K_t\hat{B}_q + \bar{\boldsymbol{\mu}}_q\bar{\boldsymbol{\mu}}_q^\top\right)$$
$$+\left(I - \hat{B}_q K_t\right)diag\left(\frac{\vartheta\hat{\mathcal{F}}}{\vartheta\mathbf{s}_q}\right)\left(I - \hat{B}_q K_t\right)^\top\Big]\frac{\vartheta K_t}{\vartheta\theta_x}\Big]$$
$$+\left(\frac{\vartheta\hat{\mathcal{F}}(\boldsymbol{\mu}_q)}{\vartheta\boldsymbol{\mu}_q}\right)^\top\frac{\vartheta K_t}{\vartheta\theta_x}\bar{\boldsymbol{\mu}}_q \tag{16}$$

where $\hat{B}_q = \Lambda_q^{\frac{1}{2}}\widetilde{B}_q^{-1}\Lambda_q^{\frac{1}{2}}$. and $\tilde{B}_q = I + \Lambda_q^{\frac{1}{2}}K_t\Lambda_q^{\frac{1}{2}}$. Note that by using this $\tilde{B}_q$ matrix (which has eigenvalues bounded below by one) we have an expression which, when implemented, leads to more numerically stable computations, as explained in [**?**] page 45-46.

3 Predictions

3.1 Predictions only given 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

$$p(Y_*|Y) = \int p(Y_*, F_*, X_*|Y_*, Y)\text{d}F_*\text{d}X_* = \int p(Y_*|F_*)p(F_*|X_*, Y)p(X_*|Y)\text{d}F_*\text{d}X_*. \tag{17}$$

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

$$q(F_*|X_*) = \int \prod_{d\in D} p(\mathbf{f}_{*,d}|\mathbf{u}_d, X_*)q(\mathbf{u}_d)\mathrm{d}\mathbf{u}_d = \prod_{d\in D} q(\mathbf{f}_{*,d}|X_*), \tag{18}$$

where $q(\mathbf{f}_{*,d}|X_*)$ is a Gaussian that can be computed analytically.The term $p(X_*|Y)$ in eq. (17) is approximated by a Gaussian variational distribution $q(X_*)$,

$$p(X_*|Y) \approx \int p(X_*|X)q(X)\mathrm{d}X = \langle p(X_*|X)\rangle_{q(X)} = q(X_*) = \prod_{q=1}^{Q} q(\mathbf{x}_{*,q}), \tag{19}$$

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

$$\mathbf{x}_{*,q} = \boldsymbol{\alpha}\mathbf{x}_q + \boldsymbol{\epsilon}, \tag{20}$$

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 (19) is a Gaussian and by taking expectations over $q(X)$ in the r.h.s. of (20) we find the mean and covariance of $q(X_*)$. Substituting for the equivalent forms of $\boldsymbol{\mu}_q$ and S_q from section **??** we obtain the final solution

$$\mu_{x_{*,q}} = \mathbf{k}_{*N}\bar{\mu}_q \tag{21}$$

$$\mathrm{var}(x_{*,q}) = k_{**} - \mathbf{k}_{*N}(K_t + \Lambda_q^{-1})^{-1}\mathbf{k}_{N*}. \tag{22}$$

(17) can then be written as:

$$p(Y_*|Y) = \int p(Y_*|F_*)q(F_*|X_*)q(X_*)\mathrm{d}F_*\mathrm{d}X_* = \int p(Y_*|F_*)\left\langle q(F_*|X_*)\right\rangle_{q(X_*)}\mathrm{d}F_* \tag{23}$$

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

$$\mathbb{E}(F_*) = B^\top\Psi_1^* \tag{24}$$

$$\mathrm{Cov}(F_*) = B^\top\left(\Psi_2^* - \Psi_1^*(\Psi_1^*)\top\right)B + \Psi_0^* I - \mathrm{Tr}\left[\left(K_{MM}^{-1} - \left(K_{MM} + \beta\Psi_2\right)^{-1}\right)\Psi_2^*\right]I, \tag{25}$$

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 (23) is just computed as: $\mathbb{E}(Y_*) = \mathbb{E}(F_*)$ and $\mathrm{Cov}(Y_*) = \mathrm{Cov}(F_*) + \beta^{-1}I_{N_*}$.

3.2 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 3.1 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 3.1; 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:

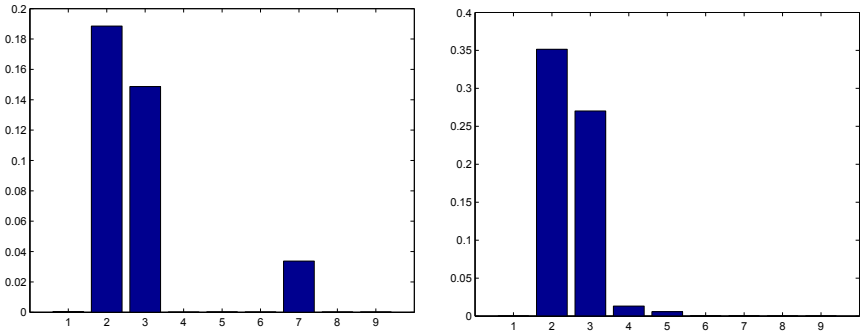
$$\begin{aligned} p(Y_*^p, Y) &= \int p(Y_*^p, Y|X_*, X)p(X_*, X)\mathrm{d}X_*\mathrm{d}X \\ &= \int p(Y^m|X)p(Y_*^p, Y^p|X_*, X)p(X_*, X)\mathrm{d}X_*\mathrm{d}X, \end{aligned}$$

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

$$\begin{aligned} &\int q(X_*, X)\log\frac{p(Y^m|X)p(Y_*^p, Y^p|X_*, X)p(X_*, X)}{q(X_*, X)}\mathrm{d}X_*\mathrm{d}X \\ &= \int q(X)\log p(Y^m|X)\mathrm{d}X + \int q(X_*, X)\log p(Y_*^p, Y^p|X_*, X)\mathrm{d}X_*\mathrm{d}X \\ &- \mathrm{KL}[q(X_*, X)||p(X_*, X)] \end{aligned} \tag{26}$$

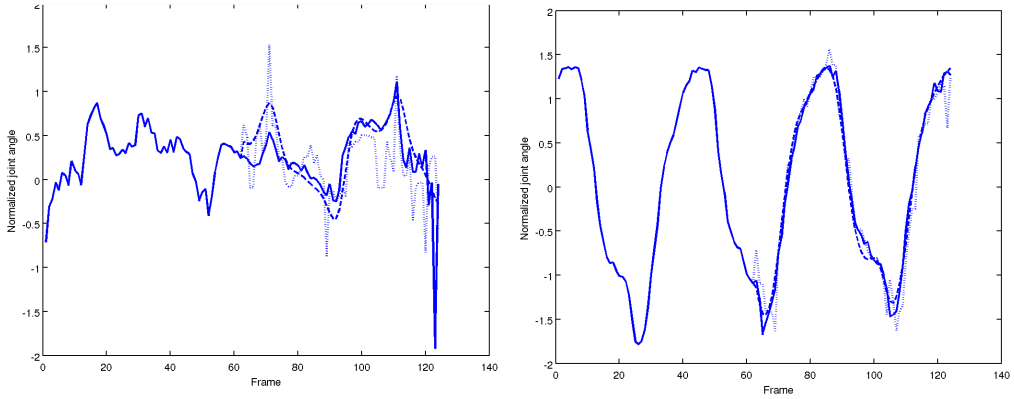
This quantity can now be maximized in the same manner as for the bound of the training phase. Unfortunately, this means that the variational parameters that are already optimised from the training procedure cannot be used here because X and X_* are coupled in $q(X_*, X)$. A much faster but less accurate 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 (26) would break into terms containing X, X_* or both. The ones containing only X could then be treated as constants.

4 Additional results from the experiments



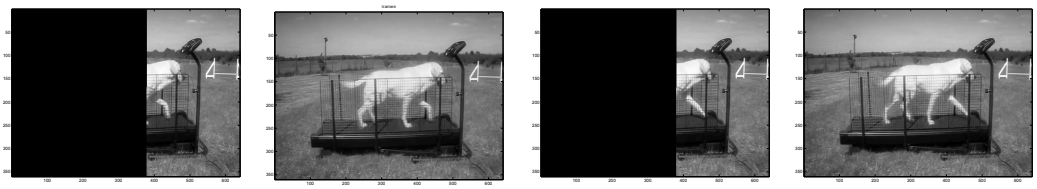
(a) (b)

Figure 1: The values of the scales of the ARD kernel after training on the motion capture dataset using the RBF (fig: (a)) and the Matern (fig: (b)) kernel 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 2: The prediction for two of the test angles for the body (fig: 2(a)) and for the legs part (fig: 2(a)). Continuous line is the original test data, dotted line is nearest neighour in scaled space, dashed line is VGPDS (using the RBF kernel for the body reconstruction and the Matern for the legs).



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

Figure 3: 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 succesfully reconstruct them, as can be seen in (b) and (d).