DP-GP-LVM: A Bayesian Non-Parametric Model for Learning Multivariate Dependency Structures – Supplemental Material

Andrew R. Lawrence [1] **Carl Henrik Ek** [2] **Neill D. F. Campbell** [1]

A. Derivation of the Lower Bound

The joint distribution for DP-GP-LVM is defined in (11) in § 3 and repeated here:

$$p(\mathbf{Y}, \mathbf{F}, \mathbf{X}, \mathbf{Z}, \mathbf{V}, \alpha, \boldsymbol{\Theta}, \boldsymbol{\beta}) = p(\mathbf{Y}|\mathbf{F}, \boldsymbol{\beta}, \mathbf{Z})\, p(\mathbf{F}|\mathbf{X}, \boldsymbol{\Theta}, \mathbf{Z})\, p(\boldsymbol{\Theta}, \boldsymbol{\beta})\, p(\mathbf{X})\, p(\mathbf{Z}|\mathbf{V})\, p(\mathbf{V}|\alpha)\, p(\alpha)\ . \tag{1}$$

Given this factorization of the joint distribution, the following describes the generative process of the DP-GP-LVM for observed data $\mathbf{Y}$ with N observations and D dimensions:

1. Draw latent inputs $\mathbf{X}$ from prior: $\mathbf{X} \sim p(\mathbf{X})$
2. Draw concentration parameter α from prior: $\alpha \sim p(\alpha)$
3. For an infinite number of draws $t = \{1, 2, \cdots\}$:
   (a) Draw v_t given α: $v_t \sim \mathrm{Beta}(1, \alpha)$
   (b) Draw kernel hyperparameter atoms $\boldsymbol{\theta}_t$ from prior: $\boldsymbol{\theta}_t \sim p(\boldsymbol{\theta})$
   (c) Draw noise precision atoms β_t from prior: $\beta_t \sim p(\beta)$
   Note: $p(\boldsymbol{\theta}, \beta)$ is the base distribution G_0 of the DP.
4. Build mixing proportions $\pi(\mathbf{V})$: $\pi_t(\mathbf{V}) = v_t \prod_{i=1}^{t-1}(1 - v_i)$
5. For each observed dimension $d \in [1, D]$:
   (a) Draw $\mathbf{z}_d$ given $\pi(\mathbf{V})$: $\mathbf{z}_d \sim \mathrm{Mult}(\pi(\mathbf{V}))$
   (b) Draw function $\mathbf{f}_d$ given $\mathbf{z}_d, \mathbf{X}$, and $\boldsymbol{\Theta}$: $\mathbf{f}_d \sim \mathcal{GP}\left(\mathbf{0}, k\left(\mathbf{X}, \mathbf{X}; \prod_{t=1}^{\infty} \boldsymbol{\theta}_t{}^{[\mathbf{z}_d = t]}\right)\right)$
   (c) Draw data $\mathbf{y}_d$ given $\mathbf{f}_d, \mathbf{z}_d$, and $\boldsymbol{\beta}$: $\mathbf{y}_d \sim \mathcal{N}\left(\mathbf{f}_d, \left[\prod_{t=1}^{\infty} \beta_t{}^{-[\mathbf{z}_d = t]}\right] \mathbf{I}_N\right)$.

As discussed in § 3.2, the joint distribution (1) is supplemented with pseudo inputs to and outputs from the GP. Therefore, the joint distribution for DP-GP-LVM becomes:

$$\begin{aligned}
&p(\mathbf{Y}, \mathbf{F}, \mathbf{X}, \mathbf{U}, \mathbf{X}_\mathrm{u}, \mathbf{Z}, \mathbf{V}, \alpha, \boldsymbol{\Theta}, \boldsymbol{\beta}) \\
&= p(\mathbf{Y}|\mathbf{F}, \boldsymbol{\beta}, \mathbf{Z})\, p(\mathbf{F}|\mathbf{U}, \mathbf{X}, \mathbf{X}_\mathrm{u}, \boldsymbol{\Theta}, \mathbf{Z})\, p(\mathbf{U}|\mathbf{X}_\mathrm{u}, \boldsymbol{\Theta}, \mathbf{Z})\, p(\boldsymbol{\Theta}, \boldsymbol{\beta})\, p(\mathbf{X})\, p(\mathbf{Z}|\mathbf{V})\, p(\mathbf{V}|\alpha)\, p(\alpha) && (2)\\
&= \prod_{d=1}^{D} \Big[p(\mathbf{y}_d|\mathbf{f}_d, \boldsymbol{\beta}, \mathbf{z}_d)\, p(\mathbf{f}_d|\mathbf{u}_d, \mathbf{X}, \mathbf{X}_\mathrm{u}, \boldsymbol{\Theta}, \mathbf{z}_d)\, p(\mathbf{u}_d|\mathbf{X}_\mathrm{u}, \boldsymbol{\Theta}, \mathbf{z}_d)\, p(\mathbf{z}_d|\mathbf{V})\Big] p(\boldsymbol{\Theta}, \boldsymbol{\beta})\, p(\mathbf{V}|\alpha)\, p(\alpha)\, p(\mathbf{X}) && (3)\\
&= \prod_{d=1}^{D} \Big[p(\mathbf{y}_d|\mathbf{f}_d, \boldsymbol{\beta}, \mathbf{z}_d)\, p(\mathbf{f}_d|\mathbf{u}_d, \mathbf{X}, \mathbf{X}_\mathrm{u}, \boldsymbol{\Theta}, \mathbf{z}_d)\, p(\mathbf{u}_d|\mathbf{X}_\mathrm{u}, \boldsymbol{\Theta}, \mathbf{z}_d)\, p(\mathbf{z}_d|\mathbf{V})\Big] \prod_{t=1}^{\infty} \Big[p(\boldsymbol{\theta}_t, \beta_t)\, p(v_t|\alpha)\Big] p(\alpha)\, p(\mathbf{X}) && (4)
\end{aligned}$$

[1]Dept. of Computer Science, University of Bath, UK [2]Dept. of Computer Science, University of Bristol, UK. Correspondence to: Andrew R. Lawrence <A.R.Lawrence@bath.ac.uk>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

We introduce fully factorized variational distributions for all the latent variables we are marginalizing out:

$$\Omega = \{\mathbf{F}, \mathbf{U}, \mathbf{X}, \mathbf{Z}, \mathbf{V}, \alpha\} \tag{5}$$
$$q(\Omega) = q(\mathbf{F}, \mathbf{U}, \mathbf{X}, \mathbf{Z}, \mathbf{V}, \alpha) \tag{6}$$
$$= p(\mathbf{F}|\mathbf{U}, \mathbf{X}, \mathbf{X}_\mathrm{u}, \boldsymbol{\Theta}, \mathbf{Z}) q(\mathbf{U}) q(\mathbf{X}) q(\mathbf{Z}) q(\mathbf{V}) q(\alpha) \tag{7}$$
$$= q(\mathbf{X}) q(\alpha) \prod_{d=1}^{D} p(\mathbf{f}_d|\mathbf{u}_d, \mathbf{X}, \mathbf{X}_\mathrm{u}, \boldsymbol{\Theta}, \mathbf{z}_d) \prod_{d=1}^{D} q(\mathbf{u}_d) \prod_{d=1}^{D} q(\mathbf{z}_d) \prod_{t=1}^{T} q(v_t) \tag{8}$$

The evidence lower bound for DP-GP-LVM is defined in (13) in § 3.2 and repeated here:

$$\log p(\mathbf{Y}, \boldsymbol{\Theta}, \boldsymbol{\beta}) = \log \int q(\boldsymbol{\Omega}) \frac{p(\mathbf{Y}, \boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{\Omega})}{q(\boldsymbol{\Omega})} \, \mathrm{d}\boldsymbol{\Omega} \tag{9}$$
$$\geq \mathbb{E}_q \left[\log p(\mathbf{Y}, \boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{\Omega})\right] - \mathbb{E}_q \left[\log q(\boldsymbol{\Omega})\right] := \mathcal{L}. \tag{10}$$

$$\mathcal{L} = \mathbb{E}_q \left[\log p(\mathbf{Y}, \boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{\Omega})\right] - \mathbb{E}_q \left[\log q(\boldsymbol{\Omega})\right] \tag{11}$$
$$= \sum_{d=1}^{D} \mathbb{E}_{p(\mathbf{f}_d|\mathbf{u}_d, \mathbf{X}, \mathbf{X}_\mathrm{u}, \boldsymbol{\Theta}, \mathbf{z}_d) q(\mathbf{u}_d) q(\mathbf{X}) q(\mathbf{z}_d)} \left[\log p(\mathbf{y}_d|\mathbf{f}_d, \boldsymbol{\beta}, \mathbf{z}_d)\right] + \mathbb{E}_{q(\mathbf{X})} \left[\log p(\mathbf{X})\right]$$
$$+ \sum_{d=1}^{D} \mathbb{E}_{q(\mathbf{u}_d) q(\mathbf{z}_d)} \left[\log p(\mathbf{u}_d|\mathbf{X}_\mathrm{u}, \boldsymbol{\Theta}, \mathbf{z}_d)\right] + \sum_{d=1}^{D} \mathbb{E}_{q(\mathbf{z}_d) q(\mathbf{V})} \left[\log p(\mathbf{z}_d|\mathbf{V})\right] + \mathbb{E}_{q(\mathbf{V}) q(\alpha)} \left[\log p(\mathbf{V}|\alpha)\right]$$
$$+ \mathbb{E}_{q(\alpha)} \left[\log p(\alpha)\right] + \sum_{t=1}^{T} \log p(\boldsymbol{\theta}_t) + \sum_{t=1}^{T} \log p(\beta_t) + \mathbb{H}[q(\mathbf{U})] + \mathbb{H}[q(\mathbf{X})]$$
$$+ \mathbb{H}[q(\mathbf{Z})] + \mathbb{H}[q(\mathbf{V})] + \mathbb{H}[q(\alpha)] \tag{12}$$

As discussed in § 3.2, the optimal $q(\mathbf{U})$ is found to be Gaussian through variational calculus (Damianou et al., 2016) and marginalized out in closed form. Additionally, the other GP terms can be rearranged into a free energy term, defined by (18) in § 3.2, and the KL divergence between $q(\mathbf{X})$ and $p(\mathbf{X})$ (Damianou et al., 2016). Therefore, the lower bound can be written as the following:

$$\mathcal{L} = \sum_{d=1}^{D} \mathcal{F}_d - \mathrm{KL}\big(q(\mathbf{X}) \,\|\, p(\mathbf{X})\big) + \sum_{d=1}^{D} \mathbb{E}_{q(\mathbf{z}_d) q(\mathbf{V})} \left[\log p(\mathbf{z}_d|\mathbf{V})\right] + \mathbb{E}_{q(\mathbf{V}) q(\alpha)} \left[\log p(\mathbf{V}|\alpha)\right]$$
$$+ \mathbb{E}_{q(\alpha)} \left[\log p(\alpha)\right] + \sum_{t=1}^{T} \log p(\boldsymbol{\theta}_t) + \sum_{t=1}^{T} \log p(\beta_t) + \mathbb{H}[q(\mathbf{Z})] + \mathbb{H}[q(\mathbf{V})] + \mathbb{H}[q(\alpha)] \tag{13}$$

$\mathcal{F}_d$ is defined by (18) in § 3.2 and a stable calculation is provided in § B. The KL divergence term is defined by (32) in § B. All the terms related to the DP are defined in § D. Finally, the priors for the kernel hyperparameters and the noise precision are defined by (9) in § 3.

B. Stable Calculation of the GP Lower Bound

As noted by Damianou et al. (2016), care must be taken when calculating the free energy term in (18) in § 3.2. We provide a derivation for stable calculation of the free energy.

$$\mathcal{F}_d = \log \left[\frac{\beta^{N/2} |\mathbf{K}_{\mathrm{uu}}|^{1/2}}{(2\pi)^{N/2} |\beta \boldsymbol{\Psi}_2 + \mathbf{K}_{\mathrm{uu}}|^{1/2}} \exp \left[-\frac{1}{2} \boldsymbol{y}_d^T W \boldsymbol{y}_d \right] \right] - \frac{1}{2} \beta \psi_0 + \frac{1}{2} \beta \mathrm{Tr} \left[\mathbf{K}_{\mathrm{uu}}^{-1} \boldsymbol{\Psi}_2 \right] \tag{14}$$
$$= \frac{N}{2} \log[\beta] + \frac{1}{2} \log |\mathbf{K}_{\mathrm{uu}}| - \frac{N}{2} \log[2\pi] - \frac{1}{2} \log |\beta \boldsymbol{\Psi}_2 + \mathbf{K}_{\mathrm{uu}}| - \frac{1}{2} \boldsymbol{y}_d^T W \boldsymbol{y}_d - \frac{1}{2} \beta \psi_0 + \frac{1}{2} \beta \mathrm{Tr} \left[\mathbf{K}_{\mathrm{uu}}^{-1} \boldsymbol{\Psi}_2 \right] \tag{15}$$

where $W = \beta I_N - \beta^2 \boldsymbol{\Psi}_1 \left(\beta \boldsymbol{\Psi}_2 + \mathbf{K}_{\mathrm{uu}} \right)^{-1} \boldsymbol{\Psi}_1^T$.

Let $L_u = \text{chol}\left[\mathbf{K}_{\text{uu}}\right]$ such that $\mathbf{K}_{\text{uu}} = L_u L_u^T$.

$$\Rightarrow \mathcal{F}_d = \frac{N}{2}\log[\beta] + \frac{1}{2}\log|L_u L_u^T| - \frac{N}{2}\log[2\pi] - \frac{1}{2}\log|\beta\mathbf{\Psi}_2 + L_u L_u^T| \tag{16}$$

$$- \frac{1}{2}\boldsymbol{y}_d^T\left[\beta I_N - \beta^2\mathbf{\Psi}_1\left(\beta\mathbf{\Psi}_2 + L_u L_u^T\right)^{-1}\mathbf{\Psi}_1^T\right]\boldsymbol{y}_d - \frac{1}{2}\beta\psi_0 + \frac{1}{2}\beta\text{Tr}\left[(L_u L_u^T)^{-1}\mathbf{\Psi}_2\right] \tag{17}$$

$$= -\frac{N}{2}\log[2\pi] + \frac{N}{2}\log[\beta] + \frac{1}{2}\log|L_u||L_u^T| - \frac{1}{2}\log|L_u||\beta L_u^{-1}\mathbf{\Psi}_2 L_u^{-T} + I_M||L_u^T| \tag{18}$$

$$- \frac{1}{2}\boldsymbol{y}_d^T\left[\beta I_N - \beta^2\mathbf{\Psi}_1\left(L_u(\beta L_u^{-1}\mathbf{\Psi}_2 L_u^{-T} + I_M)L_u^T\right)^{-1}\mathbf{\Psi}_1^T\right]\boldsymbol{y}_d - \frac{1}{2}\beta\psi_0 \tag{19}$$

$$+ \frac{1}{2}\beta\text{Tr}\left[L_u^{-1}\mathbf{\Psi}_2 L_u^{-T}\right] \tag{20}$$

$$= -\frac{N}{2}\log[2\pi] + \frac{N}{2}\log[\beta] - \frac{1}{2}\log|\beta L_u^{-1}\mathbf{\Psi}_2 L_u^{-T} + I_M| \tag{21}$$

$$- \frac{1}{2}\boldsymbol{y}_d^T\left[\beta I_N - \beta^2\mathbf{\Psi}_1\left(L_u(\beta L_u^{-1}\mathbf{\Psi}_2 L_u^{-T} + I_M)L_u^T\right)^{-1}\mathbf{\Psi}_1^T\right]\boldsymbol{y}_d - \frac{1}{2}\beta\psi_0 \tag{22}$$

$$+ \frac{1}{2}\beta\text{Tr}\left[L_u^{-1}\mathbf{\Psi}_2 L_u^{-T}\right] \tag{23}$$

Let $A = \beta L_u^{-1}\mathbf{\Psi}_2 L_u^{-T} + I_M$ and $L_A = \text{chol}\left[A\right]$ such that $A = L_A L_A^T$.

$$\Rightarrow \mathcal{F}_d = -\frac{N}{2}\log[2\pi] + \frac{N}{2}\log[\beta] - \frac{1}{2}\log|A| - \frac{1}{2}\beta\psi_0 + \frac{1}{2}\beta\text{Tr}\left[L_u^{-1}\mathbf{\Psi}_2 L_u^{-T}\right] \tag{24}$$

$$- \frac{1}{2}\boldsymbol{y}_d^T\left[\beta I_N - \beta^2\mathbf{\Psi}_1\left(L_u A L_u^T\right)^{-1}\mathbf{\Psi}_1^T\right]\boldsymbol{y}_d \tag{25}$$

$$= -\frac{N}{2}\log[2\pi] + \frac{N}{2}\log[\beta] - \log|L_A| - \frac{1}{2}\beta\psi_0 + \frac{1}{2}\beta\text{Tr}\left[L_u^{-1}\mathbf{\Psi}_2 L_u^{-T}\right] - \frac{1}{2}\beta\boldsymbol{y}_d^T\boldsymbol{y}_d \tag{26}$$

$$+ \frac{1}{2}\boldsymbol{y}_d^T\left[\beta^2\mathbf{\Psi}_1 L_u^{-T}L_A^{-T}L_A^{-1}L_u^{-1}\mathbf{\Psi}_1^T\right]\boldsymbol{y}_d \tag{27}$$

$$= -\frac{N}{2}\log[2\pi] + \frac{N}{2}\log[\beta] - \log|L_A| - \frac{1}{2}\beta\psi_0 + \frac{1}{2}\beta\text{Tr}\left[L_u^{-1}\mathbf{\Psi}_2 L_u^{-T}\right] - \frac{1}{2}\beta\boldsymbol{y}_d^T\boldsymbol{y}_d \tag{28}$$

$$+ \frac{1}{2}\boldsymbol{y}_d^T\left[\beta^2 C^T C\right]\boldsymbol{y}_d \tag{29}$$

where $C = L_A^{-1}L_u^{-1}\mathbf{\Psi}_1^T$.

Now $\mathcal{L}_{\mathcal{GP}} = \mathcal{F} - \text{KL}\big(q(\mathbf{X})\,\|\,p(\mathbf{X})\big)$ where $\mathcal{F} = \sum_{d=1}^{D}\mathcal{F}_d$.

$$\Rightarrow \mathcal{F} = -\frac{ND}{2}\log[2\pi] + \frac{ND}{2}\log[\beta] - D\log|L_A| - \frac{D}{2}\beta\psi_0 + \frac{D}{2}\beta\text{Tr}\left[L_u^{-1}\mathbf{\Psi}_2 L_u^{-T}\right] \tag{30}$$

$$- \frac{1}{2}\beta\text{Tr}\left[\mathbf{Y}\mathbf{Y}^T\right] + \frac{1}{2}\beta^2\text{Tr}\left[\mathbf{Y}^T C^T C\mathbf{Y}\right] \tag{31}$$

and

$$\text{KL}\big(q(\mathbf{X})\,\|\,p(\mathbf{X})\big) = \frac{1}{2}\sum_{n=1}^{N}\text{Tr}\left[\mu_n\mu_n^T + \Sigma_n - \log(\Sigma_n)\right] - \frac{NQ}{2} \tag{32}$$

C. Kernel Expectations

We provide details of the kernel expectations used in § 3.2. We assume expectations with respect to $\mathbf{Z}$ have already been taken for β and the covariance kernels.

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sigma^2 \exp\Big(-\frac{1}{2}\sum_{q=1}^{Q} \gamma_q \big\| x_{i,q} - x_{j,q} \big\|^2 \Big) \tag{33}$$

$$\psi_0 = \sum_{i=1}^{N} \int \kappa(\boldsymbol{x}_i, \boldsymbol{x}_i)\, \mathcal{N}(\boldsymbol{x}_i | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) d\boldsymbol{x}_i \tag{34}$$

$$\big[\boldsymbol{\Psi}_1\big]_{i,j} = \int \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j^{\mathrm{u}})\, \mathcal{N}(\boldsymbol{x}_i | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) d\boldsymbol{x}_i \tag{35}$$

$$= \sigma^2 \prod_{q=1}^{Q} \frac{1}{(\gamma_q \Sigma_{i,q} + 1)^{1/2}} \exp\left[\frac{-\gamma_q (\mu_{i,q} - x_{j,q}^{\mathrm{u}})^2}{2(\gamma_q \Sigma_{i,q} + 1)} \right] \tag{36}$$

$$= \sigma^2 \frac{1}{\prod_{q=1}^{Q} (\gamma_q \Sigma_{i,q} + 1)^{1/2}} \exp\left[-\sum_{q=1}^{Q} \frac{\gamma_q (\mu_{i,q} - x_{j,q}^{\mathrm{u}})^2}{2(\gamma_q \Sigma_{i,q} + 1)} \right] \tag{37}$$

$$\log\Big(\big[\boldsymbol{\Psi}_1\big]_{i,j}\Big) = \log(\sigma^2) - \frac{1}{2}\sum_{q=1}^{Q} \log\big(\gamma_q \Sigma_{i,q} + 1\big) - \sum_{q=1}^{Q} \frac{\gamma_q (\mu_{i,q} - x_{j,q}^{\mathrm{u}})^2}{2(\gamma_q \Sigma_{i,q} + 1)} \tag{38}$$

$$\big[\boldsymbol{\Psi}_2\big]_{j,j'} = \sum_{i=1}^{N} \int \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j^{\mathrm{u}})\, \kappa(\boldsymbol{x}_{j'}^{\mathrm{u}}, \boldsymbol{x}_i)\, \mathcal{N}(\boldsymbol{x}_i | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) d\boldsymbol{x}_i \tag{39}$$

$$\log\Big(\big[\boldsymbol{\Psi}_2\big]_{j,j'}\Big) = \sum_{i=1}^{N} 2\log(\sigma^2) - \frac{1}{2}\sum_{q=1}^{Q} \log\big(2\gamma_q \Sigma_{i,q} + 1\big) - \frac{1}{4}\sum_{q=1}^{Q} \gamma_q \left(x_{j,q}^{\mathrm{u}} - x_{j',q}^{\mathrm{u}}\right)^2 \tag{40}$$

$$- \sum_{q=1}^{Q} \frac{\gamma_q}{2\gamma_q \Sigma_{i,q} + 1} \left(\mu_{i,q} - \frac{1}{2}(x_{j,q}^{\mathrm{u}} + x_{j',q}^{\mathrm{u}}) \right)^2 \tag{41}$$

D. Details of the DP Lower Bound

We provide further details about the expressions in the DP lower bound of (24).

$$\mathbb{E}_q[\log p(\mathbf{z}_d | \mathbf{V})] = \sum_{t=1}^{T-1} \left[\phi_{d,t} \left(\Psi(a_t) - \Psi(a_t + b_t) \right) + \sum_{t'=t+1}^{T} \phi_{d,t'} \left(\Psi(b_t) - \Psi(a_t + b_t) \right) \right] \tag{42}$$

$$\mathbb{E}_q[\log p(\mathbf{V} | \alpha)] = (T-1)\left[\Psi(w_1) - \log(w_2) \right] + \sum_{t=1}^{T-1} \left[\left(\frac{w_1}{w_2} - 1 \right) \left(\Psi(b_t) - \Psi(a_t + b_t) \right) \right] \tag{43}$$

$$\mathbb{E}_q[\log p(\alpha)] = -\log\Gamma(s_1) + s_1 \log(s_2) - s_2 \left(\frac{w_1}{w_2} \right) + (s_1 - 1)\left[\Psi(w_1) - \log(w_2) \right] \tag{44}$$

$$\mathbb{H}[q(\mathbf{V})] = \sum_{t=1}^{T} \log B(a_t, b_t) - (a_t - 1)\Psi(a_t) - (b_t - 1)\Psi(b_t) + (a_t + b_t - 2)\Psi(a_t + b_t) \tag{45}$$

$$\mathbb{H}[q(\mathbf{Z})] = -\sum_{d=1}^{D}\sum_{t=1}^{T} \phi_{d,t} \log(\phi_{d,t}) \tag{46}$$

$$\mathbb{H}[q(\alpha)] = w_1 - \log(w_2) + \log\Gamma(w_1) + (1 - w_1)\Psi(w_1) \tag{47}$$

We have used $\Psi(\cdot)$ as the digamma function and $B(\cdot, \cdot)$ as the beta function.

E. Priors on the Latent Variables X

In § 3 we allocated an fully factorized prior over the latent representation $\mathbf{X}$ in (1). We note, however, that the prior only appears in the KL term in (17). This allows for the possibility of using a variety of more informative priors for dealing with specific data sets. As noted by Damianou et al. (2016), it is straight forward to extend this model to include a dynamical prior for sequential data.

Dynamic Prior Motion capture and pose tracking data sets include time information. The DP-GP-LVM can be extended to include this additional information. The only necessary change is to the prior on the latent variables $\mathbf{X}$. We can define each latent dimension as a temporal latent function drawn from a GP. Therefore,

$$\mathbf{x}_q(\mathbf{t}) \sim \mathcal{GP}(\mathbf{0}, k_{\mathrm{x}}(\mathbf{t}, \mathbf{t}')) \tag{48}$$

$$p(\mathbf{X}|\mathbf{t}) = \prod_{q=1}^{Q} \mathcal{N}(\mathbf{x}_q|\mathbf{0}, \mathbf{K}_{\mathrm{x}}), \tag{49}$$

where $\mathbf{K}_{\mathrm{x}} = k_{\mathrm{x}}(\mathbf{t}, \mathbf{t})$.

The variational distribution $q(\mathbf{X})$ is then defined as

$$q(\mathbf{X}) = \prod_{q=1}^{Q} \mathcal{N}(\mathbf{x}_q|\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q), \tag{50}$$

where $\mathbf{x}_q$ and $\boldsymbol{\mu}_q$ and N-length vectors and $\boldsymbol{\Sigma}_q$ is a full $N \times N$ matrix.

F. Variational Parameters

We provide a comparison of the variational free parameters between the three models discussed: Bayesian GP-LVM, MRD, and our proposed DP-GP-LVM.

The total number of free parameters being fit when maximising the ELBO of DP-GP-LVM defined with the ARD covariance function is $(2NQ + MQ + DT + QT + 4T)$, where N is the number of observations, M is the number of inducing points, where $M \ll N$, D is the observed dimensionality, Q is the latent dimensionality, where $Q \ll D$, and T is the truncation level.

The Bayesian GP-LVM requires less $(2NQ + MQ + Q + 2)$, but it is not as expressive of a model as all the observed dimensions share the same mapping function. MRD requires $(2NQ + TMQ + QT + 2T)$, where T is the number of views, and the fully independent variant of MRD (fi-MRD) requires $(2NQ + DMQ + QD + 2D)$.

A key difference between DP-GP-LVM and MRD is that DP-GP-LVM only requires one set of inducing input points $\mathbf{X}_u$, while MRD requires one set per view; therefore, the TMQ term for MRD will grow linearly with additional views. With respect to fi-MRD, this term (DMQ) is likely large as we are studying high-dimensional data. This is the critical complexity saving for the DP-GP-LVM.

Another important difference between MRD and DP-GP-LVM is that the allocation variable $\mathbf{Z}$ is observed in MRD. Therefore, DP-GP-LVM has DT extra terms for the variational parameters for $q(\mathbf{Z})$. The number of free parameters in MRD is highly dependent on the data set being studied. For the experiments presented in this paper, it was always observed that DP-GP-LVM trained faster than MRD.

G. Additional Results

Missing Data Configuration For all the missing data experiments, the data was **randomly permuted** across n and across groups of $2d$ (so as not to break up corresponding (x, y) pairs). Each missing data scenario was run ten times with a different random seed to start. Fig. 1 illustrates how the data was removed after the random permutations; the shaded region is the missing data block removed from the full $[N \times D]$ data set. As a block was removed, we know the true values of the data when comparing against prediction. The models were first trained on a training set with all dimensions. The latent factorization (i.e., ARD weights) and latent inputs ($\mathbf{X}$) were kept fixed during testing. Using the partially observed test set (with dimensionality less than D), new latent input locations were inferred ($\mathbf{X}^*$), given the partial test observations and latent input test locations, the missing data can be predicted. The models provide a predictive mean and covariance for the missing data, which allows us to calculate the log-likelihood of the ground truth values given this predictive posterior.

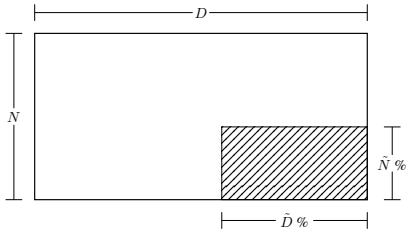


Figure 1. Missing data scenario.

Resolving Ambiguities on CMU MoCap One of the benefits of a factorized latent space is that it allows one to model scenarios that are inherently ambiguous (Ek et al., 2008). To evaluate the performance of the DP-GP-LVM in this scenario we modified a data set of subject 35 from the CMU motion capture database (CMU Graphics Lab, 2002) walking such that for each upper-body pose two different configurations of the lower-body existed (arms swinging in and out of phase with the legs). Fig. 2 shows the latent dimensions found by our model is shown. DP-GP-LVM learns a three-dimensional latent representation, two representing the cyclic structure of the walk and a third dimensions which effectively parametrises the ambiguity, providing two separate latent coordinates for each position in the walking cycle.

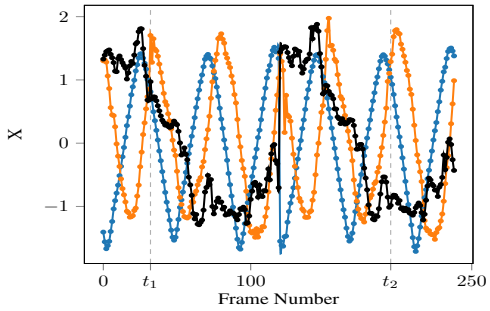
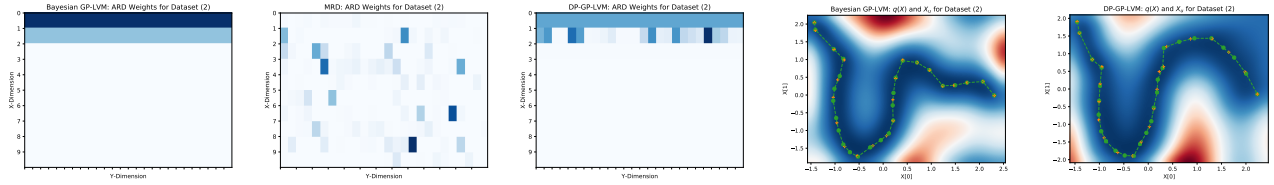


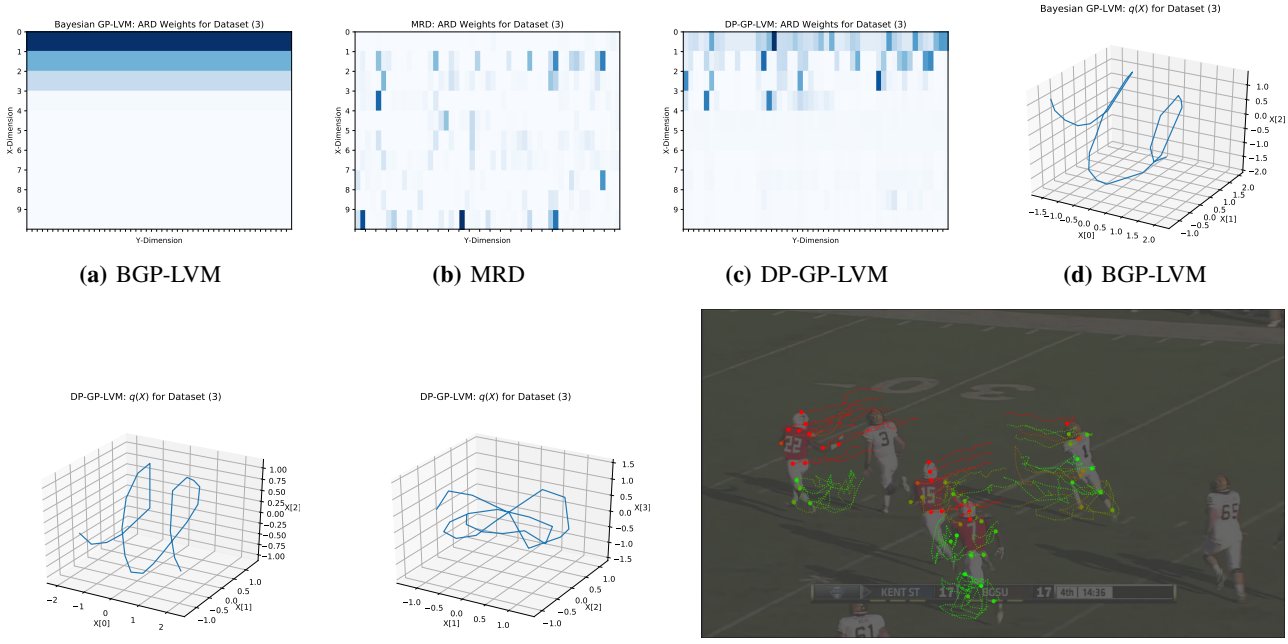
Figure 2. The three active latent dimensions for the walking sequence with leg ambiguities. Each dimension value is plotted against the frames in the sequence. The blue and the orange dimensions encode the cyclic motion while the black encodes a signal which disambiguates the two configurations of the lower body for each upper body configuration. An example of this is frame, t_1 and t_2 where the blue and the orange curve takes the same value, while the black disambiguates the pose by changing location.

Extended Results We include expanded versions of experimental results on the PoseTrack data in Fig. 3 and Fig. 4, the horse motion capture sequence in Fig. 5, and a challenging synthetic example in Fig. 6.



(a) BGP-LVM **(b)** MRD **(c)** DP-GP-LVM **(d)** BGP-LVM **(e)** DP-GP-LVM

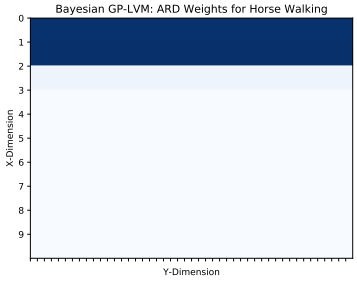
Figure 3. Results for PoseTrack data set with two individuals. (a)-(c) show the ARD weights returned for the three models. BGP-LVM result shares only two dimensions, failing to capture the full diversity of the data. MRD captures these changes but fails to find the appropriate correlations, creating unnecessary independence. DP-GP-LVM capture both the shared correlation and the subtle independences. (d) and (e) show the manifold for the first two shared latent dimensions for BGP-LVM and DP-GP-LVM indicating that they both capture the smooth structure.



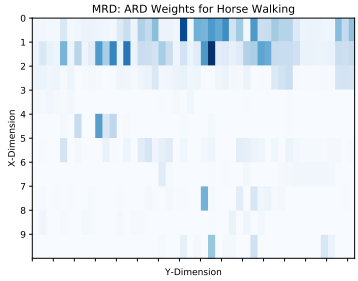
(a) BGP-LVM **(b)** MRD **(c)** DP-GP-LVM **(d)** BGP-LVM

(e) DP-GP-LVM X_1-X_3 **(f)** DP-GP-LVM X_2-X_4 **(g)** DP-GP-LVM Inferred Groups

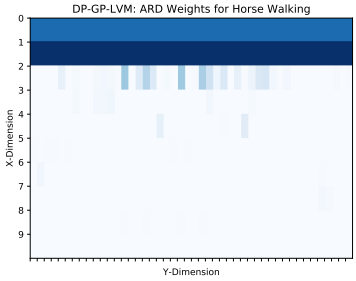
Figure 4. Results for PoseTrack dataset with four individuals. (a)-(c) the ARD weights found for the three models. BGP-LVM increases to three latent dimensions but without the independence structure captured by DP-GP-LVM. (d) and (e) show the smooth, interpretable manifolds from DP-GP-LVM with subtle structure that extends into the fourth latent dimension (f). (g) Here we show the learned groups (posterior over $\mathbf{Z}$) as the color of the points overlaid on the first frame of the image. The path traces out the future motion. We observe the separation of the clusters into translation and periodic motions.



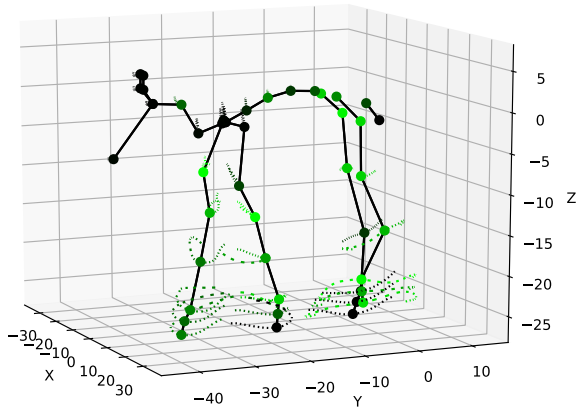
(a) BGP-LVM



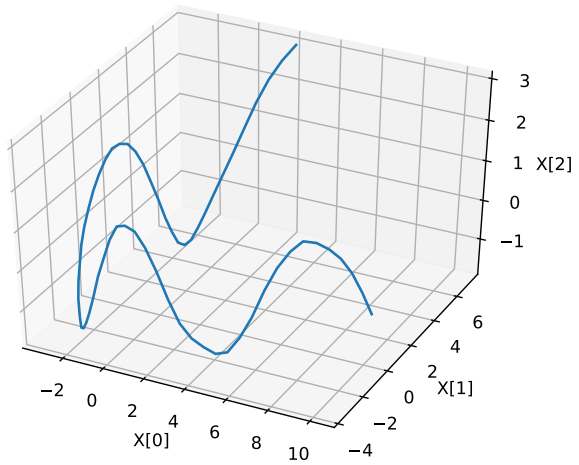
(b) MRD



(c) DP-GP-LVM

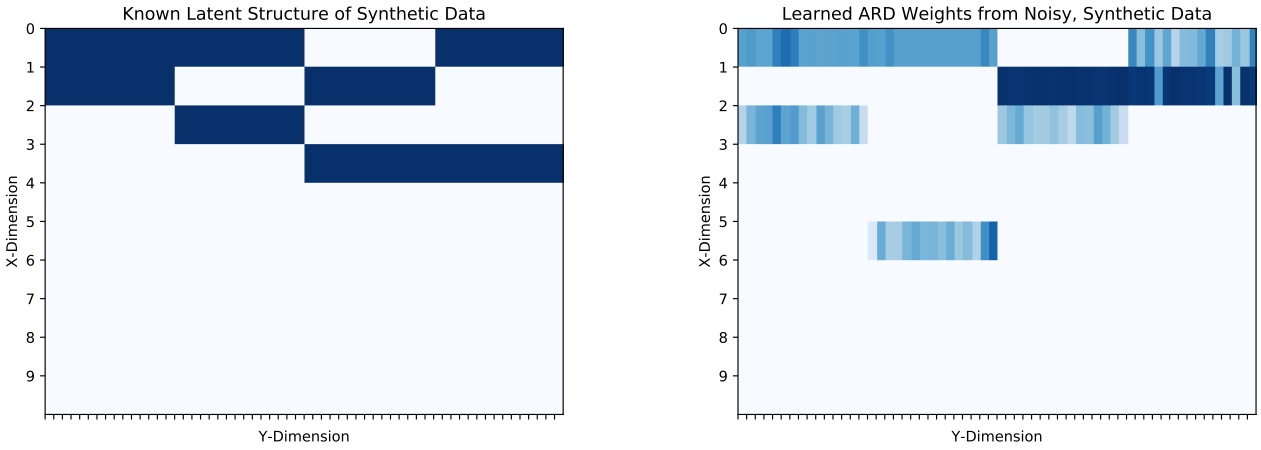


(d) Effect of X_3



(e) DP-GP-LVM X_1-X_3

Figure 5. Horse Walking Dataset. (a)-(c) ARD weights for the three models. BGP-LVM can model the data well with two latent dimensions while MRD is unable to learn a reasonable factorization of the latent space. DP-GP-LVM uses three dimensions with the first two shared and the third encoding the oscillating motion at certain joints when the horse is walking. (d) The position of the horse at a frame in the middle of the sequence. Each joint has a line showing the position of the joint in past and future frames with the color modulated by the third ARD weight. Joints in the legs are a function of X_3, while the spine and head are not, due to the legs oscillating as the horse is walking. (e) The path in the latent space for DP-GP-LVM is periodic.



(a) Ground Truth

(b) Latent Factorization

Figure 6. DP-GP-LVM latent factorization of a difficult, noisy synthetic data set. Sixty dimensional data were generated from four unique mapping function, each with different shared/private latent variables. Figure (a) shows the ground truth factorization while Figure (b) indicates the inferred latent factorization by DP-GP-LVM where a row indicates the latent dimensions in $\mathbf{X}$ allocated to represent the corresponding observed dimension $\mathbf{Y}$. We observe that the DP-GP-LVM correctly recovers the latent structure (4 distinct functions) underlying the data. Note the actual indices of the latent variables are not important due to the interchangeable characteristic of the DP prior.

References

CMU Graphics Lab. Carnegie mellon university - cmu graphics lab - motion capture library, 2002. URL http://mocap.cs.cmu.edu/.

Damianou, A. C., Titsias, M. K., and Lawrence, N. D. Variational inference for latent variables and uncertain inputs in Gaussian processes. *Journal of Machine Learning Research*, 17(1):1425–1486, January 2016.

Ek, C. H., Rihan, J., Torr, P. H. S., Rogez, G., and Lawrence, N. D. Ambiguity modeling in latent spaces. *Int. Conference on Machine Learning for Multimodal Interaction*, 2008.