Variational Inference for Latent Variable Modelling of Correlation Structure

Mark van der Wilk
University of Cambridge
mv310@cam.ac.uk

Andrew G. Wilson
Carnegie Mellon University
andrewgw@cs.cmu.edu

Carl E. Rasmussen
University of Cambridge
cer54@cam.ac.uk

1 Introduction

Latent variable models have played an important part in unsupervised learning, where the goal is to capture the structure of some complicated observed data in a set of variables that are somehow simpler. PCA or Factor Analysis, for example, models high dimensional data using lower dimensional and uncorrelated latent variables. The value of the latent variable represents some underlying unobserved explanation of the observation. Often, the latent variables can be interpreted as having some kind of meaning, which is useful for exploratory data analysis. Many models follow the same principle of explaining the value of an observation by mapping from a "simple" latent space to complicated observations. Much work has been done to increase the flexibility of these mappings. The GPLVM (Lawrence and Hyvrinen, 2005), for example, uses a non-linear function with a Gaussian process prior to map from the latent space.

Instead of the latent variables determining the *value* of an observation, we could instead let them determine the *correlations* within the data. This would give the features found in the latent space a different meaning, allowing different kinds of features to be captured. For example, in a study of people's preference for drinks, we can imagine someone's ability to perceive the taste "bitter" to cause correlations in their enjoyment of bitter drinks. A latent variable model that explains correlations would be able to distinguish between individuals on the basis of whether or not the enjoyment of bitter drinks are correlated, while factor analysis would not. Additionally, modelling correlations can allow us to assign latent variables to whole data*sets* rather than data points. This allows us to judge the proximity or similarity of datasets according to their covariance structure. Investigating the difference has been termed "contrastive learning" and has been investigated in the context of mixture models by Zou et al. (2013). Our model should be able to expand on this by investigating correlation structure. Here we introduce the Wishart process latent variable model, a latent variable for covariances.

2 Model

In order to model correlations, we start with the Wishart process (Wilson and Ghahramani, 2011), which can be seen as a stochastic process over covariance matrices, indexed by some covariate such as time or space. Wishart processes produce a sequence of correlated covariance matrices and have been applied to the regression of fluctuating correlations between prices of goods in financial markets. We propose to do unsupervised learning by inferring the input, as is done with the GPLVM. In this case, we assume that we (indirectly) observe the covariance matrices from a Wishart process, but not the input that they are correlated by. We then infer the input. This gives us a latent representation where dissimilar covariance matrices are placed in distant regions in the latent space.

2.1 Wishart processes

The Wishart distribution (Wishart, 1928) gives a probability density over D dimensional positive definite matrices. It has two free parameters, the number of degrees of freedom $\nu \geq D$ and the scale parameter $V \in \mathbb{R}^{D \times D}$ (positive definite). Samples from $\mathcal{W}(\nu, I)$ can be drawn by taking the sum of ν outer products of vectors drawn from $\mathcal{N}(0, I_D)$:

$$\mathbf{u}_n \sim \mathcal{N}(0, I_D) \qquad S = \sum_{k=1}^{\nu} \mathbf{u}_k \mathbf{u}_k^\intercal \tag{1}$$

$$\therefore S \sim \mathcal{W}(\nu, I_D) \qquad LSL^\intercal \sim \mathcal{W}\left(\nu, LL^\intercal\right) \tag{2}$$

The Wishart process allows correlated matrices to be drawn by replacing the elements in the $\mathbf{u}_n$ vectors by Gaussian processes. We now have $D \times \nu$ GPs, each with N points. A single covariance matrix is obtained by the outer product of the matrix formed by all the GP values at the nth time point. Each $S(\mathbf{x}_n)$ from the Wishart process will be marginally Wishart distributed.

$$u_{dk}(\mathbf{x}_n) \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}')) \qquad S(\mathbf{x}_n) = \sum_{k=1}^{\nu} \mathbf{u}_k(\mathbf{x}_n)\mathbf{u}_k(\mathbf{x}_n)^\intercal = U_n U_n^\intercal \tag{3}$$

Notation-wise, we collect all the GP values in a large $\mathbb{R}^{N \times D \times \nu}$ tensor. Points or sub-vectors are referred to using a Matlab-like slicing notation. E.g. U_{ndk} refers to th nth time point of the dkth GP, while $U_{:dk}$ referres to the vector of all values of the dkth GP.

2.2 Wishart Process Latent Variable Model

Two modifications to the Wishart process are needed to turn it into the Wishart process latent variable model. Firstly, the inputs $\mathbf{x}_n$ become latent and are given a prior. Secondly, we do not observe the covariance matrices directly, but only through data. Here we shall consider observing Gaussian distributed data with the covariance given by the Wishart process:

$$\mathbf{x}_n \sim \mathcal{N}(0, I_Q) \qquad n \in \{1 \dots N\} \tag{4}$$

$$u_{dk}(\mathbf{x}) \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}')) \qquad d \in \{1 \dots D\}, k \in \{1 \dots \nu\} \tag{5}$$

$$\Sigma_n = U_n U_n^\intercal \tag{6}$$

$$\mathbf{y}_{ni} \sim \mathcal{N}(0, \Sigma_n) \qquad i \in \{1 \dots I_n\} \tag{7}$$

Here we perform inference on a slightly modified model, which is approximately equal to the model above, but allows for a tractable variational inference method. Instead of explicitly representing the covariance matrix Σ_n, we generate the data y_{ni} by multiplying $\epsilon_{ni} \sim \mathcal{N}(0, I_\nu)$ by U_n, as is common when sampling correlated Gaussians. We replace the last two lines with:

$$\boldsymbol{\epsilon}_{ni} \sim \mathcal{N}(0, I_\nu) \tag{8}$$

$$\mathbf{y}_{ni} = \mathcal{N}\left(U_n \boldsymbol{\epsilon}_{ni}, \sigma^2 I_D\right) \qquad i \in \{1 \dots I_n\} \tag{9}$$

This model will have $\mathrm{Cov}[\mathbf{y}_{ni}] = U_n U_n^\intercal + \sigma^2 I_D$, which is approximately marginally Wishart distributed if σ^2 is small and $\nu \geq D$. This formulation also allows us to constrain ourselves to factor analysis style covariance matrices, instead of considering full-rank matrices.

3 Variational inference

We consider the adapted model from the previous section because any variational distribution over $\mathcal{U}$ in the original model would involve taking an expectation over $\log |U_n U_n^\intercal|$. Explicitly representing the $\boldsymbol{\epsilon}_{ni}$s removes this problem. This adapted model is a special case of a Gaussian Process Regression Network (Wilson et al., 2012) for which Nguyen and Bonilla (2013) present a variational inference scheme. Our variational inference scheme is similar, but modified using the sparse variational approximation as in Titsias and Lawrence (2010) to allow integration over the latent Xs.

We derive a lower bound for the log marginal likelihood, as usual. As in Titsias and Lawrence (2010), we augment each GP with inducing points, the input being $\mathbf{z}_n \in \mathbb{R}^Q$, and the output ζ_{dk}. The trick from the GPLVM is to choose a particular variational distribution over $\mathcal{U}$ so that the terms with the problematic K_{XX}^{-1} terms are removed. If we choose $q(u_{dk})$ to be the GP conditional

distribution given ζ_{dk}, we cancel out any terms containing the problematic K_{XX}^{-1}.

$$\begin{aligned}
\log p(\mathcal{Y}) &= \log \int p(\mathcal{Y}|\mathcal{U},\epsilon)p(\epsilon)p(\mathcal{U}|X)p(X)\mathrm{d}X\mathrm{d}\epsilon\mathrm{d}\mathcal{U} \\
&\geq \int q(\epsilon)q(X)q(\zeta)p(\mathcal{U}|\zeta,X,Z)\log\frac{p(\mathcal{Y}|\mathcal{U},\epsilon)p(\epsilon)\cancel{p(\mathcal{U}|\zeta,X,Z)}p(\zeta|Z)p(X)}{q(\epsilon)q(X)q(\zeta)\cancel{p(\mathcal{U}|\zeta,X,Z)}}\mathrm{d}X\mathrm{d}\epsilon\mathrm{d}\mathcal{U}\mathrm{d}\zeta \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{E}\left[\log p(\mathcal{Y}|\mathcal{U},\epsilon)\right]_{q(\epsilon)p(\mathcal{U}|\zeta,X)}\right]_{q(X)} + \log\frac{p(\zeta)}{q(\zeta)}\right]_{q(\zeta)} \\
&\quad - \mathrm{KL}\left(q(\epsilon)||p(\epsilon)\right) - \mathrm{KL}\left(q(X)||p(X)\right) \qquad (10)
\end{aligned}$$

We choose a fully factorised Gaussian distribution for $q(\epsilon)$ and a Gaussian $q(X)$ which is factored over each latent point, with a diagonal covariance. The optimal form for $q(\zeta)$ will be derived and will turn out to be Gaussian. This gives the overall variational distribution:

$$\begin{aligned}
q(\epsilon,\mathcal{U}) &= \prod_{nik} q(\epsilon_{nik}) \cdot \prod_{dk} p(U_{:dk}|\zeta_{:dk},X) \qquad (11) \\
&= \left[\prod_{n=1}^{N}\prod_{i=1}^{I_n}\prod_{k=1}^{\nu}\mathcal{N}\left(\epsilon_{nik};\overset{\epsilon}{\mu}_{nik};\overset{\epsilon}{\sigma}{}^2_{nik}\right)\right]\cdot\left[\prod_{d=1}^{D}\prod_{k=1}^{\nu}\mathcal{N}\left(\mathbf{u}_{dk};M_{:dk},\Sigma\right)\right] \qquad (12) \\
q(X) &= \prod_{n}\mathcal{N}\left(x_n;\mathbf{m},\mathrm{diag}\left(\mathbf{v}\right)\right) \qquad (13)
\end{aligned}$$

M_{ndk} and Σ contain the GP conditional mean and covariance respectively (the covariance depends on the inputs only, which are the same for all $D \times \nu$ GPs).

$$M_{:dk} = K_{XZ}K_{ZZ}^{-1}\zeta_{:dk} \qquad \Sigma = K_{XX} - K_{XZ}K_{ZZ}^{-1}K_{ZX} \qquad \sigma_n^2 = \Sigma_{nn} \qquad (14)$$

The expectations are now, in principle, all Gaussian expectations over quadratic forms, which are analytically tractable.

$$\begin{aligned}
\mathbb{E}\left[\log p(\mathcal{Y}|\mathcal{U},\epsilon)\right]_{q(\epsilon,\mathcal{U})} &= \int\prod_{nik}q(\epsilon_{nik})\prod_{dk}p(U_{:dk}|\zeta_{:dk},X)\left[\log\prod_{ni}\mathcal{N}\left(\mathbf{y}_{ni};U_n\boldsymbol{\epsilon}_{ni},\sigma^2 I\right)\right]\mathrm{d}(\mathcal{U},\epsilon) \\
&= -\frac{DT}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{ni}\Bigg[\left(\mathbf{y}_{ni} - M_n\overset{\epsilon}{\boldsymbol{\mu}}_{ni}\right)^{\mathsf{T}}\left(\mathbf{y}_{ni} - M_n\overset{\epsilon}{\boldsymbol{\mu}}_{ni}\right) \\
&\qquad + D\sigma_n^2\overset{\epsilon}{\boldsymbol{\mu}}{}_{ni}^{\mathsf{T}}\overset{\epsilon}{\boldsymbol{\mu}}_{ni} + \mathrm{Tr}\left(M_n^{\mathsf{T}}M_n\overset{\epsilon}{\Sigma}_{ni}\right) + D\sigma_n^2\sum_k\overset{\epsilon}{\Sigma}_{nik}\Bigg] \qquad (15)
\end{aligned}$$

The expectation over $q(X)$ is done by unfolding M_n using Kronecker products ($\boldsymbol{\zeta} = \mathrm{vec}\,\zeta_{:::}$). Due to linearity, the integral reduces to taking expectations of the kernel matrices with respect to $q(\mathbf{x}_n)$, denoted using angled brackets[1].

$$\begin{aligned}
\mathbf{y}_{ni}^{\mathsf{T}}\left\langle M_n\right\rangle\overset{\epsilon}{\boldsymbol{\mu}}_{ni} &= \left(\overset{\epsilon}{\boldsymbol{\mu}}_{ni}\otimes\left(I\otimes K_{ZZ}^{-1}\left\langle\mathbf{k}_{Zx}\right\rangle\right)\mathbf{y}_{ni}\right)^{\mathsf{T}}\boldsymbol{\zeta} \qquad (16) \\
\overset{\epsilon}{\boldsymbol{\mu}}{}_{ni}^{\mathsf{T}}\left\langle M_n^{\mathsf{T}}M_n\right\rangle\overset{\epsilon}{\boldsymbol{\mu}}_{ni} &= \boldsymbol{\zeta}^{\mathsf{T}}\left(\overset{\epsilon}{\boldsymbol{\mu}}_{ni}\overset{\epsilon}{\boldsymbol{\mu}}{}_{ni}^{\mathsf{T}}\otimes\left(I\otimes K_{ZZ}^{-1}\left\langle\mathbf{k}_{Zx}\mathbf{k}_{Zx}^{\mathsf{T}}\right\rangle K_{ZZ}^{-1}\right)\right)\boldsymbol{\zeta} \qquad (17) \\
\mathrm{Tr}\left(\left\langle M_n^{\mathsf{T}}M_n\right\rangle\overset{\epsilon}{\Sigma}_{ni}\right) &= \boldsymbol{\zeta}^{\mathsf{T}}\left(\overset{\epsilon}{\Sigma}_{ni}\otimes\left(I\otimes K_{ZZ}^{-1}\left\langle\mathbf{k}_{Zx}\mathbf{k}_{Zx}^{\mathsf{T}}\right\rangle K_{ZZ}^{-1}\right)\right)\boldsymbol{\zeta} \qquad (18) \\
\left\langle\sigma_n^2\right\rangle &= \left\langle k_{xx}\right\rangle - \mathrm{Tr}\left(K_{ZZ}^{-1}\left\langle\mathbf{k}_{Zx}\mathbf{k}_{Zx}^{\mathsf{T}}\right\rangle\right) \qquad (19)
\end{aligned}$$

Next, the optimal form of $q(\boldsymbol{\zeta})$ can be derived by taking the functional derivative of our lower bound, which gives us:

$$\log q(\boldsymbol{\zeta}) = \mathbb{E}\left[p\left(\mathcal{Y}|\mathcal{U},\epsilon\right)\right]_{q(\mathcal{U},\epsilon,X)} + \log p(\zeta) + \lambda - 1 = -\frac{1}{2}(\boldsymbol{\zeta}-\boldsymbol{\mu}_\zeta)^{\mathsf{T}}\Sigma_\zeta^{-1}(\boldsymbol{\zeta}-\boldsymbol{\mu}_\zeta) + c \qquad (20)$$

[1]These are referred to as the Ψ statistics in Titsias and Lawrence (2010). Detailed derivations can be found in Gal and van der Wilk (2014).

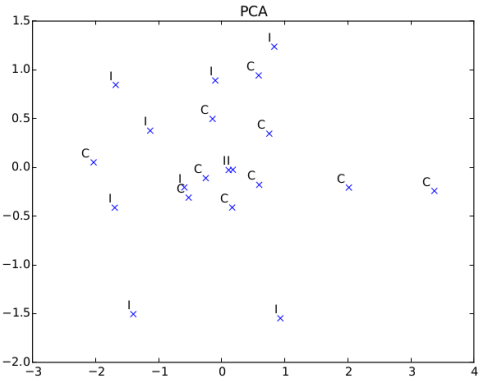

Figure 1: Visualisation using PCA. See figure 3 for a description of the labels.

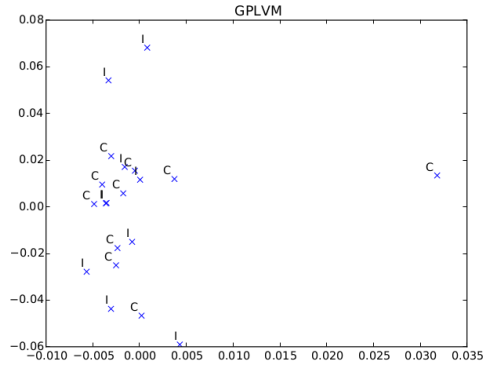

Figure 2: Visualisation using GPLVM. See figure 3 for a description of the labels.

$$\Sigma_\zeta^{-1} = I \otimes K_{ZZ}^{-1} + \frac{1}{\sigma^2} \sum_{ni} \left(\left(\overset{\epsilon}{\boldsymbol{\mu}}_{ni} \overset{\epsilon}{\boldsymbol{\mu}}_{ni}^{\mathsf{T}} + \overset{\epsilon}{\Sigma}_{ni} \right) \otimes \left(I \otimes K_{ZZ}^{-1} \left\langle \mathbf{k}_{Zx} \mathbf{k}_{Zx}^{\mathsf{T}} \right\rangle K_{ZZ}^{-1} \right) \right) \tag{21}$$

$$\boldsymbol{\mu}_\zeta = \Sigma_\zeta \frac{1}{\sigma^2} \sum_{ni} \left(\overset{\epsilon}{\boldsymbol{\mu}}_{ni} \otimes \left(I \otimes K_{ZZ}^{-1} \left\langle \mathbf{k}_{Zx} \right\rangle \right) \mathbf{y}_{ni} \right) \tag{22}$$

Finally, we take the final expectation over $q(\zeta)$, giving our lower bound:

$$\begin{aligned} \mathcal{L}\left(\mathcal{Y}\right) = &-\frac{DT}{2} \log 2\pi\sigma^2 - \frac{1}{2} \left(\left(\frac{1}{\sigma^2} \sum_{ni} \mathbf{y}_{ni}^{\mathsf{T}} \mathbf{y}_{ni} \right) - \boldsymbol{\mu}_\zeta^{\mathsf{T}} \Sigma_\zeta^{-1} \boldsymbol{\mu}_\zeta + \mathrm{Tr}\left(I_{\nu DM} \right) \right) + \mathcal{H}\left(q(\boldsymbol{\zeta}) \right) \\ &+ \frac{D}{2\sigma^2} \sum_{ni} \left\langle \sigma_n^2 \right\rangle \left[\overset{\epsilon}{\boldsymbol{\mu}}_{ni}^{\mathsf{T}} \overset{\epsilon}{\boldsymbol{\mu}}_{ni} + \sum_k \overset{\epsilon}{\Sigma}_{nik} \right] + \mathrm{KL}\left(q(\epsilon) || p(\epsilon) \right) + \mathrm{KL}\left(q(X) || p(X) \right) \end{aligned} \tag{23}$$

4 Preliminary results

We first show how the latent space of the WPLVM extracts different features from the data, compared to existing methods like PCA or the GPLVM. We generated data from two 3D Gaussians[2], one isotropic and one with a strong correlation. The latent variable of interest determines which Gaussian a data point is generated from. PCA and the GPLVM (figures 1 & 2) can not separate the points based on their correlation, while the WPLVM (figure 3) manages to cluster points from the correlated distribution together while forcing away those from the isotropic distribution.

Alternatively, the WPLVM can be used to visualise the difference in correlations between datasets. For this case we generated 15 groups of points from Gaussians. The first 5 get have an increasing correlation in one direction, the following 5 in another and the final 5 have an increasing correlation in both directions. Figure 4 shows that the WPLVM finds a representation where datasets are ordered by increasing correlation and by the direction of correlation.

5 Conclusions & Future work

We argue that latent variable models can be used in different ways to capture interesting properties of data such as correlation, which current models like Factor Analysis or the GPLVM do not. We show using toy datasets that the WPLVM can cluster data based on its internal correlations. Additionally, we give an example of the WPLVM finding a useful latent space representing the difference in correlations between several datasets.

The WPLVM as presented here can be extended in several ways. If we maintain a GP prior over the ϵ variables, we are essentially inferring the inputs to a GPRN. This would allow us to do contrastive learning using both the mean and covariance of datasets. There is also still work to be done on initialising the model properly in order to improve visualisations like in figure 3.

[2]Giving 6 degrees of freedom in the covariance matrix.

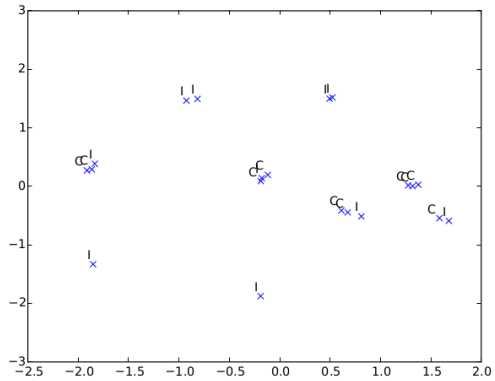

Figure 3: Visualisation using WPLVM. The labels "I" and "C" indicate that points come from an isotropic or correlated Gaussian, respectively. Note how the WPLVM forces most C clusters away from the I clusters.

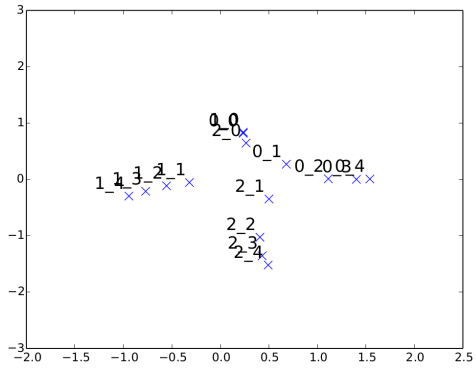

Figure 4: Latent space produced by the WPLVM. The first number in the label indicates which direction the correlation grows in (0 for one direction, 1 for another and 2 for both), while the second number indicates the strength (0 for no correlation).

References

Yarin Gal and Mark van der Wilk. Variational inference in the Gaussian process latent variable model and sparse GP regression – a gentle tutorial. *arXiv:1402.1412*, 2014.

Neil Lawrence and Aapo Hyvrinen. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *Journal of Machine Learning Research*, 6:1783–1816, 2005.

T. V. Nguyen and E. V. Bonilla. Efficient variational inference for gaussian process regression networks. *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS) 2013*, 2013.

M. K. Titsias and N. D. Lawrence. Bayesian gaussian process latent variable model. *Journal of Machine Learning Research - Proceedings Track*, 9:844–851, 2010.

Andrew G. Wilson and Zoubin Ghahramani. Generalised wishart processes. In Fabio Cozman and Avi Pfeffer, editors, *Proceedings of the Twenty-Seventh Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-11)*, Barcelona, Spain, 2011. AUAI Press.

Andrew G. Wilson, David A. Knowles, and Zoubin Ghahramani. Gaussian process regression networks. In J. Langford and J. Pineau, editors, *Proceedings of the 29th International Conference on Machine Learning (ICML)*, Edinburgh, June 2012. Omnipress.

John Wishart. The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, 20A(1/2):pp. 32–52, 1928. ISSN 00063444. URL `http://www.jstor.org/stable/2331939`.

James Y. Zou, Daniel Hsu, David C. Parkes, and Ryan Prescott Adams. Contrastive learning using spectral methods. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013.*, pages 2238–2246, 2013. URL `http://papers.nips.cc/paper/5007-contrastive-learning-using-spectral-methods`.