Latent Correlation Gaussian Processes

Sami Remes Markus Heinonen Samuel Kaski

Helsinki Institute of Information Technology HIIT,
Department of Computer Science, Aalto University

February 28, 2017

Abstract

We introduce a novel kernel that models input-dependent couplings across multiple latent processes. The pairwise kernel measures covariance both along inputs and across different latent signals in a mutually-dependent fashion. The latent correlation Gaussian process (LCGP) model combines these non-stationary latent components into multiple outputs by an input-dependent mixing matrix. Probit classification and support for multiple observation sets are derived by Variational Bayesian inference. Results on several datasets indicate that the LCGP model can recover the correlations between latent signals while simultaneously achieving state-of-the-art performance. We highlight the latent covariances with an EEG classification dataset where latent brain processes and their couplings simultaneously emerge from the model.

1 Introduction

Gaussian processes (GP) are Bayesian non-parametric models that explicitly characterize the uncertainty in the learned model by describing distributions over functions (Rasmussen and Williams, 2006). These models assume a prior over functions, and subsequently the function posterior given the data can be derived. The prior covariance plays the key roles of both regularising the model by determining its smoothness properties, and characterising how the underlying function varies in the input space.

Recently, there has been interest in deriving non-stationary covariance kernels, where the general signal variances or the intrinsic kernel parameters – such as the lengthscales in the squared exponential or Matérn kernels – are input-dependent (Adams and Stegle, 2008; Gibbs, 1997; Heinonen et al., 2016; Paciorek and Schervish, 2004; Tolvanen et al., 2014). For instance, in geostatistical applications, a non-stationary kernel can both model a difference in the covariance along or across geological formations (Goovaerts, 1997). Input-dependent, heteroscedastic noise models have also been studied in single-task (Goldberg et al., 1997; Kersting et al., 2007; Lazaro-Gredilla and Titsias, 2011;

Le et al., 2005; Quadrianto et al., 2009; Wang and Neal, 2012) and in multi-task settings (Rakitsch et al., 2013).

In multi-task learning Gaussian processes are utilized by modeling the output covariances between possibly several latent functions (Álvarez et al., 2012; Alvarez et al., 2010; Bonilla et al., 2007; Yu et al., 2005). In latent function models[1] the outputs are linear combinations of multiple underlying latent functions (Schmidt, 2009; Teh and Seeger, 2005). In Gaussian Process Regression Networks (GPRN) the mixing coefficients of multiple independent latent signals are input-dependent Gaussian processes as well, leading to a general multi-task framework that adaptively combines latent signals into outputs along the input space (Wilson et al., 2012).

In this paper we extend the GPRN framework by a non-stationary covariance function for the latent signals. We explicitly model the input-dependent couplings between the latent processes. To such end, the main contribution of this paper is to introduce a structured *latent correlation kernel* (LCK) that combines a covariance structure between the latent signals that depends on the inputs, with an input kernel that depends on the latent signals. The signal and input kernels are interdependent, conditional on each other. The LCK generalizes Wishart processes (Wilson and Ghahramani, 2011) into cross-covariances for input-dependent correlation structure, and a non-stationary Gaussian kernel (Gibbs, 1997) for measuring input-space correlations at specific latent signals.

Furthermore, the proposed latent correlation Gaussian process (LCGP) incorporates multiple latent signals that are linearly combined into multiple outputs in an input-dependent fashion. The latent signals have a structured LCK model that leads to non-stationary signal variances. We account for both regression, and Probit-based classification. Finally, the model is extended for multiple observation sets, where each observation set is modeled by a separate latent model with shared latent correlations. In such a model, the latent correlations effectively regularize the latent models of each observation. Variational Bayes approximate inference with whitened gradients is derived for scalable implementation.

We highlight the model with several datasets where interesting latent signal covariance models emerge, while retaining or improving the state-of-the-art regression and classification performance. Multi-observation classification is demonstrated on EEG data from a large set of scalp measurements from several subjects, where the model is able to learn the covariance model between the underlying brain processes. In simulation studies, we show that our model is capable of accurately learning the latent variable correlations.

2 Latent Correlation Gaussian Process

We consider M-dimensional observations $\mathbf{y}(x) \in \mathbb{R}^M$ over N data points $(x_1, \ldots, x_N)$. We denote vectors with boldface symbols, matrices with capital symbols and

[1] Latent models are coined as *linear models of coregionalisation* (LCM) in geostatistics literature (Goovaerts, 1997)

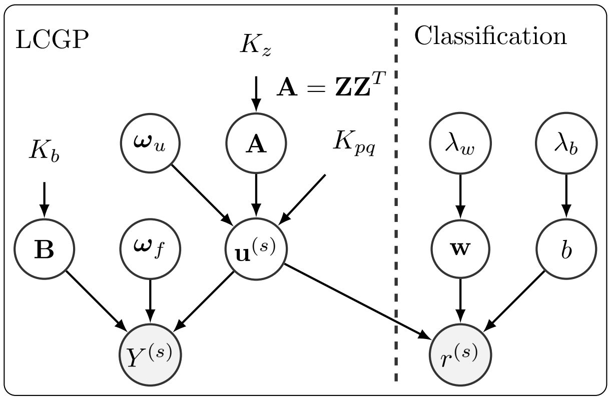


Figure 1: Graphical model of the LCGP.

block matrices with boldface capital symbols.

2.1 Multi-output regression

Following Wilson et al. (2012), we model the M-dimensional outputs as an input-dependent mixture of Q latent signals $\mathbf{u}(x) \in \mathbb{R}^Q$,

$$\begin{aligned} \mathbf{y}(x) &= \mathbf{f}(x) + \mathbf{e} \\ &= B(x)\big(\mathbf{u}(x) + \boldsymbol{\varepsilon}\big) + \mathbf{e}, \end{aligned} \tag{1}$$

where $\mathbf{e} = \mathbf{e}(x)$ is zero-mean M-dimensional observation noise and $\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}(x)$ is zero-mean Q-dimensional latent noise

$$\begin{aligned} \mathbf{e} &\sim \mathcal{N}(\mathbf{0}, \omega_f^{-1}\mathbf{I}), \quad \omega_f \sim \text{Gamma}(\alpha_f, \beta_f) \\ \boldsymbol{\varepsilon} &\sim \mathcal{N}(\mathbf{0}, \omega_u^{-1}\mathbf{I}), \quad \omega_u \sim \text{Gamma}(\alpha_u, \beta_u). \end{aligned}$$

The mixing matrix $B(x)$ is an $M \times Q$ matrix of independent Gaussian processes over outputs i and latent signals p,

$$B_{ip}(x) \sim \mathcal{GP}(0, K_b(x, x')). \tag{2}$$

The kernel $K_b(x, x')$ between two input points x and x' determines how mixing of latent signals into outputs evolves along the input space. For instance, with temporal data the mixing matrix allows time-dependent linear combinations of the outputs.

In the following, we seek to infer the mixing matrix $B(x)$, noise precisions ω_f and ω_u, latent signals $\mathbf{u}(x)$ and their underlying input-dependent covariance structure $\mathbf{cov}[\mathbf{u}(x), \mathbf{u}(x')] = C_{pq}(x, x')$ (see Figure 1).

2.2 Latent correlation kernel

The latent signals $u_p(x)$ are functions of the signal p and input x. We propose to encode the latent signals $\mathbf{u}(x)$ as *mutually dependent* Gaussian processes over pairs of inputs x,x' and signals p,q,

$$u_p(x) \sim \mathcal{GP}\left(0, A_{xx'}(p,q)K_{pq}(x,x')\right), \tag{3}$$

such that the joint covariance $\mathbf{cov}[u_p(x), u_q(x')] = A_{xx'}(p,q)K_{pq}(x,x')$ is a product of signal and input similarities. Both similarities depend on each other to produce a non-stationary joint covariance.

The pairwise *latent correlation kernel* $C_{pq}(x,x') = A_{xx'}(p,q)K_{pq}(x,x')$ encodes a rich similarity between input x of latent signal p and x' of latent signal q as the product of the two conditional kernels. The kernel $A_{xx'}(p,q)$ encodes signal similarity between inputs x and x', while the kernel $K_{pq}(x,x')$ denotes input similarity at latent signals p and q. Since the two kernels depend on each other, a simple model such as Kronecker kernel product (Stegle et al., 2011) is not suitable. Both kernels can be interpreted as cross-covariances.

For instance, in EEG data the kernels could signify correlations $A_{tt'}(p,q)$ between latent brain processes p and q at two time points t and t', while $K_{pq}(t,t')$ is a smooth temporal kernel that connects events that occur at similar time points. In geospatial applications, the correlations $A_{\mathbf{xx'}}(p,q)$ can encode similarity between two latent ore functions p and q at two locations $\mathbf{x},\mathbf{x'} \in \mathbb{R}^2$, for instance between cadmium and zinc concentrations (Goovaerts, 1997). The location kernel $K_{pq}(\mathbf{x},\mathbf{x'})$ could encode a smooth spatial proximity.

We start forming the joint kernel by considering a non-stationary Gaussian kernel for the inputs x Gibbs (1997),

$$K_{pq}(x,x') = \sqrt{\frac{2\ell_p\ell_q}{\ell_p^2+\ell_q^2}} \exp\left(-\frac{(x-x')^2}{\ell_p^2+\ell_q^2}\right), \tag{4}$$

which encodes specific lengthscales $l_1, \ldots, l_Q$ for each latent signal. The kernel is a smooth Gaussian similarity function between inputs associated with these lengthscales. The kernel reduces to standard Gaussian kernel for $l_p = l_q$.

We base our construction of the mutually dependent covariance structure $A_{xx'}(p,q)$ on Wishart processes. A Generalized Wishart Process (GWP) prior on a covariance matrix, that depends on a single variable x, is (Wilson and Ghahramani, 2011)

$$A(x) = \sum_{r=1}^{\nu} L\mathbf{z}_r(x)\mathbf{z}_r(x)^T L^T \sim \mathcal{GWP}(V, \nu, K_z),$$

where $V = LL^T$ and all $z_{pr}(x) \sim \mathcal{GP}(0, K_z(x,x'))$ are independent Gaussian processes for $p = 1, \ldots, Q$ and $r = 1, \ldots, \nu$. The kernel K_z determines the change of $A(x)$ in the input space. From this formulation we define our joint kernel, such that we preserve the GWP marginal for $A(x)$ by extending the GWP into *cross-covariances* of two variables, as

$$A_{xx'}(p,q) = \mathbf{z}_p(x)^T\mathbf{z}_q(x'), \tag{5}$$

where we have for each element of $\mathbf{z}_p(x) \in \mathbb{R}^\nu$ a GP prior. With this choice the prior expectation of the covariance is the identity matrix.

The resulting covariance of $u_p(x)$ is then a product of covariance between inputs at signals p and q, and a covariance between signals at inputs x and x'. This covariance can be seen marginally from two perspectives,

$$\begin{aligned} \mathbf{u}_p &\sim \mathcal{N}_N(\mathbf{0}, A(p,p) \circ K_{pp}) \\ \mathbf{u}(x) &\sim \mathcal{N}_Q(\mathbf{0}, A(x) \circ P), \end{aligned}$$

where $\mathbf{u}_p \in \mathbb{R}^N$ is a single latent signal that follows a Normal distribution weighted by variances $A(p,p)$, and $\mathbf{u}(x) \in \mathbb{R}^Q$ contains all Q latent signals at input x and follows a Normal distribution with generalized Wishart process prior scaled by the matrix $P_{pq} = \sqrt{\frac{2\ell_p \ell_q}{\ell_p^2 + \ell_q^2}}$. The element-wise, or Hadamard, product of x and y is denoted by $x \circ y$.

The joint covariance over the concatenated column vector of all latent signals $\mathbf{u} \in \mathbb{R}^{QN}$ is a block matrix

$$\begin{aligned} \mathbf{cov}(\mathbf{u}, \mathbf{u}) &= \left(Z_i Z_j^T \circ K_{ij} \right)_{i,j=1}^N + \mathbf{\Omega}_u \\ &= \mathbf{Z}\mathbf{Z}^T \circ \mathbf{K}_Q + \mathbf{\Omega}_u \\ &= \mathbf{A}_Q \circ \mathbf{K}_Q + \mathbf{\Omega}_u \\ &= \mathbf{C}_Q + \mathbf{\Omega}_u = \mathbf{K}_u, \end{aligned}$$

where Z_i is a $(Q \times \nu)$-sized matrix, and $\mathbf{Z} = (Z_1, \ldots, Z_N)^T$ is a $(NQ \times \nu)$-sized concatenation of N blocks of Z_i. Thus the block matrix $\mathbf{A}_Q = \mathbf{Z}\mathbf{Z}^T$ contains $(N \times N)$ covariance blocks of size $(Q \times Q)$. The kernel $K_{ij} = (K_{pq}(x_i, x_j))_{p,q=1}^Q$ is a $(Q \times Q)$ matrix of nonstationary kernels between inputs, the block matrix $\mathbf{K}_Q = (K_{ij})_{i,j=1}^N$ contains $(N \times N)$ blocks of kernel values, and finally the noise matrix is $\mathbf{\Omega}_u = \omega_u^{-1} \mathbf{I}_{QN}$, introducing the latent noise directly into the covariance of the $\mathbf{u}$'s. The resulting joint input-output covariance $\mathbf{C}_Q = \mathbf{A}_Q \circ \mathbf{K}_Q$ consists of $N \times N$ block matrices of size $Q \times Q$.

The kernel matrix $\mathbf{A}_Q = \mathbf{Z}\mathbf{Z}^T$ is positive semi-definite (PSD) as an outer product, and the non-stationary Gaussian kernel is PSD as well (Gibbs, 1997). The Hadamard product $\mathbf{C}_Q$ retains this property.

The proposed latent correlation Gaussian process (LCGP) model is a flexible Bayesian regression model that simultaneously infers the latent signals and their mixings matching the output processes, while learning the underlying correlation structure kernel of the latent space. The latent correlations are parameterized by two terms that characterize the input and signal similarities with Gaussian and Wishart functions, respectively. A key feature of the model is the ability adaptively couple and decouple latent processes along the input space.

3 Classification with multiple observations

We further suppose that we have S observations or samples $\mathbf{y}^{(s)}(x)$ associated with a class label, or response $r^{(s)}$, and assume that all these observations share their latent space. We then learn separate latent functions $\mathbf{u}^{(s)}$ for each sample, while keeping the mixing model $B(x)$, latent correlations $\mathbf{A}(x)$ and $\mathbf{K}_Q$, and the noise precisions ω_f and ω_u shared. The noiseless sample is then reconstructed as

$$\mathbf{f}^{(s)}(x) = B(x)\mathbf{u}^{(s)}(x),$$

which results in the same likelihood as in eq. (1).

We build a classifier in the latent signal space as a Probit classification model over all latent signals $\mathbf{w}^T\mathbf{u}^{(s)}$ with Gaussian-Gamma priors, where $\mathbf{w} \in \mathbb{R}^{NQ}$ is a concatenated column vector of linear weights $\mathbf{w}_p \in \mathbb{R}^N$ for the Q latent signals. The classifier is then

$$\begin{aligned} r^{(s)} \mid \mathbf{w}, \mathbf{u}^{(s)} &\sim \text{Bernoulli}(\Phi(\mathbf{w}^T\mathbf{u}^{(s)} + b)), \\ \mathbf{w}_p \mid \lambda_w &\sim \mathcal{N}(0, \lambda_w^{-1}),\ \lambda_w \sim \text{Gamma}(\alpha_w, \beta_w) \\ b \mid \lambda_b &\sim \mathcal{N}(0, \lambda_b^{-1}),\ \lambda_b \sim \text{Gamma}(\alpha_b, \beta_b), \end{aligned} \tag{6}$$

where we index the observations with s, and $\mathbf{w}$ and b are the classifier weights and bias, respectively. We additionally assume, for notational clarity, that all data are observed at the same input points $x_1, \ldots, x_N$. To accommodate varying input points a GP prior for $\mathbf{w}$ could be defined.

Essentially, our model now has two likelihoods for the two types of data, one defined for the output data in eq. (1) and one for the class labels related to the outputs in eq. (6).

4 Inference

4.1 Variational Bayes

For inference in our Bayesian model we adopt the Variational Bayesian approach (Attias, 1999), which is based on maximising a lower bound on the log marginal likelihood of the data with respect to a distribution $q(\Theta)$, where Θ represents all model parameters. The lower bound is of an easier form than the intractable true posterior distribution $p(\Theta|\mathcal{D})$, where $\mathcal{D} = (Y^{(s)}, r^{(s)})_{s=1}^S$ and $Y^{(s)} \in \mathbb{R}^{M \times N}$. The lower bound is obtained by Jensen's inequality as

$$\begin{aligned} \log p(\mathcal{D}) &= \log \int q(\Theta) \frac{p(\mathcal{D}, \Theta)}{q(\Theta)} d\Theta \\ &\geq \int q(\Theta) \log \frac{p(\mathcal{D}, \Theta)}{q(\Theta)} d\Theta = \text{Ł}(q). \end{aligned}$$

Typically, a factorised approximation

$$q(\Theta) = \prod_i q(\theta_i)$$

is used, where θ_i are some disjoint subsets of the variables Θ. It can be shown that the optimal solution that maximizes $\text{Ł}(q)$ is

$$q(\theta_i) \propto \exp(\langle \log p(\Theta, \mathcal{D}) \rangle_{\theta_{-i}}),$$

in which the expectation is taken with respect to all variables except θ_i. The VB algorithm consists of iterating through updating each factor $q(\theta_i)$.

4.2 Derivation of classifier with multiple observations

We employ the following factorization

$$\begin{aligned} q(\Theta) &= \prod_s q(\mathbf{u}^{(s)}) q(h^{(s)}) \prod_m q(\mathbf{B}_m) q(\omega_f) q(\mathbf{w}, b) \\ &\times q(\lambda_w) q(\lambda_b) q(\mathbf{Z}), \end{aligned}$$

where $q(\mathbf{B}_m)$ factorizes the mixing matrix $\mathbf{B}$ row-wise. Most factors have standard distributions:

$$\begin{aligned}
q(\mathbf{u}^{(s)}) &= \mathcal{N}(\mathbf{u}^{(s)} \mid \boldsymbol{\mu}_u^{(s)}, \boldsymbol{\Sigma}_u) \\
\boldsymbol{\Sigma}_u^{-1} &= \mathbf{K}_u^{-1} + \langle \omega_f \rangle \left\langle \mathbf{B}^T \mathbf{B} \right\rangle + \left\langle \mathbf{w}\mathbf{w}^T \right\rangle \\
\boldsymbol{\mu}_u^{(s)} &= \boldsymbol{\Sigma}_u \left(\left[\left\langle h^{(s)} \right\rangle - \langle b \rangle \right] \langle \mathbf{w} \rangle + \langle \omega_f \rangle \left\langle \mathbf{B}^T \right\rangle \mathbf{y}^{(s)} \right) \\
q(\mathbf{B}_m) &= \mathcal{N}(\mathbf{B}_m \mid \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_b) \\
\boldsymbol{\Sigma}_b^{-1} &= \mathbf{K}_b^{-1} + \langle \omega_f \rangle \sum_s \left\langle \tilde{\mathbf{u}}^{(s)T} \tilde{\mathbf{u}}^{(s)} \right\rangle \\
\boldsymbol{\mu}_m &= \boldsymbol{\Sigma}_b \langle \omega_f \rangle \sum_s \left\langle \tilde{\mathbf{u}}^{(s)T} \mathbf{y}_m^{(s)} \right\rangle \\
q(\mathbf{w}, b) &= \mathcal{N}\left(\begin{pmatrix} \mathbf{w} \\ b \end{pmatrix} \mid \boldsymbol{\mu}_{w,b}, \boldsymbol{\Sigma}_{w,b} \right) \\
\boldsymbol{\Sigma}_{w,b}^{-1} &= \begin{pmatrix} \left\langle \mathbf{u}\mathbf{u}^T \right\rangle + \operatorname{diag} \langle \lambda_w \rangle & \langle \mathbf{u} \rangle \mathbf{1} \\ \mathbf{1}^T \langle \mathbf{u} \rangle^T & S + \langle \lambda_b \rangle \end{pmatrix} \\
\boldsymbol{\mu}_{w,b} &= \boldsymbol{\Sigma}_{w,b} \begin{pmatrix} \langle \mathbf{u} \rangle \langle \mathbf{h} \rangle \\ \mathbf{1}^T \langle \mathbf{h} \rangle \end{pmatrix} \\
q(\lambda_w) &= \text{Gamma}(a_w, b_w) \\
a_w &= \alpha_w + \frac{1}{2} NQ \\
b_w &= \beta_w + \frac{1}{2} \left\langle ||\mathbf{w}||^2 \right\rangle \\
q(\omega_f) &= \text{Gamma}(a_{\omega_f}, b_{\omega_f}) \\
a_{\omega_f} &= \alpha_f + \frac{1}{2} NMS \\
b_{\omega_f} &= \beta_f + \frac{1}{2} \left\langle ||\mathbf{y} - \mathbf{B}\mathbf{u}||_2^2 \right\rangle,
\end{aligned}$$

where $\tilde{\mathbf{u}}$ collects $\mathbf{u}(x_i)$ into a block diagonal matrix. The $q(\lambda_b)$ is analogous to $q(\lambda_w)$, replacing only $\mathbf{w}$ by b.

We introduced auxiliary variables h to make the variational inference tractable for Probit classification (Albert and Chib, 1993),

$$h \mid \mathbf{w}, \mathbf{u} \sim \mathcal{N}(\mathbf{w}^T \mathbf{u} + b, 1). \tag{7}$$

Class labels are then generated based on h as

$$r \mid h \sim \delta(rh > 0) \quad \text{with} \quad r \in \{-1, 1\}. \tag{8}$$

Integrating out h recovers the Probit likelihood

$$p(r|\mathbf{w}, \mathbf{u}) = \text{Bernoulli}(r|\Phi(\mathbf{w}^T \mathbf{u} + b)). \tag{9}$$

Factors $q(h)$ are truncated Gaussian (Albert and Chib, 1993), which has analytical formulas for first and second moments. The approximate posterior is given by

$$\begin{aligned} q(h^{(s)}) &= \begin{cases} \mathcal{TN}_{[-\infty,0]}(h^{(s)} \mid g^{(s)}, 1), & \text{if } r^{(s)} = -1 \\ \mathcal{TN}_{[0,\infty]}(h^{(s)} \mid g^{(s)}, 1), & \text{if } r^{(s)} = +1 \end{cases} \\ g^{(s)} &= \left\langle \mathbf{w}^T \right\rangle \left\langle \mathbf{u}^{(s)} \right\rangle + \langle b \rangle . \end{aligned}$$

The final term $q(\mathbf{Z})$ in the factorisation is updated by optimising the variational lower bound with gradients. The relevant part of the bound is given by

$$\begin{aligned} Ł(\mathbf{Z}) &= \langle \sum_s \log p(\mathbf{u}^{(s)}|\mathbf{Z}) \rangle + \log p(\mathbf{Z}) \\ &= -\frac{S}{2} \log |\mathbf{K}_u| - \frac{1}{2} \sum_s \langle \mathbf{u}^{(s)T} \mathbf{K}_u^{-1} \mathbf{u}^{(s)} \rangle \\ &\quad - \frac{1}{2} \operatorname{Tr}(\mathbf{Z}^T \mathbf{K}_z^{-1} \mathbf{Z}) \end{aligned}$$

and its gradient by

$$\begin{aligned} \frac{\partial 2Ł}{\partial \mathbf{Z}_{ij}} &= -S \operatorname{Tr}\left(\mathbf{K}_u^{-1} \frac{\partial \mathbf{K}_u}{\partial \mathbf{Z}_{ij}} \right) \\ &\quad + \sum_s \left\langle \mathbf{u}^{(s)T} \mathbf{K}_u^{-1} \frac{\partial \mathbf{K}_u}{\partial \mathbf{Z}_{ij}} \mathbf{K}_u^{-1} \mathbf{u}^{(s)} \right\rangle - \frac{1}{2} [\mathbf{K}_z^{-1} \mathbf{Z}]_{ij} \\ &= \operatorname{Tr}\left(\left[\mathbf{K}_u^{-1} (\sum_s \langle \mathbf{u}^{(s)} \mathbf{u}^{(s)T} \rangle) \mathbf{K}_u^{-1} - S \mathbf{K}_u^{-1} \right] \frac{\partial \mathbf{K}_u}{\partial \mathbf{Z}_{ij}} \right) \\ &\quad - \frac{1}{2} [\mathbf{K}_z^{-1} \mathbf{Z}]_{ij}, \end{aligned}$$

where

$$\frac{\partial \mathbf{K}_u}{\partial \mathbf{Z}_{ij}} = (\mathbf{Z} \mathbf{1}_{ij}^T + \mathbf{1}_{ij} \mathbf{Z}^T) \circ \mathbf{K}_Q,$$

and $\mathbf{K}_z = K_z \otimes I_Q$ is a block matrix of full size $(QN \times QN)$. We optimise the $\mathbf{Z}$ variables using the conjugate gradient method while additionally making a change of variables by whitening the variables by their prior as

$$\hat{\mathbf{Z}}_{ij} = \mathbf{L}^{-1}\mathbf{Z}_{ij}$$

$$\frac{\partial \text{\L}}{\partial \hat{\mathbf{Z}}_{ij}} = \frac{\partial L}{\partial \mathbf{Z}_{ij}} \frac{\partial \mathbf{Z}_{ij}}{\partial \hat{\mathbf{Z}}_{ij}} = \mathbf{L}^T \frac{\partial \text{\L}}{\partial \mathbf{Z}_{ij}}$$

to make the optimization more efficient using the Cholesky decomposition of the kernel $\mathbf{K}_z = \mathbf{L}\mathbf{L}^T$ (Heinonen et al., 2016; Kuss and Rasmussen, 2005).

5 Related Work

In semiparametric latent factor models (SLFM) the signal $\mathbf{f}(x) = B\mathbf{u}(x)$ over M outputs is a linear combination of Q independent latent Gaussian process signals $\mathbf{u}(x)$ with a fixed mixing matrix $B \in \mathbb{R}^{M \times Q}$, with appropriate hyperparameter learning (Teh and Seeger, 2005). A Gaussian process regression network (GPRN) (Wilson et al., 2012) extends this model by considering a mixing matrix $B(x)$ where each element $B_{pi}(x)$ is an independent Gaussian process along x.

In geostatistics vector-valued regression with Gaussian processes is called *cokriging* (Álvarez et al., 2012). In *linear coregionalization models* (LCM) latent Gaussian processes are mixed from latent signals $u_p(x)$ and $u_q(x')$ that are independent. In contrast to SLFM, each signal $u_p(x)$ is an additional mixture of R_Q signals with separate shared covariances $K_q(x, x')$. In the intrinsic coregionalization model (ICM) only a single ($Q = 1$) latent mixture with a single shared kernel is used, while in SLFM there are multiple latent singleton ($R_Q = 1$) signals. In spatially varying LCMs (SVLCM) the mixing matrices are input-dependent, similar to GPRNs (Gelfand et al., 2004). Vargas-Guzmán et al. (2002) used non-orthogonal latent signals $u_p(x)$ and $u_q(x')$ with fixed covariances.

Multi-task Gaussian processes employ structured covariances that combine a task covariance with an input covariance. Simple Kronecker products between the covariances assume that task and input covariances are independent functions (Bonilla et al., 2007; Rakitsch et al., 2013; Stegle et al., 2011). This is computationally efficient (Flaxman et al., 2015), but it does not take into account interactions between the tasks and inputs.

In Generalised Wishart Processes (Wilson et al., 2012) an input-dependent covariance matrix $\Sigma(x) = \sum_{n=1}^{\nu} \mathbf{z}_n(x)\mathbf{z}_n(x)^T$ is a sum of ν outer products and follows a generalized Wishart process. The random variables $z_{ni}(x) \sim \mathcal{GP}(0, K(x, x'))$ are all independent Gaussian processes. Copula processes also describe dependencies of random variables by Gaussian processes (Wilson and Ghahramani, 2010). In Bayesian nonparametric covariance regression, covariances of multiple predictors share a common dictionary of Gaussian processes (Fox and Dunson, 2015).

Table 1: Results on all datasets. Boldface numbers indicate better method performance. MAE and MSE refer to the mean absolute and squared errors, respectively. AUC refers to the area under the ROC curve statistic. For the EEG simulation study, the Fisher's method is used to combine p-values from the simulations.

`JURA`	Average MAE	Average MSE
LCGP	**0.686**	0.804
GPRN	0.693	**0.801**
`EEG: classif.`	Average AUC	
LCGP	**0.830**[1]	
RLDA	0.826	
`EEG: simulation`	Mean score	p-value
$Q = 2$	0.87	9.62e-10
$Q = 3$	0.84	4.21e-08
$Q = 4$	0.87	1.11e-15

[1] Paired t-test, $p < 0.05$.

Finally, Gaussian process dynamical or state-space systems are a general class of discrete-time state-space models that combine the latent state into time-dependent outputs as Markov processes (Damianou et al., 2011; Deisenroth and Mohamed, 2012; Frigola et al., 2014; Wang et al., 2005). In Gaussian process factor analysis the outputs are described as factors that have GP priors, however not modeling the factor dependencies (Lawrence, 2004; Luttinen and Ilin, 2009).

6 Experiments

In the first experiment we show that our model can recover the true latent correlations in a simple simulated-data case, and compare our method with GPRNs, which is a state-of-the-art multi-output Gaussian process regression model (Wilson et al., 2012). We employ the mean-field variational inference implementation of GPRN by Nguyen and Bonilla (2013). Second, we apply our method to the Jura geospatial dataset to elucidate latent ore concentration process couplings. Finally, we demonstrate our full modelling framework on an EEG single-trial classification task, outperforming state-of-the-art regularised LDA in classification and additionally recovering an interesting latent representation that we further evaluate in a simulation study. Results from Jura and EEG experiments are summarised in Table 1.

6.1 Simulated-Data Experiment

We use simple toy data to show that we are able to recover known latent correlation structure. We generated data with a varying number of latent components $Q = 2, \ldots, 5$ and amount of samples $S = 1, \ldots, 20$. The mixing matrix was

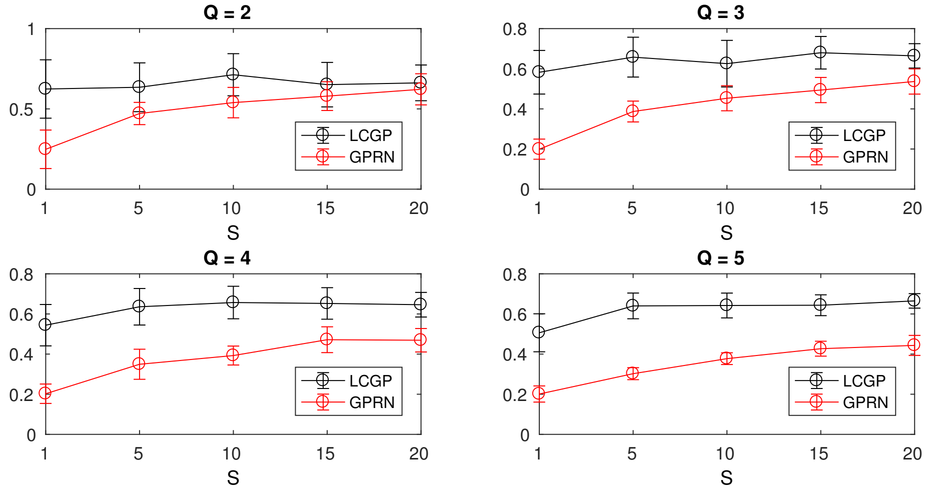


Figure 2: The true latent covariances can be recovered more accurately with our method than with GPRN. The curves show correlation to the true correlation matrix, over the elements of the recovered correlation matrix, as a function of the (very small) sample size S. The Q is the number of latent signals.

binary such that one output corresponded to a single latent variable. In this experiment we only consider LCGP without the classifier.

To assess the accuracy, we measured the correlation between the elements of the true covariance matrix to the one estimated. With GPRN we computed the empirical covariance $\hat{\mathbf{\Sigma}} = \sum_s \mathbf{u}^{(s)}\mathbf{u}^{(s)\,T}$ of the latent variables. As the order of the recovered latent variables is not identifiable, we computed the correlations over all permutations and report the best. Rotations of the latent space are not accounted for, however. The results in Figure 2 show that our model can recover the true underlying latent covariance with high correlations.

6.2 Jura

The Jura dataset[2] consists of measurements of cadmium, nickel and zinc concentrations in a region of the Swiss Jura (Goovaerts, 1997). For training we are given the concentrations measured at 259 locations and for validation the measurements at 100 additional locations. We set hyperparameters for both our model and GPRN as $\ell_u = 0.5$ and $\ell_b = 1$, and for our model the parameter for the latent correlation lengthscale to $\ell_z = 1$. We learned the models with $Q = 2$ latent variables, which resulted in the best model performance. We report both the mean squared and absolute errors for the predicted concentrations in Table 1. Our model performs at the same level as the state-of-the-art competitor GPRN, with slightly better performance in absolute errors.

[2] Data available at `https://sites.google.com/site/goovaertspierre/pierregoovaertswebsite/publications/book`.

Figure 3 shows the inferred model. The latent variables are 2D spatial surfaces on which the measurement points are indicated as black points. The two latent variables learn different geological processes that have an interesting two-pronged correlation pattern that indicates two kinds of negative correlations (the scatter plot). By explicitly modelling the latent covariance, we are able to see the regions of the input space that contribute to this pattern; the latent covariances indicate the combined covariance $C_{pq}(\mathbf{x}, \mathbf{x}') = A_{\mathbf{x}\mathbf{x}'}(p, q)K_{pq}(\mathbf{x}, \mathbf{x}')$. The diagonal plots of Figure 3a show the variances of the two latent signals, while the off-diagonal covariance plot indicates the two-pronged negative correlation model between the geological processes. Finally, the mixing matrices of the two latent components reconstruct the three ore observation surfaces.

6.3 EEG

Our main motivation for developing the present model was in modelling EEG data. We demonstrate our full model on data from a P300 study (Vareka et al., 2014), where the subjects were shown either a target or non-target stimulus, specifically a green or a red LED flashing, respectively. The classification task is to classify the stimulus based on the brain measurements. Additionally, we evaluate our modelling approach in a simulation study.

6.3.1 Classification Results

We evaluate the classification performance using a Monte Carlo cross-validation scheme where in each fold we randomly sample training and test sets of $S = 1000$ trials from the full dataset consisting in total of 7351 trials from 16 subjects. A single trial is the continuous voltage measurement of $M = 19$ channels in an EEG cap for 800ms with $N = 89$ after filtering and downsampling the time series (Hoffmann et al., 2008). We report the average area under the ROC curve (AUC) statistic over 100 folds, and compare our method to the state-of-the-art regularised LDA method implemented in the BBCI toolbox (Blankertz et al., 2010). Results in Table 1 show that our method performs statistically significantly ($p < 0.05$) better than RLDA.

An example visualization of the model from one of the cross-validation folds is depicted in Figure 4 for the three first latent signals. Panel (a) indicates the shared variances and covariances of the latent signals along time. The first and third latent signals have a monotonically increasing covariance coupling, while the first and second latent signals have a periodicity in the covariance. The average latent variables of the target and non-target trials are shown in panel (b). The third latent variable captures a strong dynamic between time points $[0.3, 0.5]$, which coincides with the expected P300 activity approximately 300 ms after the stimulus representation. The first two variables show peaks also at approximately 300 ms. In general the positive trials have a remarkably different latent representations than the negative trials. Panel (c) shows the classifier weights $\mathbf{w}_p$ for the three latent signals with the average classification plotted. Finally, a subset of the EEG channels are shown in panel (d), highlighting the differences in the channel dynamics. In panel (e) the components are found to be discriminative also when plotted on the scalp map.

6.3.2 Finding the Latent Correlations

In addition to the classification results, we evaluated our modelling approach in a simulation study to test whether we can find the latent correlations correctly using data that resembles the real EEG as closely as possible, but where we know the ground truth. To this end, we simulated datasets from the fitted models, and learn a new model on the simulated data. We repeated this for varying number of latent variables ($Q = 2, 3, 4$) with 10 simulations done for each value of Q. For each simulation we computed the empirical p-value with the hypothesis that the accuracy of our model is greater than using randomly simulated covariances from the model. The p-values from the simulations were combined using the Fisher's method (Fisher, 1925), with results reported in Table 1. We use again the correlation-based score for assessing the accuracy as with the toy data case.

7 Discussion

The latent correlation Gaussian process is a flexible framework for multi-task learning. We demonstrate in two experiments that our model can robustly learn the latent variable correlations. The model also achieves state-of-the-art performance in both regression and classification. The added modeling of the correlations of the underlying latent processes both improves model interpretability, and regularises the model especially with multiple observations. The novel latent correlation kernel encodes mutually-dependent covariances between latent signals and inputs in a parameterised way without being overly flexible.

In place of the non-stationary Gaussian kernel other non-stationary kernels are possible. Paciorek and Schervish (2006) propose a class of non-stationary convolution kernels containing, for instance, a non-stationary Matérn kernel. For future work coupling the spectral kernels (Wilson and Adams, 2013) with Wishart correlations is another highly interesting avenue for a general family of dependent, structured kernels.

Acknowledgements

This work has been partly supported by the Academy of Finland (Finnish Centre of Excellence in Computational Inference Research COIN), the Finnish Funding Agency for Innovation (project Re:Know), and the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no 611570. We also acknowledge the computational resources provided by the Aalto Science-IT project.

References

R. P. Adams and O. Stegle. Gaussian process product models for nonparametric nonstationarity. In *ICML*, volume 25, pages 1–8. ACM, 2008.

J. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679, 1993.

M. Álvarez, L. Rosasco, and N. D. Lawrence. Kernels for vector-valued functions: A review. *Foundations and Trends in Machine Learning*, 4(3):195–266, 2012.

M. A. Alvarez, D. Luengo, M. K. Titsias, and N. D. Lawrence. Efficient multi-output Gaussian processes through variational inducing kernels. In *AISTATS*, volume 9, pages 25–32, 2010.

H. Attias. Inferring parameters and structure of latent variable models by variational bayes. In *UAI*, pages 21–30. Morgan Kaufmann Publishers Inc., 1999.

B. Blankertz, M. Tangermann, C. Vidaurre, S. Fazli, C. Sannelli, S. Haufe, C. Maeder, L. Ramsey, I. Sturm, and G. Curio. The Berlin brain–computer interface: non-medical uses of BCI technology. *Frontiers in neuroscience*, 4 (198), 2010.

E. Bonilla, K. Chai, and C. Williams. Multi-task Gaussian process prediction. In *NIPS*, volume 20, pages 153–160, 2007.

A. Damianou, M. K. Titsias, and N. D. Lawrence. Variational Gaussian process dynamical systems. In *NIPS*, volume 24, pages 2510–2518, 2011.

M. Deisenroth and S Mohamed. Expectation propagation in Gaussian process dynamical systems. In *NIPS*, volume 25, pages 2609–2617, 2012.

R. A. Fisher. *Statistical Methods For Research Workers*. Oliver and Boyd, Edinburgh, 1925.

S. Flaxman, A. G. Wilson, D. Neill, H. Nickisch, and A. Smola. Fast kronecker inference in Gaussian processes with non-Gaussian likelihoods. In *ICML*, volume 2015, 2015.

E. B. Fox and D. B. Dunson. Bayesian nonparametric covariance regression. *Journal of Machine Learning Research*, 16:2501–2542, 2015.

R. Frigola, Y. Chen, and C. Rasmussen. Variational Gaussian process state-space models. In *NIPS*, volume 27, pages 3680–3688, 2014.

A. Gelfand, A. Schmidt, S. Banerjee, and C. Sirmans. Nonstationary multivariate process modeling through spatially varying coregionalization. *Test*, 13(2): 263–312, 2004.

M. Gibbs. *Bayesian Gaussian Processes for Regression and Classification.* PhD thesis, University of Cambridge, 1997.

P. Goldberg, C. Williams, and C. Bishop. Regression with input-dependent noise: A Gaussian process treatment. In *NIPS*, volume 10, pages 493–499, 1997.

P. Goovaerts. *Geostatistics for natural resources evaluation.* Oxford University Press, 1997.

M. Heinonen, H. Mannerström, J. Rousu, S. Kaski, and H. Lähdesmäki. Nonstationary Gaussian process regression with Hamiltonian Monte Carlo. In *AISTATS*, volume 51, pages 732–740, 2016.

U. Hoffmann, J. Vesin, T. Ebrahimi, and K. Diserens. An efficient p300-based brain–computer interface for disabled subjects. *Journal of Neuroscience methods*, 167(1):115–125, 2008.

K. Kersting, C. Plagemann, P. Pfaff, and W. Burgard. Most likely heterscedatic Gaussian process regression. In *ICML*, volume 24, pages 393–400, 2007.

M. Kuss and C. E. Rasmussen. Assessing approximate inference for binary Gaussian process classification. *Journal of Machine Learning Research*, 6: 1679–1704, 2005.

N. D. Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. In *NIPS*, volume 16, pages 329–336, 2004.

M. Lazaro-Gredilla and M. K. Titsias. Variational heteroscedatic Gaussian process regression. In *ICML*, pages 841–848, 2011.

Q. Le, A. Smola, and S. Canu. Heteroscedastic Gaussian process regression. In *ICML*, pages 489–496, 2005.

J. Luttinen and A. Ilin. Variational Gaussian-process factor analysis for modeling spatio-temporal data. In *NIPS*, pages 1177–1185, 2009.

T. Nguyen and E. Bonilla. Efficient variational inference for Gaussian process regression networks. In *AISTATS*, pages 472–480, 2013.

C. Paciorek and M. Schervish. Nonstationary covariance functions for Gaussian process regression. In *NIPS*, pages 273–280, 2004.

C. Paciorek and M. Schervish. Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*, 17(5):483–506, 2006.

N. Quadrianto, K. Kersting, M. Reid, T. Caetano, and W. Buntine. Kernel conditional quantile estimation via reduction revisited. In *IEEE International Conference on Data Mining*, pages 938–943, 2009.

B. Rakitsch, C. Lippert, K. Borgwardt, and O. Stegle. It is all in the noise: Efficient multi-task Gaussian process inference with structured residuals. In *NIPS*, pages 1466–1474, 2013.

C. E. Rasmussen and C. Williams. *Gaussian processes for machine learning*. MIT Press, 2006.

M. Schmidt. Function factorization using warped Gaussian processes. In *ICML*, pages 921–928. ACM, 2009.

O. Stegle, C. Lippert, J. Mooij, N. D. Lawrence, and K.M. Borgwardt. Efficient inference in matrix-variate Gaussian models with iid observation noise. In *NIPS*, pages 630–638, 2011.

Y. Teh and M. Seeger. Semiparametric latent factor models. In *AISTATS*, 2005.

V. Tolvanen, P. Jylänki, and A. Vehtari. Expectation propagation for nonstationary heteroscedastic Gaussian process regression. In *Machine Learning for Signal Processing (MLSP), 2014 IEEE International Workshop on*, pages 1–6. IEEE, 2014.

L. Vareka, P. Bruha, and R. Moucek. Event-related potential datasets based on a three-stimulus paradigm. *GigaScience*, 3(1):1, 2014.

J. A. Vargas-Guzmán, A.W. Warrick, and D.E. Myers. Coregionalization by linear combination of nonorthogonal components. *Mathematical Geology*, 34 (4):405–419, 2002.

C. Wang and R. Neal. Gaussian process regression with heteroscedastic or non-Gaussian residuals. Technical report, University of Toronto, 2012. https://arxiv.org/abs/1212.6246.

J. Wang, A. Hertzmann, and D. Blei. Gaussian process dynamical models. In *NIPS*, pages 1441–1448, 2005.

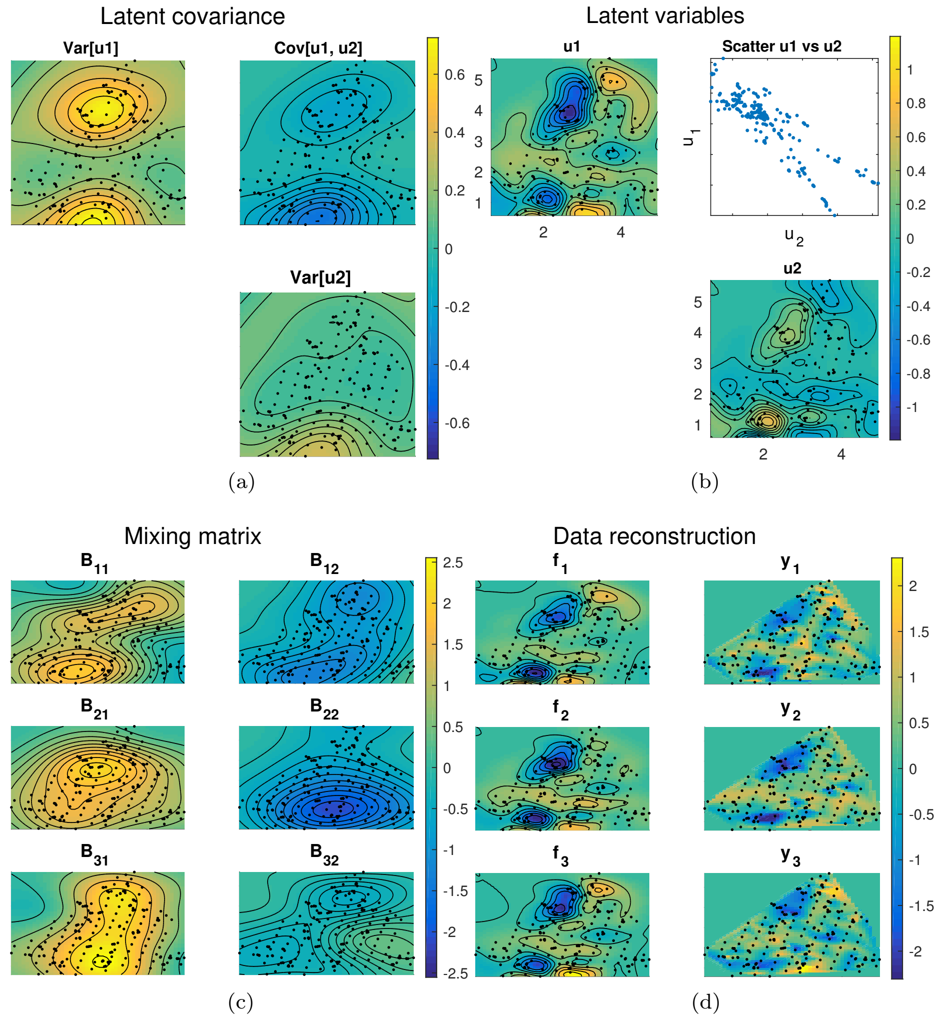
A. G. Wilson and R. Adams. Gaussian process kernels for pattern discovery and extrapolation. In *ICML*, 2013.

A. G. Wilson and Z. Ghahramani. Copula processes. In *NIPS*, pages 2460–2468, 2010.

A. G. Wilson and Z. Ghahramani. Generalised Wishart processes. In *Uncertainty in Artificial Intelligence*, 2011.

A. G. Wilson, D. Knowles, and Z. Ghahramani. Gaussian process regression networks. In *ICML*, 2012.

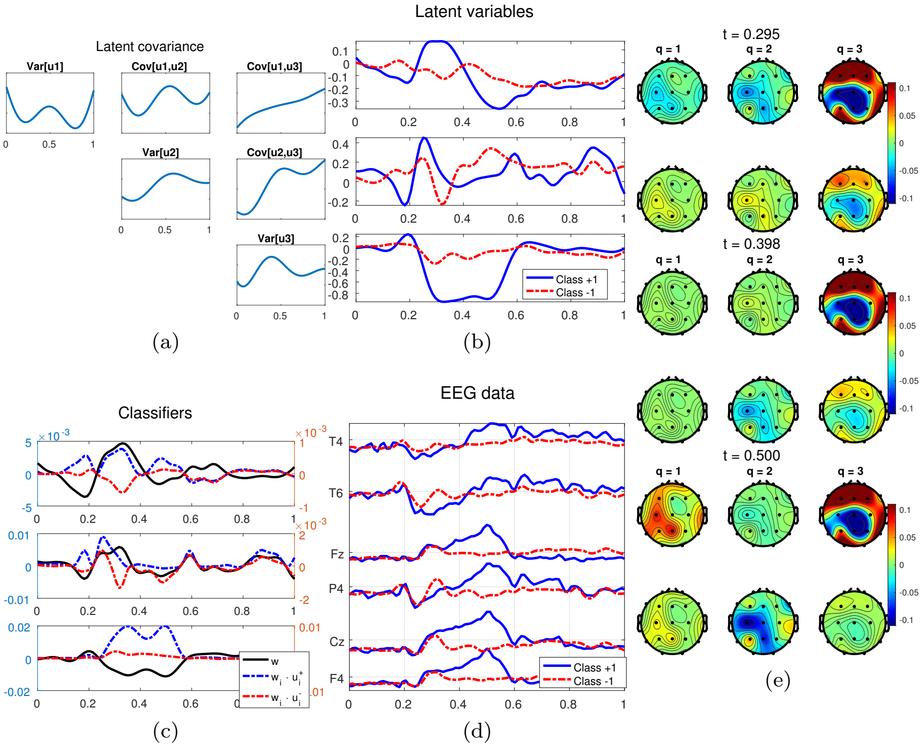
K. Yu, V. Tresp, and A. Schwaighofer. Learning Gaussian processes from multiple tasks. In *ICML*, pages 1012–1019. ACM, 2005.



(a) (b)

(c) (d)

Figure 3: Results on the Jura data. **(a)**: The learned latent covariance matrix shows an overall negative correlation over the input space. **(b)**: The learned latent variables clearly show a negative correlation. **(c)**: The learned input-dependent mixing matrix. **(d)**: LCGP reconstructs accurately the original data, which is plotted using a simple linear interpolation of the observed points for comparison.



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

Figure 4: **(a)**: Latent covariances. **(b)**: Latent variables, averaged within each class. **(c)**: Classifier weights and class-wise average scores. **(d)** Class-wise averaged EEG data from a subset of 6 channels. **(e)** The latent variables mapped onto the EEG scalp map at several time points t. The top figures shows the target class (+1) and bottom the non-target (-1), for each time point.