Non-stationary Gaussian process discriminant analysis with variable selection for high-dimensional functional data

Weichang Yu
Melbourne Centre for Data Science
University of Melbourne
Parkville, VIC 3010, Australia
weichang.yu@unimelb.edu.au

Sara Wade
School of Mathematics
University of Edinburgh
Edinburgh, EH9 3FD, United Kingdom
sara.wade@ed.ac.uk

Howard D. Bondell
Melbourne Centre for Data Science
University of Melbourne
Parkville, VIC 3010, Australia
howard.bondell@unimelb.edu.au

Lamiae Azizi
School of Mathematics and Statistics
University of Sydney
Camperdown, NSW 2006, Australia
lamiae.azizi@gmail.com

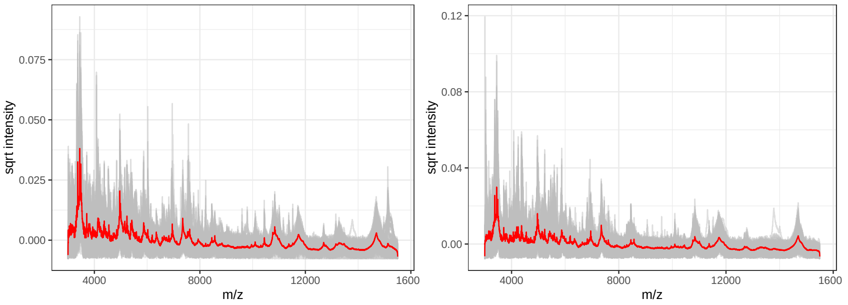
Abstract

High-dimensional classification and feature selection tasks are ubiquitous with the recent advancement in data acquisition technology. In several application areas such as biology, genomics and proteomics, the data are often functional in their nature and exhibit a degree of roughness and non-stationarity. These structures pose additional challenges to commonly used methods that rely mainly on a two-stage approach performing variable selection and classification separately. We propose a novel Gaussian process discriminant analysis (GPDA) that combines these steps in a unified framework. Our model is a two-layer non-stationary Gaussian process coupled with an Ising prior to identify differentially-distributed locations. Scalable inference is achieved via developing a variational scheme that exploits advances in the use of sparse inverse covariance matrices. We demonstrate the performance of our methodology on simulated datasets and two proteomics datasets: breast cancer and SARS-CoV-2. Our approach distinguishes itself by offering explainability as well as uncertainty quantification in addition to low computational cost, which are crucial to increase trust and social acceptance of data-driven tools.

1 Motivating example

We consider the context of predicting phenotypes and identifying biomarkers based on mass spectrometry (MS) data. MS technology measures the mixtures of proteins/peptides of tissues or fluids and produces an MS spectrum (Cruz-Marcelo et al., 2008).The resulting experimental data consists of discretely observed functional spectra, with typically tens of thousands of observed locations and just a few hundred samples. Moreover, data at neighboring locations tends to be highly correlated, with

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.



(a) Controls

(b) SARS-CoV-2

Figure 1: Illustration of the SARS-CoV-2 data (Nachtigall et al., 2020), with the square root intensities of the preprocessed spectra for (a) healthy controls and (b) SARS-CoV-2 positive patients (class-average given in red).

the strength of such correlations varying across the mass-to-charge (m/z) range. In addition, MS data tends to be noisy due to chemical noise, misalignment, calibration and other issues (Cruz-Marcelo et al., 2008). As an example, Figure 1 depicts a set of processed MS spectra for healthy controls and SARS-CoV-2 positive patients (Nachtigall et al., 2020), with the x-axis indicating the mass-to-charge ratio (m/z) and the y-axis indicating the intensity of the protein or peptide ions.

To address challenges in the analysis of such datasets, we propose a novel Bayesian discriminant analysis (DA) which performs variable selection and classification jointly, by combining recent developments in deep Gaussian processes (Dunlop et al., 2018) to flexibly model the functional inputs and incorporating Ising priors to identify differentially-distributed locations within a unified model framework. To ameliorate the computational burden of posterior inference for high-dimensional Gaussian processes (GPs), we develop a scalable inference algorithm that utilizes the link between GPs and stochastic partial differential equations (SPDEs) to construct sparse precision matrices (Lindgren et al., 2011; Grigorievskiy et al., 2017) and combine various variational inference schemes.

2 Model

Consider a set of n functional inputs $\{x_i(t)\}_{i=1}^n$ defined on the domain $\mathcal{D} \subset \mathbb{R}$ and their corresponding class labels $\{y_i\}_{i=1}^n$, where $x_i(t) \in \mathbb{R}$, $y_i \in \mathcal{Y}$, and $\mathcal{Y}$ is a set of class labels. We propose the following model that performs the variable selection and the classification steps simultaneously on the entire functional trajectory:

$$x_i(t) = \mu_k(t) + z_i(t) + \epsilon_{k,i}(t), \tag{1}$$

$$z_i \mid \psi_i \sim \mathrm{GP}(0, K_{\psi_i}) \tag{2}$$

where $y_i = k \in \{0, 1\}$ refers to the class label; $\mu_k(t) \in \mathbb{R}$ is the group-specific mean function; K_{ψ_i} is a covariance function (or kernel) with observation-specific parameters ψ_i; $\epsilon_{k,i}(t) \in \mathbb{R}$ is a white-noise process with class- and location-dependent variance $\sigma_k^2(t) > 0$; and $\mathrm{GP}(0, K_\psi)$ denotes a Gaussian process with zero mean and covariance kernel K_ψ. We allow for variable selection in our proposed model by defining a binary signal process $\gamma(t) \in \{0, 1\}$ such that

$$\mu_k(t) = \gamma(t)\widetilde{\mu}_k(t) + (1 - \gamma(t))\widetilde{\mu}_\emptyset(t) \;\text{ and }\; \sigma_k^2(t) = \gamma(t)\widetilde{\sigma}_k^2(t) + (1 - \gamma(t))\widetilde{\sigma}_\emptyset^2(t),$$

where $\widetilde{\mu}_k$ and $\widetilde{\sigma}_k^2$ are the group-specific mean and noise variance processes at discriminative locations and $\widetilde{\mu}_\emptyset$ and $\widetilde{\sigma}_\emptyset^2$ are the common mean and noise variance processes at non-discriminative locations. In the context of MS data, γ allows for detection of relevant m/z values within the classification model.

Two-level non-stationary Gaussian processes. To account for this varying correlation along the spectrum, we assign a non-stationary covariance kernel (Paciorek and Schervish, 2003) for K_{ψ_i}. Specifically, the kernel parameter, $\psi_i = (\tau, \nu_i)$, consists of the magnitude $\tau > 0$ and a location-varying log length-scale process ν_i, i.e., $K_{\psi_i} = K_{NS;\tau,\nu_i}$, and hence we may write

$$z_i|\tau, \nu_i \sim \mathrm{GP}(0, K_{NS;\tau,\nu_i}).$$

At the second level, we place Gaussian process priors on the log length-scale processes with

$$\nu_i(t) = R(t) + \zeta_i, \qquad \text{and} \qquad R \sim \mathrm{GP}(\mu_\nu, K_{S;\tau_2,\lambda}),$$

where $K_{S;\tau_2,\lambda}$ is a stationary covariance kernel with marginal scale τ_2 and length scale λ. Here, each observation-specific log length-scale process has been decomposed into a common component $R(t)$ to account for the location-varying covariance structure common across all observed functions, and an observation-specific perturbation $\zeta_i \in \mathbb{R}$ to allow for between-spectra variation in smoothness across the entire domain.

Motivated by the link between GPs and SPDEs (Lindgren et al., 2011; Monterrubio-Gómez et al., 2020), we employ an SDE representation of the nonstationary processes (Zhao et al., 2021):

$$dz_i = -\frac{1}{\exp(\nu_i)} z_i dt + \sqrt{\frac{2\tau}{\exp(\nu_i)}} d\omega_1, \tag{3}$$

$$dR = -\frac{1}{\lambda} R dt + \sqrt{\frac{2\tau_2}{\lambda}} d\omega_2, \tag{4}$$

where $\nu_i(t) = R(t) + \zeta_i$ and ω_1 and ω_2 are Wiener processes. From this representation, the induced posterior precision matrices for discretized z_i and R are tridiagonal which facilitates computational shortcuts such as Thomas' algorithm and the sparse inverse subset algorithm (Durrande et al., 2019).

2.1 Choice of priors

To reflect our prior belief that the underlying variable selection process γ is smooth, we assign a linear chain Ising prior (Li and Zhang, 2010). In particular, the conditional distribution of $\gamma(t)$ given its corresponding set of neighbors with locations in $\mathcal{N}_t \subset \{t_1, \ldots, t_T\}$ is

$$\mathbb{P}(\gamma(t) = 1 \mid \{\gamma(t')\}_{t' \in \mathcal{N}_t}) = \operatorname{expit}\left\{-\alpha + \sum_{t' \in \mathcal{N}_t} \beta(t, t')\gamma(t')\right\},$$

where $\operatorname{expit}(x) = \{1 + \exp(-x)\}^{-1}$, $\alpha \in \mathbb{R}$, and $\beta(t, t') > 0$. Here, a larger value of α corresponds to more sparsity in γ, whereas β controls the correlation and smoothness between the values of γ at neighboring locations. In our context, we define $\mathcal{N}_t = \{t - 1, t + 1\}$ and $\beta(t, t') = \beta$.

We assign a hierarchical GP priors for the mean functions

$$\begin{aligned} \widetilde{\mu}_k(t) \mid \widetilde{\tau}_k, \widetilde{\nu}_k &\sim \mathrm{GP}(0, K_{NS;\widetilde{\tau}_k,\widetilde{\nu}_k}), \\ \widetilde{\nu}_k(t) \mid \widetilde{\eta}, \widetilde{\lambda} &\sim \mathrm{GP}(\mu_{\widetilde{\nu}}, K_{S;\widetilde{\eta},\widetilde{\lambda}}), \end{aligned}$$

with hyperpriors $\widetilde{\tau}_k \sim \mathrm{InvGa}(A_{\widetilde{\tau}}, B_{\widetilde{\tau}})$ for $k = 0, 1, \emptyset$; $\widetilde{\eta} \sim \mathrm{InvGa}(A_{\widetilde{\eta}}, B_{\widetilde{\eta}})$; and $\widetilde{\lambda} \sim \mathrm{LogN}(\mu_{\widetilde{\lambda}}, \sigma^2_{\widetilde{\lambda}})$.

3 Posterior inference

Let $\boldsymbol{\theta}$ denote the vector of all model parameters (excluding the hyperparameters ζ_i, λ, τ_2, $\widetilde{\lambda}$, $\widetilde{\eta}$, α, and β). We specify the mean-field family for the approximate posterior:

$$q(\boldsymbol{\theta}) = q(\tau) q(\boldsymbol{R}) \prod_{i=1}^{n} \{q(\boldsymbol{z}_i)\} \times \prod_{k \in \{\emptyset, 0, 1\}} \left\{ q(\widetilde{\boldsymbol{\mu}}_k) q(\widetilde{\boldsymbol{\nu}}_k) q(\widetilde{\tau}_k) \prod_{j=1}^{T} q(\widetilde{\sigma}^2_{kj}) \right\} \times \prod_{j=1}^{T} \{q(\gamma_j)\}.$$

The DAG for our proposed model is provided in Figure 2. Note that parameters in white, red, and blue fill are updated with coordinate-ascent variational inference (CAVI), stochastic variational Bayes (SVB) and maximum a posteriori (MAP) respectively, whereas gray fill denotes observed quantities.

3.1 Classification

Upon convergence of the variational parameters in the posterior inference phase, we proceed to derive a classification rule for a new process $x_{n+1}(t)$ that follows the distribution as described in equation

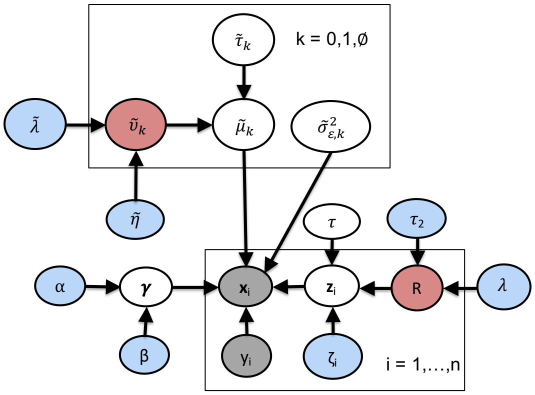


Figure 2: DAG representation of the proposed model. The colour fills denote the inference approach adopted: blue for MAP; red for SVB; white for CAVI; and gray denotes observed data.

(1). This requires the predictive distribution $p(y_{n+1}, \boldsymbol{z}_{n+1} \,|\, \mathcal{D}, \boldsymbol{x}_{n+1})$, where $\mathcal{D}$ denotes all observed data. To simplify computations, we make the following mean field approximation for the predictive distribution of y_{n+1} and $\boldsymbol{z}_{n+1}$:

$$p(\boldsymbol{z}_{n+1}, y_{n+1} \,|\, \mathcal{D}, \boldsymbol{x}_{n+1}) \approx q(\boldsymbol{z}_{n+1})q(y_{n+1}).$$

We adopt a CAVI update for $\boldsymbol{z}_{n+1}$ and y_{n+1}. Since the approximate posterior for R has been computed in the posterior inference phase, we only need an estimate for the perturbation ζ_{n+1}. This may be computed via MAP estimation.

4 Numerical results

We compare our `GPDA` model with seven other competing methods - variational nonparametric DA (`VNPDA`, Yu et al., 2020a), penalized linear DA with fused lasso penalty (`penLDA-FL`, Witten and Tibshirani, 2011), random forest, sparse linear DA (`SparseLDA`, Clemmensen et al., 2011), variational linear DA (`VLDA`, Yu et al., 2020b), and both the L_2-regularized and L_1-regularized support vector machine (`SVM`) with linear kernels (Cortes and Vapnik, 1995; Fan et al., 2008). For the proteomics datasets, we also compare with the traditional two-stage algorithm, involving peak detection and followed by linear DA and quadratic DA.

4.1 Simulation study

In **Simulation 1**, we have a large proportion (40%) of the locations have weak predictive power, whereas the rest of the locations do not have any predictive power. The `GPDA` model is correctly specified, i.e., the covariance function of the i-th observation is $\boldsymbol{\Sigma}^\star_{ik} = D^\star_{\epsilon,k} + Q^{-1}_{NS;\tau^\star,\boldsymbol{\nu}^\star_i}$, and $\boldsymbol{R}^\star \sim$ N$(\mathbf{0}, Q^{-1}_{S;\tau^\star_2,\lambda^\star})$. For **Simulation 2**, we consider a similar scenario to Simulation 1 but with a much smaller proportion (5%) of the locations having strong predictive power, whereas the rest of the locations do not have any predictive power. For **Simulation 3**, we assess the performance of the methods when the locations are mutually independent, the noise variances are equal between groups, and a small proportion (10%) of the locations are weak signals, i.e. `VLDA` is correctly specified. This is a boundary case whereby the true log length scale $\boldsymbol{\nu}_i \to -\infty$. Lastly, **Simulation 4** allows us to assess the performance of the methods when the GPDA model is misspecified. In particular, the true covariance matrix has a uniform structure with all diagonal entries equal 1 and the off-diagonal entries equal 0.95. A small proportion (10%) of the locations have strong predictive power.

4.2 Proteomics datasets

SARS-CoV-2: To improve COVID-19 testing capacity in countries that lack resources to handle large-scale PCR testing, the SARS-CoV-2 dataset was collected using equipment and expertise commonly found in clinical laboratories in developing countries. The dataset contains samples from

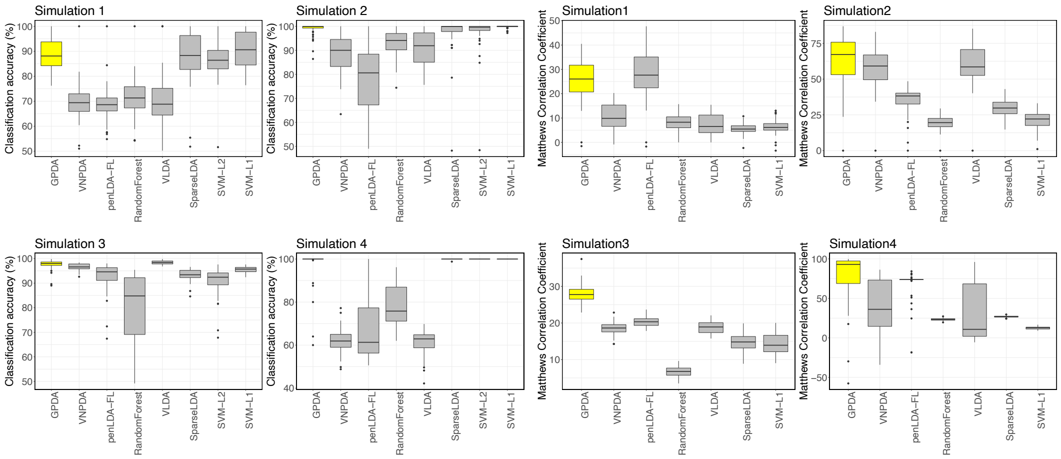


Figure 3: Classification accuracy (left) and Matthews correlation coefficient (right) for Simulations 1 to 4. GPDA refers to our proposed model.

362 individuals, of which 211 were SARS-CoV-2 positive and 151 were negative by PCR testing. The processed spectra contain $T = 25,001$ variables.

Breast cancer: This dataset was collected to investigate and identify markers from plasma that discriminate between controls and breast cancer patients. The processed spectra contain $T = 10,451$ variables. Due to heterogeneity in breast cancers, in the following, we focus on discriminating between healthy controls and HER2 (with $n = 119$).

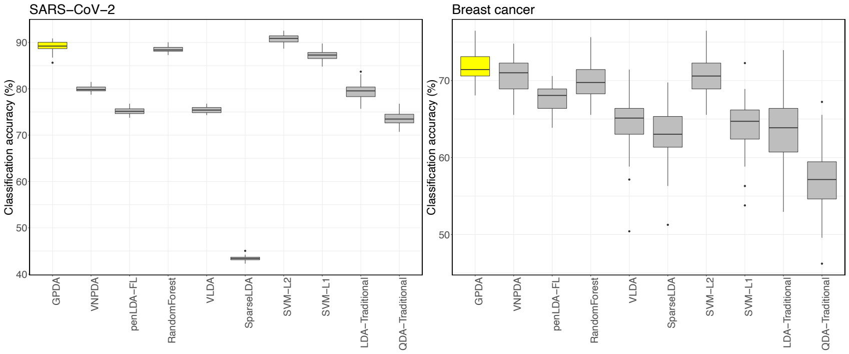


Figure 4: Classification accuracy (%) for SARS-CoV-2 (left) and breast cancer (right).

Supporting documents

The full paper and supporting R codes may be downloaded from `https://github.com/weichangyu10/GPDAPublic`.

References

Clemmensen, L., Hastie, T., Witten, D., and Ersboll, B. (2011). Sparse discriminant analysis. *Technometrics*, 53(4):406–413.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20:273–297.

Cruz-Marcelo, A., Guerra, R., Vannucci, M., Li, Y., Lau, C. C., and Man, T.-K. (2008). Comparison of algorithms for pre-processing of SELDI-TOF mass spectrometry data. *Bioinformatics*, 24(19):2129–2136.

Dunlop, M. M., Girolami, M. A., Stuart, A. M., and Teckentrup, A. L. (2018). How deep are deep Gaussian processes? *Journal of Machine Learning Research*, 19(1):2100–2145.

Durrande, N., Adam, V., Bordeaux, L., Eleftheriadis, S., and Hensman, J. (2019). Banded matrix operators for Gaussian Markov models in the automatic differentiation era. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, volume 89.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Grigorievskiy, A., Lawrence, N., and Särkkä, S. (2017). Parallelizable sparse inverse formulation Gaussian processes (SpInGP). In *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE.

Li, F. and Zhang, N. R. (2010). Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Journal of the American Statistical Association*, 105(491):1202–1214.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B*, 73(4):423–498.

Monterrubio-Gómez, K., Roininen, L., Wade, S., Damoulas, T., and Girolami, M. (2020). Posterior inference for sparse hierarchical non-stationary models. *Computational Statistics & Data Analysis*, 148.

Nachtigall, F. M., Pereira, A., Trofymchuk, O. S., and Santos, L. S. (2020). Detection of SARS-CoV-2 in nasal swabs using MALDI-MS. *Nature Biotechnology*, 38(10):1168–1173.

Paciorek, C. J. and Schervish, M. J. (2003). Nonstationary covariance functions for Gaussian process regressions. In *Advances in Neural Information Processing Systems*.

Witten, D. M. and Tibshirani, R. (2011). Penalized classification using Fisher's linear discriminant. *Journal of the Royal Statistics Society: Series B*, 73(5):753–772.

Yu, W., Azizi, L., and Ormerod, J. T. (2020a). Variational nonparametric discriminant analysis. *Computational Statistics & Data Analysis*, 142:106817.

Yu, W., Ormerod, J. T., and Stewart, M. (2020b). Variational discriminant analysis with variable selection. *Statistics and Computing*, 30:933–951.

Zhao, Z., Emzir, M., and Särkkä, S. (2021). Deep state-space Gaussian processes. *Statistics and Computing*.