Scalable Feature Extraction in Confounded Data

Florian Wenzel[*]
TU Kaiserslautern, Germany

Stephan Mandt
UCI, USA

Marius Kloft
TU Kaiserslautern, Germany
USC, USA

Abstract

We present a novel scalable inference algorithm for sparse feature selection in binary classification where the training data show spurious correlations, e.g., due to confounding. The approach builds on the sparse probit linear mixed model which was limited to datasets of a few hundred points. Our algorithm potentially scales to datasets with millions of points.

1 Introduction

In genetic association studies the goal is to find causal associations between high-dimensional vectors of *genotypes*, such as single nucleotide polymorphisms (SNPs), and observable outcomes. Genetic associations can be spurious, unreliable, and unreproducible when the data are subject to spurious correlations due to confounding [1, 2, 3].

Recently, the sparse probit linear mixed model [4] was proposed. I was shown that this approach is very successful in finding a sparse set of relevant features while correcting for confounding. However, inference in this model is challenging and the application is limited to datasets of a few hundred samples.

In this work we propose a new Gaussian process based view on the model and propose a scalable inference algorithm which scales to datasets with million of points.

2 The Gaussian Process Linear Mixed Model

Let $X = (x_1, \dots, x_n) \in \mathbb{R}^{d \times n}$ be the d-dimensional training points with labels $y = (y_1, \dots, y_n) \in \{-1, 1\}^n$. The score for each data point is given by

$$s_i = f(x_i) + \beta^\top x_i, \tag{1}$$

which consists of two contributions. First, a linear term, parameterized by a *sparse weight vector* β which models the true underlying fixed effect. We place as Laplace prior over $\beta \sim \text{Laplace}(\lambda_0^{-1})$. This corresponds to the LASSO [5] and results in a sparse weight vector, i.e. only a small number of important features are selected.

The second contribution is $f(x_i)$ which models confounding by means of a *correlated noise term*. The more similar two data points x_i and x_j are, the higher is the correlation of their noise contributions $f(x_i)$, $f(x_j)$. We place a Gaussian process prior over $f \sim \text{GP}(0, k)$, whereby the kernel function k encodes similarity with respect to a potential confounder.

In order to obtain a likelihood suitable for classification, the score values have to be pushed through an activation function mapping them to probabilities. We use the logit link function and obtain the likelihood of the labels $p(y|f, X) = \prod_{i=1}^n \sigma(y_i s_i)$, where $\sigma(z) = (1 + \exp(-z))^{-1}$. The joint distribution of the labels y, the sparse weight vector β and the latent GP f is

$$p(y, \beta, f|X) = p(y|\beta, f, X)p(f)p(\beta), \tag{2}$$

[*] contact: wenzelfl@hu-berlin.de

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

For the sake of clarity we omit the conditioning on X in the following.

3 Inference

The goal is to compute the maximum-a-posterior (MAP) estimate of β.

$$\beta^* = \arg\max_\beta \log p(\beta|y) = \arg\max_\beta \log p(y|\beta) + \lambda_0||\beta||_1.$$

Since computing the marginal likelihood term $p(y|\beta) = \int p(y|\beta, f)p(f)df$ is intractable, we employ a *variational expectation maximization approach* [6].

To this end, we use a variational approximation $p(f|y,\beta) \approx q(f|m,S) = \mathcal{N}(f|m,S)$, where m, S are variational parameters. We obtain an approximation to the solution of the original problem by optimizing the variational lower bound

$$\log p(y,\beta) \geq \mathbb{E}_{q(f|m,S)}[\log p(y,f|\beta) - \log q(f|m,S)] + \lambda_0||\beta||_1. \tag{3}$$

We alternate between optimization steps in β and updating the variational parameters m, S.

3.1 Scalability

The variational maximization step is very similar to computing posterior of a Gaussian process classification model. Naive inference in a Gaussian process classification model scales cubicly in the number of data points and former work on the sparse probit linear mixed model [4] was limited to datasets of a few hundred data points.

In this work, we scale inference to million of data points building on recent work on scalable Gaussian process classification [7]. The approach is based on inducing points and Pólya-Gamma data augmentation making the model applicable to big datasets.

4 Preliminary Experiments

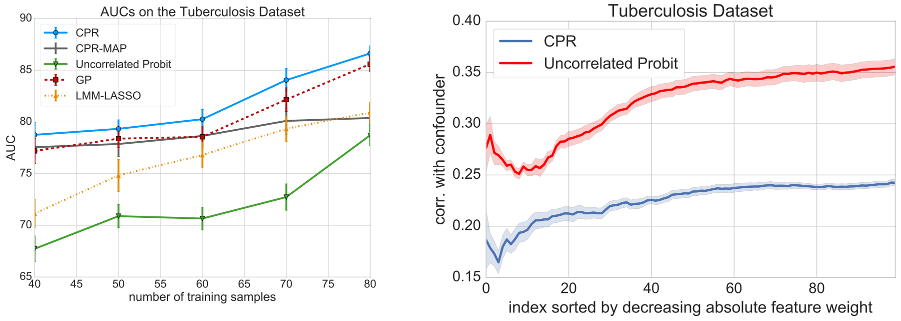

Figure 1: LEFT: Average AUC values with respect to the training set size. RIGHT: Correlation of the of top features sorted by descending absolute weights. Light-red/light-blue areas indicate standard errors.

We show preliminary results based on the inference algorithm presented in former work [4]. We compare our model against three competing methods, including sparse Probit regression, GP classification and the LMM-Lasso. We also include a MAP approximation of our model, which point estimates f instead of integrating it out. We predict the outcome of Tuberculosis from gene expression levels on the dataset from [8].

We observe that Probit-LMM achieves a consistent improvement over sparse Probit regression (by up to 12 percentage points), GP classification (by up to 3 percentage points), LMM-Lasso (by up to 7 percentage points) and its MAP approximation (by up to 7 percentage points). Moreover, the features that our model find are less affected by spurious correlations induced by population structure.

In future work we plan to apply our algorithm to much bigger dataset to empirically demonstrate its superior scalability.

References

[1] Guido W Imbens and Donald B Rubin. Causal Inference in Statistics, Social, and Biomedical Sciences. 2015.

[2] Judea Pearl et al. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009.

[3] Stephen L Morgan and Christopher Winship. *Counterfactuals and Causal Inference*. Cambridge University Press, 2014.

[4] S. Mandt, F. Wenzel, S. Nakajima, J. P. Cunningham, C. Lippert, and M. Kloft. Sparse Probit Linear Mixed Model. *Machine Learning Journal*, 2017.

[5] Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[6] Radford Neal and Geoffrey E. Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*, pages 355–368. Kluwer Academic Publishers, 1998.

[7] Florian Wenzel, Théo Galy-Fajou, Christian Donner, Marius Kloft, and Manfred Opper. Efficient gaussian process classification using polya-gamma data augmentation. *Conference on Artificial Intelligence (AAAI)*, 2019.

[8] Matthew PR Berry, Christine M Graham, Finlay W McNab, Zhaohui Xu, Susannah AA Bloch, Tolu Oni, Katalin A Wilkinson, Romain Banchereau, Jason Skinner, Robert J Wilkinson, et al. An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. *Nature*, 466(7309):973–977, 2010.