Finding Sparse Features in Strongly Confounded Medical Binary Data

S. Mandt	F. Wenzel	S. Nakajima	J. Cunningham	C. Lippert	M. Kloft
Columbia U	HU Berlin	TU Berlin	Columbia U	Human Longevity	HU Berlin

Abstract

A typical task in statistical genetics is to find a sparse linear relation between genotypes with phenotypes, but often the data are confounded by age, ethnicity or population structure. We generalize the linear mixed model (LMM) Lasso approach for feature selection under confounding to the case of binary labels. This case is much more involved, as marginalization over the correlated noise leads to an intractable integral. We can overcome this problem with approximate inference techniques. We demonstrate on synthetic and real-world data that the sparse features that our method finds are less correlated with the top confounders.

Introduction Genetic association studies have emerged as an important field of statistical genetics [1, 2]. In this class of problems, we associate high dimensional vectors of *genotypes*, such as SNPs or gene expression levels, with observable outcomes or *phenotypes*. These outcomes may be binary, such as the risk of getting a certain disease. For various diseases such as type 2 diabetes [3], the sparse linear effects that relate genotypes and phenotypes are largely undetected, which is why these missing associations have been entitled the *The Dark Matter of Genomic Associations* [4].

The problem is that these sparse signals can be spurious due to confounders that induce spurious non-causal correlations between genotypes and phenotypes. Confounding can stem from varying experimental conditions and demographics such as age, ethnicity or gender [5]. The perhaps most important type of confounding in statistical genetics arises due to population structure [6], which is due to the relatedness between the samples [7, 5, 8]. Ignoring such confounders can often lead to spurious false positive findings that cannot be replicated on independent data [9]. Correcting for such confounding dependencies is considered to be one of the greatest challenges in statistical genetics [10].

In this paper, we propose an algorithm for feature selection in binary classification in the presence of confounding. Our goal is to eliminate the confounder as well as possible and find a sparse weight vector that best captures causal relations.

Model Our model builds on the LMM-Lasso [11], an important method of statistical genetics to limit the impact of confounding. While the LMM-lasso relies on linear regression, we generalize this approach to the much more involved classification setup, where the target values are binary. Let $X \in \mathbb{R}^{d\times n}$ be the matrix of n observed data points. The corresponding labels $y \in \{-1,+1\}^n$ are assumed to be realized according to the following model,

$$y = \mathrm{sign}(X^\top w + \epsilon), \qquad \epsilon \sim \mathcal{N}(0,\Sigma), \tag{1}$$

where $\Sigma \in \mathbb{R}^{n\times n}$ is a fixed covariance matrix and the model parameter $w \in \mathbb{R}^d$ is unknown. As in the LMM-lasso, the noise covariance Σ captures similarities between the samples that offer an alternative explanation of the observed labels. This way, the sparse weight vector w focuses on strong sparse signals that can not be well explained in terms of correlated noise. We choose $\Sigma = \lambda_1\mathbf{I} + \lambda_2 X^\top X + \lambda_3\Sigma_{side}$, where Σ_{side} is constructed from side information such as age and geographical location. The weights λ_i are cross-validated.

In order to train the model we aim to maximize the marginal likelihood in the presence of a ℓ_1-norm regularizer (Lasso). This leads to the objective function

$$\mathcal{L}(w) \quad = \quad -\log \int_{\mathbb{R}^n_+} \mathcal{N}(\epsilon;\mu(w),\hat{\Sigma})\, d^n\epsilon \;+\; \lambda_0 ||w||_1^1, \tag{2}$$

where $\mu(w) := \mathrm{diag}(y)X^\top w$ and $\hat{\Sigma} := \mathrm{diag}(y)\Sigma\mathrm{diag}(y)$. The central computational problems of minimizing the objective function is that first, it contains an intractable, high-dimensional integral, and second, the ℓ_1-norm regularizer is not everywhere differentiable.

Our solution to this problem is based on a combination of the alternating direction method of multipliers (ADMM) [12] and expectation propagation (EP) [13] to approximate the integration over the truncated Gaussian distribution. Our model connects to sparse probit regression [14, 15] when we omit the off-diagonal parts of the noise covariance. It connects to Gaussian process (GP) classification [16] when we leave out the linear effect. Capturing both methods as limiting cases, our model therefore naturally outperforms both approaches in terms of prediction performance.

Empirical Analysis and Applications In the following we will apply our model to synthetic and real world data and show that it outperforms LMM-Lasso, GP classification and sparse probit regression in terms of prediction performance and feature quality.

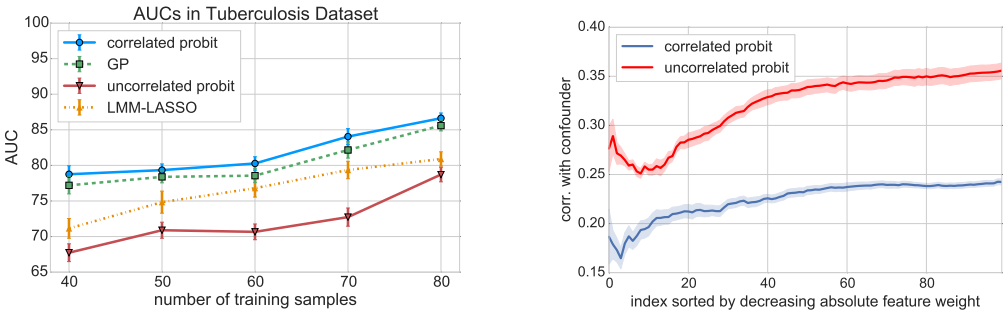

Figure 1: TBC: Results of the tuberculosis experiment. LEFT: Average AUC. RIGHT: Feature correlation with first principal component ($\widehat{=}$ population structure), where the x-axis is sorted by descending absolute weights.

Tuberculosis Disease Outcome Prediction From Gene Expression Levels: We obtained the dataset by [17] from the National Center for Biotechnology Information website[1], which includes 40 blood samples from patients with active tuberculosis as well as 103 healthy controls, together with the transcriptional signature of blood samples measured in a microarray experiment with 48,803 gene expression levels, which serve as features for our purposes. Also available is the age of subjects when the blood sample was taken, from which we compute Σ_{side}. All competing methods are trained for various training set sizes $n \in [40, 80]$. We report on the area under the ROC curve (AUC) and present the results in Figure 1. We furthermore computed the empirical correlation of the weight vector with the first principle component of the linear kernel (population structure). We found that the features that our model finds show much less correlation with population structure (confounding) than the features found by probit regression. This is because population structure was built into our model as a source of correlated noise.

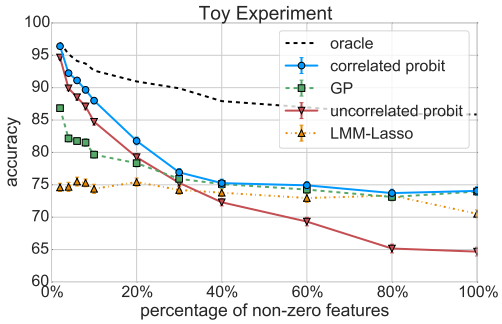

Figure 2: TOY: Average accuracies as a function of the number of true non-zero features in the generating model.

Simulated Data:

We evaluate our algorithm on synthetic data that we generate as follows. We generate a weight vector $w \in \mathbb{R}^{50}$ with $k \leq 50$ entries being 1, and else 0 and create a random covariance matrix Σ. Then we generate data according to our model (1) and train the competing methods and measure the performance on a held-out dataset. As a benchmark we introduce the oracle classifier, where we use the true underlying w for prediction. In Fig. 2 we report on the so-achieved accuracies as a function of the sparsity of the true underlying model parameter w.

[1] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE19491

Acknowledgments We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research. This work was partly funded by the German Research Foundation (DFG) award KL 2698/2-1.

References

[1] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, *et al.*, "Finding the missing heritability of complex diseases," *Nature*, vol. 461, no. 7265, pp. 747–753, 2009.

[2] S. Vattikuti, J. J. Lee, C. C. Chang, S. D. Hsu, and C. C. Chow, "Applying compressed sensing to genome-wide association studies," *GigaScience*, vol. 3, no. 1, p. 10, 2014.

[3] N. Craddock, M. E. Hurles, N. Cardin, *et al.*, "Genome-wide association study of cnvs in 16,000 cases of eight common diseases and 3,000 shared controls," *Nature*, vol. 464, no. 7289, pp. 713–720, 2010.

[4] T. N. H. G. R. Institute, "Proceedings of the workshop on the dark matter of genomic associations with complex diseases: Explaining the unexplained heritability from genome-wide association studies," 2009.

[5] L. Li, B. Rakitsch, and K. M. Borgwardt, "ccsvm: correcting support vector machines for confounding factors in biological data classification," *Bioinformatics*, vol. 27, no. 13, pp. 342–348, 2011.

[6] W. Astle and D. J. Balding, "Population structure and cryptic relatedness in genetic association studies," *Statistical Science*, pp. 451–471, 2009.

[7] C. Lippert, J. Listgarten, Y. Liu, C. Kadie, R. Davidson, and D. Heckerman, "Fast linear mixed models for genome-wide association studies," *Nature Methods*, vol. 8, pp. 833–835, October 2011.

[8] N. Fusi, O. Stegle, and N. D. Lawrence, "Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical studies," *PLoS comp. bio.*, vol. 8, no. 1, 2012.

[9] P. Kraft, E. Zeggini, and J. P. Ioannidis, "Replication in genome-wide association studies," *Statistical Science: A review journal of the Institute of Mathematical Statistics*, vol. 24, no. 4, p. 561, 2009.

[10] B. J. Vilhjálmsson and M. Nordborg, "The nature of confounding in genome-wide association studies," *Nature Reviews Genetics*, vol. 14, no. 1, pp. 1–2, 2013.

[11] B. Rakitsch, C. Lipprt, O. Stegle, and K. Borgwardt, "A lasso multi-marker mixed model for association mapping with population structure correction," *Bioinformatics*, vol. 29, no. 2, pp. 206–214, 2013.

[12] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the ADMM," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[13] J. P. Cunningham, P. Hennig, and S. Lacoste-Julien, "Gaussian probabilities and expectation propagation," *arXiv preprint arXiv:1111.6832*, 2011.

[14] C. I. Bliss, "The method of probits," *Science*, vol. 79, no. 2037, pp. 38–39, 1934.

[15] L. Fahrmeir, T. Kneib, S. Lang, and B. Marx, *Regression*. Springer, 2013.

[16] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: MIT Press, 2006.

[17] M. P. Berry, C. M. Graham, F. W. McNab, Z. Xu, S. A. Bloch, T. Oni, K. A. Wilkinson, R. Banchereau, J. Skinner, R. J. Wilkinson, *et al.*, "An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis," *Nature*, vol. 466, no. 7309, pp. 973–977, 2010.