Probit Regression with Correlated Label Noise: An EM-EP approach

Stephan Mandt
Institute for Data Sciences and Engineering
Columbia University, USA
sm3976@columbia.edu

Florian Wenzel
Department of Computer Science
HU Berlin, Germany
wenzel@math.hu-berlin.de

John Cunningham
Department of Statistics
Columbia University, USA
jpc2181@columbia.edu

Marius Kloft
Department of Computer Science
HU Berlin, Germany
kloft@hu-berlin.de

Abstract

Probit regression and logistic regression are well-known models for classification. In contrast to logistic regression, probit regression has a canonical generalization that allows us to model correlations between the labels. This is a way to include metadata into the model that correlate the noisy observation process. We show that the approach leads to the mathematical problem of integrating a high-dimensional Gaussian density over the positive orthant. We derive a novel parameter estimation algorithm for this correlated probit regression model. We interpret the noise as a latent variable, which leads to a natural formulation of our algorithm as an expectation-maximization (EM) scheme. Each partial M-step is a gradient step, and we can express the gradient in terms of moments of the truncated multivariate Gaussian. Calculating these moments - the E-step - is expensive using traditional methods. Instead, we use a recent application of expectation propagation (EP) to Gaussian densities. The resulting EM-EP scheme is much faster and thus allows us to treat large data sets.

1 Introduction

Logistic regression is one of the most prominent and widely used classification models in statistics and machine learning [1]. The model assumes that, given n observed inputs $X = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) \in \mathbb{R}^{d\times n}$, the corresponding labels $\boldsymbol{y} = (y_1, \dots, y_n)^\top \in \{-1, +1\}^n$ are realized from n *independent* random variables. In various applications in the sciences and technology, the assumption of independent label variables is violated. In statistical genetics, for example, predictions might be based on a vector $\boldsymbol{x}_i$ of gene expression levels, and the corresponding label variable y_i might indicate the presence of absence of a certain disease. These labels, however, can show dependencies caused by varying experimental conditions and confounding factors such as age, ethnicity, gender, and population structure [2, 3, 4, 5]. We plan to model these effects in terms of correlated label noises. Here we give the technical details for a corresponding algorithm.

Extending the logistic model to correlated multiple dimensions is non-straightforward [6, 7]. In this paper we study the probit regression model, which is closely related to the logistic model. The probit model has a particular advantage over the logistic model: it naturally extends to the scenario of correlated noise. However, although the i.i.d. probit and logistic regression models are standard in statistics and machine learning, the correlated probit model has rarely been analyzed or used in practice [8]. Probit regression with correlated noise is a hard problem, because it is inherently con-

nected to the classical problem in mathematics of integrating n-dimensional multivariate Gaussian distributions over the positive orthant $\mathbb{R}^n_+$. This problem was until recently only solvable for small n. In this paper, we explore a version of expectation propagation [9] to approximate this computation. The proposed approach is much faster than previous approaches based on Genz methods [10].

2 Correlated Probit Regression

Let $X = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$ be the $d \times n$ matrix of data points and $y \in \{+1, -1\}^n$ the observed labels. We assume the following generative process for the vector of random labels $Y = (Y_1, \cdots, Y_n)^\top$:

$$Y = \operatorname{sign}\left(X^\top w + \tilde{\epsilon}\right), \qquad \tilde{\epsilon} \sim \mathcal{N}(0, \tilde{\Sigma}). \tag{1}$$

In this paper we consider the correlation matrix $\tilde{\Sigma}$ as given, but it can be also parametrized and learned from the data. We want to find the vector $w \in \mathbb{R}^d$ according to the maximum likelihood principle. We consider the data as fixed (we always implicitly condition on X). The likelihood function is

$$\mathbb{P}(Y = y|w) \quad = \quad \mathbb{P}(y \circ (X^\top w + \tilde{\epsilon}\,) > 0). \tag{2}$$

The symbol $\circ$ denotes element-wise multiplication with $y_i = \pm 1$. We can introduce a new Gaussian random noise variable $\epsilon = y \circ \tilde{\epsilon}$. The new variable satisfies $\epsilon \sim \mathcal{N}(0, \Sigma)$, where we defined $\Sigma = \operatorname{diag}(y) \cdot \tilde{\Sigma} \cdot \operatorname{diag}(y)$. This simplifies the likelihood,

$$\mathbb{P}(Y = y|w) \;=\; \mathbb{P}(\,\mu(w) + \epsilon > 0) = \int_{\mathbb{R}^n_+} \mathcal{N}(\epsilon; \mu(w), \Sigma)\, d^n\epsilon. \tag{3}$$

We defined $\mu(w) = y \circ X^\top w$ for notational convenience. The likelihood is therefore an integral of the multivariate Gaussian over the positive orthant. Our goal is to minimize the negative log likelihood. To avoid overfitting, we add a quadratic regularizer,

$$\mathcal{L}(w) \quad = \quad -\log \int_{\mathbb{R}^n_+} \mathcal{N}(\epsilon; \mu(w), \Sigma)\, d^n\epsilon \;+\; \frac{1}{2}\lambda w^\top w. \tag{4}$$

In many applications, the dimensionality of the data d is much larger than the number of samples n. We therefore work in the dual kernel representation where we optimize over parameters α_i, where $w = \sum_{i=1}^n \alpha_i \boldsymbol{x}_i$. We furthermore introduce the following probability distribution,

$$p(\epsilon|\mu, \Sigma) \quad = \quad \frac{\mathbb{1}[\epsilon \in \mathbb{R}^n_+]\, \mathcal{N}(\epsilon; \mu, \Sigma)}{\int_{\mathbb{R}^n_+} \mathcal{N}(\epsilon; \mu, \Sigma)\, d^n\epsilon}, \tag{5}$$

where $\mathbb{1}[\cdot]$ is the indicator function. This is just the multivariate Gaussian, truncated and normalized to the positive orthant. For reasons to become clear below, we will refer to it as the *posterior*. The objective function can also be expressed as a function of α,

$$\mathcal{L}(\alpha) \quad = \quad -\log \int_{\mathbb{R}^n_+} \mathcal{N}(\epsilon; \mu(\alpha), \Sigma) \;+\; \frac{1}{2}\lambda \alpha^\top K \alpha, \tag{6}$$

where $K = X^\top X$ is the kernel matrix. The gradient can be easily computed:

$$\nabla_\alpha \mathcal{L}(\alpha) \;=\; -\frac{\int_{\mathbb{R}^n_+} \nabla_\mu^\top \mathcal{N}(\epsilon; \mu, \Sigma)}{\int_{\mathbb{R}^n_+} \mathcal{N}(\epsilon; \mu, \Sigma)} \frac{\partial \mu}{\partial \alpha} + \lambda K \alpha \;=\; -\left[\int_{\mathbb{R}^n} (\epsilon - \mu)^\top p(\epsilon|\mu, \Sigma) d^n\epsilon\right] \frac{\partial \mu}{\partial \alpha} + \lambda K \alpha. \tag{7}$$

This involves the Hessian matrix $\frac{\partial \mu}{\partial \alpha} = \operatorname{diag}(y) \cdot K$, and we used the identity $\nabla_\mu^\top \mathcal{N}(\epsilon; \mu, \Sigma) = (\epsilon - \mu)^\top \mathcal{N}(\epsilon; \mu, \Sigma)$ and the definition of the posterior $p(\epsilon|\mu, \Sigma)$. Note that the gradient involves the first moment of the posterior. This results in the following gradient descent scheme, involving the learning rate ρ_t:

$$\begin{aligned} \alpha_{t+1} \quad &= \quad \alpha_t - \rho_t \nabla_\alpha \mathcal{L}(\alpha_t), \\ \nabla_\alpha \mathcal{L}(\alpha_t) \quad &= \quad -\left(\mathbb{E}_{p(\epsilon|\mu(\alpha_t), \Sigma)}[\epsilon] - \mu(\alpha_t)\right)^\top \frac{\partial \mu}{\partial \alpha} \;+ \lambda K \alpha_t. \end{aligned} \tag{8}$$

Latent variable model and EM algorithm Our model can be interpreted as a latent variable model. Let us consider w and ϵ as latent variables. Let $p(w) = \mathcal{N}(0, 2\mathbf{I}/\lambda)$ be a prior for the vectors w, where $\mathbf{I}$ is the identity matrix. We also introduce a prior on the latent variables $\tilde{\epsilon}$ according to $p(\tilde{\epsilon}) = \mathcal{N}(0, \tilde{\Sigma})$. As our model assumes that $Y|w = \mathrm{sign}(X^\top w + \tilde{\epsilon})$, we have

$$p(Y|w, \tilde{\epsilon}) \quad = \quad \mathbb{1}[Y = \mathrm{sign}(\mathrm{X}^\top \mathrm{w} + \tilde{\epsilon})]. \tag{9}$$

Now we observe $Y = y$ and condition on y. We want to calculate a MAP estimate of the distribution $p(w|y) = p(y|w)p(w)/p(y)$. Because $p(y|w) = \int p(y|w, \tilde{\epsilon})p(\tilde{\epsilon})d\tilde{\epsilon}$, applying logarithms yields

$$\log p(w|y) \quad = \quad \log \int p(y|w, \tilde{\epsilon})p(\tilde{\epsilon})d\tilde{\epsilon} + \log p(w) - \log p(y). \tag{10}$$

We now define $\mathcal{L}(w) = -\log p(w|y)$ as our objective and simplify:

$$\begin{aligned}
\mathcal{L}(w) &= -\log \int \mathbb{1}[y = \mathrm{sign}(\mathrm{X}^\top \mathrm{w} + \tilde{\epsilon})]\, \mathcal{N}(\tilde{\epsilon}; 0, \tilde{\Sigma}) d\tilde{\epsilon} \;+\; \frac{1}{2}\lambda w^\top w \;+\; \log p(y) \\
&= -\log \int \mathbb{1}[\mathrm{sign}(\epsilon) = 1]\, \mathcal{N}(\epsilon; y \circ X^\top w, \Sigma) d\epsilon \;+\; \frac{1}{2}\lambda w^\top w \;+\; \log p(y) \\
&= -\log \int_{\mathbb{R}^n_+} \mathcal{N}(\epsilon; y \circ X^\top w, \Sigma) d\epsilon \;+\; \frac{1}{2}\lambda w^\top w \;+\; \log p(y).
\end{aligned} \tag{11}$$

In the second line, we again substituted $\epsilon = y \circ \tilde{\epsilon}$ and also shifted the integration variable. The final result is exactly (up to a constant) what we obtained earlier. The formulation as a latent variable model has the advantage that it allows us to formulate an EM algorithm. This algorithm is constructed according to the standard scheme,

$$\begin{aligned}
\mathrm{E-step}: \quad Q(w, w_t) &= \mathbb{E}_{p(\epsilon|y,w_t)}[\log p(y, w, \epsilon)] \\
\mathrm{M-step}: \quad w_{t+1} &= w_t + \rho \nabla_w Q(w, w_t)|_{w=w_t}.
\end{aligned} \tag{12}$$

One can show that this algorithm is exactly the same as our previous one (Eq. 8, but in w-space). To show the equivalence, we proved that $\nabla\mathcal{L}(w) = -\nabla_{w'} Q(w', w)|_{w=w'}$ (not shown here). Note that in slight deviation from usual EM, we use a partial maximization as an M-step (a gradient step).

From an EM to an EM-EP algorithm. Without further approximation, the EM algorithm is still not tractable on large data sets: it involves the first moment of the posterior in each gradient step. We therefore use a recent application of expectation propagation (EP) [9] to carry out the E-step approximately. This leads to an EM-EP algorithm. In a different context, such a combined EM-EP algorithm has been suggested in [11].

EP approximates moments of the posterior $p(\epsilon|\mu, \Sigma)$ in terms of a variational distribution $q(\epsilon)$, approximately minimizing the Kullback-Leibler divergence,

$$q(\epsilon) \quad = \quad \arg\min_q \left(\int_{\mathbb{R}^d} p(\epsilon|\mu, \Sigma) \log p(\epsilon|\mu, \Sigma) - \int_{\mathbb{R}^d} p(\epsilon|\mu, \Sigma) \log q(\epsilon) \right). \tag{13}$$

$q(\epsilon)$ is another Gaussian, characterized by the variational parameters μ_q and Σ_q:

$$q(\epsilon; \mu_q, \Sigma_q) \quad = \quad \mathcal{N}(\epsilon; \mu_q, \Sigma_q). \tag{14}$$

We approximate the mean of the posterior p in terms of the variational distribution,

$$\mathbb{E}_{p(\epsilon|\mu,\Sigma)}[\epsilon] \;\approx\; \mathbb{E}_{q(\epsilon)}[\epsilon] \;=\; \mu_q. \tag{15}$$

EM-EP does sequential updates on the variational parameters by iterating over the coordinates. We call N the number of coordinate loops of the algorithm. Typically, N is unconstrained; the algorithm is run until convergence. We will later study the properties of our algorithm where we fix N.

Warm-starting the algorithm with the variational distribution of the previous step makes the number of necessary loops N considerably smaller. We therefore pass the parameters of the previous variational distribution as an argument to EM-EP,

$$q^{\mathrm{new}}(\epsilon) \leftarrow \text{EM-EP}\,(\mu, \Sigma, q(\epsilon)^{\mathrm{old}}). \tag{16}$$

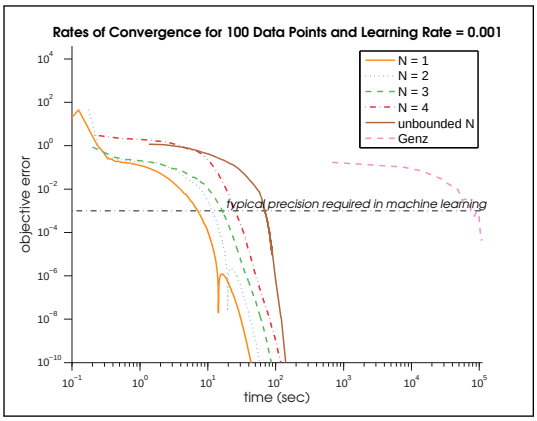

Figure 1: Objective error of EM-EP. We compare the unconstrained algorithm to versions where we constrain N. We argue that running EP until convergence is wasteful for small learning rates ρ.

3 Empirical Results

We generated 100 data points from our model. The correlation matrix was chosen to be a weighted sum of the unit matrix and a random positive definite symmetric matrix. We tested EM-EP on this artificial data set. Fig. 1 shows the objective error of this optimization process. For a fixed learning rate of $\rho = 0.001$, different versions of the algorithm are shown: we compared the unconstrained algorithm against a version where we fixed N.

Our experimental findings suggest that running EP until convergence in each gradient step is wasteful. For small learning constant rates ρ, it is better to run a single loop of coordinate updates in EM-EP after each gradient step, i.e. $N = 1$. As a benchmark, we also compare our method to a gradient descent scheme where the integral is computed using the Genz method, and the gradient is computed numerically. EM-EP is about 3 orders of magnitude faster in this experiment.

The learning curves for $N = 1$ and $N = 2$ show an interesting feature, namely the presence of spikes. Note that the optimization problem is convex, and hence the spikes cannot be associated with real local optima. Also, note that the algorithm gets only stuck temporarily. We explain this effect as follows. The algorithm can only converge to fake local optima if the gradient is computed not precisely enough. When the computed gradient is zero, the algorithm stops moving in parameter space. Due to the warm start of EP we keep running multiple loops of EM-EP for the same value of μ. This, in turn, will lead to a better and better computation of the local gradient, and finally it will deviate from zero, as we are not in a true optimum of the objective.

To conclude, we found that our learning algorithm using expectation propagation is fast and is likely to be applicable to large datasets of several thousands of data points, which we plan to test in the future.

4 Conclusion

We presented a new algorithm that solves the problem of probit regression with correlated noise. We used variational techniques to approximately compute the moments of a multivariate Gaussian distribution, truncated to the positive orthant. Our algorithm has a natural interpretation as an EM-EP algorithm when considering the correlated noise as a latent variable. In the future, we plan to apply the method to large biological data sets.

Acknowledgments

We thank Manfred Opper, Shinichi Nakajima, Tian Jiang, Mehryar Mohri, David Blei and Gunnar Rätsch for stimulating discussions. SM acknowledges the support of the U.S. National Science Foundation I2CAM International Materials Institute Award, Grant DMR-0844115.

References

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

[2] 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.

[3] J. Listgarten, C. Lippert, and D. Heckerman, "FaST-LMM-Select for addressing confounding from spatial structure and rare variants," *Nature Genetics*, pp. 470–471, April 2013.

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

[5] N. Fusi, O. Stegle, and N. D. Lawrence, "Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies," *PLoS computational biology*, vol. 8, no. 1, p. e1002330, 2012.

[6] A. Ragab, "On multivariate generalized logistic distribution," *Microelectronics Reliability*, vol. 31, no. 2, pp. 511–519, 1991.

[7] H. J. Malik and B. Abraham, "Multivariate logistic distributions," *The Annals of Statistics*, vol. 1, pp. 588–590, 05 1973.

[8] Y. Ochi and R. L. Prentice, "Likelihood inference in a correlated probit regression model," *Biometrika*, vol. 71, no. 3, pp. 531–543, 1984.

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

[10] A. Genz and F. Bretz, "Numerical computation of multivariate t-probabilities with application to power calculation of multiple contrasts," *Journal of Statistical Computation and Simulation*, vol. 63, no. 4, pp. 103–117, 1999.

[11] H.-C. Kim and Z. Ghahramani, "Bayesian gaussian process classification with the em-ep algorithm," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 12, pp. 1948–1959, 2006.