JEFFREYS PRIOR REGULARIZATION FOR LOGISTIC REGRESSION

Tam Nguyen, Raviv Raich, and Phung Lai

School of EECS, Oregon State University, Corvallis, OR, 97331-5501, USA
nguyeta4@oregonstate.edu, raich@eecs.oregonstate.edu, laith@oregonstate.edu

ABSTRACT

Logistic regression is a statistical model widely used for solving classification problems. Maximum likelihood is used train the model parameters. When data from two classes is linearly separable, maximum likelihood is ill-posed. To address this problem as well as to handle over-fitting issues, regularization is commonly considered. A regularization coefficient is used to control the tradeoff between model complexity and data fit and cross-validation is applied to determine the coefficient. In this paper, we develop a regularization framework for logistic regression using Jeffreys prior, which is free of any tuning parameters. Our experiments show that the proposed regularization outperforms other well-known regularization approaches.

Index Terms— Logistic Regression, Jeffreys Prior, Fisher Information Matrix, Regularization, Maximum A-Posteriori

1. INTRODUCTION

Logistic regression (LR) is the statistical model used in classification. Due to its conditional form, it is viewed as discriminative, i.e, it is used to model label dependence on features rather than feature distribution. Despite the many alternatives, LR is still frequently used in data science as stand-alone or as a building block in more complex probabilistic graphical models due of its efficiency and simplicity. As with other classification algorithms, training a LR model can become ill-posed when no regularization is considered.

Several regularization methods were developed to handle over-fitting in the logistic regression model. A Gaussian prior or l_2-regularization [1–4] encourages a small sum of the squares of the parameters and promotes a smooth parameter vector. A Laplacian prior or l_1-regularization alternative [3–6] employs a penalty term which encourages the sum of the absolute values of the parameters to be small. It has frequently been observed that l_1-regularization promotes sparseness in the parameter vector. These methods depend on a regularization parameter [7] that controls the trade-off between model complexity and data fit. Larger value of the regularization parameter limits over-fitting but introduces greater bias. Small value of the regularization parameter leads to a better data fit at the expense of over-fitting. To determine an appropriate value of the regularization parameter, cross-validation is considered. To eliminate the need for parameter tuning, [7] proposes to marginalize of the the scale parameter of a Laplacian prior using the non-informative Jeffreys prior on the parameter.

In this paper, we propose a tuning-free approach for regularization in LR using the Jeffreys prior. We obtain an explicit expression for the Fisher information matrix (FIM) for the LR model under the assumption of Gaussian distributed observations that depends only on the model parameters. We proceed with the derivation of the maximum-a-posteriori (MAP) procedure under the assumption that the model parameters follow Jeffreys prior. Numerical experiments that compare the proposed approach with l_1 and l_2 regularization tuned with cross-validation demonstrate that our approach is not only computationally efficient since no parameter tuning is required but is also highly competitive in terms of estimation accuracy.

The rest of this paper is organized as follows. Section 2 discusses previous related work. Section 3 reviews background on the LR model, presents a novel derivation of the FIM and CRLB under the assumption of Gaussian distributed observations, and introduces the Jeffreys prior based MAP estimation procedure. Experimental results and discussion are provided in Section 4. Finally, Section 5 summarizes our contribution.

2. RELATED WORK

Regularization for the LR model has been explored in prior work. In [8], the problem of estimating a graph structure with a discrete Markov random field for high-dimensional data is addressed. For each node, the neighborhood is estimated using an l_1-regularized LR. In [6], Lasso penalized (i.e., l_1-regularized) logistic regression is deployed for in case-control disease gene mapping with a large number of single nucleotide polymorphisms predictors which exceeds the number of observations. In other work, a fast hybrid algorithm was integrated into l_1-regularized LR to reduce computation time for the large-scale high dimensional data [9]. The l_2-regularized LR is adopted in [2] to lean gene-gene and gene-environment interaction models. An analysis of l_1- and l_2-regularized LR in the presence of many irrelevant features was presented in [3] illustrating that for l_2-regularization, sample complexity grows logarithmically in the number of irrelevant features. In [7], a Bayesian approach for regularization eliminates the regularization parameter entirely by marginalizing a Laplace prior using an uninformative Jeffreys prior. Elastic net, a combination of l_1 and l_2 approach for regularization of the LR model is presented in [10].

3. PROPOSED METHOD

In the following, we present a review of the logistic regression model, maximum likelihood (ML) parameter estimation, and the FIM for the model. Additionally, we present a novel closed-form expression for the FIM and the Cramér-Rao lower bound (CRLB) under the assumption of zero-mean Gaussian feature distribution with homoscedastic covariance. Based on the FIM, we derive Jeffreys prior and present the MAP estimation procedure using the proposed prior.

This work is partially supported by the National Science Foundation grant CCF-1254218.

978-1-4673-7802-4/16/$31.00 ©2016 IEEE

3.1. Logistic Regression

The logistic regression model is commonly used as means of training a linear classifier given a set of observations $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ where $\mathbf{x}_i$ denotes the d-dimensional ith feature vector and $y_i \in \{0, 1\}$ denotes its label. Assume $(\mathbf{x}_i, y_i)$ for $i = 1, 2, \ldots, n$ are $i.i.d$ measurements. The ith measurement $(\mathbf{x}_i, y_i)$ is drawn from the joint distribution

$$p(\mathbf{x}, y|\mathbf{w}) = p(y|\mathbf{x}; \mathbf{w}) f(\mathbf{x}), \tag{1}$$

where $f(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$ is an arbitrary distribution and $p(y|\mathbf{x}; \mathbf{w})$ is the logistic regression probability model given by

$$p(y|\mathbf{x}; \mathbf{w}) = \frac{e^{(\mathbf{w}^\top \mathbf{x}+\beta)y}}{1 + e^{\mathbf{w}^\top \mathbf{x}+\beta}}, \tag{2}$$

where $\mathbf{w} \in \mathbb{R}^d$ is an unknown parameter vector. Using the $i.i.d$ assumption and the joint distribution of $(\mathbf{x}_i, y_i)$ in (1)-(2), the joint probability model of the observations is

$$p(\mathbf{X}, \mathbf{y}|\mathbf{w}) = \prod_{i=1}^{n} \frac{e^{(\mathbf{w}^\top \mathbf{x}_i+\beta)y_i}}{1 + e^{\mathbf{w}^\top \mathbf{x}_i+\beta}} f(\mathbf{x}_i), \tag{3}$$

where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$ denotes the collection of all feature vectors and $\mathbf{y} = [y_1, y_2, \ldots, y_n]^\top$ denotes the collection of all instance labels. Consequently, the log-likelihood for the model is given by:

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left[y_i(\mathbf{w}^\top \mathbf{x}_i + \beta) - \log(1 + e^{\mathbf{w}^\top \mathbf{x}_i+\beta}) \right] + c(\boldsymbol{\phi}), \tag{4}$$

where $\boldsymbol{\theta} = [\mathbf{w}^\top, \beta]^\top$ and $c(\boldsymbol{\phi}) = \sum_i \log f(\mathbf{x}_i|\boldsymbol{\phi})$ where $\boldsymbol{\phi}$ is the parameter vector of $f(\mathbf{x})$. Note that the ML estimation of $\boldsymbol{\theta}$ does not require $f(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n|\boldsymbol{\phi})$ and can be replaced with the maximization of the conditional log-likelihood $\log p(\mathbf{y}|\mathbf{X})$, i.e., the first term on the RHS of (4). Estimating $\boldsymbol{\phi}$ can be done based on the $\mathbf{x}_i$'s only and will not affect the estimation of $\boldsymbol{\theta}$. For some settings, regularization is critical [5]. As discussed in Section 2, l_2- [1] and l_1-regularization [5] terms are routinely considered for training a LR model. In this paper, we are interested in deriving a tuning-free alternative using Jeffreys prior. We proceed with the derivation of the FIM and the CRLB.

3.2. CRLB Analysis

Following a similar derivation as in [11], we compute the FIM using $\text{FIM} = -\mathbb{E}[\frac{d^2 l(\boldsymbol{\theta})}{d\boldsymbol{\theta} d\boldsymbol{\theta}^\top}]$ and obtain

$$\text{FIM} = n \begin{bmatrix} \mathbb{E}\left[\frac{e^{\mathbf{w}^\top \mathbf{x}+\beta}}{1+(e^{\mathbf{w}^\top \mathbf{x}+\beta})^2} \mathbf{x}\mathbf{x}^\top\right] & \mathbb{E}\left[\frac{e^{\mathbf{w}^\top \mathbf{x}+\beta}}{1+(e^{\mathbf{w}^\top \mathbf{x}+\beta})^2} \mathbf{x}\right] \\ \mathbb{E}\left[\frac{e^{\mathbf{w}^\top \mathbf{x}+\beta}}{1+(e^{\mathbf{w}^\top \mathbf{x}+\beta})^2} \mathbf{x}^\top\right] & \mathbb{E}\left[\frac{e^{\mathbf{w}^\top \mathbf{x}+\beta}}{1+(e^{\mathbf{w}^\top \mathbf{x}+\beta})^2}\right] \end{bmatrix}. \tag{5}$$

A closed-form expression for FIM for arbitrary $f(\mathbf{x})$ is hard to find and is commonly replaced with an empirical evaluation of the expectation in (5) by setting $f(\mathbf{x}) = \frac{1}{T}\sum_{t=1}^{T} \delta(\mathbf{x}-\mathbf{x}_t)$, where $\mathbf{x}_t \sim f(\mathbf{x})$ [11]. This approach limits the usefulness of the FIM for further applications. Additionally, if the dimension of $\mathbf{x}_i$, d is larger than T, then the approximation of the FIM is singular and its determinant is zero [11]. To derive a closed-form FIM, which is non-singular, we assume $\mathbf{x} \sim N(0, \sigma^2\mathbf{I})$. The FIM for this case is

$$\text{FIM} = n \begin{bmatrix} \sigma^2[(\alpha_2 - \alpha_0)\mathbf{u}_1\mathbf{u}_1^\top + \alpha_0\mathbf{I}] & \sigma\mathbf{u}_1\alpha_1 \\ \sigma\alpha_1\mathbf{u}_1^\top & \alpha_0 \end{bmatrix}, \tag{6}$$

where $\mathbf{u}_1 = \mathbf{w}/\|\mathbf{w}\|$, $\alpha_k = \alpha_k(\sigma\|\mathbf{w}\|, \beta)$ and

$$\alpha_k(a, b) = \mathbb{E}\left[\frac{e^{az+b}}{(1 + e^{az+b})^2} z^k\right], \text{ with } z \sim \mathcal{N}(0, 1). \tag{7}$$

We use $\|\cdot\|$ to denote the l_2-norm $\|\cdot\|_2$. The detailed derivation is provided in the Appendix. Using the matrix inversion lemma, the CRLB given by $\text{CRLB} = \text{FIM}^{-1}$ can be expressed as

$$\frac{1}{n} \begin{bmatrix} \frac{1}{\sigma^2}\frac{1}{\alpha_0}[\mathbf{I} - \frac{\alpha_2\alpha_0 - \alpha_0^2 - \alpha_1^2}{\alpha_2\alpha_0 - \alpha_1^2}\mathbf{u}_1\mathbf{u}_1^\top] & -\frac{1}{\sigma}\frac{\alpha_1}{\alpha_0}[1 - \frac{\alpha_2\alpha_0 - \alpha_0^2 - \alpha_1^2}{\alpha_2\alpha_0 - \alpha_1^2}]\mathbf{u}_1 \\ -\frac{1}{\sigma}\frac{\alpha_1}{\alpha_0}\mathbf{u}_1^\top[1 - \frac{\alpha_2\alpha_0 - \alpha_0^2 - \alpha_1^2}{\alpha_2\alpha_0 - \alpha_1^2}] & [\frac{\alpha_2}{\alpha_2\alpha_0 - \alpha_1^2}] \end{bmatrix}. \tag{8}$$

We can use the CRLB to bound the mean squared error (MSE) of unbiased estimation of $[\mathbf{w}^\top, \beta]^\top$:

$$\begin{aligned} \mathbb{E}(\|\hat{\mathbf{w}} - \mathbf{w}\|^2) &\geq \sum_{i=1}^{d} \text{CRLB}_{ii} = \frac{1}{n\sigma^2\alpha_0}(d - 1 + \frac{\alpha_0^2}{\alpha_2\alpha_0 - \alpha_1^2}), \\ \mathbb{E}(\|\hat{\beta} - \beta\|^2) &\geq \text{CRLB}_{d+1,d+1} = \frac{1}{n}\frac{\alpha_2}{\alpha_2\alpha_0 - \alpha_1^2}. \end{aligned} \tag{9}$$

Based on the FIM in (6), we derive Jeffreys prior for MAP estimation for LR.

3.3. Jeffreys Prior for MAP Estimation

In MAP estimation, the model parameters are obtained by solving

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}). \tag{10}$$

Our goal is to solve the maximization in (10) when Jeffreys prior is considered. According to Jeffreys prior, the probability of the prior is proportional to the square root of the determinant of the FIM:

$$p(\boldsymbol{\theta}) \propto \sqrt{\det(\text{FIM})}. \tag{11}$$

The determinant of the FIM of the LR model in (6) is given by:

$$\det(\text{FIM}) = \sqrt{(n\sigma^2)^d \alpha_0^{d-1}(\alpha_2\alpha_0 - \alpha_1^2)}. \tag{12}$$

This result can be obtained using the property of the determinant. Let $\text{FIM} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}$ where $\mathbf{D}$ is a 1×1 matrix, $\mathbf{B}$ and $\mathbf{C}$ are column and row vectors respectively. The determinant of FIM is computed as follows:

$$\det(\text{FIM}) = (\mathbf{D} - 1)\det(\mathbf{A}) + \det(\mathbf{A} - \mathbf{B}\mathbf{C}). \tag{13}$$

Substituting (12) into (10) and removing terms which are independent of $\boldsymbol{\theta}$, we obtain:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \log l(\boldsymbol{\theta}) + \frac{1}{2}[(d - 1)\log\alpha_0 + \log(\alpha_2\alpha_0 - \alpha_1^2)]. \tag{14}$$

As can be seen in (7), α_0, α_1 and α_2 depend on 3 parameters: $\sigma, \|\mathbf{w}\|, \beta$. Given X, σ^2 is estimated using $\hat{\sigma}^2 = \frac{1}{nd}\sum_{i=1}^{n}\|\mathbf{x}_i\|^2$. Since σ^2 is already determined, (14) is solved for $\mathbf{w}$ and β only. To that end, we adopt a Gradient ascent approach [12]:

$$\begin{aligned} \hat{\mathbf{w}}^{new} &= \hat{\mathbf{w}}^{old} + \eta\frac{\partial \log p(\mathbf{X}, \mathbf{y}|\boldsymbol{\theta})}{\partial \mathbf{w}}, \\ \hat{\beta}^{new} &= \hat{\beta}^{old} + \eta\frac{\partial \log p(\mathbf{X}, \mathbf{y}|\boldsymbol{\theta})}{\partial \beta}. \end{aligned} \tag{15}$$

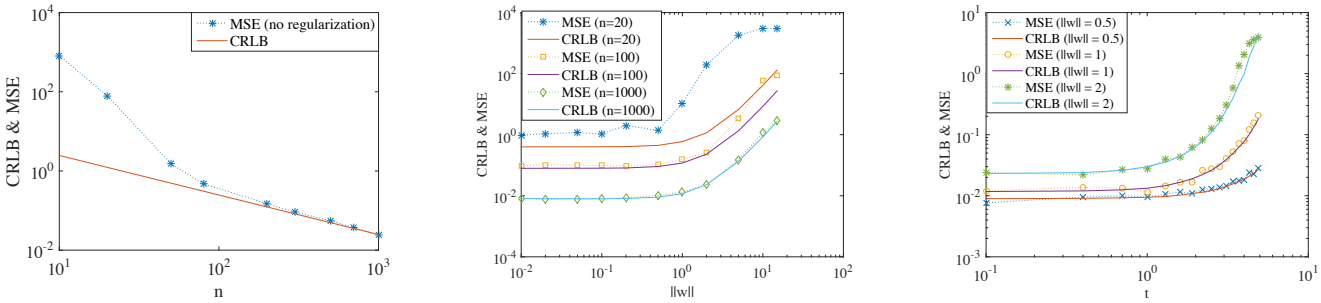

Fig. 1: CRLB and MSE as a function of (a) n for $\mathbf{w} = 5/\sqrt{2}.[1,1]^\top$ and $\beta = 1$, (b) $\|\mathbf{w}\|$ for $\beta = 1$ and $n \in [20, 100, 1000]$, and (c) t ($\beta = \sigma\|\mathbf{w}\|t$) for $n = 1000$ and $\|\mathbf{w}\| \in [0.5, 1, 2]$.

According to (14), the derivative of the $\log p(\mathbf{X}, \mathbf{y}|\boldsymbol{\theta})$ is

$$\frac{d \log p(\mathbf{X}, \mathbf{y}|\boldsymbol{\theta})}{d\boldsymbol{\theta}} = \frac{d \log l(\boldsymbol{\theta})}{d\boldsymbol{\theta}} + \frac{d \log p(\boldsymbol{\theta})}{d\boldsymbol{\theta}} \tag{16}$$

where the derivative of $\log l(\boldsymbol{\theta})$ w.r.t. $\mathbf{w}$ and β is obtained as in [13] with:

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \mathbf{w}} = \sum_{i=1}^{n} [y_i \mathbf{x}_i - \frac{\mathbf{x}_i}{1 + e^{\mathbf{w}^\top \mathbf{x}_i + \beta}}], \frac{\partial l(\boldsymbol{\theta})}{\partial \beta} = \sum_{i=1}^{n} \left[y_i - \frac{1}{1 + e^{\mathbf{w}^\top \mathbf{x}_i + \beta}} \right], \tag{17}$$

and the derivative of $\log p(\boldsymbol{\theta})$ w.r.t. $\mathbf{w}$ and β is given by

$$\begin{aligned} \frac{\partial \log p(\boldsymbol{\theta})}{\partial \mathbf{w}} &= \frac{1}{2} \left[(d-1) \frac{\frac{\partial \alpha_0}{\partial \|\mathbf{w}\|}}{\alpha_0} + \frac{\frac{\partial \alpha_2}{\partial \|\mathbf{w}\|}\alpha_0 + \alpha_2 \frac{\partial \alpha_0}{\partial \|\mathbf{w}\|} - 2\alpha_1 \frac{\partial \alpha_1}{\partial \|\mathbf{w}\|}}{\alpha_2 \alpha_0 - \alpha_1^2} \right] \frac{\mathbf{w}}{\|\mathbf{w}\|}, \\ \frac{\partial \log p(\boldsymbol{\theta})}{\partial \beta} &= \frac{1}{2} \left[(d-1) \frac{\frac{\partial \alpha_0}{\partial \beta}}{\alpha_0} + \frac{\frac{\partial \alpha_2}{\partial \beta}\alpha_0 + \alpha_2 \frac{\partial \alpha_0}{\partial \beta} - 2\alpha_1 \frac{\partial \alpha_1}{\partial \beta}}{\alpha_2 \alpha_0 - \alpha_1^2} \right], \end{aligned} \tag{18}$$

where $\alpha_k = \alpha_k(\sigma\|\mathbf{w}\|, \beta)$ as in (7) and

$$\frac{\partial \alpha_k}{\partial \|\mathbf{w}\|} = \sigma \mathbb{E}\left[z^{k+1}(p - 3p^2 + 2p^3) \right], \frac{\partial \alpha_k}{\partial \beta} = \mathbb{E}\left[z^k (p - 3p^2 + 2p^3) \right], \tag{19}$$

with $p = \frac{e^{\|\mathbf{w}\|\sigma z + \beta}}{1 + e^{\|\mathbf{w}\|\sigma z + \beta}}$.
From (18) and (19), the derivative of $\log p(\boldsymbol{\theta})$ depends on the α_k's and their derivatives which are in a univariate integral forms.

4. EXPERIMENTAL RESULTS

In this section, we provide numerical simulations to (i) evaluate our expression for the CRLB by comparison to the MSE of ML estimator across a range of parameter values and (ii) to analyze the performance of the Jeffreys Prior MAP estimator and compare to the two well-known regularization methods l_2 and l_1.

General settings We conduct Monte-Carlo simulations to compute the MSE of all estimators under consideration. For each of the 200 Monte-Carlo runs, we generate n feature vectors $\mathbf{x} \sim N(0, \sigma^2 \mathbf{I})$ with $\sigma^2 = 1$ and their corresponding label $y|\mathbf{x}$ according the model in (2). Values of n, $\|\mathbf{w}\|$, and β are provided below. We compute the CRLB using (9).

4.1. CRLB Evaluation

To evaluate the CRLB, we change the values of n, $\|\mathbf{w}\|$, and β in turns and compare the CRLB to the MSE of ML estimators.

4.1.1. CRLB and MSE as a function of the data size n

To examine the dependence of the CRLB and the MSE on the number of data points, n, we change n in $[10, 20, 50, 80, 200, 300, 500, 700, 1000]$ and $\mathbf{w} = 5/\sqrt{2} \cdot [1,1]^\top$ and $\beta = 1$.

Figure 1(a) shows a decreasing trend for the CRLB and the MSE as a function of n in the range of 10 to 1000. Starting with a small value of n, the gap between CRLB and MSE is quite large. As n increases, the gap shrinks and MSE reaches the CRLB at $n \approx 200$. As expected, the ML estimator becomes efficient (i.e., its MSE approaches the CRLB) for sufficiently large n.

4.1.2. CRLB and MSE as a function of $\|\mathbf{w}\|$

In this scenario, we change the norm of $\mathbf{w}$ increasingly in $[0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10, 15]$ and the data size $n \in [20, 100, 1000]$ and fix β to 1. The purpose of this scheme is to investigate the effect of the sharpness of classifier (as represented by $\|\mathbf{w}\|$) on the MSE of ML estimator.

The result are demonstrated in the Fig. 1(b). The figure shows the MSE of the ML estimator and the CRLB as function of $\|\mathbf{w}\| \in [0.01, 15]$ using a logarithmic scale for both MSE and $\|\mathbf{w}\|$. We observe that as we increase the value of $\|\mathbf{w}\|$, both CRLB and MSE increase. By comparing the curves that are associated with different values of n, we observe that an increase in n decreases both MSE and CRLB. Additionally, we observe a gap between the MSE and the CRLB for a large value of $\|\mathbf{w}\|$. In general, to achieve a reasonable accuracy in estimating the parameters of the LR model when $\|\mathbf{w}\|$ is large, the data size n should be increased significantly.

4.1.3. CRLB and MSE as a function of β

To evaluate the relation between the CRLB and MSE and the value of the bias term β, we fix the data size to $n = 1000$. The bias term β is modified as a function $\mathbf{w}$ according to $\beta = \sigma\|\mathbf{w}\|t$ with $t \in [10^{-1}, 5]$ for the three value of $\|\mathbf{w}\|$ in $[0.5, 1, 2]$.

Figure 1(c) shows the agreement between CRLB and MSE for varying values of the bias term β and $\|\mathbf{w}\|$. For each of the three values of $\|\mathbf{w}\|$, a solid line is shown for the CRLB and a dashed line for

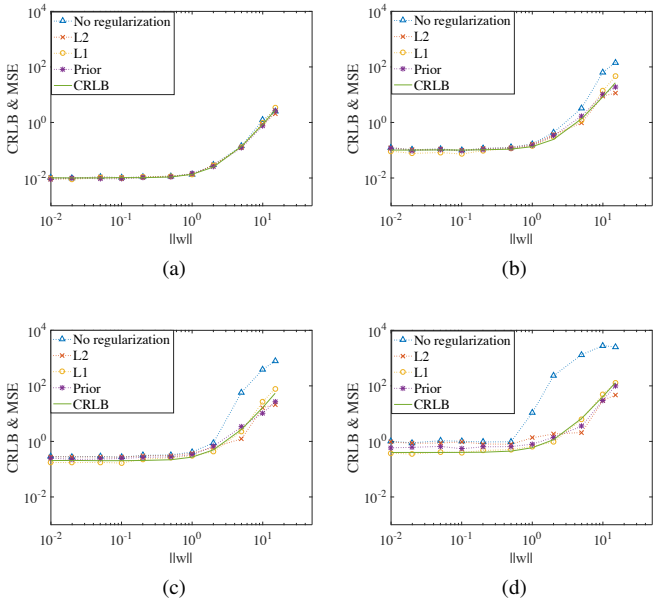

Fig. 2: CRLB and MSE of ML, l_1-regularized ML, l_2-regularized ML and Jeffreys prior MAP with $d = 2$ for (a) $n = 1000$, (b) $n = 100$, (c) $n = 50$, and (d) $n = 20$.

the MSE. We observe that both MSE and CRLB are monotonically increasing in both t (or β) and $\|\mathbf{w}\|$.

4.2. CRLB and MSE for MAP Estimation

In this section, we evaluate our tuning-free regularization approach and compare it with the l_1- and l_2-regularization approaches over the same setting as in Section 4.1.2 with the exception of $n \in [20, 50, 100, 1000]$. For l_1, we add $\lambda\|\mathbf{w}\|_1$ to $l(\boldsymbol{\theta})$ in (4) and for l_2, we add $\lambda\|w\|_2^2$. For each of the regularization approaches, the parameter λ is determined by a ten-fold cross-validation over the grid of $\lambda = [0, 0.01, 0.02, 0.025, 0.05, 0.07, 0.1, 0.3, 0.5, 0.7, 1, 5, 10]$.

Figure 2 presents the CRLB and MSE as a function of $\|\mathbf{w}\|$ when the dimension of data is $d = 2$. With a large data size $n = 1000$, there is no gap between the MSE and CRLB. The MSE of the three regularization methods agrees with the MSE of the ML estimator and with the CRLB. When the data size n is 20, the MSE of the ML estimator increases rapidly with $\|\mathbf{w}\|$ forming a large gap with the CRLB. However, the l_1- and l_2- regularized ML estimators and the proposed Jeffreys prior MAP estimator yield a lower MSE than that of the ML estimator that approaches the CRLB.

5. CONCLUSION

In this paper, we introduced a tuning-free method for regularizing the likelihood of the logistic regression model using Jeffreys prior. A novel derivation of the FIM and CRLB under the assumption of Gaussian distributed observations is presented, and MAP estimation which relies on the Jeffreys prior is considered. As compared to the l_1 and l_2 regularized ML, the Jeffreys prior approach requires no tuning and therefore is computationally more efficient. Experimental results illustrate that the proposed approach is on par with l_1 and l_2 regularized ML.

6. APPENDIX

6.1. Second derivative

The second derivative of log-likelihood function in (4) w.r.t. $\mathbf{w}$ and β is given by

$$-\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \mathbf{w}\mathbf{w}^\top} = -\sum_{i=1}^{n} \frac{e^{\mathbf{w}^\top \mathbf{x}_i+\beta}}{(1+e^{\mathbf{w}^\top \mathbf{x}_i+\beta})^2}\mathbf{x}_i\mathbf{x}_i^\top, \tag{20}$$

$$-\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \mathbf{w}\partial \beta} = -\sum_{i=1}^{n} \frac{e^{\mathbf{w}^\top \mathbf{x}_i+\beta}}{(1+e^{\mathbf{w}^\top \mathbf{x}_i+\beta})^2}\mathbf{x}_i, \tag{21}$$

$$-\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \beta^2} = -\sum_{i=1}^{n} \frac{e^{\mathbf{w}^\top \mathbf{x}_i+\beta}}{(1+e^{\mathbf{w}^\top \mathbf{x}_i+\beta})^2}. \tag{22}$$

Substituting (20)-(22) into FIM $= -\mathbb{E}[\frac{d^2 l(\boldsymbol{\theta})}{d\boldsymbol{\theta} d\boldsymbol{\theta}^\top}]$, we obtain (5).

6.2. FIM transformation

Let $\mathbf{u}_1 = \frac{\mathbf{w}}{\|\mathbf{w}\|}$, $c = \|\mathbf{w}\|$, and $\mathbf{U} = [\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_d}]$ where $\mathbf{u}_2, \mathbf{u}_3, \ldots, \mathbf{u}_d$ can be arbitrarily chosen to satisfy $\mathbf{U}\mathbf{U}^\top = \mathbf{I}$. We begin by multiplying $\mathbf{U}\mathbf{U}^\top$ on both size of the top-left term of the FIM, on the left of the top-right term, and the right of the bottom-left term.

$$\text{FIM} = n\begin{bmatrix} \mathbf{U}\mathbb{E}\left[\frac{e^{c\mathbf{u}_1^\top \mathbf{x}+\beta}}{(1+e^{c\mathbf{u}_1^\top \mathbf{x}+\beta})^2}\mathbf{U}^\top \mathbf{x}\mathbf{x}^\top \mathbf{U}\right]\mathbf{U}^\top & \mathbf{U}\mathbb{E}\left[\frac{e^{c\mathbf{u}_1^\top \mathbf{x}+\beta}}{(1+e^{c\mathbf{u}_1^\top \mathbf{x}+\beta})^2}\mathbf{U}^\top \mathbf{x}\right] \\ \mathbb{E}\left[\frac{e^{c\mathbf{u}_1^\top \mathbf{x}+\beta}}{(1+e^{c\mathbf{u}_1^\top \mathbf{x}+\beta})^2}\mathbf{x}^\top \mathbf{U}\right]\mathbf{U}^\top & \mathbb{E}\left[\frac{e^{c\mathbf{u}_1^\top \mathbf{x}+\beta}}{(1+e^{c\mathbf{u}_1^\top \mathbf{x}+\beta})^2}\right] \end{bmatrix} \tag{23}$$

Let $\mathbf{z} = \mathbf{U}^\top \mathbf{x}/\sigma$ and note that $\mathbf{z} \sim \mathcal{N}(0, I_d)$. Replacing $\mathbf{U}^\top \mathbf{x}$ with $\sigma\mathbf{z}$ and $\mathbf{u}_1^\top \mathbf{x} = \sigma\mathbf{z}_1$ into (23), we obtain

$$\text{FIM} = n\begin{bmatrix} \sigma^2\mathbf{U}\mathbb{E}\left[\frac{e^{c\sigma\mathbf{z}_1+\beta}}{(1+e^{c\sigma\mathbf{z}_1+\beta})^2}\mathbf{z}\mathbf{z}^\top\right]\mathbf{U}^\top & \sigma\mathbf{U}\mathbb{E}\left[\frac{e^{c\sigma\mathbf{z}_1+\beta}}{(1+e^{c\sigma\mathbf{z}_1+\beta})^2}\mathbf{z}\right] \\ \sigma\mathbb{E}\left[\frac{e^{c\sigma\mathbf{z}_1+\beta}}{(1+e^{c\sigma\mathbf{z}_1+\beta})^2}\mathbf{z}^\top\right]\mathbf{U}^\top & \mathbb{E}\left[\frac{e^{c\sigma\mathbf{z}_1+\beta}}{(1+e^{c\sigma\mathbf{z}_1+\beta})^2}\right] \end{bmatrix}. \tag{24}$$

We can abbreviate (24) using

$$\text{FIM} = n\begin{bmatrix} \sigma^2\mathbf{U}\mathbf{A}\mathbf{U}^\top & \sigma\mathbf{U}\mathbf{V} \\ \sigma\mathbf{V}^\top\mathbf{U}^\top & \alpha_0 \end{bmatrix}, \tag{25}$$

where $\mathbf{A} = \mathbb{E}[\frac{e^{c\sigma\mathbf{z}_1+\beta}}{(1+e^{c\sigma\mathbf{z}_1+\beta})^2}\mathbf{z}\mathbf{z}^\top]$, $\mathbf{V} = \mathbb{E}[\frac{e^{c\sigma\mathbf{z}_1+\beta}}{(1+e^{c\sigma\mathbf{z}_1+\beta})^2}\mathbf{z}]$, and α_0 is as in (7). Using the independence among the $\mathbf{z}_i$'s, we can further simplify $\mathbf{A}$ and $\mathbf{V}$ as $\mathbf{A} = \text{diag}\begin{bmatrix}\alpha_2 & \alpha_0 & \ldots & \alpha_0\end{bmatrix}$ and $\mathbf{V} = \begin{bmatrix}\alpha_1 & 0 & \ldots & 0\end{bmatrix}^\top$ where α_k is as in (7). Let $\mathbf{e}_i$ be the canonical vector with all zero entries except for 1 at the ith entry. Replacing $\mathbf{A} = (\alpha_2 - \alpha_0)\mathbf{e}_1\mathbf{e}_1^\top + \alpha_0\mathbf{I}$ and $\mathbf{V} = \mathbf{e}_1\alpha_1$ into (25), we obtain

$$\text{FIM} = n\begin{bmatrix} \sigma^2\mathbf{U}[(\alpha_2 - \alpha_0)\mathbf{e}_1\mathbf{e}_1^\top + \alpha_0\mathbf{I}]\mathbf{U}^\top & \sigma\mathbf{U}\mathbf{e}_1\alpha_1 \\ \sigma(\mathbf{U}\mathbf{e}_1)^\top\alpha_1 & \alpha_0 \end{bmatrix}. \tag{26}$$

Substituting $\mathbf{U}\mathbf{e}_1 = \mathbf{u}_1$ into (26), we obtain (6).

7. REFERENCES

[1] T. Zhang and F. Oles, "Text categorization based on regularized linear classifiers," *Information Retrieval*, vol. 4, pp. 531, 2001.

[2] Mee Young Park and Trevor Hastie, "Penalized logistic regression for detecting gene interactions," *Biostatistics*, pp. 30–50, 2008.

[3] Andrew Y Ng, "Feature selection, l 1 vs. l 2 regularization, and rotational invariance," *Proceedings of the twenty-first international conference on Machine learning. ACM*, 2004.

[4] Srikanth Ryali, Kaustubh Supekar, Daniel A. Abrams, and Vinod Menon, "Sparse logistic regression for whole-brain classification of fmri data," *NeuroImage*, pp. 752–764, 2010.

[5] Su-In Lee, Honglak Lee, Pieter Abbeel, and Andrew Y. Ng, "Efficient l 1 regularized logistic regression," *Proceedings of the National Conference on Artificial Intelligence*, vol. 21, 2006.

[6] Tong Tong Wu, Yi Fang Chen, Trevor Hastie, Eric Sobel, and Kenneth Lange, "Genome-wide association analysis by lasso penalized logistic regression," *Bioinformatics*, pp. 714–721, June 2009.

[7] Gavin C. Cawley and Nicola LC Talbot, "Gene selection in cancer classification using sparse logistic regression with bayesian regularization," *Bioinformatics*, pp. 2348–2355, 2006.

[8] John D. Lafferty Wainwright, Martin J. and Pradeep K. Ravikumar, "High-dimensional graphical model selection using ell_1-regularized logistic regression," *Advances in neural information processing systems*, 2006.

[9] Jianing Shi, Wotao Yin, Stanley Osher, and Paul Sajda, "A fast hybrid algorithm for large-scale l_1-regularized logistic regression," *J. Mach. Learn. Res.*, vol. 11, pp. 713–741, 2010.

[10] Hui Zou and Trevor Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[11] T. Zhang and F. J. Oles, "A probability analysis on the value of unlabeled data for classification problems," *ICML*, pp. 1191–1198, 2000.

[12] Stephen Boyd and Lieven Vandenberghe, "Convex optimization," *Cambridge university press*, 2004.

[13] Thomas P Minka, "Algorithms for maximum-likelihood logistic regression," 2003.