Optimizing the regularization parameters selection in sparse modeling

Victoria Peterson  
Instituto de Matemática Aplicada del Litoral, IMAL-UNL-CONICET, Santa Fe, Argentina  
Facultad de Ingeniería, FI-UNER, Oro Verde, Argentina  
vpeterson@santafe-conicet.gov.ar

Rubén D. Spies  
Instituto de Matemática Aplicada del Litoral, IMAL-FIQ-UNL-CONICET, Santa Fe, Argentina  
rspies@santafe-conicet.gov.ar

1 Introduction

Linear regression seeks to approximate a response variable $y \in \mathbb{R}^n$ given a set of samples $X \in \mathbb{R}^{n \times p}$, by a linear combination of each predictor (feature) vector $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$

$$\hat{y} = \sum_{j=1}^{p} x_{i,j} \beta_j,$$

where the solution vector $\beta = (\beta_1, \beta_2, \ldots, \beta_p) \in \mathbb{R}^p$ denotes the model weights.

This inverse problem typically suffers from ill-posedness in the Hadamard sense [1]. Regularization methods are mathematical tools designed to restore numeral stability in such ill-posed inverse problems. Tikhonov regularization [2] is the most widely used regularization strategy, in which the following functional is minimized:

$$J_{\lambda, \psi}(\beta) = \phi(y, X\beta) + \lambda \cdot \psi(\beta),$$

where $\phi(\cdot)$ is so-called fidelity term, which measures the discrepancy between the true observations and its estimate, $\psi(\cdot) = (\psi_1, \ldots, \psi_m)^T$ are a set of $m$ regularization terms, which induces different penalties in the solution, and $\lambda = (\lambda_1, \ldots, \lambda_m)^T$ are positive constants, called regularization parameters, which balance fidelity with penalties. Different choices of both the fitting and the regularizer functional lead to different solutions. For instance, in linear regression the squared distance between the observation vector $y$ and its estimation given by $X\beta$ is the obvious natural choice for the fidelity term. This leads to the so-called ordinary least-squares (OLS) estimation [3].

Although the OLS estimation given by $\hat{\beta} = (X^T X)^{-1} X y$ is easy to find, it presents large variance and lack of interpretability. Different solutions have been proposed in order to improve such estimator, most of them based on $\ell_p-\text{norm}$ regularization strategies. For instance, by introducing the $\ell_2$-norm term into OLS formulation, the ridge regression estimator is then obtained, which, by adding positive elements to its diagonal, seeks to alleviate the near-singular problem of the matrix $X^T X$ [4]. For better model interpretability, a subset of relevant predictors (features) must then be identify. Thus, for doing so, the lasso estimator can be used, which by means of the $\ell_1$-norm of the solution, the OLS regression coefficients are shrinkage towards zero [5]. Adding these two penalizers in a convex form is also possible, leading to the e-net estimator [6].

Besides a proper selection of the penalizer term $\psi$ is crucial, the efficiency of a regularization method also strongly depends on the accurate selection of the regularization parameters. Although different authors have proposed several approaches for proper regularization parameter selection (e.g.  

most of them are not suitable for multi-parameter Tikhonov regularization, such as e-net regression. The well-known L-curve method for estimating the λ value in ridge regression (Tikhonov regularization) has been extended for the multiple penalty scenario. This extension, named as to L-hypersurface [12], has already been evaluated in a mixed-norm sparse discriminative approach [13]. Recently, the balancing principle [14] has been proposed for choosing an appropriate (vector-value) regularization parameter for multi-parameter Tikhonov regularization. This method not only provides strong mathematical formulations but also shows promising numerical results. In this work, we present our preliminary results on the impact analysis of the regularization parameters estimation in a binary classification framework based on the e-net formulation.

2 Experiments and Results

Numerical experiments were made with a 25-subjects database consisted of electroencephalography (EEG) signals acquired at 10 channels (Fz, C3, Cz, C4, P3, P4, P7, P8, Oz) with a sampling rate of 256 Hz. Each subject participated in a P300-based Brain-Computer Interface (BCI) experiment, in which different words had to be spelled using the oddball paradigm [15]. The P300-BCI, from the patter recognition point of view, is a binary classification problem in which the 16.6% of the EEG records contain an unconscious brain response (namely, P300 wave) to an external stimulus. In this work, the EEG records were filtered from 0.1 Hz to 12 Hz by a 4th order forward-backward Butterworth band-pass filter and 1000 ms segments were extracted from the EEG records at the beginning of each stimulus and then a downsampled at 32 Hz was implemented. A total of 3780 EEG trials (630 of them being target) of dimension of 10 × 320, conforms each subject’s database.

For classification we tested the Sparse Discriminant Analysis (SDA) [16] as well as it generalized version based on Kullback-Leibler divergence, named GSDA [13]. Both methods make use of the e-net formulation, in which two regularization parameter, named here as λ_1 and λ_2, balanced the contribution of the ℓ_1−norm and ℓ_2−norm in the solution, respectively. For its numerical implementation the LARS-EN [17] algorithm was used, in which an upper bound of the ℓ_1−norm penalizing term was used for early stopping. We tested here the balancing principle for proper estimation of the λ_2 parameter in GSDA (GSDA_{bm}), and compared it performance by using the L-hypersurface approach (GSDA) when i) the stop parameter is fixed and ii) when it is updated at each GSDA iteration by 0.1 ||β_{OLS}||_1 (GSDA_{bmstop} and GSDA_{stop}). Table 1 shows the average classification results evaluated by means of the area under the receiver operator characteristic curve (AUC) [18] in a 10-fold cross-validation procedure, for each tested method. The best classification performance (p−value < 0.05) were achieved by the stop implementations.

Table 1: Overall classification results (mean ± standard deviation) over the 25 subjects yielded by each tested method in a 10-fold cross-validation procedure. Best classification performances are in bold.

<table>
<thead>
<tr>
<th></th>
<th>SDA</th>
<th>GSDA</th>
<th>GSDA_{bm}</th>
<th>GSDA_{bmstop}</th>
<th>GSDA_{stop}</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.86 ± 0.02</td>
<td>0.87 ± 0.02</td>
<td>0.87 ± 0.02</td>
<td><strong>0.90 ± 0.02</strong></td>
<td>0.90 ± 0.02</td>
</tr>
</tbody>
</table>

3 Discussions

In this preliminary work we analyze the impact on classification performance of the different optimizing techniques for the tuning regularization parameters in a mixed-term regularized discriminative framework. Although for optimizing λ_2 value the balancing principle seems to be competitive with the L-hypersurface approach, the former is easily extendable to multi-parameter estimation. In addition, we proposed here an a stop update procedure based on the ℓ_1−norm of the OLS solution, which lead to data-driven subject-specific estimations. This simple update in the upper bound of the ℓ_1−norm of the solution vector β increments average classification performance for up to 3%, with no additional computational cost. Future work involves the analysis of balancing principle for simultaneous optimization of the regularization parameters, comparative analysis with cross-validation and other Bayes theorem-based principles [13][20], as well as extending the numerical results in other and different multi-parameter Tikhonov regularization applications.
References


