Optimizing the regularization parameters selection in sparse modeling

Anonymous Author(s)
Affiliation
Address
email

1 Introduction

Linear regression seeks to approximate a response variable $\mathbf{y} \in \mathbb{R}^n$ given a set of samples $\mathbf{X} \in \mathbb{R}^{n \times p}$, by a linear combination of each predictor (feature) vector $\mathbf{x}_i = (x_1, x_2, \dots, x_p)$

$$\hat{\mathbf{y}} = \sum_{j=1}^{p} \mathbf{x}_{i,j} \boldsymbol{\beta}_j,$$

where the solution vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p) \in \mathbb{R}^p$ 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:

$$\mathcal{J}_{\boldsymbol{\lambda},\boldsymbol{\psi}}(\beta) = \phi(\mathbf{y}, \mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\lambda} \cdot \boldsymbol{\psi}(\boldsymbol{\beta}),$$

where $\phi(\cdot)$ is so-called fidelity term, which measures the discrepancy between the true observations and its estimate, $\boldsymbol{\psi}(\cdot) = (\psi_1, \dots, \psi_m)^T$ are a set of m regularization terms, which induces different penalties in the solution, and $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_m)^T$ 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 $\mathbf{y}$ and it estimation given by $\mathbf{X}\boldsymbol{\beta}$ 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 $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}\mathbf{y}$ 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 ℓ_p-norm regularization strategies. For instance, by introducing the ℓ_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 $\mathbf{X}^T\mathbf{X}$ [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 ℓ_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 $\boldsymbol{\psi}$ 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. [9, 10, 11]), 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, Pz, P4, PO7, PO8, 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 4^{th} 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 $0.1\|\boldsymbol{\beta}_{\text{OLS}}\|_1$ ($\text{GSDA}_{\text{bmstop}}$ and $\text{GSDA}_{\text{stop}}$). 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.

SDA	GSDA	GSDA_{bm}	$\text{GSDA}_{\text{bmstop}}$	$\text{GSDA}_{\text{stop}}$
0.86 ± 0.02	0.87 ± 0.02	0.87 ± 0.02	$\mathbf{0.90 \pm 0.02}$	$\mathbf{0.90 \pm 0.02}$

3 Discussions

In this preliminary work we analyze the impact on classification performance of the different optimazing techniques for the tunning 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 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 $\boldsymbol{\beta}$ 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 [19, 20], as well as extending the numerical results in other and different multi-parameter Tikhonov regularization applications.

References

[1] Per Christian Hansen. *Rank-deficient and discrete ill-posed problems: numerical aspects of linear inversion*, volume 4. SIAM, 2005.

[2] Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

[3] Thomas Kailath et al. *Linear least-squares estimation*. Dowden, Hutchinson & Ross Stroudsburg, PA, 1977.

[4] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

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

[6] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[7] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[8] Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

[9] Ghadban Khalaf and Ghazi Shukur. Choosing ridge parameter for regression problems. 2005.

[10] Per Christian Hansen. Analysis of discrete ill-posed problems by means of the l-curve. *SIAM review*, 34(4):561–580, 1992.

[11] Thomas Bonesky. Morozov's discrepancy principle and tikhonov-type functionals. *Inverse Problems*, 25(1):015015, 2008.

[12] Murat Belge, Misha E Kilmer, and Eric L Miller. Simultaneous multiple regularization parameter selection by means of the L-hypersurface with applications to linear inverse problems posed in the wavelet transform domain. In *SPIE's International Symposium on Optical Science, Engineering, and Instrumentation*, pages 328–336. International Society for Optics and Photonics, 1998.

[13] Victoria Peterson, Hugo Leonardo Rufiner, and Ruben Daniel Spies. Generalized sparse discriminant analysis for event-related potential classification. *Biomedical Signal Processing and Control*, 35:70–78, 2017.

[14] Kazufumi Ito, Bangti Jin, and Tomoya Takeuchi. Multi-parameter tikhonov regularization. *arXiv preprint arXiv:1102.1173*, 2011.

[15] Lawrence Ashley Farwell and Emanuel Donchin. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and Clinical Neurophysiology*, 70(6):510–523, 1988.

[16] Line Clemmensen, Trevor Hastie, Daniela Witten, and Bjarne Ersbøll. Sparse discriminant analysis. *Technometrics*, 53:406–413, 2012.

[17] Karl Sjöstrand, Line Harder Clemmensen, Rasmus Larsen, and Bjarne Ersbøll. SpaSM: A matlab toolbox for sparse statistical modeling. *Journal of Statistical Software Accepted for publication*, 2012.

[18] Andrew P Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

[19] Qing Li, Nan Lin, et al. The bayesian elastic net. *Bayesian analysis*, 5(1):151–170, 2010.

[20] Deirel Paz-Linares, Mayrim Vega-Hernandez, Pedro A Rojas-Lopez, Pedro A Valdes-Hernandez, Eduardo Martinez-Montes, and Pedro A Valdes-Sosa. Spatio temporal EEG source imaging with the hierarchical bayesian elastic net and elitist lasso models. *Frontiers in neuroscience*, 11:635, 2017.