ACADEMIC World Academic Union

**ISSN** 1746-7659, England, UK
**Journal of Information and Computing Science**
Vol. 6, No. 4, 2011, *pp*. 295-302

# A Novel Sparse Learning Method: Compressible Bayesian Elastic Net Model

Ke-Yang Cheng[1, 2], Qi-rong Mao[2],Xiao-yang Tan[1],Yong-zhao Zhan[2]

[1] School of Information Science & Technology, Nanjing University of aeronautics & astronautics, Nanjing, Jiangsu, China,210016

[2] School of Computer Science & Telecommunications Engineering, Jiangsu University, Zhenjiang, Jiangsu, China,212013

(*Received February 10, 2011, accepted May 22, 2011*)

**Abstract.** In this paper, we study the combination of compression and Bayesian elastic net. By including a compression operation into the ℓ1 and ℓ2 regularization, the assumption on model sparsity is relaxed to compressibility: model coefficients are compressed before being penalized, and sparsity is achieved in a compressed domain rather than the original space. We focus on the design of compression operations, by which we can encode various compressibility assumptions and inductive biases. We show that use of a compression operation provides an opportunity to leverage auxiliary information from various sources. The compressible Bayesian elastic net has another two major advantages. Firstly, as a Bayesian method, the distributional results on the estimates are straightforward, making the statistical inference easier. Secondly, it chooses the two penalty parameters simultaneously, avoiding the "double shrinkage problem" in the elastic net method. We conduct extensive experiments on braincomputer interfacing, handwritten character recognition and text classification. Empirical results show clear improvements in prediction performance by including compression in Bayesian elastic net. We also analyze the learned model coefficients under appropriate compressibility assumptions, which further demonstrate the advantages of learning compressible models instead of sparse models.

**Keywords:** Sparse Learning, compression operation, Bayesian elastic net

## 1. Introduction

Regularization was initially proposed to solve ill-posed problems (Tikhonov & Arsenin, 1977)[1]. In statistical learning, regularization is widely used to control model complexity and prevent overfitting (Hastie et al., 2001)[2]. Regularization seeks a trade-off between fitting the observations and reducing the model complexity, which is justified by the minimum description length (MDL) principle in information theory (Rissanen, 1978)[3] and the bias-variance dilemma in statistics (Sullivan, 1986)[4]. Since the introduction of lasso (Tibshirani, 1996)[5], ℓ1-regularization has become very popular for learning in high-dimensional spaces. A fundamental assumption of ℓ1-regularization is the sparsity of model parameters, i.e., a large fraction of coefficients are zeros. While demonstrating promising performance for many problems, the lasso estimator does have some shortcomings.

Zou and Hastie (2005) [6] emphasized three inherent drawbacks of the lasso estimator. Firstly, due to the nature of the convex optimization problem, the lasso method cannot select more predictors than the sample size. But in practice there are often studies that involve much more predictors than the sample size, e.g. microarray data analysis (Guyon et al. 2002)[7]. Secondly, when there is some group structure among the predictors, the lasso estimator usually selects only one predictor from a group while ignoring others. Thirdly, when the predictors are highly correlated, the lasso estimator performs unsatisfactorily. Zou and Hastie (2005) proposed the elastic net (en) estimator to achieve improved performance in these cases. The en estimator can also be viewed as a penalized least squares method where the penalty term is a convex combination of the lasso penalty and the ridge penalty.

Another shortcoming of Lasso is that the sparsity assumption on model coefficients might be too restrictive and not necessarily appropriate in many application domains. Indeed, many signals in the real

[1] Corresponding author. *E-mail address*: kycheng@ujs.edu.cn

world (e.g., images, audio, videos, time series) are found to be compressible (i.e., sparse in certain compressed domain) but not directly sparse in the observed space. Naturally, the assumption of sparsity can be relaxed to compressibility. Inspired by the recent development of compressive sampling (or compressed sensing) (Candes, 2006[8]; Donoho, 2006[9]), we study learning compressible models: a compression on model coefficients can be included in the ℓ1 and ℓ2 penalty, and model is assumed to be sparse after compression.

The rest of this paper is organized as follows. In section 2 we will briefly introduce naïve elastic net. In Section 3 we discuss the definition, computation issues and potential benefits of learning compressible Bayesian elastic net model. In this Section, we propose some classes of model compressibility assumptions and model hierarchy distributions. In Sections 4, we empirically study some real-world problems using compressibility as a more appropriate inductive bias than sparsity. Experimental results also demonstrate the advantages of compressible Bayesian elastic net (cben) than compressible Bayesian lasso (cbl), elastic net (en) and lasso. Section 5 concludes and mentions some discussions.

## 2. Naive elastic net

Suppose that the data set has n observations with p predictors. Let $y = (y_1,...,y_n)^T$ be the response and $X = (x_1,...,x_p)$ be the model matrix, where $x_j = (x_{1j},...,x_{nj})^T, j = 1,...,p$, are the predictors. After a location and scale transformation, we can assume that the response is centred and the predictors are standardized.

$$\sum_{i=1}^{n} y_i = 0 \ \sum_{i=1}^{n} x_{ij} = 0 \text{ and } \sum_{i=1}^{n} x^2{}_{ij} = 1 \text{.for j=1,2,…,p.}$$

For any fixed non-negative λ1 and λ2, we define the naive elastic net criterion

$$L(\lambda_1, \lambda_2, \beta) = | y - X\beta |^2 + \lambda_2 | \beta |_2^2 + \lambda_1 | \beta |_1 \tag{1}$$

where

$$| \beta |_2^2 = \sum_{j=1}^{p} \beta_j^2,$$

$$| \beta |_1 = \sum_{j=1}^{p} | \beta_j |.$$

The naive elastic net estimator $\hat{\beta}$ is the minimizer of equation (1):

$$\hat{\beta} = \arg\min_{\beta} \{L(\lambda_1, \lambda_1, \beta)\}.$$

This procedure can be viewed as a penalized least squares method. Let $\alpha=\lambda_1+\lambda_2$; then solving $\hat{\beta}$ in equation (1) is equivalent to the optimization problem

$$\hat{\beta} = \arg\min_{\beta} | y - X\beta |^2, \quad \text{subject to } (1-\alpha) | \beta |_1 + \alpha | \beta |_2^2 \le t \text{ for some t.}$$

We call the function $(1-\alpha) | \beta |_1 + \alpha | \beta |_2^2$ the elastic net penalty, which is a convex combination of the lasso and ridge penalty. When α=1, the naive elastic net becomes simple ridge regression. In this paper, we consider only α<1. For all $\alpha \in [0, 1)$, the elastic net penalty function is singular (without first derivative) at 0 and it is strictly convex for all α>0, thus having the characteristics of both the lasso and ridge regression. Note that the lasso penalty (α=0) is convex but not strictly convex. These arguments can be seen clearly from Fig. 1.