Regularizers for structured sparsity

Charles A. Micchelli · Jean M. Morales · Massimiliano Pontil

Received: 29 August 2011 / Accepted: 10 October 2011 / Published online: 24 November 2011
© Springer Science+Business Media, LLC 2011

Abstract We study the problem of learning a sparse linear regression vector under additional conditions on the structure of its sparsity pattern. This problem is relevant in machine learning, statistics and signal processing. It is well known that a linear regression can benefit from knowledge that the underlying regression vector is sparse. The combinatorial problem of selecting the nonzero components of this vector can be “relaxed” by regularizing the squared error with a convex penalty function like the $\ell_1$ norm. However, in many applications, additional conditions on the structure of the regression vector and its sparsity pattern are available. Incorporating this information into the learning method may lead to a significant decrease of the estimation error. In this paper, we present a family of convex penalty functions, which encode prior knowledge on the structure of the vector formed by the absolute values of the regression coefficients. This family subsumes the $\ell_1$ norm and is flexible enough to include different models of sparsity patterns, which are of

Communicated by Lixin Shen.
C. A. Micchelli
Department of Mathematics, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon Tong, Hong Kong, People’s Republic of China

C. A. Micchelli
Department of Mathematics and Statistics, State University of New York, The University at Albany, 1400 Washington Avenue, Albany, NY 12222, USA

J. M. Morales · M. Pontil
Department of Computer Science, University College London, Gower Street, London WC1E, England, UK
e-mail: m.pontil@cs.ucl.ac.uk

J. M. Morales
e-mail: j.morales@cs.ucl.ac.uk
practical and theoretical importance. We establish the basic properties of these penalty functions and discuss some examples where they can be computed explicitly. Moreover, we present a convergent optimization algorithm for solving regularized least squares with these penalty functions. Numerical simulations highlight the benefit of structured sparsity and the advantage offered by our approach over the Lasso method and other related methods.

Keywords Convex optimization · Feature selection · LASSO · Linear regression · Regularization · Sparse estimation

1 Introduction

The problem of sparse estimation is becoming increasingly important in statistics, machine learning and signal processing. In its simplest form, this problem consists in estimating a regression vector $\beta^* \in \mathbb{R}^n$ from a set of linear measurements $y \in \mathbb{R}^m$, obtained from the model

$$y = X\beta^* + \xi$$  \hspace{1cm} (1.1)

where $X$ is an $m \times n$ matrix, which may be fixed or randomly chosen and $\xi \in \mathbb{R}^m$ is a vector which results from the presence of noise.

An important rational for sparse estimation comes from the observation that in many practical applications the number of parameters $n$ is much larger than the data size $m$, but the vector $\beta^*$ is known to be sparse, that is, most of its components are equal to zero. Under this sparsity assumption and certain conditions on the data matrix $X$, it has been shown that regularization with the $\ell_1$ norm, commonly referred to as the Lasso method [23], provides an effective means to estimate the underlying regression vector, see for example [5, 7, 16, 24] and references therein. Moreover, this method can reliably select the sparsity pattern of $\beta^*$ [16], hence providing a valuable tool for feature selection.

In this paper, we are interested in sparse estimation under additional conditions on the sparsity pattern of the vector $\beta^*$. In other words, not only do we expect this vector to be sparse but also that it is structured sparse, namely certain configurations of its nonzero components are to be preferred to others. The prior knowledge that we consider in this paper is that the vector $|\beta^*|$, whose components are the absolute value of the corresponding components of $\beta^*$, should belong to some prescribed convex subset $\Lambda$ of the positive orthant. For certain choices of $\Lambda$ this implies a constraint on the sparsity pattern as well. For example, the set $\Lambda$ may include vectors with some desired monotonicity constraints, or other constraints on the “shape” of the regression vector. Unfortunately, the constraint that $|\beta^*| \in \Lambda$ is nonconvex and its implementation is computationally challenging. To overcome this difficulty, we propose a family of penalty functions, which are based on an extension of the $\ell_1$ norm used by the Lasso method and involves the solution of a smooth