Signal Processing 151 (2018) 107–118

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Hyperparameter selection for group-sparse regression: A probabilistic approach[☆]

Ted Kronvall[*], Andreas Jakobsson

Department of Mathematical Statistics, Lund University, Sweden

ARTICLE INFO

Article history:
Received 3 October 2017
Revised 26 March 2018
Accepted 24 April 2018
Available online 5 May 2018

Keywords:
Sparse estimation
Group-LASSO
Regularization
Model order estimation
Sparsistency
Extreme value analysis

ABSTRACT

This work analyzes the effects on support recovery for different choices of the hyper- or regularization parameter in LASSO-like sparse and group-sparse regression problems. The hyperparameter implicitly selects the model order of the solution, and is typically set using cross-validation (CV). This may be computationally prohibitive for large-scale problems, and also often overestimates the model order, as CV optimizes for prediction error rather than support recovery. In this work, we propose a probabilistic approach to select the hyperparameter, by quantifying the type I error (false positive rate) using extreme value analysis. From Monte Carlo simulations, one may draw inference on the upper tail of the distribution of the spurious parameter estimates, and the regularization level may be selected for a specified false positive rate. By solving the e group-LASSO problem, the choice of hyperparameter becomes independent of the noise variance. Furthermore, the effects on the false positive rate caused by collinearity in the dictionary is discussed, including ways of circumventing them. The proposed method is compared to other hyperparameter-selection methods in terms of support recovery, false positive rate, false negative rate, and computational complexity. Simulated data illustrate how the proposed method outperforms CV and comparable methods in both computational complexity and support recovery.

1. Introduction

Estimating the sparse parameter support for a high-dimensional regression problem has been the focus of much scientific attention during the past two decades, as this methodology has shown its usefulness in a wide array of applications, ranging from spectral analysis [1–3], array- [4–6] and audio processing [7–9], to biomedical modeling [10], magnetic resonance imaging [11,12], and more. For many of these and for other applications, the retrieved data may be well explained using a highly underdetermined regression model, in which only a small subset of the explanatory variables are required to represent the data. The approach is typically referred to as sparse regression; the individual regressors are called atoms, and the entire regressor matrix the dictionary, which is typically customized for a particular application. The common approach of inferring sparsity on the estimates is to solve a regularized regression problem, i.e., appending the fit term with a regularization term that increases as variables become active (or non-zero). Much of the work in the research area springs from extensions on the seminal work by Tibshirani et al., wherein the least absolute selection and shrinkage operator (LASSO) [13] was introduced. The LASSO is a regularized regression problem where an ℓ_1-norm on the variable vector is used as regularizer, which in signal processing is also referred to as the basis pursuit denoising (BPDN) method [14]. Another early alternative to the LASSO problem is the penalized likelihood problem, introduced in [15].

In this paper, we focus on a generalization of the sparse regression problem, wherein the atoms of the dictionary exhibit some form of grouping behavior which is defined *a priori*. This follows the notion that a particular data feature is modeled not only using a single atom, but instead by a group of atoms, such that each atom has an unknown (and possibly independent) response variable, but where the entire group is assumed to be either present or not present in the data. This is achieved in the group-LASSO [16] by utilizing an ℓ_1/ℓ_2-regularizer, but other approaches have also been successful, such as in, e.g., [9,10]. Being a generalization of the LASSO, the group-LASSO reverts back to the standard LASSO when the group sizes in the dictionary all have size one. Typically, results which hold for the group-LASSO thus also hold for the LASSO. One reason behind the success of LASSO-like approaches is that these are typically cast as a convex optimization problems, for which there exists strong theoretical results for convergence and recovery guarantees (see, e.g., [17–19], and the refer-

[☆] This work was supported in part by the Swedish Research Council, Carl Trygger's foundation, and the Royal Physiographic Society in Lund.

[*] Corresponding author.
E-mail address: ted@maths.lth.se (T. Kronvall).

https://doi.org/10.1016/j.sigpro.2018.04.021
0165-1684/© 2018 Elsevier B.V. All rights reserved.

ences therein). For convex problems, there also exist user-friendly scientific software for simple experimentation and investigation of new regularizers [20].

The sparse regression problems described here, being a subset of the regularized regression problems, have in common the requirement of selecting one or several hyperparameters, which have the role of controlling the degree of sparsity in the solution by adjusting the level of regularization in relation to the fit term. Thus, sparsity is subject to user control, and must therefore be chosen adequately for each problem. From the perspective of model order selection, one may note that there is no currently consistent approach to finding a correct model order (see, e.g., [21]). Still, as an implicit agent of model order selection in regularization problems, one may distinguish three main methodologies of selecting the regularization level. Firstly, and perhaps most commonly are the data-driven, or post-model selection, methods, where the performance of a number of candidate models are compared in some user-selected metric. To that end, the least angle regression (LARS) algorithm [22] calculates the entire (so called) path of solutions on an interval of values for the hyperparameter of a LASSO-like problem, and at a computational cost similar to solving the LASSO for a single value of the hyperparameter. However, by using warm-starts, a solution path may also be calculated quickly using some appropriate implementation of the group-LASSO. A single point on the solution path is then chosen based on user preference; most commonly prediction performance, i.e., using cross-validation (CV), as was done, for instance, in [23] for the multi-pitch estimation problem. However, due to the computationally burdensome process of CV, one often instead reverts to using heuristic data-driven approaches, or choosing the hyperparameter based on some information criteria (see, e.g., [24]). Another interesting contribution was made in [25], wherein a covariance test statistic was used to determine whether to include every new regressor along a path of regularization values. Bayesian techniques offer another common approach to the model order selection problem, wherein the joint posterior distribution of the regression variables and the hyperparameter are utilized, under the assumption on statistical priors on these, such as in, e.g., [26]. The third main group of approaches may be considered to be probabilistic in the sense that they make assumptions on only the noise statistics of the measured signal, and not the regression variables. Among these, the approach suggested in [27] might be most prominent (here, for simplicity, referred to as CDS), where in order to suppress the noise components from propagating into the estimate, an upper-endpoint of the distribution is used, such that for independent Gaussian regressors (i.e., orthogonal dictionaries), the largest interfering noise components in the limit grows in proportion to some quantity. Under these assumptions, CDS is ostensibly blind, but will by construction set the regularization level high enough to guarantee noise suppression, and might thereby also suppress the signal-of-interest. In applications containing atoms with a high degree of collinearity, thereby violating the orthogonal assumption, this will result in overshooting of the regularization level. To simplify the selection of regularization level, the scaled LASSO [28] reparametrizes the hyperparameter by introducing an auxiliary variable describing the standard deviation of the model residual. This has the effect that the regularization level may be selected (somewhat) independently of the noise variance, which is useful for the probabilistic approaches.

Another method of selecting the regularization level that might fall into the probabilistic category is the sparse iterative covariance-based estimation (SPICE) method, which yields a relatively sparse parameter support by matching the observed covariance matrix and a covariance matrix parametrized by a dictionary. The method has been shown to work well for a variety of applications, especially those pertaining to estimation of line spectra and directions-of-arrival (see, e.g., [29]). In subsequent publications (see, e.g., [29,30]), SPICE was shown to be equivalent to either the least absolute deviation (LAD) LASSO under a heterscedastic noise assumption, or the square root (SR) LASSO under a homoscedastic noise assumption, both for particular choices of the hyperparameter. It may be shown that the SR LASSO and the scaled LASSO are equivalent, and we conclude that SPICE is a robust (and possibly heuristic) approach of fixing the hyperparameter (somewhat) independently of the noise level. In a recent effort, the SPICE approach was extended for group sparsity [31], showing promising results, e.g., for multi-pitch estimation, but also illustrating how the fixed hyperparameter yields estimates which are not as sparse as one may typically expect. A valid argument in defence of the SPICE approach is that the measure of 'good' in sparse estimation is not entirely straightforward, and not sparse enough may still be good enough.

Borrowing some terminology from detection theory [32], one way of measuring performance is to calculate the false negatives (FNs), i.e., whether atoms pertaining to the true support of the signal (those atoms of which the data is truly composed) are estimated as zero for some choice of the hyperparameter. As the SPICE regularization level is typically set too low, the possibility of FNs is consequently also low, which for some applications may be the focus. Conversely, for some applications, the focus may be to eliminate the false positives (FPs), i.e., when noise components are falsely set to be non-zero while not being in the true support set. The FPs and FNs are also sometimes referred to as the type I and type II errors, respectively. In addition, a metric called sparsistency is sometimes used, measuring the binary output of whether the estimated and the true supports are identical, which is the complement of the union between FN and FP [33]. Sparsistency might also be unobtainable for a certain problem; avoiding FPs requires selecting the hyperparameter so large that FNs will arise, and similarly avoiding FNs will result in more FPs. Model order estimation can thus be seen as prioritizing between FPs and FNs, which is also referred to as the bias-variance trade off, and has a long history in the literature. Typically, model order estimation can be formulated as a series of hypothesis tests, subsequently tested using, e.g., F-test statistics for some specified significance level [34].

In this paper, we further this development, formulating a probabilistic method for hyperparameter selection using hypothesis testing. By analyzing how the noise components propagate into the parameter estimates for different estimators and different choices of the hyperparameters, we seek to increase the sparsistency of the group-LASSO estimate by means of optimizing the FP rate. By making assumptions on the noise distribution and then sampling from the corresponding extreme value distribution using the Monte Carlo method, the hyperparameter is chosen as an appropriate quantile of the largest anticipated noise components. Avoiding FPs can never be guaranteed without maximizing the regularization level, thereby setting the entire solution to zero, but the risk may be quantified. By specifying the type I error, the sparsistency rate is also indirectly controlled, whenever this is feasible. Furthermore, for Gaussian noise, we show that the distribution for the maximum noise components follows a type I extreme value distribution (Gumbel), from which a parametric quantile may be obtained at a low computational cost.

For coherent dictionaries, i.e., where there is a high degree of collinearity between the atoms, many of the theoretical guarantees for sparse estimation will fail to hold, along with a few of the methods themselves. The effects on the estimates for the collinear atoms are difficult to discern; depending on the problem either all of them, or just a few of them, become non-zero. Coherence therefore typically results in FPs, if the regularization level is not increased, which in turn might yield FNs. There exists some approaches of dealing with coherent dictionaries. The elastic net uses