Sparse Logistic Regression utilizing Cardinality Constraints and Information Criteria

Gabriel Urrutia*, Ramón Delgado‡, Rodrigo Carvajal*, Dimitrios Katselis†, and Juan C. Agüero*,‡
*Electronics Engineering Department, Universidad Técnica Federico Santa María, Chile.
email addresses: gabriel.urrutia@alumnos.usm.cl, {rodrigo.carvajalg, juan.aguero}@usm.cl
†Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, IL, USA.
email address: katselis@illinois.edu
‡School of Electrical Engineering and Computer Science, The University of Newcastle, Australia.
email addresses: ramon.delgado@uon.edu.au, juan.aguero@newcastle.edu.au

***Abstract*— In this paper we address the problem of estimating a sparse parameter vector that defines a logistic regression. The problem is then solved using two approaches: i) inequality constrained Maximum Likelihood estimation and ii) penalized Maximum Likelihood which is closely related to Information Criteria such as AIC. For the promotion of sparsity, we utilize a nonlinear constraint based on the ℓ_0 (pseudo) norm of the parameter vector. The corresponding optimization problem is solved using an equivalent representation of the problem that is simpler to solve. We illustrate the benefits of our proposal with an example that is inspired by a gene selection problem in DNA microarrays.**

I. Introduction

Sparse estimation problems are of great interest in the scientific community. Several applications that consider sparse estimation are approached by incorporating a regularization/penalty term as a mean of inducing sparsity, such as the ℓ_1-norm in the classical Lasso [1] and an $\ell_1-\ell_2$ norm combination in the Elastic Net [2]. However, it is usually difficult to give those penalties a physical meaning. On the other hand, the ℓ_0-(pseudo)norm can also be used to promote (induce) sparsity, see e.g [3], [4], [5], and its interpretation is based on understanding that the ℓ_0-(pseudo)norm is a cardinality function that outputs the number of *active elements*, i.e. the non-zero ones. On the other hand, sparse estimation problems can be defined as an inequality constrained optimization problem, see e.g. [6]. A traditional approach to solve ℓ_0-constrained optimization problems is to utilize the so-called Greedy Algorithms [7], which provide a suboptimal solution to the optimization problem. The solution is computed in several iterations, where, on each iteration, the element of the parameter vector that has the most impact on the objective function is included in the solution. The iterations are repeated until the ℓ_0-norm constraint is satisfied.

Classification algorithms such as logistic regression and support vector machine are tools to perform fault detection (see e.g. [8], [9]). Logistic regression is applied when the output is discrete valued and can only take a finite number of values, such as $\{1, 0\}$ or "on–off" [10]. Logistic regressions are based on the *logistic function*, which is defined by a parameter vector that, in turn, defines the *boundaries* of the regions in the space associated with the discrete output values [10], [11].

In this paper we focus on the estimation of sparse parameter vector using an inequality constrained Maximum Likelihood (ML) approach based on the ℓ_0-(pseudo)norm in a logistic regression, in contrast to the more common practice of inducing sparsity by introducing a penalty term. In [12], a sparse logistic regression problem is solved by the use of $\ell_{1/2}$-norm penalization. The $\ell_{1/2}$-norm can be thought as a balance between the ℓ_1-norm and the ℓ_q-norm with q close to 0. In that sense, the $\ell_{1/2}$-norm solution is better than the ℓ_1-norm solution in terms of sparseness, while it is also better than the ℓ_q-(pseudo)norm (q close to 0) in terms of convergence [13]. A similar approach was considered in [14], where a Group Lasso penalized logistic regression model was considered. The Group Lasso [15] has the advantage of doing variable selection on grouped variables in linear regression models. This approach is capable to induce sparsity in the solution of high dimensional problems. In order to obtain sparse estimates, it is common to choose a *bound* or *threshold* to turn into zero the elements of $\boldsymbol{\theta}$ whose estimates exhibit an absolute value that is less than the *bound*. On the other hand, for $0 < q \leq 1$, it is common to express the ℓ_q-norm of the parameter $\boldsymbol{\theta}$ as $\lambda\|\boldsymbol{\theta}\|_q^q$. However, the choice of the hyperparameter λ and q are of great importance, since they define the maximum number of zero entries of $\boldsymbol{\theta}$ that can be identified [16].

Here, we consider the recently published approach for solving a general class of problems [5], where ℓ_0 inequality constrained optimization problems lie in. The approach in [5] accounts for a reinterpretation of the original problem, obtaining an equivalent optimization that is simpler to solve than the original problem. The main goal of our proposal is the attainment of low complexity models, which can be achieved utilizing the ℓ_0-(pseudo)norm. On the other hand, a popular method model slection is Akaike's information Criterion (AIC) [17], which we also consider for comparison purposes in our numerical examples.

Notation: Bold lowercase letters denote vectors and uppercase letters matrices. Calligraphic letters denote sets.

The work of R. Carvajal was supported by Chile's National Fund for Scientific and Technological Development (FONDECYT) through its Postdoctoral Fellowship Programme 2014 – Grant No 3140054. The work of J. C. Agüero was supported by FONDECYT through grant No 1150954.

This work was partially supported by the Advanced Centre for Electrical and Electronic Engineering (AC3E, Proyecto Basal FB0008), Chile.

Peer-reviewed author's copy of:
G. Urrutia, R.A. Delgado, R. Carvajal, D. Katselis, and J.C. Agüero. **Sparse Logistic Regression utilizing Cardinality Constraints and Information Criteria**. In *IEEE Conference on Control Applications (CCA 2016), part of 2016 IEEE Multi-Conference on Systems and Control*, Buenos Aires, Argentina, 2016.
Available at http://doi.org/10.1109/CCA.2016.7587916

Rank($\boldsymbol{A}$) denotes the rank of a matrix and tr($\boldsymbol{A}$) denotes its trace. For a vector $\boldsymbol{x}$, diag($\boldsymbol{x}$) denotes the diagonal matrix with diagonal entries the elements of $\boldsymbol{x}$ and $|\boldsymbol{x}|$ denotes the vector with entries the absolute values of the entries of $\boldsymbol{x}$. $\mathbb{R}^n$ is the set of $x \times 1$ real vectors and $\mathbb{R}^{m\times n}$ is the set of $m \times n$ real matrices. $\mathbb{S}^n$ stands for the set of $n \times n$ symmetric matrices and $\mathbb{S}^n_+$ denotes the set of $n \times n$ positive-semidefinite matrices. For the $n \times n$ matrices A, $B \in \mathcal{S}^n$, $A \geq B$ denotes the Löwner partial ordering, i.e., $A - B \in \mathbb{S}^n_+$. Finally, the superscript ^{T} stands for the transposition operator.

II. Maximum Likelihood Estimation of Logistic Regression Models

A. Logistic Regression

Logistic regression is a classification algorithm that models the output of a categorical dependent variable, that is, the output is discrete valued (taking, in general, a limited and fixed number of possible values) and can express the presence or absence of a given attribute or characteristic (known as class). The model considered in logistic regressions is based on the *sigmoid* or *logistic function*

$$c(z) = \frac{1}{1 + e^{-z}}, \tag{1}$$

where $z \in \mathbb{R}$ is the independent variable used for predictions in the logistic regression model [10]. Notice that $\forall z \in \mathbb{R}$, $0 < c(z) < 1$, which allows for the *logistic function* to be interpreted as an estimation of the probability that a given attribute is present or not. In general, the variable z is defined in terms of a regression and a parameter vector $\boldsymbol{\theta}$. In turn, *training data* (measurements, surveys, etc.) is used to estimate $\boldsymbol{\theta}$. In particular, for a linear classifier, the variable z is modelled as

$$z(\boldsymbol{x}) = \boldsymbol{a}^T\boldsymbol{x} + b, \tag{2}$$

where $\boldsymbol{a} \in \mathbb{R}^M$, $\boldsymbol{x} \in \mathbb{R}^M$, $b \in \mathbb{R}$. Thus, $\boldsymbol{\theta} = [\boldsymbol{a}^T\, b]^T$. The definition of $z(\boldsymbol{x})$ in (2) maps an M dimensional space to the probability of a given class or feature [18].

The resulting model that describes the probability of a given set of attributes belonging to a certain class is given by

$$p(Y = y_i|\boldsymbol{X} = \boldsymbol{x}_i) = c(\boldsymbol{x}_i;\boldsymbol{\theta}) = \frac{1}{1 + e^{-(\boldsymbol{x}_i^T\boldsymbol{a}+b)}}, \tag{3}$$

where $\boldsymbol{X}$ is the set of characteristics or attributes that are related to a given class through the probability of Y for a given $\boldsymbol{x}_i$. For example, if the model considers an "on-off" (or "present-absent") characteristic, then $Y = \{y_1, y_2\}$ represents the set of classes and $\boldsymbol{X}$ is the set of grouped attributes. Then $p(Y = y_1|\boldsymbol{X} = \boldsymbol{x}_i) = 1 - p(Y = y_2|X = \boldsymbol{x}_i)$.

B. Logistic Regression ML Estimation

For a logistic regression, the *likelihood function* is defined in terms of the individual probabilities of each possible value of the output variable. Thus, for a collection of N samples, we have

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{N} p(Y = y_i|\boldsymbol{X} = \boldsymbol{x}_i), \tag{4}$$

where $\boldsymbol{y} = [y_1, ..., y_N]^T$. Hence, the *log-likelihood* function is defined as

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log p(Y = y_i|\boldsymbol{X} = \boldsymbol{x}_i). \tag{5}$$

For a binomial logistic regression, with data set $\{z_i, y_i\}_1^N$, where $y_i \in \{0, 1\}$ and $z_i = z(\boldsymbol{x}_i)$, the *likelihood function* is defined as [18]

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{N} \kappa_i^{z_i}(1 - \kappa_i)^{1-z_i}, \tag{6}$$

with

$$\kappa_i = p(Y = y_i|\boldsymbol{X} = \boldsymbol{x}_i) = \frac{1}{1 + e^{-(\boldsymbol{x}_i^T\boldsymbol{a}+b)}}.$$

In case of indexing the classes as $y_i \in \{-1, 1\}$, after a mathematical arrangement, the probability of the pair $(\boldsymbol{x}_i, y_i)$ can be written in the following general form

$$P(Y = y_i|X = \boldsymbol{x}_i) = \frac{1}{1 + e^{-y_i(\boldsymbol{x}^T\boldsymbol{a}+b)}} \tag{7}$$

Using (5), the negative log-likelihood can be written as follows

$$\ell(\boldsymbol{a}, b|\mathcal{S}) = \sum_{i=1}^{N} lse(0, -y_i(\boldsymbol{x}_i^T\boldsymbol{a} + b) \tag{8}$$

where $lse(\boldsymbol{w}) = log(\sum_{i=1}^{n} e^{w_i})$ is the LogSumExp function.

III. Sparse Parameter Estimation in Logistic Regression Systems

A. Constrained ML Estimation in Logistic Regressions

A sparse estimate for the parameter vector $\boldsymbol{\theta}$ can be obtained by solving the following cardinality-constrained optimization problem:

$$\begin{aligned} \mathcal{P}: \min_{\boldsymbol{\theta}} \quad & \ell(\boldsymbol{y}|\boldsymbol{\theta}) \\ s.t. \quad & ||\boldsymbol{\theta}||_0 \leq r \end{aligned} \tag{9}$$

where the cardinality of $\boldsymbol{\theta}$ is constrained by $||\boldsymbol{\theta}||_0 \leq r$, limiting the complexity of the model. In general, the optimization problem (9) involves a high computational cost. On the other hand, the wrong choice of r might result in an increase of the bias in the estimator (9). To avoid this problem, we consider the alternative and equivalent (has the same global optimum) optimization problem (see Apendix B, Corollary 1)

$$\begin{aligned} \mathcal{P}_{\text{eq}}: \min_{\boldsymbol{\theta},\boldsymbol{w}} \quad & \ell(\boldsymbol{y}|\boldsymbol{\theta}) \\ s.t. \quad & \theta_i w_i = 0 \; i = 1, ..., M+1 \\ & 0 \leq w_i \leq 1, \; i = 1, ..., M+1 \\ & \sum_{i=1}^{M+1} w_i = M + 1 - r \end{aligned} \tag{10}$$

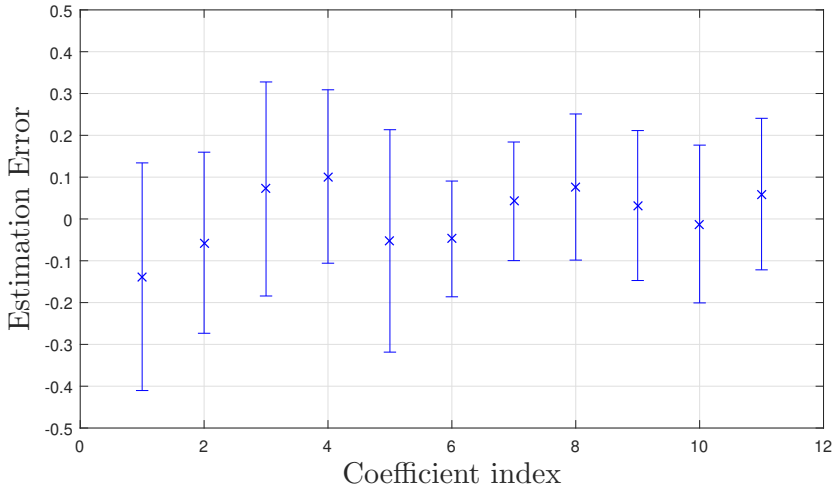

Fig. 1. Parameter estimation via Maximum Likelihood.

where $\boldsymbol{w} = [w_1, ..., w_{M+1}]^T$ is a *latent variable* that allows for this representation of (9) based on the incorporation of a linear and a bilinear constraint. Here, $\ell(\boldsymbol{y}|\boldsymbol{\theta})$ denotes a negative log-likelihood function.

B. AIC Applied to Logistic Regressions

Another way to induce sparsity in a model description is by considering the complexity and the goodness of a model (i.e. the number of elements) in the cost function, as in Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC) [17], [19]. In particular, AIC considers the number of parameters and the likelihood of these parameters fitting the model, and gives a measure of balance between them. The smaller the AIC number, the better the choice of parameters. For a system model defined by a parameter vector $\boldsymbol{\theta}$, AIC is given by

$$AIC(\boldsymbol{\theta}) = 2r - 2\ell(\boldsymbol{y}|\boldsymbol{\theta}) \tag{11}$$

where r is the number of parameters used and $\ell(\boldsymbol{y}|\boldsymbol{\theta})$ is the log-likelihood function of the model. For different model structures, the selection of the best model is carried out simultaneously by the following minimization problem [20]

$$\mathcal{P}_{\text{AIC}}: \ \arg\min_{\boldsymbol{\theta}} \ \ 2(||(\boldsymbol{\theta})||_0) + 2\ell(\boldsymbol{y}|\boldsymbol{\theta}) \tag{12}$$

where the cardinality of $\boldsymbol{\theta}$ is expressed in terms of its ℓ_0-(pseudo) norm. *Remark:* Note that problems (9) and (12) are similar. However in (9) cardinality of the parameter vector is imposed as a hard constraint, whereas in (12) the penalty term promotes sparsity. In the same way, traditional sparsity problems can be rewritten as in (9).

In a similar way to $\mathcal{P}_{\text{eq}}$ in (10), $\mathcal{P}_{\text{AIC}}$ can be reformulated as (see Apendix C):

$$\begin{aligned} \mathcal{P}_{\text{AIC,eq}}: \ \min_{\boldsymbol{\theta},\boldsymbol{w}\in\mathbb{R}^{M+1}} \ & 2(M+1-\sum_{i=1}^{M+1} w_i) + 2\ell(\boldsymbol{\theta}|\mathcal{S}) \\ s.t. \ & \theta_i w_i = 0 \ i = 1, ..., M+1 \\ & 0 \le w_i \le 1, \ i = 1, ..., M+1, \end{aligned} \tag{13}$$

where the term $M + 1 - \sum_{i=1}^{M+1} w_i$ is an upper bound of $||\boldsymbol{\theta}||_0$.

The optimization problems $\mathcal{P}_{eq}$ and $\mathcal{P}_{AIC,eq}$ defined in (10) and (13) are equivalent to the problems $\mathcal{P}$ and $\mathcal{P}_{AIC}$ in (9) and (12) respectively in the sense that they have the same global optima. However, the equivalent problems in (10) and (13) have more local minima than the original ones [21], [22], [5].

The main benefit of these equivalent representations is that they might help to reduce the computational load corresponding to the combinatorial nature of the original problems and have recently been analyzed in [23]. In addition, these problems can be solved by a standard nonlinear programming software, such as BARON [24], [25].

Note that the problems (9) and (12) have non-convex functionals, which implies that traditional optimization toolboxes such as CPLEX and CVX cannot be directly applied.

IV. Numerical Example

In this section we investigate the performance of the proposed strategy to the solution of the sparse logistic regression. Our motivation comes from a specific application that has attracted a lot of interest in recent years in the bioinformatics community, namely the gene selection based on DNA microarrays for the diagnosis of cancer. The main goal is to identify the gene biomarkers so that different types of cancer are easily classified and predicted with high accuracy. The corresponding mathematical problem for gene selection involves an appropriate regularization step to adequately deal with the high dimensionality and ill-conditioning of the selection process. This is due to the fact that from a biological perspective, only a small subset of genes is strongly indicative of a targeted disease and most of the genes are irrelevant to the classification of different types of cancer [12]. Irrelevant selected genes may reduce the accuracy and distort the process of classification.

For our problem, let us suppose that we have N samples $\mathcal{S} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), ..., (\boldsymbol{x}_N, y_N)\}$, where $\boldsymbol{x}_i =$

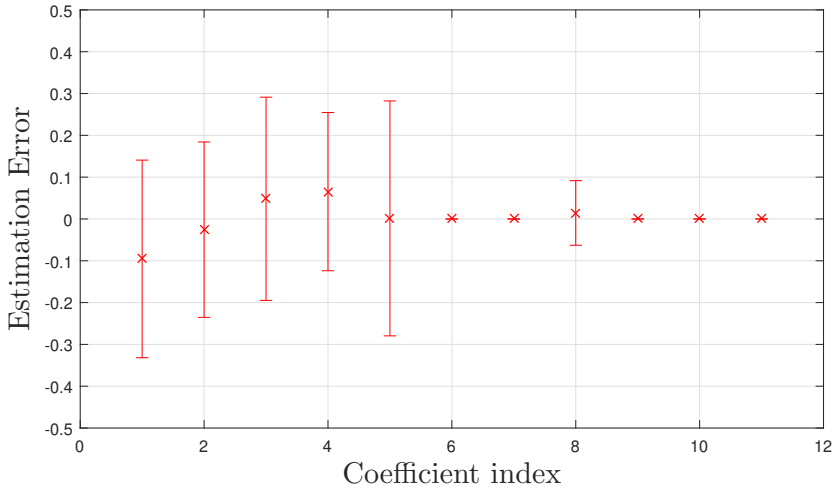

Fig. 2. Parameter estimation via ℓ_0-norm constrained Maximum Likelihood.

$[x_{i1}, x_{i2}, ..., x_{iM}]^T$ is the ith input pattern, denoting the M features/genes and $y_i \in \{-1, 1\}$ indicates the class of the corresponding sample with respect to a specific cancer type.

The classifier for any input/output pair $(\boldsymbol{x}_i, y_i)$ is selected to be the logistic regression model shown in (3). Therefore, for any gene set $\boldsymbol{x}_i$, the classifier $c(\boldsymbol{x}_i; \boldsymbol{\theta})$ predicts the corresponding class of cancer type y_i. Thus the probability of cancer type $Y = 1$ for a given set of attributes can be written as:

$$P(Y = 1|X = \boldsymbol{x}_i) = c(\boldsymbol{x}_i; \boldsymbol{a}, b). \tag{14}$$

which corresponds to the probability that a given gene pattern belongs to a class of a specific cancer type, given a certain linear combination of the predictors (genes).

The simulation setup is as follows:

i) a set of vectors $\boldsymbol{f}_{i0}, \boldsymbol{f}_{i1}, ..., \boldsymbol{f}_{ip}$, $i = 1, 2, ..., N$ is drawn according to the standard normal distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$,

ii) given a correlation coefficient $\rho \in [0, 1]$, the entries of the regressors are generated according to the relationship $x_{ij} = f_{ij}\sqrt{1-\rho} + f_{i0}\sqrt{\rho}$, $j = 1, 2, ..., M$ [26],

iii) the simulated data is generated according to the logistic model $z(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{a} + b$,

iv) additive noise is included in order to account for unobserved differences that may correspond to modelling error or measurement noise, obtaining

$$\tilde{c}(\boldsymbol{x}_i; \boldsymbol{\theta}) = \frac{1}{1 + e^{-(\boldsymbol{a}^T\boldsymbol{x}_i + b + \nu_i)}}, \tag{15}$$

with $\nu_i \sim \mathcal{N}(0, \sigma^2)$, $\sigma = 0.2$.

For the simulations, $M = 10$, $N = 200$, $b = 0$, $\boldsymbol{a} = [1\,1\,-1\,-1\,1\,0\,0\,0\,0\,0]^T$, and the cardinality of the solution (r) is set to be equal to the number of parameters used in (15), that is $||\boldsymbol{a}||_0 = 5$ (since $b = 0$). We also consider 30 Monte Carlo simulations. Note that $\boldsymbol{\theta} = (\boldsymbol{a}, b)$.

In order to compare the results, we calculate the average error and standard deviation of the estimates for each coefficient. The average error $\bar{\epsilon}_i$ for an estimated coefficient is taken as follows

$$\bar{\epsilon}_i = \frac{1}{N_{exp}} \sum_{j=1}^{N_{exp}} \theta_i^0 - \theta_i^{[j]} \tag{16}$$

where N_{exp} is the number of Monte Carlo simulations, θ_i^0 is the true parameter and $\theta_i^{[j]}$ is the i-th parameter estimate corresponding to the j-th experiment.

The standard deviation σ_i for each estimate is calculated as follows

$$\sigma_i = \sqrt{\frac{1}{N_{exp}} \sum_{j=1}^{N_{exp}} (\theta_i^{[j]} - \bar{\epsilon}_i)^2} \tag{17}$$

The results of the numerical examples are shown in Figs. 1, 2, and 3, where the average error and standard deviation of the estimates are plotted. In Fig. 1 we show the estimation error and standard deviation for ML estimation. The average number of non-zero elements used in the solutions is 11, thus the estimates are clearly not sparse. In Fig. 2 the estimation error and deviation of the estimates of problem $\mathcal{P}_{\text{eq}}$ are shown. Clearly, the estimation error is smaller than the one obtained by using ML, and the standard deviation is even much smaller for the coefficients that are zero in the original model. Finally, in Fig. 3 we show the corresponding error estimates and deviation of the estimates for AIC. The estimation error is bigger compared to the ones obtained from $\mathcal{P}_{\text{eq}}$, but smaller in comparison with the ML approach. The standard deviations of the last six coefficients are also smaller than the ML approach, showing that the most plausible model only included the first 5 nonzero elements of $\boldsymbol{a}$ and $b = 0$.

V. Conclusion

In this paper we addressed a sparse logistic regression estimation problem. We have studied two approaches: i)

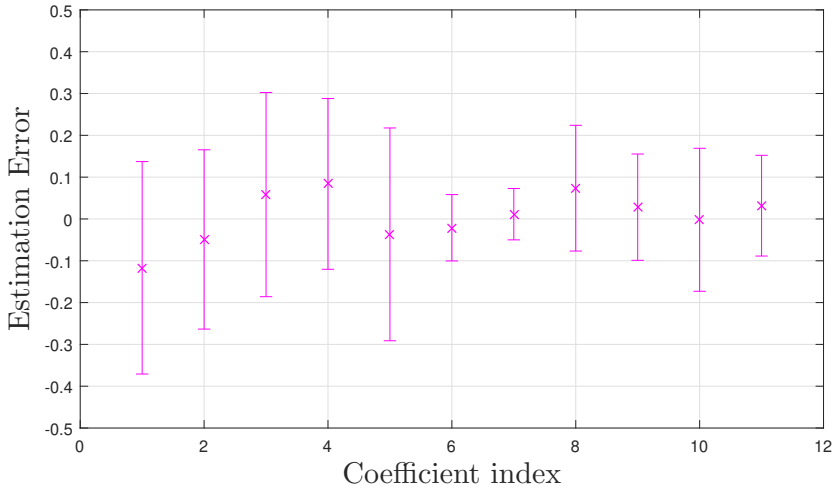

Fig. 3. Parameter estimation via Akaike's Information Criteria (AIC).

penalized Maximum Likelihood via AIC and ii) constrained ML. The inequality constraint is obtained by utilizing the ℓ_0-(pseudo) norm of the parameter vector, which accounts for its cardinality. The corresponding optimization problem is then rewritten in an equivalent form, yielding a less computationally demanding problem. We compared our solutions against standard Maximum Likelihood estimation estimates. The simulations show that constrained ML performs better than AIC and ML when the number of features is known.

REFERENCES

[1] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. R. Stat. Soc. B*, vol. 58, pp. 267–288, 1996.

[2] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. R. Stat. Soc. B*, vol. 67, no. 2, pp. 301–320, 2005.

[3] R. P. Aguilera, R. Delgado, D. Dolz, and J. C. Agüero, "Quadratic mpc with l0-input constraint," in *19th World Congress The International Federation of Automatic Control*, 2014.

[4] R. A. Delgado, J. C. Agüero, and G. C. Goodwin, "A rank-constrained optimization approach: Application to factor analysis," in *19th IFAC World Congress*, 2014.

[5] ——, "A novel representation of rank constraints for real matrices," *Linear Algebra and its Applications*, vol. 496, pp. 452–462, 2016.

[6] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *Information Theory, IEEE Transactions on*, vol. 53, no. 12, pp. 4655–4666, 2007.

[7] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, 3rd ed. The MIT Press, 2009.

[8] C. Batur, L. Zhou, and C.-C. Chan, "Support vector machines for fault detection," in *Decision and Control, 2002, Proceedings of the 41st IEEE Conference on*, vol. 2. IEEE, 2002, pp. 1355–1356.

[9] L. Bako, G. Mercere, S. Lecoeuche, and M. Lovera, "Recursive subspace identification of hammerstein models based on least squares support vector machines," *IET Control Theory and Applications*, vol. 3, no. 9, pp. 1209–12, 2009.

[10] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference and prediction*, 2nd ed. Springer, 2009.

[11] D. R. Cox, "The regression analysis of binary sequences," *J. R. Stat. Soc. B*, vol. 20, no. 2, pp. 215–242, 1958.

[12] Y. Liang, C. Liu, X.-Z. Luan, K.-S. Leung, T.-M. Chan, Z.-B. Xu, and H. Zhang, "Sparse logistic regression with a l 1/2 penalty for gene selection in cancer classification," *BMC bioinformatics*, vol. 14, no. 1, p. 1, 2013.

[13] Z. Xu, H. Zhang, Y. Wang, X. Chang, and Y. Liang, "L1/2 regularization," *Science China Information Sciences*, vol. 53, no. 6, pp. 1159–1169, 2010.

[14] L. Meier, S. Van De Geer, and P. Bühlmann, "The group lasso for logistic regression," *J. R. Stat. Soc. B*, vol. 70, no. 1, pp. 53–71, 2008.

[15] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. R. Stat. Soc. B*, vol. 68, no. 1, pp. 49–67, 2006.

[16] X. Chen, F. Xu, and Y. Ye, "Lower bound theory of nonzero entries in solutions of ℓ_2-ℓ_p minimization," *SIAM J. Sci. Comput.*, vol. 32, no. 5, pp. 2832–2852, 2010.

[17] H. Akaike, "Automatic data structure search by the maximum likelihood," in *Computers in Biomedicine, A Supplement to the Proceedings of the Fifth Hawaii International Conference on Systems Sciences*, 1972, pp. 99–101.

[18] C. M. Bishop, *Pattern Recognition and Machine Learning*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[19] G. Schwarz *et al.*, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[20] L. Ljung, *System Identification: Theory for the User*, 2nd ed., ser. 2nd Edition. Prentice Hall, 1999.

[21] O. Burdakov, C. Kanzow, and A. Schwartz, "On a reformulation of mathematical programs with cardinality constraints," in *Advances in Global Optimization*. Springer, 2015, pp. 3–14.

[22] M. Feng, J. E. Mitchell, J.-S. Pang, X. Shen, and A. Wächter, "Complementarity formulations of l0-norm optimization problems," Technical report, Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY, Tech. Rep., 2013.

[23] O. P. Burdakov, C. Kanzow, and A. Schwartz, "Mathematical programs with cardinality constraints: Reformulation by complementarity-type conditions and a regularization method," *SIAM Journal on Optimization*, vol. 26, no. 1, pp. 397–425, 2016.

[24] M. Tawarmalani and N. V. Sahinidis, "A polyhedral branch-and-cut approach to global optimization," *Mathematical Programming*, vol. 103, pp. 225–249, 2005.

[25] N. V. Sahinidis, *BARON 14.3.1: Global Optimization of Mixed-Integer Nonlinear Programs,* User's Manual, 2014.

[26] I. Sohn, J. Kim, S.-H. Jung, and C. Park, "Gradient lasso for cox proportional hazards model," *Bioinformatics*, vol. 25, no. 14, pp. 1775–1781, 2009.

[27] M. C. Grant and S. P. Boyd, "Graph implementations for nonsmooth convex programs," in *Recent advances in learning and control*. Springer, 2008, pp. 95–110.

APPENDIX

In this appendix we include the following results [5] for clarity and completeness of the presentation.

A. Rank-Constrained Optimization

First we show a result applied in rank-constrained optimization.

Theorem 1: Let $G \in \mathbb{R}^{m\times n}$ then the following expressions are equivalent:

1) $rank(G) \leq r$
2) $\exists W_R \in \Phi_{n,r}$, such that $GW_R = 0_{m\times n}$
3) $\exists W_L \in \Phi_{m,r}$, such that $W_L G = 0_{m\times n}$

where

$$\Phi_{n,r} = \{W \in \mathbb{S}^n, 0 \leq W \leq I, trace(W) = n - r\} \tag{18}$$

Proof: For a detailed proof see [5].

□□□

Consider the following general rank-constrained optimization problem over a convex set

$$\begin{aligned} \mathcal{P}_{rco}: \min_{\boldsymbol{x}\in\mathbb{R}^p} \quad & f(\boldsymbol{x}) \\ s.t. \quad & \boldsymbol{x} \in \Omega \\ & rank(G(\boldsymbol{x})) \leq r \end{aligned} \tag{19}$$

Now consider the following optimization problem that incorporates bilinear constraints

$$\begin{aligned} \mathcal{P}_{rco,equiv}: \min_{\boldsymbol{x}\in\mathbb{R}^p, W\in\mathbb{S}^n} \quad & f(\boldsymbol{x}) \\ s.t. \quad & \boldsymbol{x} \in \Omega \\ & G(\boldsymbol{x})W = 0_{m\times n} \\ & 0 \leq W \leq I_n \\ & trace(W) = n - r \end{aligned} \tag{20}$$

where $\Omega \subset \mathbb{R}^p$ is a constraining set, the cost function $f : \mathbb{R}^p \to \mathbb{R}$ is the objective function and $G : \mathbb{R}^p \to \mathbb{R}^{m\times n}$.

Considering Theorem 1, the two optimization problems $\mathcal{P}_{rco}$ and $\mathcal{P}_{rco,equiv}$ stated before are equivalent.

Theorem 2: A vector $\boldsymbol{x}^\star \in \mathbb{R}^p$ is a global solution of $\mathcal{P}_{rco}$ if and only if there exists a $W^\star$ such that the pair $(\boldsymbol{x}^\star, W^\star)$ is a global solution of $\mathcal{P}_{rco,equiv}$.

Proof: For a detailed proof, see [5].

□□□

B. Cardinality-Constrained Optimization

The problem $\mathcal{P}_{rco}$ can also be used to solve ℓ_0-norm constrained problems (i.e cardinality-constrained optimization problem) instead of those involving the $rank()$ of a matrix. The cardinality problem can be stated from $\mathcal{P}_{rco}$ by considering $G(\boldsymbol{x}) = diag(\boldsymbol{x})$. By this definition, the cardinality of $\boldsymbol{x}$ is expressed by means of the rank of a matrix, $||\boldsymbol{x}||_0 = rank(G(\boldsymbol{x}))$.

This particular case arises the following corollary:

Corollary 1: [5] Let $\boldsymbol{x} \in \mathbb{R}^n$, then the cardinality of $\boldsymbol{x}$ is $||\boldsymbol{x}||_0 \leq r$ if and only if there exist a $\boldsymbol{w} \in \{w \in \mathbb{R}^n | 0 \leq w_i \leq 1, i = 1, ..., n; \sum_{n=1}^{n} w_i = n - r\}$, such that $x_i w_i = 0$, $\forall i = 1, .., n$.

Next we consider the following ℓ_0-norm constrained optimization problem

$$\begin{aligned} \mathcal{P}_{\ell_0 co}: \min_{\boldsymbol{x}\in\mathbb{R}^n} \quad & f(\boldsymbol{x}) \\ s.t. \quad & \boldsymbol{x} \in \Omega \\ & ||\boldsymbol{x}||_0 \leq r \end{aligned} \tag{21}$$

From Corollary 1, problem $\mathcal{P}_{\ell_0 co}$ can be reformulated as a optimization problem subject to bilinear constraints as follows

$$\begin{aligned} \mathcal{P}_{\ell_0 co,equiv}: \min_{\boldsymbol{x}\in\mathbb{R}^n, \boldsymbol{w}\in\mathbb{R}^n} \quad & f(\boldsymbol{x}) \\ s.t. \quad & w_i x_i = 0, \ i = 1, ..., n \\ & 0 \leq w_i \leq 1 \ i = 1, ..., n \\ & \sum_{i=1}^{n} w_i = n - r \\ & \boldsymbol{x} \in \Omega \end{aligned} \tag{22}$$

C. Cost function involving ℓ_0-norm

Theorem 1 can also be used in problems that incorporate the ℓ_0-norm in its cost functional. This is done by using the epigraph notation [27] as follows

$$\begin{aligned} \mathcal{P}_{rm}: \min_{\boldsymbol{\theta}\in\mathbb{R}^p} \quad & r \\ s.t. \quad & \boldsymbol{\theta} \in \Omega \\ & rank(G(\boldsymbol{\theta})) \leq r \end{aligned} \tag{23}$$

where $G(\boldsymbol{x}) = diag(\boldsymbol{x})$. Problem $\mathcal{P}_{rm}$ is equivalent to $\mathcal{P}_{rm,eq}$

$$\begin{aligned} \mathcal{P}_{rm,eq}: \min_{\boldsymbol{\theta}\in\mathbb{R}^p, W\in\mathbb{S}^n} \quad & n - trace(W) \\ s.t. \quad & \boldsymbol{\theta} \in \Omega \\ & G(\boldsymbol{\theta})W = 0_{m\times n} \\ & 0 \leq W \leq I_n \end{aligned} \tag{24}$$