Sparse identification of posynomial models

Automatica, 2015, in press

Sparse Identification of Posynomial Models

G.C. Calafiore[a], L. El Ghaoui[b], C. Novara[a]

[a] *DAUIN, Politecnico di Torino, Italy.*
Email: giuseppe.calafiore@polito.it, carlo.novara@polito.it

[b] *EECS and IEOR, UC Berkeley, CA, USA.*
Email: elghaoui@berkeley.edu

Abstract

Posynomials are nonnegative combinations of monomials with possibly fractional and both positive and negative exponents. Posynomial models are widely used in various engineering design endeavors, such as circuits, aerospace and structural design, mainly due to the fact that design problems cast in terms of posynomial objectives and constraints can be solved efficiently by means of a convex optimization technique known as geometric programming (GP). However, while quite a vast literature exists on GP-based design, very few contributions can yet be found on the problem of identifying posynomial models from experimental data. Posynomial identification amounts to determining not only the coefficients of the combination, but also the exponents in the monomials, which renders the identification problem numerically hard. In this paper, we propose an approach to the identification of multivariate posynomial models, based on the expansion on a given large-scale basis of monomials. The model is then identified by seeking coefficients of the combination that minimize a mixed objective, composed by a term representing the fitting error and a term inducing sparsity in the representation, which results in a problem formulation of the "square-root LASSO" type, with nonnegativity constraints on the variables. We propose to solve the problem via a sequential coordinate-minimization scheme, which is suitable for large-scale implementations. A numerical example is finally presented, dealing with the identification of a posynomial model for a NACA 4412 airfoil.

1 Introduction

A posynomial model is defined by a function ψ of the form $\psi(w) = \sum_{i=1}^{n_c} c_i w^{\alpha_i}$, where $w \in \mathbb{R}^{n_w}_{++}$ (the positive orthant), $\psi(w) \in \mathbb{R}$, $c_i \geq 0$ are coefficients, $\alpha_i = [\alpha_{i1} \cdots \alpha_{in_w}]^\top \in \mathbb{R}^{n_w}$ are vectors of exponents with $\alpha_{ij} \in \mathbb{R}$, and w^{α_i} is defined as $w^{\alpha_i} \doteq \prod_{j=1}^{n_w} w_j^{\alpha_{ij}}$. The term $c_i w^{\alpha_i}$ is called a *monomial*. Note that, while in polynomial models the exponents α_{ij} are nonnegative integers, in posynomial models these exponents may also be negative and/or noninteger.

Posynomial models are of great importance in many fields of technology, ranging from structural design, network flow, optimal control (see (Beightler & Phillips 1976, Wilde 1978)), to aerospace system design (Hoburg & Abbeel 2012), circuit design (Boyd, Kim, Patil & Horowitz 2005, Daems, Gielen & Sansen 2003, Sapatnekar, Rao, Vaidya & Kang 1993), antennas (Babakhani, Lavaei, Doyle & Hajimiri 2010) and communication systems (Chiang 2005). The interest in posynomials is motivated by the fact that they lead to computationally efficient geometric programming models for optimal system design, see, e.g., (Duffin, Peterson & Zener 1967, Beightler & Phillips 1976, Wilde 1978).

Despite the fact that a consistent number of papers is available in the literature where posynomial models and geometric programming are used for design purposes, very few works can be found to date addressing the relevant problem of identifying a posynomial model from experimental data; see (Daems et al. 2003) for such an exception. Typically, the model is assumed known (i.e., the coefficients c_i and the exponents α_{ij} are assumed known), and then it is processed by geometric programming to obtain an optimal design. However, in most real-world applications, the model is *not* known a priori, and it has to be identified from experimental data. Note that both the coefficients c_i and the exponents α_{ij} have to be estimated, and this task is quite hard, making the identification problem significantly more difficult than a linear regression problem.

Identification of posynomial models can be performed following the standard approach used for polynomials. In this approach, an heuristic search finalized at finding a viable model structure, i.e., a suitable set of exponent vectors $\{\alpha_i\}$ is first carried out. Once the exponent vector set has been chosen, the coefficients c_i are estimated by means of least-squares or other convex optimization algorithms, see, e.g., (Spinelli, Piroddi & Lovera 2006, Pulecchi & Piroddi 2007, Daems et al. 2003, Novara, Vincent, Hsu, Milanese & Poolla 2011). A critical issue in this approach is that the model structure search may be extremely time consuming and in most cases leads only to approximate model structures. An alternative approach is to assume (or estimate by means of some heuristic) a value $\hat{n}_c$ for the basis cardinality n_c, and then estimate c_i and α_i by means of nonlinear programming algorithms. However, these kind of algorithms are non-convex and thus do not ensure convergence to the optimal parameter estimate. A third approach, which overcomes the issues of the other two, consists in considering an over-parametrized model and inserting in the optimization problem a sparsity promoting term, given by the ℓ_1-norm of the coefficient vector. This term allows one to efficiently select the model structure and, at the same time, to avoid the problem of overfitting. This approach is based on the well-known LASSO (least absolute shrinkage and selection operator) or other similar algorithms, see, e.g., (Tibshirani 1996, Kukreja, Lofberg & Brenner 2006, De Mol, De Vito & Rosasco 2009, Bonin, Seghezza & Piroddi 2010, Novara 2012). The optimization problem is in this case convex but, due to the over-parametrization, it typically involves a very large number of decision variables.

In this paper, we propose a novel posynomial identification method, based on this latter approach: we minimize a convex objective, defined as the sum of a regularized accuracy term based on the ℓ_2-norm of the estimation residual, and a sparsity-inducing term given by a weighted ℓ_1-norm of the coefficient vector. We name this method *nonnegative regularized square-root LASSO* or nnrsqrt-LASSO. Within this method, we provide three main contributions. The first one is an optimization technique for nonnegative constrained sqrt-LASSO problems, which exploits an a-priori condition, called *feature elimination*. This condition, if satisfied for a certain monomial, guarantees that this monomial does not appear in the representation (i.e., it has a null coefficient). The condition is very easy to check, and can thus be applied a-priori, in a pre-optimization phase, to eliminate all the monomials which are not needed to explain the data. For the optimization phase, a large-scale-capable algorithm is proposed to solve the nnrsqrt-LASSO problem. The algorithm is based on a sequential coordinate minimization scheme where, at each step, a univariate optimization problem is solved analytically (and thus very efficiently). The sequential scheme is shown to converge to an optimal solution. The second main contribution is the idea of using such a computational framework for efficient identification of posynomial models. The third contribution consists in the application of the technique to a problem of interest in the aerospace filed. In particular, we show through a numerical example that the method is effective for identifying reliable posynomial models for airfoils.

2 Identification of posynomial models

2.1 Model setup

Consider a posynomial $\psi^o(w) = \sum_{i=1}^{n_c} c_i^o w^{\alpha_i^o}$, where the coefficients c_i^o, the exponent vectors α_i^o and the expansion cardinality n_c are not known. Suppose that a set of noise-corrupted measurements is available: $\mathfrak{D} = \{y(k), w(k)\}_{k=1}^m$, where $y(k) = \psi^o(w(k)) + e(k)$, and $e(k) \in \mathbb{R}$ is a noise term. The problem considered in this paper is to estimate from these data the unknown parameters c_i^o, α_i^o, $i = 1, \ldots, n_c$, and the cardinality n_c. To this end, we define an over-parametrized posynomial family

$$\psi(w) = \sum_{i=1}^{n} x_i w^{\alpha_i} \tag{1}$$

where $n \gg n_c$. In real-world situations, this over-parametrization can be obtained from the available prior information on the exponents α_{ij}^o. Formally, we assume that the following prior information is available on the exponents: $\alpha_{ij} \in Q_j$, $i = 1, \ldots, n$, where Q_j is a set of exponents which, on the basis of the available prior information, can be considered reasonable for the variable w_j. Then, the set of exponent vectors $\{\alpha_i\}_{i=1}^n$ defining the over-parametrization (1) can be constructed as $S_\alpha \doteq \{\alpha_i\}_{i=1}^n = \prod_{j=1}^{n_w} Q_j$ (the Cartesian product of the Q_js). Note that this approach can be adopted also if an exponent is known to belong to a continuous (finite) interval, in which case the set Q_j can be obtained by properly discretizing the interval. If the a-priori information is correct, then S_α is guaranteed to contain the true exponent vectors: $S_\alpha \supset S_{\alpha^o} \doteq \{\alpha_i^o\}_{i=1}^{n_c}$.

2.2 Square-root LASSO formulation

Model identification is here performed by minimizing with respect to the coefficients x_i in the expansion (1) an objective function defined as the sum of an accuracy objective and a sparsity-promoting term, allowing us to select, in the over-parametrized family, a parsimonious model structure. Define $y = [y(1) \cdots y(m)]^\top$, $x = [x_1 \cdots x_n]^\top$, and $\Phi = \begin{bmatrix} w(1)^{\alpha_1} & \cdots & w(1)^{\alpha_n} \\ \vdots & \ddots & \vdots \\ w(m)^{\alpha_1} & \cdots & w(m)^{\alpha_n} \end{bmatrix}$.

The objective we consider is of the form

$$f(x) \doteq \left\| \begin{bmatrix} \Phi x - y \\ \sigma x \end{bmatrix} \right\|_2 + \lambda^\top |x|, \tag{2}$$

where $\sigma \geq 0$, $\lambda \in \mathbb{R}^n$ with $\lambda \geq 0$ (component-wise), and $|x|$ denotes a vector whose entries are the absolute values of the entries in x. We define, for notational compactness, $\tilde{\Phi} \doteq \begin{bmatrix} \Phi \\ \sigma I \end{bmatrix}$, $\tilde{y} \doteq \begin{bmatrix} y \\ 0 \end{bmatrix}$, $\tilde{\phi}_i \doteq \begin{bmatrix} \phi_i \\ \sigma e_i \end{bmatrix}$, where $\tilde{\phi}_i$ denotes the ith column of $\tilde{\Phi}$, and e_i is the i-th vector of the standard basis of $\mathbb{R}^n$. The objective thus becomes

$$f(x) \doteq \|\tilde{\Phi}x - \tilde{y}\|_2 + \lambda^\top |x|. \tag{3}$$

Note that $\lambda^\top |x|$ is a weighted ℓ_1-norm. Vector λ is thus a penalty factor which quantifies the tradeoff between the accuracy objective $\|\tilde{\Phi}x - \tilde{y}\|_2$ and the term $\lambda^\top |x|$, which is a proxy for sparsity in the solution, see (Donoho, Elad & Temlyakov 2006, Candes & Tao 2006). Clearly, for $\lambda = \gamma \mathbf{1}$ (where $\mathbf{1}$ is a vector with all entries equal to one), and $\sigma = 0$, the rsqrt-LASSO problem coincides with the standard sqrt-LASSO. The use of the sparsity promoting term $\lambda^\top |x|$ instead of the standard term $\gamma\|x\|_1$ allows for more flexibility, in problems where the entries of x have different scales, see, e.g., (Daubechies, Defrise & Mol 2004, Zou 2006, Carvajal, Godoy, Aguero & Goodwin 2012). The regularization parameter $\sigma \geq 0$ is introduced to improve the numerical conditioning of the problem, guaranteeing (if $\sigma > 0$) that $\tilde{\Phi}$ has full rank, and that the ℓ_2 term of the objective remains differentiable for all x, if $y \neq 0$.

We hence consider the following two optimization problems, which we name regularized square-root LASSO (rsqrt-LASSO)

$$p^* \doteq \min_{x \in \mathbb{R}^n} f(x), \tag{4}$$

and nonnegative regularized square-root LASSO (nnrsqrt-LASSO)

$$p^*_+ \doteq \min_{x \in \mathbb{R}^n_+} f(x), \tag{5}$$

where $\mathbb{R}^n_+ \doteq \{x \in \mathbb{R}^n : x \geq 0\}$ (the inequality is component-wise). The first model can be used for polynomial model identification, and the second one for posynomial model identification, which is the main focus of this paper.

Remark 1 The proposed regularized square-root LASSO and nnsquare-root LASSO models are quite different from the standard LASSO, (Tibshirani 1996), the Elastic-Net, (Zou & Hastie 2005, De Mol et al. 2009), and the LAR (Least Angle Regression), (Efron, Hastie, Johnstone & Tibshirani 2004), models. In fact, these latter formulations do not allow natively for a-priori feature elimination (see Remark 2 below), and, moreover, the corresponding algorithms are not tailored for dealing with the non-negativity constraints on x that we need for our posynomial identification application. Further, although we do not develop here a specific statistical analysis of our model, we remark that it has been observed in, e.g., (Belloni, Chernozhukov & Wang 2011, Babu & Stoica 2014), that using a plain ℓ_2 term instead of a squared one may lead to "pivotal" estimates that are less sensitive to the noise level in the data, and hence to the choice of the regularization parameter λ. □

In the following sections, we describe a simple scheme for solving both the unconstrained and the constrained versions of the regularized sqrt-LASSO problem, based on a two-phase procedure: problem reduction using feature elimination, followed by a coordinate-minimization (CM) scheme applied to the reduced problem. There clearly exist other general algorithms that can potentially be used for the numerical solution of our problem, such as, for instance, the FISTA algorithm (Beck & Teboulle 2009), or the forward-backward splitting method of (Duchi & Singer 2009); see also (Combettes & J.-C-Pesquet 2007). A brief discussion on the numerical performance of these methods compared to the CM method on a numerical example is reported in Section 6.

We shall assume throughout that $y \neq 0$, since for $y = 0$ the optimal solution of both problems (4), (5) is trivially $x^* = 0$.

3 Dual formulations and feature elimination

We next derive dual formulations of the rsqrt-LASSO and nnrsqrt-LASSO problems, and then show how a feature elimination condition is obtained from these dual formulations.

3.1 Dual of the rsqrt-LASSO problem

We here derive a dual formulation for problem (4). We use the fact that $\|\tilde{\Phi}x - \tilde{y}\|_2 = \max_{\|u\|_2 \leq 1} u^\top(\tilde{\Phi}x - \tilde{y})$, and $\lambda^\top |x| = \sum_{i=1}^n \lambda_i |x_i| = \max_{|v| \leq \lambda} v^\top x$. We thus rewrite problem (4) as

$$p^* = \min_{x \in \mathbb{R}^n} \max_{\|u\|_2 \leq 1, |v| \leq \lambda} u^\top(\tilde{\Phi}x - \tilde{y}) + v^\top x.$$

Then, a standard saddle-point result (see, for instance, Sion's theorem, (Komiya 1988, Sion 1958)), prescribes that we may exchange the order of min and max in the previous expression without changing the optimal value, whence

$$p^* = \max_{\|u\|_2 \leq 1, |v| \leq \lambda} \min_{x \in \mathbb{R}^n} u^\top(\tilde{\Phi}x - \tilde{y}) + v^\top x.$$

Notice further that the infimum over $x \in \mathbb{R}^n$ of the term $(u^\top \tilde{\Phi} + v^\top)x$ is $-\infty$, unless the coefficient $u^\top \tilde{\Phi} + v^\top$ is zero, hence

$$\begin{aligned} p^* = \max_{u,v} \quad & -u^\top \tilde{y} \\ \text{s.t.:} \quad & \tilde{\Phi}^\top u + v = 0, \ \|u\|_2 \leq 1, \ |v| \leq \lambda. \end{aligned}$$

Eliminating the v variable, we obtain the following formulation for the dual of problem (4)

$$p^* = \max_u \quad -u^\top \tilde{y} \tag{6}$$

$$\begin{aligned} \text{s.t.:} \quad & \|u\|_2 \leq 1 \\ & |\tilde{\phi}_i^\top u| \leq \lambda_i, i = 1, \ldots, n. \end{aligned} \tag{7}$$

3.2 Dual of the nnrsqrt-LASSO problem

The derivation of the dual for the nnrsqrt-LASSO problem (5) follows similar lines, noticing that, for $x \geq 0$, we have $\lambda^\top |x| = \lambda^\top x$, hence

$$p^*_+ = \max_{\|u\|_2 \leq 1} \min_{x \geq 0} u^\top(\tilde{\Phi}x - \tilde{y}) + \lambda^\top x,$$

and the infimum over $x \geq 0$ of the term $(u^\top \tilde{\Phi} + \lambda^\top)x$ is $-\infty$, unless $u^\top \tilde{\Phi} + \lambda^\top \geq 0$, thus

$$p^*_+ = \max_u \quad -u^\top \tilde{y} \tag{8}$$

$$\begin{aligned} \text{s.t.:} \quad & \|u\|_2 \leq 1 \\ & \tilde{\phi}_i^\top u + \lambda_i \geq 0, i = 1, \ldots, n. \end{aligned} \tag{9}$$

3.3 Safe feature elimination

In this section we analyze the dual formulations of problems (4), (5), in order to derive a simple sufficient condition that permits one to predict when an entry x_i is zero at optimum, and hence to eliminate a priori some features (i.e., columns of $\tilde{\Phi}$) from the problem. This type of condition, first introduced by (El Ghaoui, Viallon & Rabbani 2012) in the context of the standard LASSO problem, is named *safe feature elimination*. Observe that $\max_{\|u\|_2\leq 1} |\tilde{\phi}_i^\top u| = \|\tilde{\phi}_i\|_2$. Therefore, if for some $i \in \{1,\ldots,n\}$ it holds that $\|\tilde{\phi}_i\|_2^2 = \|\phi_i\|_2^2 + \sigma^2 < \lambda_i^2$, then the corresponding constraint in (7), as well as in (9), will certainly be satisfied with strict inequality, that is, it will be *inactive* at the optimum. This means that it can be safely eliminated from the dual optimization problem, without changing the optimal objective value. Such reduced dual is associated with a reduced primal problem where the columns $\tilde{\phi}_i$ corresponding to the inactive dual constraints are removed from matrix $\tilde{\Phi}$. This in turn implies that the corresponding variables x_i in the primal problem are set to zero, i.e.,

$$\|\phi_i\|_2^2 + \sigma^2 < \lambda_i^2 \quad \Rightarrow \quad x_i^* = 0. \tag{10}$$

Remark 2 Feature elimination permits one to eliminate some features (and thus to reduce the variable size) *before* we actually run the algorithm, by just checking the condition in Eq. (10). This is not true for the LASSO and, e.g., for the problem studied in (Hale, Yin & Zhang 2008), where only an a-posteriori elimination condition is obtained, i.e., a condition checkable only *after* the algorithm has been run. □

4 Univariate solutions

Consider the following rsqrt-LASSO problem with a single scalar variable x:

$$\min_{x\in\mathbb{R}} f(x) \doteq \left\| \begin{bmatrix} \phi x - y \\ \sigma e x - \xi \end{bmatrix} \right\|_2 + \lambda |x|, \tag{11}$$

where $\lambda \geq 0$ (now a scalar), $\sigma \geq 0$, $\phi \in \mathbb{R}^m, y \in \mathbb{R}^m$, $\xi \in \mathbb{R}^n$ are given, and e is a vector of all zeros, except for an entry in generic position i, which is equal to one, and correspondingly we postulate that $\xi_i = 0$, thus it holds that $e^\top \xi = 0$. We set for convenience

$$\tilde{\phi} \doteq \begin{bmatrix} \phi \\ \sigma e \end{bmatrix}, \quad \tilde{y} \doteq \begin{bmatrix} y \\ \xi \end{bmatrix}. \tag{12}$$

Thus, problem (11) rewrites to

$$\min_{x\in\mathbb{R}} f(x) \doteq \|\tilde{\phi} x - \tilde{y}\|_2 + \lambda |x|. \tag{13}$$

We assume that $\tilde{y} \neq 0$ and $\tilde{\phi} \neq 0$, otherwise the optimal solution is simply $x = 0$. Let us define

$$x_{\text{ls}} \doteq \frac{\tilde{\phi}^\top \tilde{y}}{\|\tilde{\phi}\|_2^2} = \frac{\phi^\top y}{\|\phi\|_2^2 + \sigma^2},$$

which corresponds to the solution of the problem for $\lambda = 0$. The following theorem holds.

Theorem 1 *Let $\tilde{y} \neq 0$, $\tilde{\phi} \neq 0$, $\lambda \geq 0$.*

(1) $x^* = 0$ is an optimal solution for (13) if and only if

$$|\tilde{\phi}^\top \tilde{y}| \leq \lambda \|\tilde{y}\|_2$$

(notice, in particular, that if $\|\tilde{\phi}\|_2 \leq \lambda$, then the above condition is certainly satisfied, hence $x^* = 0$).

(2) If $|\tilde{\phi}^\top \tilde{y}| > \lambda \|\tilde{y}\|_2$ (hence $\|\tilde{\phi}\|_2 > \lambda$), then the optimal solution of (13) is given by

$$x^* = x_{\text{ls}} - \text{sgn}\,(x_{\text{ls}}) \frac{\lambda}{\|\tilde{\phi}\|_2^2} \sqrt{\frac{\|\tilde{\phi}\|_2^2 \|\tilde{y}\|_2^2 - (\tilde{\phi}^\top \tilde{y})^2}{\|\tilde{\phi}\|_2^2 - \lambda^2}}. \tag{14}$$

Proof. The problem is convex but nonsmooth, hence we write the optimality conditions in terms of the subdifferential of the objective: $0 \in \partial f(x) = \partial \|\tilde{\phi} x - \tilde{y}\|_2 + \lambda \partial |x|$, where

$$\partial \|\tilde{\phi} x - \tilde{y}\|_2 = \begin{cases} \dfrac{\tilde{\phi}^\top (\tilde{\phi} x - \tilde{y})}{\|\tilde{\phi} x - \tilde{y}\|_2} & \text{if } \tilde{\phi} x - \tilde{y} \neq 0 \\ \{\tilde{\phi}^\top g : \|g\|_2 \leq 1\} & \text{if } \tilde{\phi} x - \tilde{y} = 0, \end{cases}$$

$$\partial |x| = \begin{cases} \text{sgn}\,(x) & \text{if } x \neq 0 \\ \{v : |v| \leq 1\} & \text{if } x = 0. \end{cases}$$

For point 1 we thus check under what conditions 0 is contained in the subdifferential of f at $x = 0$, that is

$$x^* = 0 \text{ is optimal}$$
$$\Updownarrow$$
$$0 \in \partial f(0) = \left\{ \tfrac{\tilde{\phi}^\top \tilde{y}}{\|\tilde{y}\|_2} + \lambda v, \ |v| \leq 1 \right\}.$$

Since the term λv may take any value in the interval $[-\lambda, \lambda]$, it follows that the above condition is satisfied if and only if $|\tilde{\phi}^\top \tilde{y}| / \|\tilde{y}\|_2 \leq \lambda$, which proves the first part of the theorem. Also, since by the Cauchy-Schwartz inequality it holds that $|\tilde{\phi}^\top \tilde{y}| \leq \|\tilde{\phi}\|_2 \|\tilde{y}\|_2$, it is clear that $\|\tilde{\phi}\|_2 \leq \lambda$ implies $|\tilde{\phi}^\top \tilde{y}| \leq \lambda \|\tilde{y}\|_2$, hence the optimal solution is certainly zero when $\|\tilde{\phi}\|_2 \leq \lambda$.

Consider next the case in point 2, where $|\tilde{\phi}^\top \tilde{y}| > \lambda \|\tilde{y}\|_2$, thus $\|\tilde{\phi}\|_2 > \lambda$ and the solution is nonzero. We initially assume for simplicity that $\tilde{\phi}$ and $\tilde{y}$ are not collinear, so that $\tilde{\phi} x - \tilde{y} \neq 0$ for all x; later we show that the derived solution is still valid if this assumption is lifted. With this assumption, and since $x \neq 0$, we have that

$$x \text{ is optimal} \quad \Leftrightarrow \quad 0 = \partial f(x) = \frac{\tilde{\phi}^\top (\tilde{\phi} x - \tilde{y})}{\|\tilde{\phi} x - \tilde{y}\|_2} + \lambda\, \text{sgn}\,(x),$$

that is, since $\|\tilde{\phi} x - \tilde{y}\|_2 \neq 0$, for

$$\tilde{\phi}^\top (\tilde{\phi} x - \tilde{y}) = -\lambda \|\tilde{\phi} x - \tilde{y}\|_2 \text{sgn}\,(x). \tag{15}$$

All solution to this equation are also solutions of the squared equation

$$(\tilde{\phi}^\top \tilde{\phi} x - \tilde{\phi}^\top \tilde{y})^2 = \lambda^2 \|\tilde{\phi} x - \tilde{y}\|_2^2, \tag{16}$$

which is a quadratic equation in x, equivalent to:

$$\|\tilde{\phi}\|_2^2(\|\tilde{\phi}\|_2^2-\lambda^2)x^2-2\tilde{\phi}^\top\tilde{y}(\|\tilde{\phi}\|_2^2-\lambda^2)x+(\tilde{\phi}^\top\tilde{y})^2-\lambda^2\|\tilde{y}\|_2^2=0.$$

The roots of this equation are in

$$x_\pm = x_{\rm ls} \pm \sqrt{{x_{\rm ls}}^2 - \frac{(\tilde{\phi}^\top\tilde{y})^2-\lambda^2\|\tilde{y}\|_2^2}{\|\tilde{\phi}\|_2^2(\|\tilde{\phi}\|_2^2-\lambda^2)}}.$$

Observe that the term under the square root is nonnegative, since

$$\begin{aligned}\delta \doteq {x_{\rm ls}}^2 - \frac{(\tilde{\phi}^\top\tilde{y})^2-\lambda^2\|\tilde{y}\|_2^2}{\|\tilde{\phi}\|_2^2(\|\tilde{\phi}\|_2^2-\lambda^2)} &= \frac{(\tilde{\phi}^\top\tilde{y})^2}{\|\tilde{\phi}\|^4} - \frac{(\tilde{\phi}^\top\tilde{y})^2-\lambda^2\|\tilde{y}\|_2^2}{\|\tilde{\phi}\|_2^2(\|\tilde{\phi}\|_2^2-\lambda^2)}\\ &= \frac{\lambda^2}{\|\tilde{\phi}\|_2^2}\cdot\frac{\|\tilde{\phi}\|_2^2\|\tilde{y}\|_2^2-(\tilde{\phi}^\top\tilde{y})^2}{\|\tilde{\phi}\|_2^2(\|\tilde{\phi}\|_2^2-\lambda^2)},\end{aligned}$$

where, under the conditions of point 2, $\|\tilde{\phi}\|_2^2-\lambda^2>0$, and $\|\tilde{\phi}\|_2^2\|\tilde{y}\|_2^2-(\tilde{\phi}^\top\tilde{y})^2\geq 0$, by the Cauchy-Schwartz inequality. Further, $\delta\geq 0$ is smaller in magnitude than ${x_{\rm ls}}^2$, since the condition $|\tilde{\phi}^\top\tilde{y}|>\lambda\|\tilde{y}\|_2$ implies that ${x_{\rm ls}}^2-\delta>0$. It follows that the sign of $x_\pm = x_{\rm ls}\pm\sqrt{\delta}$ is the same sign of $x_{\rm ls}$ (since adding $\pm\sqrt{\delta}$ to $x_{\rm ls}$ cannot change its sign). Then, plugging $x \leftarrow x_\pm$ into equation (15), we have the left-hand side

$$\|\tilde{\phi}\|_2^2 x_\pm - \tilde{\phi}^\top\tilde{y} = \|\tilde{\phi}\|_2^2(x_{\rm ls}\pm\sqrt{\delta}) - \tilde{\phi}^\top\tilde{y} = \pm\sqrt{\delta}$$

and the right-hand side

$$-\lambda\|\tilde{\phi}x_\pm-\tilde{y}\|_2 {\rm sgn}\,(x_\pm) = -\lambda\|\tilde{\phi}x_\pm-\tilde{y}\|_2 {\rm sgn}\,(x_{\rm ls}).$$

Thus, sign consistency is obtained by choosing the solution with "+" when $x_{\rm ls}$ is negative, and with "-" when $x_{\rm ls}$ is positive. In conclusion, the unique solution to eq. (15) is given by

$$x^* = x_{\rm ls} - {\rm sgn}\,(x_{\rm ls})\frac{\lambda}{\|\tilde{\phi}\|_2^2}\sqrt{\frac{\|\tilde{\phi}\|_2^2\|\tilde{y}\|_2^2-(\tilde{\phi}^\top\tilde{y})^2}{\|\tilde{\phi}\|_2^2-\lambda^2}},$$

which is the expression we wished to prove.

It only remains to be proved that the above expression is still valid also when $\tilde{y}$ and $\tilde{\phi}$ are collinear. In this case, since $\|\tilde{\phi}\|_2^2\|\tilde{y}\|_2^2=(\tilde{\phi}^\top\tilde{y})^2$, eq. (14) gives $x^*=x_{\rm ls}$, and we have that $\tilde{\phi}x^*-\tilde{y}=0$. Let us check that this solution is indeed optimal. The subdifferential of f at $x^*\neq 0$ such that $\tilde{\phi}x^*-\tilde{y}=0$ is

$$\partial f(x^*) = \{\tilde{\phi}^\top g + \lambda\,{\rm sgn}\,(x^*),\ \|g\|_2\leq 1\},$$

and we see that $0\in\partial f(x^*)$ if $\|\tilde{\phi}\|_2\geq\lambda$, which is indeed the condition under which the expression (14) for x^* holds. □

4.1 Univariate solution of nnrsqrt-LASSO

The solution of the univariate nnrsqrt-LASSO problem in the scalar variable x

$$\min_{x\geq 0} f(x) \doteq \|\tilde{\phi}x-\tilde{y}\|_2+\lambda|x|, \qquad (17)$$

can be readily obtained from the solution of the corresponding unconstrained problem (13), by the following reasoning. Since (17) is a convex optimization problem in one variable and one linear inequality constraint, its optimal solution is either on the boundary of the feasible set (in this case, at $x=0$), or it coincides with the solution of the unconstrained version of the problem. Thus, we solve the unconstrained problem (13): if this solution is nonnegative, then it is also the optimal solution to (17); if it is negative, then the optimal solution to (17) is $x=0$. Since the sign of the solution of (13) is simply the sign of $\tilde{\phi}^\top\tilde{y}$, we can state the following theorem.

Theorem 2 *Let $\tilde{y}\neq 0$, $\tilde{\phi}\neq 0$, $\lambda\geq 0$.*

(1) $x^*=0$ is an optimal solution for (17) if and only if

$$\tilde{\phi}^\top\tilde{y}\leq\lambda\|\tilde{y}\|_2.$$

(2) Otherwise, the optimal solution of (17) is given by

$$x^* = x_{\rm ls} - \frac{\lambda}{\|\tilde{\phi}\|_2^2}\sqrt{\frac{\|\tilde{\phi}\|_2^2\|\tilde{y}\|_2^2-(\tilde{\phi}^\top\tilde{y})^2}{\|\tilde{\phi}\|_2^2-\lambda^2}}. \qquad (18)$$

Remark 3 For the specific structure of $\tilde{\phi}$ and $\tilde{y}$ in (12), we have that

$$\|\tilde{\phi}\|_2^2 = \|\phi\|_2^2+\sigma^2,\quad \tilde{\phi}^\top\tilde{y}=\phi^\top y,\quad \|\tilde{y}\|_2^2=\|y\|_2^2+\|\xi\|_2^2,$$

and the solutions in theorems 1 and 2 can be expressed accordingly in terms of $\phi^\top y$, $\|\phi\|_2$, $\|y\|_2$, $\|\xi\|_2$, and σ, λ. In particular, the condition for $x=0$ being optimal becomes

$$|\phi^\top y|\leq\lambda\sqrt{\|y\|_2^2+\|\xi\|_2^2},$$

which, in particular, is satisfied if $\|\phi\|_2^2+\sigma^2\leq\lambda^2$.

Notice further that $\tilde{\phi}x-\tilde{y}\neq 0$ for $x=0$, since we assumed $\tilde{y}\neq 0$, and that, for $\sigma>0$, $\tilde{\phi}x-\tilde{y}\neq 0$ also for $x\neq 0$, since the i-th entry of ξ is zero by definition. Therefore, for $\sigma>0$, the ℓ_2-norm part of the objective is always nonzero, and hence differentiable. □

5 Sequential coordinate minimization scheme

We next outline a sequential coordinate minimization scheme for the rsqrt-LASSO problem (4). Suppose all variables x_j, $j\in\{1,\dots,n\}\setminus i$, are fixed to some numerical values, and we wish to minimize the objective in (4) with respect to the scalar variable x_i. We have that

$$\begin{aligned}f_i(x_i) &\doteq \|\sum_{j=1}^n\tilde{\phi}_j x_j-\tilde{y}\|_2+\sum_{j=1}^n\lambda_j|x_j|\\ &= \|\tilde{\phi}_i x_i-\tilde{y}(i)\|_2+\lambda_i|x_i|+\sum_{j\neq i}\lambda_j|x_j|,\end{aligned}$$

where we defined $\tilde{y}(i) \doteq \tilde{y} - \sum_{j\neq i} \tilde{\phi}_j x_j$. We thus have that

$$\begin{aligned} x_i^* &\doteq \arg\min_{x_i} \; f_i(x_i) \\ &= \arg\min_{x_i} \; \|\tilde{\phi}_i x_i - \tilde{y}(i)\|_2 + \lambda_i |x_i|, \end{aligned}$$

where the minimizer x_i^* is readily computed by applying Theorem 1.

A sequential coordinate minimization scheme works by updating the variables x_i cyclically, according to the above univariate minimization criterion. The scheme of the algorithm is as follows.

(1) Initialize $x^{(0)} = 0$ (an n-vector of zeros), $k = 1$;

(2) For $i = 1, \ldots, n$, let

$$x_i^{(k)} = \arg\min_{x_i} f(x_1^{(k)}, \ldots, x_{i-1}^{(k)}, x_i, x_{i+1}^{(k-1)}, \ldots, x_n^{(k-1)});$$

(3) If some stopping criterion is met, finish and return $x^{(k)}$, else set $k \leftarrow k + 1$, and goto 2.

The same coordinate minimization scheme can be used also for solving the nnrsqrt-LASSO problem (5), by using the result in Theorem 2 for updating the i-th coordinate. The detailed data management involved in applying this scheme to our specific problem is described in Section 5.1.

Remark 4 Observe that, due to Theorem 1, all variables x_i for which $\|\tilde{\phi}_i\|_2 \leq \lambda_i$ are *never* updated by the algorithm, i.e., they remain fixed at their initial zero value. The inner loop on i can thus be sped up by considering only the indices i such that $\|\tilde{\phi}_i\|_2 > \lambda_i$, which can be determined a priori (feature elimination). □

Remark 5 As a stopping criterion, one may use a standard check on sufficient progress in objective reduction. For improved efficiency, in the numerical implementation of this algorithm, we actually used a variant of the cyclic coordinate update scheme based on a so-called *active set* convergence criterion, see Section 2.6 in (Friedman, Hastie & Tibshirani 2010). □

Convergence of the proposed scheme is established in the following theorem, which is a direct consequence of a result in (Tseng 2001).

Theorem 3 (Convergence) *For $\sigma > 0$, $y \neq 0$, the sequential coordinate minimization algorithm converges to an optimal point, for both the rsqrt-LASSO and the nnrsqrt-LASSO problems.*

Proof. The function $f(x)$ in (2) that we minimize using coordinate minimization is convex and composite, i.e., $f(x) = f_0(x) + \sum_{i=1}^n \psi_i(x_i)$, where ψ_i are convex and nonsmooth. In the unconstrained case, we have $\psi_i(x_i) = \lambda_i|x_i|$. The constrained case, where $x_i \geq 0$, can also be tackled as an unconstrained one, by considering $\psi_i(x_i) = \lambda_i|x_i| + I_+(x_i)$, where $I_+(x_i)$ is equal to zero if $x_i \geq 0$ and it is $+\infty$ otherwise. Further, the function $f_0(x) = \|\tilde{\Phi}x - \tilde{y}\|_2$ is convex and, for $\sigma > 0$ and $y \neq 0$, it is differentiable over all $x \in \mathbb{R}^n$. Since the objective we minimize satisfies the hypotheses of Theorem 5.1 in (Tseng 2001), convergence of the sequential coordinate minimization algorithm to an optimal point is guaranteed for both the rsqrt-LASSO and the nnrsqrt-LASSO problems. □

5.1 Data management and cost per iteration

We next analyze in more detail the data management and the computational cost per iteration of the coordinate minimization scheme.

5.1.1 Variable update

Suppose we have a current value of x and we want to update the i-th coordinate of x. Suppose further that the following quantities are available: $h \doteq \tilde{\Phi}^\top r$, $c \doteq \|r\|_2^2$, where $r \doteq \tilde{\Phi}x - \tilde{y}$ is the current value of the residual vector (as we shall see, we do not need to store r: only h and c need be updated). We set up the univariate minimization problem $\min_z \; \|\tilde{\phi}_i z - \tilde{y}(i)\|_2 + \lambda_i|z|$, where $\tilde{y}(i) = \tilde{y} - \sum_{j\neq i} \tilde{\phi}_j x_j = \tilde{\phi}_i x_i - (\tilde{\Phi}x - \tilde{y}) = \tilde{\phi}_i x_i - r$. Notice that all we need in order to compute the optimal coordinate z^*, by applying Theorem 1 (or Theorem 2, in the nonnegative constrained case) is the following data: $\tilde{\phi}_i^\top \tilde{y}(i) = \|\tilde{\phi}_i\|_2^2 x_i - h_i$, $\|\tilde{y}(i)\|_2^2 = \|\tilde{\phi}_i\|_2^2 x_i^2 + c - 2x_i h_i$. Therefore, we find the optimal z^*, and we update the solution x to $x_+ = x + e_i(z^* - x_i) = x + e_i\delta_i$, where $\delta_i \doteq z^* - x_i$. Also, we update the data necessary for the next iteration. Since $r_+ \doteq \tilde{\Phi}x_+ - \tilde{y} = r + \tilde{\phi}_i\delta_i$, we have that $c_+ \doteq \|r_+\|_2^2 = c + \|\tilde{\phi}_i\|_2^2\delta_i^2 + 2\delta_i h_i$, $h_+ \doteq \tilde{\Phi}^\top r_+ = h + \tilde{\Phi}^\top \tilde{\phi}_i \delta_i$. Then, we let $i \leftarrow i + 1$, $h \leftarrow h_+$, $c \leftarrow c_+$, $x \leftarrow x_+$ and iterate. The whole process is initialized with $x = 0$, $h = -\tilde{\Phi}^\top \tilde{y}$, $c = \|\tilde{y}\|_2^2$.

5.1.2 Storage and computational cost per iteration

Let us define the *kernel* matrix $\tilde{K} \in \mathbb{R}^{n,n}$ and the projected response vector $q \in \mathbb{R}^n$: $\tilde{K} \doteq \tilde{\Phi}^\top\tilde{\Phi} = K + \sigma^2 I_n$, $q \doteq \tilde{\Phi}^\top \tilde{y} = \Phi^\top y$, where $K \doteq \Phi^\top\Phi$. Initialization of the coordinate minimization method requires $h = -q$, and $c = \|y\|_2^2$, as described previously.

For updating the i-th variable, the method does not necessarily need to store or access the whole kernel matrix $\tilde{K}$. Indeed, computing the i-th optimal update just requires access to $\|\tilde{\phi}_i\|_2^2 = \tilde{K}_{ii}$, and $O(1)$ operations. Then, the update of the h vector requires access to the i-th column of $\tilde{K}$, and then n operations for computing h_+. The storage requirement of the method is thus essentially given by keeping in memory $h \in \mathbb{R}^n$ and $x \in \mathbb{R}^n$, so it is $O(n)$, if $\tilde{K}$ is not stored. Evaluating the i-th column of the kernel matrix requires $O(mn)$ operations, unless the values of the kernel can be obtained directly (i.e., without actually performing the inner products $\phi_i^\top \phi_j$), as it is the case, for instance, for polynomial kernels.

6 Identification of an airfoil drag force

In this numerical experiment, we considered the problem of identifying a posynomial model for the drag force (per unit length) of a NACA 4412 airfoil.

This force can be evaluated as a function of the air flow density ρ, the wing chord η, the incidence angle θ and the flow velocity v, that is $F_D = \psi^o(w)$, where $w = [\rho\eta\theta v]^\top$. No analytical expression is available for this function. The values $\psi^o(w)$ can be obtained via simulations based on CFD (computational fluid dynamics), by integration of the Navier-Stokes equations. Each evaluation is numerically very costly, thus it is of interest to obtain a simple model for F_D, to be used, for instance, in a later stage of system evaluation or design. A model in posynomial form is important since this form allows the application of geometric programming algorithms, which in turn allow for efficient optimization of the airfoil characteristics, see, e.g., (Hoburg & Abbeel 2012).

In this example, we identified a posynomial model for the drag force of the airfoil, from data obtained through CFD simulations. A set $\mathfrak{D} = \{y(k) = \psi^o(w(k)), w(k)\}_{k=1}^{50}$ of 50 input-output data points has been collected, for randomly chosen values of ρ, η, θ and v in the intervals shown in Table 1.

PARAM.	Minimum	Maximum	Dimension
ρ	0.039	1.2250	[kg/m^3]
η	0.1	1	[m]
θ	-5	10	[deg]
v	0	40	[m/s]

Table 1
Parameter intervals considered in the CFD simulations.

The exponent sets $Q_j = \{-2, -1, 0, 1, 2\}$, $j = 1, \ldots, 4$, have been assumed, following the approach described in Section 2.1. This choice has been made after a preliminary trial and error process, and induces a regression matrix Φ having $n = 625$ columns.

Since a low number of data was available (as discussed above, data generation through simulation is very time-consuming), a leave-one-out (LOO) cross-validation was carried out.

First, a subset $\hat{\mathfrak{D}} \subset \mathfrak{D}$ was defined, composed of all the pairs $(y(j), w(j)) \in \mathfrak{D}$ for which $w(j)$ lies within 80% from the boundary of the the hyperrectangle defining the minimum and maximum deviation for each parameter (as defined in Table 1). This was done in order to avoid points near the boundary of the w domain, which are too close to the region not explored by the data $\mathfrak{D}$.

Then, we set for simplicity $\lambda = \gamma\mathbf{1}$, $\sigma = \gamma/10$, and we considered 20 values of γ, logarithmically spaced in the interval $[1, 10^5]$. For each pair $(y(j), w(j)) \in \hat{\mathfrak{D}}$ and for each value of γ, a posynomial model was identified from the data set $\mathfrak{D} \setminus (y(j), w(j))$, applying the safe feature elimination and solving the optimization problem (5). The identified model is then tested on the single datum $(y(j), w(j))$, and the relative error $\mathrm{ERR}(\gamma, j) = |y(j) - \hat{y}(\gamma, j)| / \|y\|_\infty$ was evaluated, where $\hat{y}(\gamma, j)$ is the output predicted by the model, and $\|y\|_\infty$ is the ℓ_∞ norm of the vector with entries $y(k)$, $k = 1, \ldots, 50$. The cardinality $\mathrm{CARD}(\gamma, j)$ of the solution x of the optimization problem (5) was also recorded, providing a measure of the model complexity (we recall that the cardinality of a vector is the number of its nonzero entries).

Figure 1 shows the obtained relative errors and cardinalities as a function of γ. For each j indexing the set $\hat{\mathfrak{D}}$, we have a light line in the first plot, representing $\mathrm{ERR}(\gamma, j)$, and a corresponding line in the second plot representing $\mathrm{CARD}(\gamma, j)$. The bold lines correspond to the following upper bounds: $\overline{\mathrm{ERR}}(\gamma) \doteq \max_j \mathrm{ERR}(\gamma, j)$, $\overline{\mathrm{CARD}}(\gamma) \doteq \max_j \mathrm{CARD}(\gamma, j)$. These two bounds are also represented in Figure 2, where $\overline{\mathrm{ERR}}(\gamma)$ is shown as a function of $\overline{\mathrm{CARD}}(\gamma)$.

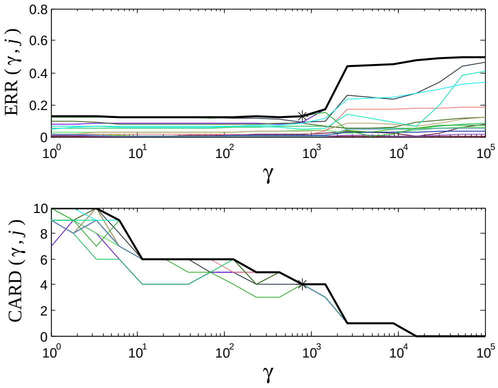

Figure 1. LOO procedure. Bold line: upper bounds $\overline{\mathrm{ERR}}(\gamma)$ and $\overline{\mathrm{CARD}}(\gamma)$. Light lines: relative errors $\mathrm{ERR}(\gamma, j)$ and cardinalities $\mathrm{CARD}(\gamma, j)$. *: best trade-off ($\gamma = 785$).

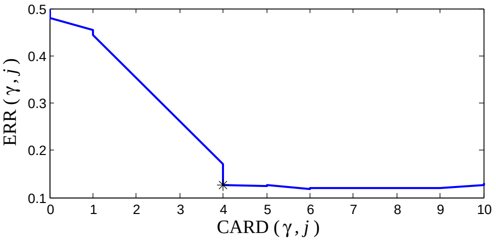

Figure 2. LOO procedure. $\overline{\mathrm{ERR}}(\gamma)$ vs $\overline{\mathrm{CARD}}(\gamma)$. *: best trade-off ($\gamma = 785$).

Based on these curves, the value $\gamma = 785$ has been chosen, since providing the best trade-off between the model complexity (measured by the cardinality of x) and its accuracy (measured by the relative error), see the marker * in Figures 1 and 2. Indeed, values of γ larger than 785 lead to models with a high relative error, whereas values smaller than 785 yield models with a higer cardinality but essentially with the same relative error. Note that, for $\gamma = 785$, the maximum relative error is $\overline{\mathrm{ERR}}(785) = 0.12$. This result is quite surprising: the model is able to approximate the unknown function with a good accuracy, even if a very small number of points (i.e., 49) are used to explore its 4-dimensional domain.

Assuming $\lambda = \gamma\mathbf{1}$, $\sigma = \gamma/10$ and $\gamma = 785$, a model was identified from the whole data set $\mathfrak{D}$, applying the safe feature elimination and solving the optimization problem (5). This model is given by

$$\psi(w) = x_{340}\eta v^2 + x_{440}\rho v^2 + x_{465}\rho\eta v^2 + x_{565}\rho^2 v^2$$

where $x_{340} = 1.2746 \times 10^{-4}$, $x_{440} = 3.5469 \times 10^{-3}$, $x_{465} = 2.8703 \times 10^{-4}$, and $x_{565} = 5.0722 \times 10^{-4}$ (the units of these coefficient can be inferred from Table 1). To this model there correspond a LOO error estimate given by $\overline{\mathrm{ERR}}(785) = 0.12 = 12\%$.

It is interesting to note that this model selected only

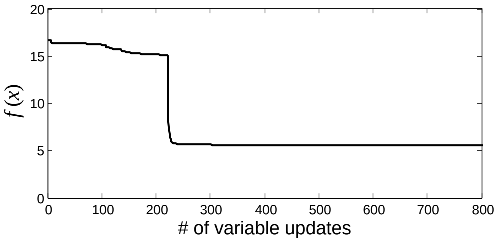

Figure 3. Plot of objective value vs. variable updates for the coordinate minimization method.

four terms out of 625. Also, the dependence of the drag force on the square velocity has been found by the algorithm and this result is consistent with the well-known drag equation. No significant dependence on the incidence angle θ has been observed. A possible interpretation for this latter result is that the range considered for θ is not sufficiently large compared to the ranges considered for ρ, η and v (see Table 1) and, consequently, the force variations due to θ are negligible with respect to those produced by the other three parameters.

The safe feature elimination discussed in Section 3.3, reduced the number of columns of Φ from 625 to 222 (this latter is the average value obtained in the LOO validation, for $\gamma = 785$), suggesting that this elimination phase can be quite useful in practical problems.

6.1 Comparison with other methods

We next briefly discuss the numerical performance of the proposed coordinate minimization (CM) method, compared with two popular algorithms for regularized regression, namely the backward-forward splitting (BFS) method described in (Duchi & Singer 2009), and the FISTA, (Beck & Teboulle 2009). In all experiments, we set an exit condition when the relative improvement of the objective value from one iteration to the next is below 10^{-7}.

The time taken for applying the safe elimination and solving the optimization problem (5) with the CM approach described in Sections 3-5 (with active set convergence) was about 0.015 seconds on a PC with a Core i7 processor and a RAM memory of 8GB (average time obtained in the LOO validations, with $\gamma = 785$). A plot of objective values vs. variable updates is shown in Figure 3. The first 220 updates in Figure 3 correspond to a full sweep over the problem variables; the subsequent abrupt decrease in objective value is due to iterations on the active set only (variables that are nonzero after the first full sweep), and this was basically enough to achieve convergence, in this example.

A plain implementation of the BFS method of (Duchi & Singer 2009) resulted in extremely slow convergence (several minutes); this method seems unsuitable for the problem at hand. Considerably better results were obtained using the FISTA: this method converged in about 8.9 seconds; a plot of objective values vs. iterations is shown in Figure 4. Anyways, the FISTA performance on our test problem resulted to be worse than the one we obtain via the CM method by almost three orders of magnitude.

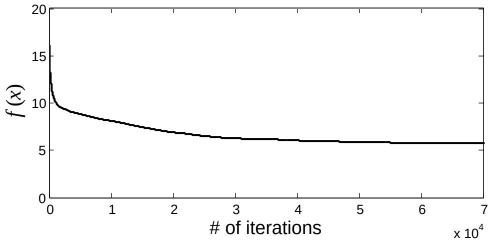

Figure 4. Plot of objective value vs. iterations for the FISTA.

Acknowledgments: We thank Valentina Dolci (Politecnico di Torino) for providing us with the fluid dynamic simulation data used in the example.

7 Conclusions

An approach for the identification of posynomial models has been presented in this paper, based on the solution of a nonnegative regularized square-root LASSO problem. In this approach, a large-scale expansion of monomials is considered, and the model is identified by seeking coefficients of the expansion that minimize an objective composed by a fitting error term and a sparsity promoting term. A sequential coordinate minimization scheme has been developed to solve the nnrsqrt-LASSO problem. This scheme guarantees convergence to a minimum of the objective function and it is suitable for large-scale implementations. An applicative example on identification of a posynomial model for a NACA 4412 airfoil demonstrates the potential effectiveness of the proposed approach.

References

Babakhani, A., Lavaei, J., Doyle, J. & Hajimiri, A. (2010), Finding globally optimum solutions in antenna optimization problems, *in* 'IEEE International Symposium on Antennas and Propagation'.

Babu, P. & Stoica, P. (2014), 'Connection between SPICE and square-root LASSO for sparse parameter estimation', *Signal Processing* **95**, 10–14.

Beck, A. & Teboulle, M. (2009), 'A fast iterative shrinkage-thresholding algorithm for linear inverse problems', *SIAM J. Imaging Sciences* **2**(1), 183–202.

Beightler, C. & Phillips, D. (1976), *Applied geometric programming*, Wiley, New York.

Belloni, A., Chernozhukov, V. & Wang, L. (2011), 'Square-root LASSO: pivotal recovery of sparse signals via conic programming', *Biometrika* **98**(4), 791–806.

Bonin, M., Seghezza, V. & Piroddi, L. (2010), 'NARX model selection based on simulation error minimisation and LASSO', *IET Control Theory and Applications* **4**(7), 1157–1168.

Boyd, S., Kim, S., Patil, D. & Horowitz, M. (2005), 'Digital circuit optimization via geometric programming', *Operation Research* **53**(6), 899–932.

Candes, E. & Tao, T. (2006), 'Near-optimal signal recovery from random projections: Universal encoding strategies?', *IEEE Transactions on Information Theory* **52**(12), 5406 –5425.

Carvajal, R., Godoy, B., Aguero, J. & Goodwin, G. (2012), Em-based sparse channel estimation in ofdm systems, *in* '13th IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC 2012)'.

Chiang, M. (2005), 'Geometric programming for communication systems', *Commun. Inf. Theory* **2**, 1–154.

Combettes, P. & J.-C-Pesquet (2007), 'Proximal thresholding algorithm for minimization over orthonormal bases', *SIAM Journal on Optimization* **18**(4), 1351–1376.

Daems, W., Gielen, G. & Sansen, W. (2003), 'Simulation-based generation of posynomial performance models for the sizing of analog integrated circuits', *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* **22**(5), 517–534.

Daubechies, I., Defrise, M. & Mol, C. D. (2004), 'An iterative thresholding algorithm for linear inverse problems with a sparsity constraint', *Communications on Pure and Applied Mathematics* **57**(11), 1413–1457.

De Mol, C., De Vito, E. & Rosasco, L. (2009), 'Elastic-net regularization in learning theory', *Journal of Complexity* **25**, 201–230.

Donoho, D., Elad, M. & Temlyakov, V. (2006), 'Stable recovery of sparse overcomplete representations in the presence of noise', *IEEE Transactions on Information Theory* **52**(1), 6 – 18.

Duchi, J. & Singer, Y. (2009), 'Efficient online and batch learning using forward backward splitting', *J. of Machine Learning Research* **10**, 2899–2934.

Duffin, R., Peterson, E. & Zener, C. (1967), *Geometric programming: theory and application*, Wiley, New York.

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004), 'Least angle regression', *The Annals of Statistics* **32**(2), 407–451.

El Ghaoui, L., Viallon, V. & Rabbani, T. (2012), 'Safe feature elimination for the LASSO and sparse supervised learning problems', *Pacific Journal of Optimization* **8**(4), 667–698.

Friedman, J., Hastie, T. & Tibshirani, R. (2010), 'Regularization paths for generalized linear models via coordinate descent', *J. of Statistical Software* **33**(1).

Hale, E., Yin, W. & Zhang, Y. (2008), 'Fixed-point continuation for l1-minimization: Methodology and convergence', *SIAM J. Optim.* **19**(3), 1107–1130.

Hoburg, W. & Abbeel, P. (2012), Geometric programming for aircraft design optimization, *in* '8th AIAA MDO Specialist Conference', Honolulu, HI, USA.

Komiya, H. (1988), 'Elementary proof of Sion's minimax theorem', *Kodai Math. J.* **11**, 5–7.

Kukreja, S., Lofberg, J. & Brenner, M. (2006), A least absolute shrinkage and selection operator (LASSO) for nonlinear system identification, *in* '14th IFAC Symp. on System Identification', Newcastle, Australia, pp. 814–819.

Novara, C. (2012), 'Sparse identification of nonlinear functions and parametric set membership optimality analysis', *IEEE Transactions on Automatic Control* **57**(12), 3236–3241.

Novara, C., Vincent, T., Hsu, K., Milanese, M. & Poolla, K. (2011), 'Parametric identification of structured nonlinear systems', *Automatica* **47**(4), 711 – 721.

Pulecchi, T. & Piroddi, L. (2007), A cluster selection approach to polynomial NARX identification, *in* 'American Control Conference', New York City, USA, pp. 852–857.

Sapatnekar, S., Rao, V., Vaidya, P. & Kang, S. (1993), 'An exact solution to the transistor sizing problem for CMOS circuits using convex optimization', *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* **12**(11), 1621–1634.

Sion, M. (1958), 'On general minimax theorems', *Pacific J. Math.* **8**, 171–176.

Spinelli, W., Piroddi, L. & Lovera, M. (2006), A two-stage algorithm for structure identfication of polynomial NARX models, *in* 'American Control Conference', pp. 2387–2392.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the Lasso', *Royal. Statist. Soc B.* **58**(1), 267–288.

Tseng, P. (2001), 'Convergence of a block coordinate descent method for nondifferentiable minimization', *J. of Optimization Theory and Applications* **109**(3), 475–494.

Wilde, D. (1978), *Globally optimal design*, Wiley interscience publication.

Zou, H. (2006), 'The adaptive lasso and its oracle properties', *JASA* **101**, 1418–1429.

Zou, H. & Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *J. R. Statist. Soc. B* **67**(2), 301–320.