Stepwise Tikhonov Regularisation: Application to the Prediction of HIV-1 Drug Resistance

Ramón A. Delgado, Zhiyong Chen, *Senior Member, IEEE,* and Richard H. Middleton, *Fellow, IEEE*

***Abstract*—This paper focuses on constructing genotypic predictors for antiretroviral drug susceptibility of HIV. To this end, a method to recover the largest elements of an unknown vector in a least squares problem is developed. The proposed method introduces two novel ideas. The first idea is a novel forward stepwise selection procedure based on the magnitude of the estimates of the candidate variables. To implement this newly introduced procedure, we revise Tikhonov regularisation from a sparse representations' perspective. This analysis leads us to the second novel idea in the paper, which is the development of a new method to recover the largest elements of the unknown vector in the least squares problem. The method implements a sequence of Tikhonov regularisation problems which aim to recover the largest of the remaining elements of the unknown vector. Additionally, we derive sufficient conditions that ensure the recovery of the largest elements of the unknown vector. We perform numerical studies using simulated data and data from the Stanford HIV resistance database. The performance of the proposed method is compared against a state-of-the-art method.**

***Index Terms*—sparse representations, least squares, exact recovery.**

I. Introduction

THE construction of genotypic predictors of HIV-1 drug resistance has received increasing attention recently, see e.g. [1], [2]. The development of these predictors is important because even when there are several antiretroviral drugs available to treat HIV-1. Once a treatment has started, the efficacy of a drug diminishes due to the development of drug-resistant variants of HIV [3]. Sadly, this drug-resistant HIV variant may infect other people, including some with limited treatment options. Understanding how certain mutations affect susceptibility to a particular drug provides a valuable tool to decide which treatment is best for a given person. The availability of such predictors may enable the analysis and development of more advanced methods as proposed in [4], [5]. Reference [2] compared several statistical methods to generate predictors of HIV-1 drug susceptibility. These methods were decision trees, neural networks, least squares regression, support vector regression and Least angle regression (LARS) [6]. In [2] it was shown that LARS is a promising tool to select the regressors in a predictor. This result motivates the work in [1] which performs an in-depth analysis of a predictor obtained with LARS.

LARS is an important tool used for exact sparse recovery. Exact sparse recovery aims to recover the indices of the non-zero elements of a sparse signal from the measurements of a linear system of equations. In some cases, this system of equations might have more unknowns than data measurements, making the system of equations under-determined. Despite this difficulty, there are many results in the literature that provide conditions that ensure the recovery of sparse signals, see [7], [8]. These results bring a new perspective to some well-studied problems, such as model selection, see [9], [10], [11]. Additionally, several results that guarantee exact sparse recovery are also applicable to non-sparse vectors, see e.g. [12].

There are two main categories of sparse recovery methods, namely, regularised methods and greedy approaches [13]. Regularised methods include the Lasso and other ℓ_1 regularised methods [14], [15]. Regularisation approaches offer great flexibility and they have been used in several applications, see e.g. [11], [10]. On the other hand, greedy algorithms have little computational cost and they are easy to implement [16]. A general greedy technique is forward stepwise selection, see e.g. [17]. This approach starts with no variables in the model. At each step, each variable that is not already in the model is evaluated for inclusion. At the end of the step, the variable that maximises a given criterion is added to the model. The steps are repeated until the resulting model is satisfactory. For linear models, algorithms based on forward stepwise selection include Orthogonal Matching Pursuit (OMP) [18], Orthogonal Least Squares (OLS) [19], and LARS.

Several tools have been used to describe necessary and/or sufficient conditions for the recovery of sparse and compressible signals. Restricted Isometry Property and the mutual coherence have been used for such purpose, see e.g. [20], [21]. For example, the paper [12] uses mutual coherence to derive sufficient conditions for the recovery of compressible signals when OMP and OLS are used.

The current paper proposes a forward stepwise procedure for the recovery of the largest elements of an unknown vector in the least squares problem. Additionally, an algorithm that implements the proposed forward stepwise procedure is developed. The algorithm implements a sequence of Tikhonov regularisation problems, and therefore name the method Stepwise Tikhonov Regularisation (STIR). We derive sufficient conditions under which STIR recovers the largest elements of an unknown vector in a least squares problem. The sufficient conditions provided here are analogous to the ones provided in [12]. However, the analysis in [12] does not apply to the STIR method, and the conditions in the current paper are not

R. A. Delgado, R. H. Middleton and Z. Chen are with the School of Electrical Engineering and Computing, The University of Newcastle, Australia, e-mail: {Ramon.Delgado, Zhiyong.Chen, Richard.Middleton}@newcastle.edu.au.

Manuscript received April 19, 2005; revised August 26, 2015.

based on the mutual coherence, but on the magnitude of the elements of an auxiliary matrix Z.

In this paper, data from the Genotype-Phenotype datasets [22], which are publicly available at the Stanford HIV resistance database [23], are used in conjunction with the proposed STIR method to construct genotypic predictors of HIV-1 drug resistance.

The remainder of the paper is organised as follows: In section II we formulate the problem. Section III introduces a novel forward selection procedure. Section IV revisits Tikhonov regularisation from the perspective of sparse representations and describes the proposed STIR method. Section V describes sufficient conditions under which STIR guarantees the recovery of the largest elements of an unknown vector. In section VI we use the proposed STIR method to construct genotypic predictors of HIV-1 drug resistance, and compare their performance against LARS. Finally, conclusions are drawn in section VII.

Notation and basic definitions: $A \succ 0$ denotes that a symmetric matrix A is positive definite, while $A \succeq 0$ denotes that A is positive semi-definite. We represent the transpose of a given matrix A as $A^\top$. The vector e_i denotes the i-th column of the identity matrix. $A_{i,j}$ denotes the element in i-th row and the j-th column of a matrix A. A_{i-row} and A_{j-col} denote the i-th row and the j-th column of A, respectively. We denote $\sigma_{min}(A)$ and $\sigma_{max}(A)$ to the smallest and largest singular values of a matrix A. $\|x\|$ denotes the Euclidean norm of a vector x. $\|x\|_0$ denotes the ℓ_0-pseudo norm which counts the number of nonzero elements in a vector x.

II. Problem formulation

In this section, we formulate the problem. Let $u \in \mathbb{R}^n$ be an unknown vector to be recovered and $\eta \in \mathbb{R}^m$ be measurement noise. We are given a matrix of regressors $A \in \mathbb{R}^{m\times n}$ and the measured signal $y \in \mathbb{R}^m$ that has been generated as follows

$$y = Au + \eta \tag{1}$$

The goal is to recover the locations of the p-largest elements (in magnitude) of u. In the sparse representations literature, most of the research has focused on the conditions that matrix A must satisfy to ensure the recovery of a sparse vector. For example, the restricted isometry property has been used to ensure recovery of a sparse vector [20]. One problem with the restricted isometry property is that it is difficult to check. In fact, it has been proved that checking the restricted isometry property is NP-hard [24]. To overcome this issue, the concept of *mutual coherence* was introduced in [25], and it has been used to guarantee exact sparse recovery of several algorithms. For example, in [12] the concept of mutual coherence has been used to derive sufficient exact recovery conditions for the OMP and OLS algorithms. An important contribution in [12], is that it takes into account information about the relative decay in the magnitude of the elements of u. This consideration allows the results in [12] to be applicable not only to sparse signals, but also to compressible signals. It is worth mentioning that sufficient conditions for exact sparse recovery the LARS algorithm are described in [26].

TABLE I
Proposed forward stepwise selection procedure.

```
1) Input: A modelling procedure $x = getmodel(\mathcal{S})$ such that $x_i = 0$ if $i \notin \mathcal{S}$.
2) Initialise:
     • Define the set of variables in the model, $\mathcal{S}^{(0)}$, as an empty set.
     • Set $k = 0$.
3) While model obtained with $\mathcal{S}^{(k)}$ is unsatisfactory, do
     a) For all $i \notin \mathcal{S}^{(k)}$ compute $v^{(k+1,i)} = getmodel(\{\mathcal{S}^{(k)} \cup \{i\}\})$
     b) Set $\ell = \arg\max_{i \notin \mathcal{S}^{(k)}} \left|v_i^{(k+1,i)}\right|$
     c) Set $\mathcal{S}^{(k+1)} = \mathcal{S}^{(k)} \cup \{\ell\}$
     d) Set $k = k + 1$
4) Return $x = getmodel(\mathcal{S}^{(k)})$.
```

Exact sparse recovery is a particular case of the class of problems that deal with the ℓ_0 pseudo-norm, $\|x\|_0$. This pseudo-norm counts the number of nonzero elements in a vector x. In general, finding a solution to an optimisation problem formulated using the $\|x\|_0$ norm leads to an NP-hard problem. Despite this difficulty, the ℓ_0-norm has been extensively used to induce sparsity in the solutions of optimisation problems, see for example [27], [28], [29], [30], [31].

III. A forward stepwise selection procedure

In this section we propose a novel forward stepwise selection procedure.

Forward stepwise selection is an iterative method to construct a model in several steps. A forward stepwise algorithm starts with no variables in the model. At each step, each variable that is not already in the model is evaluated for inclusion in the model. At the end of the step, the variable that maximises a given criterion is added to the model. The steps are repeated until the resulting model is satisfactory.

In exact sparse recovery, examples of forward stepwise algorithms are OMP and OLS. The main difference between these algorithms is the criterion to be maximised at each step by the candidate variables. In OLS the variable that results in the maximum least squares reduction is included in the model. In OMP the variable to be included in the model is the one corresponding to the regressor that is most correlated with the residual of the current model.

In the current paper, we use a different criterion. We propose to include the variable that will have the largest absolute value in the next iteration. To explain this criterion in more detail, we first denote by $\mathcal{S}^{(k)}$ the set of indices that are already included in the model at the beginning of iteration k. We also denote the current estimate (at iteration k) as $\hat{x}^{(k)}$. Thus, the proposed criterion can be described as follows. For each candidate variable $i \notin \mathcal{S}^{(k)}$ define the temporary set $\bar{\mathcal{S}}^{(k+1,i)} = \{\mathcal{S}^{(k)} \cup \{i\}\}$ and its corresponding (candidate) estimate $v^{(k+1,i)}$. The criterion requires the computation of the i-th element of $v^{(k+1,i)}$, denoted as $v_i^{(k+1,i)}$. The element, ℓ, to be included in the model is the one that has the maximum absolute value of $v_i^{(k+1,i)}$, i.e. ℓ is given by

$$\ell = \arg\max_{i \notin \mathcal{S}^{(k)}} \left|v_i^{(k+1,i)}\right| \tag{2}$$

Next, we have that $\hat{x}^{(k+1)} = v^{(k+1,\ell)}$. A simple implementation of the proposed forward stepwise selection procedure is described in Table I. Note that the implementation in Table I computes at the k-th iteration, the vector $v^{(k+1,i)}$ for all $i \notin \mathcal{S}^{(k)}$. Such an approach, in general, may have a high computational cost. However, for the linear model presented in (1) the computational cost can be considerably reduced. In the following section, we describe a method based on Tikhonov regularisation that computes i-th element $v_i^{(k+1,i)}$ for each for $i \notin \mathcal{S}^{(k)}$ without the need to compute the whole vector $v^{(k+1,i)}$.

IV. Stepwise Tikhonov Regularisation

This section revisits Tikhonov regularisation from the perspective of sparse representations and describes the proposed STIR method.

A classic approach to induce a sparsity pattern in the solution of an optimisation problem is by using Tikhonov regularisation [32]. Consider the following least squares problem with a Tikhonov regularisation

$$\mathcal{P}_{tikho}: \quad \underset{x\in\mathbb{R}^n}{\text{minimise}} \;\; f(x) + \rho x^\top W x$$

where $f(x) := (y - Ax)^\top (y - Ax)$ is the function to be minimised, $\rho > 0$ is a regularisation constant and $W \succeq 0$ is a regularisation matrix. Tikhonov regularisation induces zeros in specific elements of the solution vector $\hat{x}$. To do so, ρ must be large enough and W should be defined as a diagonal matrix with diagonal elements having value $W_{i,i} = 1$ to induce a zero in x_i and $W_{j,j} = 0$ to allow x_j to take any value.

The following Lemma describes some properties of the optimal solution of $\mathcal{P}_{tikho}$.

Lemma 1: Given $A \in \mathbb{R}^{m\times n}$, $y \in \mathbb{R}^m$, $\rho > 0$ and $W \succeq 0$. Suppose that $(A^\top A + \rho W) \succ 0$. Then, the optimal solution of $\mathcal{P}_{tikho}$ is given by

$$\hat{x} = (A^\top A + \rho W)^{-1} A^\top y \tag{3}$$

with optimal cost

$$f(\hat{x}) + \hat{x}^\top \rho W \hat{x} = y^\top y - y^\top A (A^\top A + \rho W)^{-1} A^\top y \tag{4}$$

□

Lemma 1 does not make any assumption about the measured vector y. The following Lemma describes some properties of the optimal solution of $\mathcal{P}_{tikho}$, for the case when the measured vector y is generated as in (1).

Lemma 2: Given $A \in \mathbb{R}^{m\times n}$, $\rho > 0$ and $W \succeq 0$. Suppose that $(A^\top A + \rho W) \succ 0$ and that $y = Au + \eta$, where $u \in \mathbb{R}^n$ is an unknown vector and $\eta \in \mathbb{R}^m$ is additive noise. Then, the optimal solution of $\mathcal{P}_{tikho}$ satisfies that

$$\hat{x} = u - (\frac{1}{\rho} A^\top A + W)^{-1} W u + (A^\top A + \rho W)^{-1} A^\top \eta \tag{5}$$

□

The above lemmas are well known results. In the following section, the Lemmas 1 and 2 are used to develop a novel forward stepwise selection algorithm.

A. Updating to the current solution

This section describes how to update the optimal solution of $\mathcal{P}_{tikho}$ when the matrix W is updated in a particular way.

Let $W^{(k)}$ be a diagonal matrix describing the current sparsity pattern as used in Tikhonov Regularisation. The diagonal elements of $W^{(k)}$ define the sparsity pattern in the sense that $W^{(k)}_{i,i} = 1$ if $i \notin \mathcal{S}^{(k)}$, and $W^{(k)}_{j,j} = 0$ if $j \in \mathcal{S}^{(k)}$. Assume that in the next iteration we want to define the sparsity pattern $W^{(k+1)}$, such that $\mathcal{S}^{(k+1)} = \mathcal{S}^{(k)} \cup \{\ell\}$, i.e. we want to include ℓ in the model. Then, we have that

$$W^{(k+1)} = W^{(k)} - e_\ell e_\ell^\top. \tag{6}$$

The optimal solution of $\mathcal{P}_{tikho}$ for the sparsity pattern $W^{(k)}$ and a given $\rho > 0$ is described by

$$x^{(k)} = (A^\top A + \rho W^{(k)})^{-1} A^\top y \tag{7}$$

and the optimal cost for $\mathcal{P}_{tikho}$ is given by

$$g(x^{(k)}) := f(x^{(k)}) + \rho\, (x^{(k)})^\top W^{(k)} x^{(k)} = y^\top (y - Ax^{(k)}) \tag{8}$$

The following Theorem describes some properties when we update from $x^{(k)}$, which is the optimal solution of $\mathcal{P}_{tikho}$ with regularization matrix $W^{(k)}$, to $x^{(k+1)}$ that is the optimal solution of $\mathcal{P}_{tikho}$ with regularization matrix $W^{(k+1)}$.

Theorem 1: Consider $W^{(k+1)}$ as in (6) and suppose that the assumptions in Lemma 1 hold for $W^{(k)}$ and $W^{(k+1)}$. Define

$$Z^{(k)} = \left(\frac{1}{\rho} A^\top A + W^{(k)}\right)^{-1} \tag{9}$$

then, we have that

$$x^{(k+1)} = x^{(k)} + \frac{x_\ell^{(k)}}{1 - Z^{(k)}_{\ell,\ell}} Z^{(k)}_{\ell-col} \tag{10}$$

$$Z^{(k+1)} = Z^{(k)} + \frac{1}{1 - Z^{(k)}_{\ell,\ell}} Z^{(k)}_{\ell-col} (Z^{(k)}_{\ell-col})^\top \tag{11}$$

and we also have that

$$g(x^{(k)}) - g(x^{(k+1)}) = \rho \frac{(x_\ell^{(k)})^2}{1 - Z^{(k)}_{\ell,\ell}} \tag{12}$$

□

Proof: The proof is given in the Appendix. ■

Theorem 1 provides a convenient way to update the current solution, $x^{(k)}$. Additionally, from (10) it is easy to see that

$$x_\ell^{(k+1)} = \frac{x_\ell^{(k)}}{1 - Z^{(k)}_{\ell,\ell}}. \tag{13}$$

Now, we are ready to implement the proposed forward stepwise strategy. The proposed algorithm is described in Table II.

For the proposed algorithm, the choice of ρ and W are mainly limited by the conditions of Theorem 1 that requires matrix $Z^{(k)}$ to be positive definite. In particular, if the value of ρ is too small, then it may produce that at machine precision matrix $Z^{(k)}$ is not positive definite. In our numerical experiments, we observed that values of ρ in the range $\sigma^2_{min}(A) \leq \rho \leq \sigma^2_{max}(A)$ work well in practice.

TABLE II
STEPWISE TIKHONOV REGULARISATION (STIR) ALGORITHM.

```
1) Input: $A \in \mathbb{R}^{m\times n}$, $\rho$, $y \in \mathbb{R}^m$ and $p$.
2) Initialise:
     • Set $Z^{(0)} = (\frac{1}{\rho}A^\top A + I)^{-1}$
     • Set $x^{(0)} = \frac{1}{\rho}Z^{(0)}A^\top y$
     • Set $\mathcal{S}^{(0)}$ as an empty set.
3) For $k = 0 \dots p-1$ compute:
     a) $\ell = \arg\max_{i \notin \mathcal{S}^{(k)}} \left|\frac{x_i^{(k)}}{1-Z_{i,i}^{(k)}}\right|$
     b) $\mathcal{S}^{(k+1)} = \mathcal{S}^{(k)} \cup \{\ell\}$
     c) $x^{(k+1)} = x^{(k)} + \frac{x_\ell^{(k)}}{1-Z_{\ell,\ell}^{(k)}} Z_{\ell-col}^{(k)}$
     d) $Z^{(k+1)} = Z^{(k)} + \frac{1}{1-Z_{\ell,\ell}^{(k)}} Z_{\ell-col}^{(k)} (Z_{\ell-col}^{(k)})^\top$
4) Return $x^\star = x^{(p)}$.
```

To analyse the computational complexity of the proposed method, we assume that none of the parameters m, n and p dominate over the others. The proposed update algorithm described in Table II has a computational complexity $\mathcal{O}(n^3 + n^2m + n^2p)$ where the term $\mathcal{O}(n^2m)$ represents the construction of matrix $A^\top A$, the term $\mathcal{O}(n^3)$ describes the matrix inversion required to compute $Z^{(0)}$, and $\mathcal{O}(n^2)$ is the complexity required to compute $Z^{(k+1)}$ for $k = 0,\dots,p-1$. On the other hand, a simple implementation of the proposed forward selection procedure requires solving a least squares problem for each candidate variable. Additionally, the size of each of these least squares problems grows with the number of elements in the model. Suppose that $p = t\cdot n$ with $0 < t < 1$, then a simple implementation has a computational complexity[1] $\mathcal{O}\left(\frac{n^5t^4}{4}(1-t\frac{4}{5}) + \frac{mn^3t^2}{2}(1-t\frac{2}{3})\right)$ which, for the case when n is large, is significantly higher than the complexity of the proposed update method, i.e. $\mathcal{O}(n^3 + n^2m + n^3t)$.

V. Exact Recovery Conditions

We are interested in deriving sufficient conditions for the recovery of the indices of the p-largest elements of an unknown vector u. In this section we use two strategies to ensure exact recovery for the proposed STIR algorithm. The first strategy is to look for conditions that guarantee that at each iteration, the algorithm always select the largest element of u that has not been included in the model. The second strategy is to look for conditions that ensure that at each iteration, only the elements that are in the set of the p-largest elements of u are included. These strategies to ensure exact recovery has been extensively used in the literature, see e.g. [12]. The forward selection procedure represented by (2) follows the spirit behind these strategies. Surprisingly, to the best of the authors' knowledge, the forward selection procedure represented by (2) has not been studied in the sparse representations literature.

The following definitions are needed for the analysis. To capture the properties of the matrix A with respect to a given sparsity pattern $W^{(k)}$ the following constant is defined.

$$\tau^{(k)} = \max_{j\notin\mathcal{S}^{(k)}} \left\| \frac{(W^{(k)} - e_j e_j^\top) Z_{j-col}^{(k)}}{1 - Z_{j,j}^{(k)}} \right\|_\infty \tag{14}$$

Additionally, to capture the properties of the noise the following definition is used.

$$\epsilon^{(k)} \geq \max_j \frac{1}{\rho\left|1 - Z_{j,j}^{(k)}\right|} \left|(Z_{j-col}^{(k)})^\top A^\top \eta\right| \tag{15}$$

Next, let $u \in \mathbb{R}^n$ such that $\|u\|_0 = \gamma$ be the vector that we want to recover, and assume that we are given a measurement $y \in \mathbb{R}^m$ that has been generated using (1). The matrix W represents the sparsity pattern that has been correctly recovered so far. Assume that $|u_\ell| = \|Wu\|_\infty$, i.e. (ℓ) is the index of the element with the largest magnitude that has not been recovered yet. The following Theorem provides sufficient conditions to ensure that the STIR algorithm always recovers the element with the largest magnitude.

Theorem 2: Let $A \in \mathbb{R}^{m\times n}$, $\rho > 0$ and $y \in \mathbb{R}^m$ generated as in Lemma 2. Suppose the unknown vector $u \in \mathbb{R}^n$ satisfies $|u_1| > |u_2| > \cdots > |u_\gamma|$ and $\|u\|_0 = \gamma \leq n$. Suppose that we are given the subset $\mathcal{S}^{(k)}$ containing the indices of the $(\ell-1)$-largest elements of u (in magnitude), with $\ell \leq \gamma$. Moreover, $\mathcal{S}^{(k)}$ has a corresponding regularisation matrix $W^{(k)}$ satisfying the conditions in Theorem 1. From $W^{(k)}$ define $Z^{(k)}$, $\tau^{(k)}$ and $\epsilon^{(k)}$ as in (9), (14) and (15), respectively.

If $Z_{i,i}^{(k)} \neq 1$ for $i \notin \mathcal{S}^{(k)}$, $\tau^{(k)} < 1$ and

$$|u_\ell| > |u_{\ell+1}| + 2\frac{\tau^{(k)}}{1-\tau^{(k)}} \sum_{j=\ell+1}^{\gamma} |u_j| + \frac{2\epsilon^{(k)}}{1-\tau^{(k)}} \tag{16}$$

Then, we have that for $\ell \notin \mathcal{S}^{(k)}$, the optimal solution of $\mathcal{P}_{tikho}$, $\hat{x}^{(k)}$, satisfies that

$$\left|\frac{\hat{x}_\ell^{(k)}}{1 - Z_{\ell,\ell}^{(k)}}\right| > \max_{\substack{j \notin \mathcal{S}^{(k)} \\ j \neq \ell}} \left|\frac{\hat{x}_j^{(k)}}{1 - Z_{j,j}^{(k)}}\right| \tag{17}$$

□

Proof: The proof is given in the Appendix. ■

If the conditions in Theorem 2 are satisfied for $k = 0,\dots,p-1$, the proposed STIR algorithm recovers the locations of the p-largest elements of u in p steps. Additionally, the order of the indices in which the elements of u are recovered is given by the magnitude of the elements of u in descending order.

Although information about the decay of the elements of u may be useful. In general, this information is not available. The following Theorem relaxes this requirement by the information about the first element that we do not want to recover, denoted as u_{p+1}.

Theorem 3: Let $A \in \mathbb{R}^{m\times n}$, $\rho > 0$ and $y \in \mathbb{R}^m$ generated as in Lemma 2. Without loss of generality, assume that the unknown vector $u \in \mathbb{R}^n$ satisfies that $|u_1| \geq |u_2| \geq \cdots \geq |u_p| > |u_{p+1}| \geq \cdots \geq |u_n|$ and $\|u\|_0 = \gamma$ with $p < \gamma \leq$

[1] The computational complexity of the simple implementation can be obtained as $\sum_{i=1}^{p}(n-i+1)(i^3+mi)$. In this last expression, the term (i^3+mi) represents the computational complexity for solving a least squares problem for each candidate variable.

n. Suppose that we are given the subset $\mathcal{S}^{(k)}$ containing q indices of the p-largest elements of u (in magnitude), with $q < p$. Denote as ℓ with $\ell \notin \mathcal{S}^{(k)}$ to be the index of the unselected element with the largest magnitude. Moreover, $\mathcal{S}^{(k)}$ has a corresponding regularisation matrix $W^{(k)}$ satisfying the conditions in Theorem 1. From $W^{(k)}$ define $Z^{(k)}$, $\tau^{(k)}$ and $\epsilon^{(k)}$ as in (9), (14) and (15), respectively.

If $Z_{i,i}^{(k)} \neq 1$ for all $i \notin \mathcal{S}^{(k)}$, $\tau^{(k)} < 1$ and

$$\begin{aligned} |u_\ell| > & |u_{p+1}| + 2\frac{\tau^{(k)}}{1-\tau^{(k)}}(|u_\ell|(p-q) + |u_{p+1}|(\gamma - p)) \\ & + 2\frac{\epsilon^{(k)}}{1-\tau^{(k)}} \end{aligned} \tag{18}$$

Then, the optimal solution of $\mathcal{P}_{tikho}$, $\hat{x}^{(k)}$, satisfies that

$$\max_{\substack{i \notin \mathcal{S}^{(k)} \\ i \leq p}} \left| \frac{\hat{x}_i^{(k)}}{1 - Z_{i,i}^{(k)}} \right| > \max_{\substack{j \notin \mathcal{S}^{(k)} \\ j > p}} \left| \frac{\hat{x}_j^{(k)}}{1 - Z_{j,j}^{(k)}} \right| \tag{19}$$

□

If the conditions in Theorem 3 are satisfied at each step, then the STIR algorithm will recover the p-largest elements of u in p steps. However, the elements are not necessarily recovered in descending order.

Finally, for the case when there is no noise, i.e. $\epsilon^{(k)} = 0$ for all $k \in \{0, \ldots, \gamma - 1\}$, the vector u is sparse with $\|u\|_0 = \gamma < n$, and we want to recover all nonzero elements of u, i.e. $p = \gamma$. Thus, we obtain the following result.

Theorem 4: Let $A \in \mathbb{R}^{m \times n}$, $\rho > 0$ and $y \in \mathbb{R}^m$ is generated as in Lemma 2 with $\eta = 0$. Without loss of generality, suppose that the unknown vector $u \in \mathbb{R}^n$ satisfies $|u_1| \geq |u_2| \geq \cdots \geq |u_\gamma| > 0$, $\|u\|_0 = \gamma$ and $0 < \gamma < n$. Moreover, suppose that all $Z^{(k)}$ for $k = 0, \ldots, \gamma - 1$ generated by the algorithm in Table II are positive definite, and also assume that $Z_{i,i}^{(k)} \neq 1$ for all $i \notin \mathcal{S}^{(k)}$ with $k = 0, \ldots, \gamma - 1$. If

$$\gamma < \frac{1}{2}\left(\frac{1}{\tau^{(0)}} - 1\right) \tag{20}$$

Then, the algorithm in Table II recovers the nonzero elements of u in γ steps. □

Remark 1: The sufficient conditions stated in Theorem 4 are obtained using a worst-case analysis. Results analogous to Theorem 4 have been obtained for OMP and LARS, see [26]. Those exact recovery conditions include a condition analogous to (20) that is given by [26]

$$\gamma \leq \frac{1}{2}\left(\frac{1}{\mu} + 1\right) \tag{21}$$

where μ is the mutual coherence which is computed as follows

$$\mu = \max_{i \neq j} \frac{|A_{i-col}^\top A_{j-col}|}{\|A_{i-col}\| \|A_{j-col}\|}. \tag{22}$$

□

VI. A Genotypic predictor of HIV-1 drug resistance

In the previous sections, we have introduced and analysed a method for exact sparse recovery. In the current section we use the proposed STIR algorithm to construct a genotypic predictor of HIV-1 drug resistance.

According to the World Health Organization, around 37 million people live with HIV at the end of 2015, and around 17 million people are receiving antiretroviral treatment worldwide [33]. Highly active antiretroviral therapy (HAART) has proven to be effective in controlling the advance of HIV infection. This therapy uses a combination of antiretroviral drugs to treat the HIV infection. However, the benefits of HAART regimes may be compromised by the development of drug resistance [3]. HIV is characterised by both a high mutation rate and a high rate of infection of healthy cells. These conditions imply that natural selection occurs at a fast pace [3]. Thus, when a genetic variant of HIV becomes tolerant to the current treatment, this genetic variant may quickly dominate the population of infected cells in an individual. In some cases, the emergence of drug resistance can occur in a matter of weeks [3]. Thus, addressing HIV drug resistance is critical for achieving viral suppression, addressing treatment failure, and preventing the need to move to more expensive and toxic second- and third-line antiretroviral therapy regimens [33].

Gaining a better understanding of HIV drug resistance is essential for an efficient use of the existing drugs [2]. This knowledge is useful to find a new treatment for a patient that has developed drug resistance in a previous treatment. This decision is particularly challenging, due to the high level of cross resistance within drug classes [3].

Several approaches have been taken to predict HIV-1 drug resistance, see e.g. [2]. In [2] a comparison of five statistical learning methods for the construction of drug susceptibility predictors. In that comparison, LARS performs very well when selecting relevant features. The paper [1] uses LARS to construct a genotypic predictor, and it also provides an in-depth analysis of the contributions of individual HIV-1 protease mutations to drug susceptibility. In part, the success of LARS may be justified by its theoretical guarantees. In [34] it has been proved that LARS offers theoretical guarantees to recover sparse solutions. The conditions for which these guarantees hold were derived by analysing a worst-case scenario. Thus, LARS may recover a sparse solution, even when the conditions described in [34] are not satisfied.

A. Data description

The Genotype-Phenotype datasets [22] are used in this section. These datasets are publicly available at the Stanford HIV resistance database [23]. For the analysis, we focus on Protease Inhibitor (PI) drugs. For protease inhibitor drugs, a high-quality Genotype-Phenotype dataset [22] is available. This dataset contains $10,935$ phenotype results from $1,808$ isolates. These results are available for the following eight PI drugs: fosamprenavir (FPV), atazanavir (ATV), indinavir (IDV), lopinavir (LPV), nelfinavir (NFV), saquinavir (SQV), tipranavir (TPV) and darunavir (DRV). However, the results

TABLE III
LIST OF THE 217 MUTATIONS CONSIDERED IN THE ANALYSIS.

4P/S/T; 6R; 10F/I/L/V; 11I/L/V; 12A/I/K/N/P/S/T; 13I/V; 14K/R; 15I/V; 16A/E/G/R/W; 18H/Q; 19I/L/P/Q/T/V; 20I/K/M/R/T/V; 21E; 22A/V; 23I; 24F/I/L; 25D; 29D; 30N; 32I; 33F/I/L/V; 34D/E/K/Q; 35D/E/G/N; 36I/L/M/V; 37C/D/E/H/N/S/T; 38L; 39P/Q/S; 41K/R; 43I/K/R/T; 45K/R; 46I/L/M/V; 47A/V; 48M/V; 50L/V; 53F/L; 54A/L/M/S/T/V; 55K/N/R; 57G/K/R; 58E/Q; 60D/E/N; 61E/H/N/Q; 62I/V; 63A/C/H/L/P/Q/S/T/V; 64I/L/M/V; 65D/E; 66F/I/V; 67C/F/S/Y; 68E; 69H/K/Q/R/Y; 70E/K/R/T; 71A/I/L/T/V; 72E/I/K/L/M/R/T/V; 73A/C/G/S/T; 74A/P/S/T; 75V; 76V; 77I/V; 79A/D/P/S; 82A/F/I/L/S/T; 83D; 84A/C/V; 85I/V; 87R; 88D/G/S; 89I/L/M/V; 90L/M; 91S; 92K/Q/R; 93I/L/M; 95F;

from a particular isolate may not be available for all eight PI drugs. For each isolate, drug susceptibility is expressed as fold resistance compared with a standard wild-type control isolate [22], [2].

The analysis in this paper focuses on the drugs for which more data is available in the Genotype-Phenotype dataset [22], namely, FPV, IDV, LPV, NFV and SQV. In this study, we consider the 217 mutations which are listed in Table III. The mutations considered in the analysis include all the mutation used for prediction in [1], and all nonpolymorphic treatment selected mutations for PI defined in [2] except by 10R, 53Y and 88T.

B. Least squares formulation

To compute a predictor of HIV-1 drugs susceptibility, we first use the information in the genotype-phenotype dataset to formulate a least squares problem for each drug. In the least-squares formulation the vector measurements $y \in \mathbb{R}^m$ is given by the logarithm, in base ten, of the n-fold drug resistance. The regression matrix $A \in \mathbb{R}^{m \times n}$ is given by a binary matrix with $A_{i,j} = 1$ if the i-th isolate has the j-th mutation, and $A_{i,j} = 0$ otherwise. Given that 217 mutations are considered, then $n = 217$. The number of isolates used for each drug are: $m = 1680$ for FPV, $m = 1729$ for IDV, $m = 1443$ for LPV, $m = 1770$ for NFV and $m = 1721$ for SQV. There are 1317 isolates that have n-fold drug resistance information for all 5 drugs considered in the analysis.

C. Data analysis

To select the predictors for each drug we use all isolates for which the n-fold resistance is available. To compare the results, we use the isolates that have n-fold drug-resistant information for all drugs considered. We compared the proposed method, denoted as STIR, against LARS. We also consider a variant of STIR in which the columns of the regression matrix are normalised to have unitary ℓ_2 norm before using STIR. This variant of the method is denoted as STIR-N. Note that STIR, STIR-N and LARS are only used to select the regressors to be used by the predictor and that the coefficients of each predictor are computed using least squares.

Figure 1 shows the 20-fold cross-validation error for the drugs FPV and SQV as a function of the cardinality of the predictor. The results show that LARS has a better performance than STIR at low cardinality and the results also show that, in general, STIR-N outperforms LARS and STIR. Figure 2 shows the 20-fold cross-validation error as a function of the cardinality of the predictor for the IDV, LPV and NFV drugs. For these drugs STIR-N also outperforms the other methods.

The poor performance of STIR predictors with low cardinality is a consequence that STIR does not take into account the ℓ_2-norm of the regressor. This means that STIR does not take into account how common a mutation is, and uncommon mutations may be selected in the first iterations of the algorithm. The proposed STIR-N method address this issue by renormalising the columns of the regression matrix to have a unitary ℓ_2 norm. This implies that the parameters to be estimated are now scaled by the ℓ_2 norm of the corresponding regressor. Note that the theoretical results in this paper also apply to STIR-N, but the values of $\tau^{(0)}$ must be computed using the normalised matrix. The STIR-N method has a tendency to select more common mutations in the first iterations of the algorithm. It is worth pointing out that LARS also normalises the columns of the regression matrix to have a unitary ℓ_2 norm.

For all drugs considered, we now consider the predictors with cardinality 46, which is the cardinality of the predictor described in [1]. Table IV shows the mutations that the predictors of a given method have in common across all drugs considered. We have highlighted the mutations that are not part of the predictor in [1]. It is observed that some mutations are common to all methods.

The predictors obtained for each drug are not included in the paper due to space limitations. The predictors produced by LARS and STIR-N agree with most of the mutations in their predictors. The largest discrepancy between these methods is 12 mutations for the SQV drug. The predictors produced by the STIR disagree with LARS predictors in at most 20 mutations, and the STIR predictors disagree with the STIR-N predictor in at most 18 mutations. The discrepancy between the predictors obtained by STIR-N and LARS is small, but this discrepancy is enough to make a difference in the prediction accuracy. The proposed STIR-N method outperforms LARS in the cases considered in this paper, but the advantage of STIR-N over LARS becomes negligible as the cardinality of the predictor increases. This fact can be seen in Figures 1 and 2.

D. Comparison under ideal conditions

In this section we compare the performance of LARS and the proposed method under ideal conditions. We consider a regression matrix generated by 1317 isolates that have n-fold drug resistance information for all five drugs considered in the analysis. We study exact recovery conditions for this particular case and evaluate the applicability of the conditions stated in Theorem 4. We perform a Monte-Carlo study using noiseless data generated for several sparse vectors, u. The cardinality of the sparse vectors ranges between one and 100, and for each of these cardinality values we generate $N_{sim} = 1000$ different sparsity patterns for the true vector, u. The magnitude of the non-zero elements of u is set to one, i.e. if $i \in \mathcal{S} \implies u_i = 1$. We consider two stop criteria for LARS, in the first criterion LARS stops when the remaining predictors have a very small

TABLE IV
MUTATIONS THAT ALL PREDICTORS WITH CARDINALITY 46 FOR EACH METHOD HAVE IN COMMON. WE HAVE HIGHLIGHTED THE MUTATIONS THAT ARE NOT IN THE PREDICTOR DESCRIBED IN [1].

Method	Number of mutations in common	Mutations in common among all drug predictors of cardinality 46
LARS	20	10F/I; **20R**; 24I; 33F; 35N; 43T; 46L; 48M/V; 50L; 54A/S/T/V; **63P**; **71V**; 84A/V; 90M;
STIR	18	10F; **22V**; 24F/I/**L**; 35N; 47A; 48M; 50L; 54A/S/T; **67F**; 76V; 82F; 84A/V; 90M;
STIR-N	18	10F/I; **20R**; 24F/I; 43T; 50L; 54A/S/T/V; **71V**; 76V; 82A/F; 84A/V; 90M;

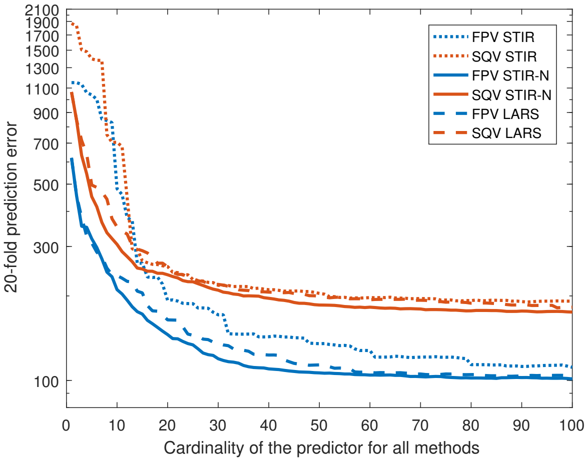


Fig. 1. Twenty-fold cross-validation error as a function of the cardinality of the predictor for FPV and SQV drugs.

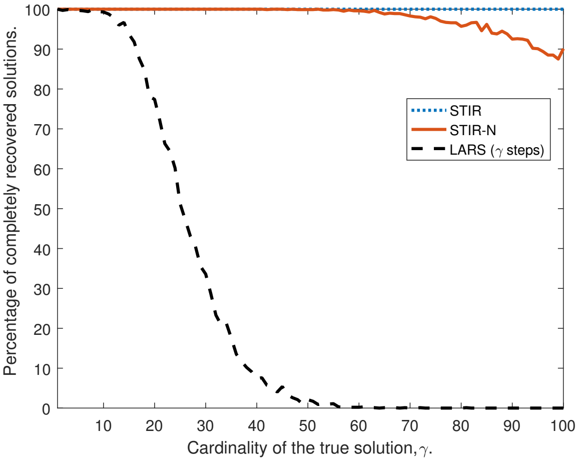


Fig. 3. Recovery rate in γ steps as a function of the cardinality of the true solution, γ.

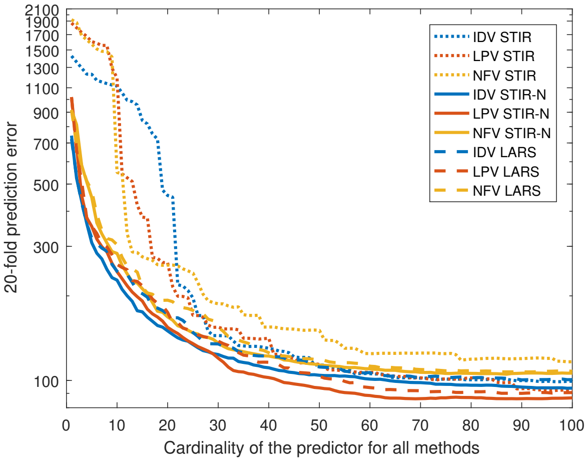


Fig. 2. Twenty-fold cross-validation error as a function of the cardinality of the predictor for IDV, LPV and NFV drugs.

correlation with the current residual. In the second criterion, LARS stops when the cardinality of the current solution matches the cardinality of the true vector, denoted as γ. Figure 3 shows the percentage of solutions that recover the true sparsity pattern for each approach. The proposed STIR method recovers the true sparsity pattern for all cases, however when LARS is restricted to recover the signal in γ steps, the recovery rate decays rapidly. The proposed STIR-N method has better performance than LARS, but it has worse performance than STIR.

For this particular regression matrix, LARS guarantees exact recovery in γ steps whenever the cardinality of the vector is less than or equal to 1.1. The proposed STIR method (using $\rho = 1$) guarantees exact recovery in γ steps when the cardinality of the vector is less than 7.47. The proposed STIR-N method (using $\rho = 1$) guarantees exact recovery for $\gamma < 0.4$, which means that it offers no guarantees at all. Note that all these bounds on the cardinality were obtained by analysing worst-case scenario of the noiseless case, thus a better performance is expected in practice.

Figure 4 shows the average number of steps required by STIR, STIR-N and LARS to recover all indices of the nonzero elements of the true solution. For LARS, the average number of steps is equal to the true cardinality of the solution only when $\gamma = 1$. This fact agrees with the theoretical guarantees for LARS. The proposed STIR method recovers the true solution in γ steps in all cases. Finally, the proposed STIR-N method takes an average number of steps that is slightly higher than the true cardinality of the solution, but such difference is small and hard to see from Figure 4. For example, when the cardinality of the true solution is $\gamma = 100$ the average number of steps required to recover all indices of the nonzero elements of the true solution are 100 steps for STIR, 100.14 steps for STIR-N, and 109.68 steps for LARS.

VII. CONCLUSIONS

In this paper we have presented a novel forward selection procedure that is based on the magnitude of the estimates of

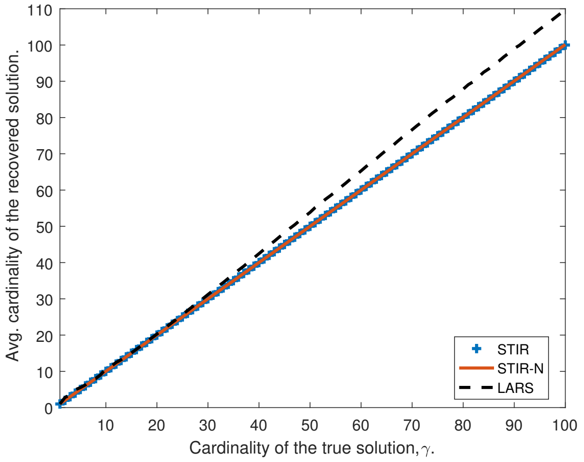


Fig. 4. Average number of steps required to recover all the indices of the nonzero elements of the true solution.

the candidate variables. We also developed a novel method for Stepwise Tikhonov regularisation (STIR) that implements the proposed forward selection procedure in least squares problems. We derive sufficient conditions under which the STIR method recovers of the largest elements (in magnitude) of an unknown vector. Finally, we use the STIR method to construct a genotypic predictor of HIV-1 drug resistance and our numerical studies show that the proposed STIR method is competitive with a state-of-the-art method such as LARS.

Future research will focus on studying computational aspects of the proposed method for the development of numerically robust and fast implementations of the STIR method.

Acknowledgement

This research was supported by the Australian Government through the Australian Research Council's Discovery Projects funding scheme (project DP130103039).

Appendix A
Technical details

This section describes technical details for the results in the current paper. First we present the proof of Theorem 1. This Theorem presents the update equations for $x^{(k)}$, $Z^{(k)}$ and the optimal cost $g(x^{(k)})$. Next, Lemma 3 introduces an upper and a lower bound for the term $\left(\frac{\hat{x}_i^{(k)}}{1-Z_{i,i}^{(k)}}\right)$. These upper and lower bounds are then used to build a proof of Theorem 2, and these bounds are also used in the proof of Theorem 3. Finally, a proof of Theorem 4 is presented.

Proof of Theorem 1: We first recall that

$$W^{(k+1)} = W^{(k)} - e_\ell e_\ell^\top$$

Next, by using the matrix inversion lemma, see e.g. [35], we have that

$$\begin{aligned}(A^\top A + \rho W^{(k+1)})^{-1} &= (A^\top A + \rho W^{(k)})^{-1}\\ &+ (A^\top A + \rho W^{(k)})^{-1}\rho e_\ell (1 - e_\ell^\top (A^\top A + \rho W^{(k)})^{-1}\rho e_\ell)^{-1}\\ &\cdot e_\ell^\top (A^\top A + \rho W^{(k)})^{-1} \end{aligned} \tag{23}$$

By using the definition $Z^{(k)}$, i.e. $Z^{(k)} := (A^\top A + \rho W^{(k)})^{-1}\rho$ we can rewrite (23) in terms of $Z^{(k)}$ as follows

$$Z^{(k+1)} = Z^{(k)} + Z^{(k)} e_\ell (1 - e_\ell^\top Z^{(k)} e_\ell)^{-1} e_\ell^\top Z^{(k)} \tag{24}$$

Since, $x^{(k+1)} = Z^{(k+1)}\frac{1}{\rho}A^\top y$, is straight forward to prove that $x^{(k+1)}$ can be computed as follows

$$\begin{aligned} x^{(k+1)} &= Z^{(k)}\frac{1}{\rho}A\top y\\ &+ Z^{(k)} e_\ell (1 - e_\ell^\top Z^{(k)} e_\ell)^{-1} e_\ell^\top Z^{(k)} \frac{1}{\rho} A^\top y\\ &= x^{(k)} + Z^{(k)} e_\ell (1 - e_\ell^\top Z^{(k)} e_\ell)^{-1} e_\ell^\top x^{(k)} \end{aligned} \tag{25}$$

Now, we turn our focus on the optimal cost. By using Lemma 1 we have that the optimal cost of $\mathcal{P}_{tikho}$ for a given W is given by

$$g(\hat{x}) = y^\top y - y^\top A(A^\top A + \rho W)^{-1} A^\top y = y^\top (y - A\hat{x})$$

Next, we compute the following cost difference

$$\begin{aligned} g(x^{(k)}) - g(x^{(k+1)}) &= y^\top A (x^{(k+1)} - x^{(k)})\\ &= y^\top A Z^{(k)} e_\ell (1 - e_\ell^\top Z^{(k)} e_\ell)^{-1} e_\ell^\top x^{(k)}\\ &= (x^{(k)}\rho)^\top e_\ell (1 - e_\ell^\top Z^{(k)} e_\ell)^{-1} e_\ell^\top x^{(k)} \end{aligned} \tag{26}$$

The final expressions are obtained by considering in (24), (25) and (26) that for a vector $u \in \mathbb{R}^n$ and for a matrix $M \in \mathbb{R}^{n\times n}$ we have that

$$u_\ell = e_\ell^\top u \tag{27}$$

$$M_{i,j} = e_i^\top M e_j \tag{28}$$

$$M_{\ell-row} = e_\ell^\top M \tag{29}$$

$$M_{\ell-col} = M e_\ell. \tag{30}$$

■

Lemma 3: Given $A \in \mathbb{R}^{m\times n}$ $\rho > 0$ and a diagonal projection matrix $W^{(k)} \succeq 0$. Suppose that the assumptions in Lemma 2 hold. Define $Z^{(k)}$ as in (9) and assume that $Z_{i,i}^{(k)} \neq 1$ for all $i \notin \mathcal{S}^{(k)}$. Consider that $|\hat{x}_\ell^{(k)}| = \|W^{(k)}u\|_\infty$ and that $e_j = W^{(k)} e_j$ if and only if $j \notin \mathcal{S}^{(k)}$. Then, for $i, \ell \notin \mathcal{S}^{(k)}$ such that $i \neq \ell$ we have

$$\begin{aligned} \left|\frac{\hat{x}_i^{(k)}}{1 - Z_{i,i}^{(k)}}\right| \leq &(1 - \tau^{(k)})|u_i| + \tau^{(k)} \|(W^{(k)} - e_\ell e_\ell^\top)u\|_1\\ &+ \tau^{(k)}\|W^{(k)}u\|_\infty + \epsilon^{(k)} \end{aligned} \tag{31}$$

$$\left|\frac{\hat{x}_\ell^{(k)}}{1 - Z_{\ell,\ell}^{(k)}}\right| \geq \|W^{(k)}u\|_\infty - \tau^{(k)}\|(W^{(k)} - e_\ell e_\ell^\top)u\|_1 - \epsilon^{(k)} \tag{32}$$

□

Proof of Lemma 3: The i-th element of $\hat{x}^{(k)}$ can be computed using Lemma 2 as $\hat{x}_i^{(k)} = e_i^\top \hat{x}^{(k)}$. This gives us

$$\begin{aligned} \hat{x}_i^{(k)} = &(1 - Z_{i,i}^{(k)})u_i - (Z_{i-col}^{(k)})^\top (W^{(k)} - e_i e_i^\top)u\\ &+ \frac{1}{\rho}(Z_{i-col}^{(k)})^\top A^\top \eta \end{aligned} \tag{33}$$

where we have used the facts $Z^{(k)}e_i = Z^{(k)}_{i-col}$ and $Z^{(k)}_{i,i} = e_i^\top Z^{(k)} e_i$. Under the assumption that $Z^{(k)}_{i,i} \neq 1$ we obtain that

$$\begin{aligned}\frac{\hat{x}^{(k)}_i}{1 - Z^{(k)}_{i,i}} =& u_i - \frac{(Z^{(k)}_{i-col})^\top}{1 - Z^{(k)}_{i,i}}(W^{(k)} - e_i e_i^\top)u \\ &+ \frac{1}{\rho(1 - Z^{(k)}_{i,i})}(Z^{(k)}_{i-col})^\top A^\top \eta \qquad (34)\end{aligned}$$

from the properties of absolute values, we have that

$$\begin{aligned}\frac{|\hat{x}^{(k)}_i|}{|1 - Z^{(k)}_{i,i}|} \leq& |u_i| + \left|\frac{(Z^{(k)}_{i-col})^\top}{1 - Z^{(k)}_{i,i}}(W^{(k)} - e_i e_i^\top)u\right| \\ &+ \left|\frac{1}{\rho(1 - Z^{(k)}_{i,i})}(Z^{(k)}_{i-col})^\top A^\top \eta\right| \qquad (35)\\ \frac{|\hat{x}^{(k)}_i|}{|1 - Z^{(k)}_{i,i}|} \geq& |u_i| - \left|\frac{(Z^{(k)}_{i-col})^\top}{1 - Z^{(k)}_{i,i}}(W^{(k)} - e_i e_i^\top)u\right| \\ &- \left|\frac{1}{\rho(1 - Z^{(k)}_{i,i})}(Z^{(k)}_{i-col})^\top A^\top \eta\right| \qquad (36)\end{aligned}$$

From the definition of $W^{(k)}$ and e_j we have that for all $j \notin \mathcal{S}^{(k)}$ is true that $(W^{(k)} - e_j e_j^\top) = (W^{(k)} - e_j e_j^\top)^2$. Moreover, Hölder's inequality states that for vectors a, b is true that $|a^\top b| \leq \|a\|_\infty \|b\|_1$. Then, we have that

$$\begin{aligned}&\left|\frac{(Z^{(k)}_{i-col})^\top}{1 - Z^{(k)}_{i,i}}(W^{(k)} - e_i e_i^\top)u\right| \\ &\leq \left\|\frac{(W^{(k)} - e_i e_i^\top)Z^{(k)}_{i-col}}{1 - Z^{(k)}_{i,i}}\right\|_\infty \|(W^{(k)} - e_i e_i^\top)u\|_1 \\ &\leq \tau^{(k)} \|(W^{(k)} - e_i e_i^\top)u\|_1 \qquad (37)\end{aligned}$$

Next, by combining the definition of $\epsilon^{(k)}$, (37) and (36) we obtain

$$\frac{|\hat{x}^{(k)}_i|}{|1 - Z^{(k)}_{i,i}|} \geq |u_i| - \tau^{(k)}\|(W^{(k)} - e_i e_i^\top)u\|_1 - \epsilon^{(k)} \qquad (38)$$

To obtain (32) we use $i = \ell$ in (38).

Now, we focus on proving (31). We combine (35) and (37) to obtain that for all $i \notin \mathcal{S}^{(k)}$ is true that

$$\frac{|\hat{x}^{(k)}_i|}{|1 - Z^{(k)}_{i,i}|} \leq |u_i| + \tau^{(k)}\|(W^{(k)} - e_i e_i^\top)u\|_1 + \epsilon^{(k)} \qquad (39)$$

Note that $\|(W^{(k)} - e_i e_i^\top)u\|_1 = \sum_{j \notin \mathcal{S}^{(k)}} |u_j| - |u_i|$, thus we obtain

$$\begin{aligned}\frac{|\hat{x}^{(k)}_i|}{|1 - Z^{(k)}_{i,i}|} \leq& |u_i| + \tau^{(k)}\left(\sum_{j \notin \mathcal{S}^{(k)}} |u_j| - |u_i|\right) + \epsilon^{(k)} \qquad (40)\\ =& (1 - \tau^{(k)})|u_i| + \tau^{(k)}\left(\sum_{j \notin \mathcal{S}^{(k)}} |u_j| - |u_\ell|\right) \\ &+ \tau^{(k)}|u_\ell| + \epsilon^{(k)} \qquad (41)\end{aligned}$$

where in the last step we added and subtracted $\tau^{(k)}|u_\ell|$. After considering that $|u_\ell| = \|W^{(k)}u\|_\infty$, equation (31) follows. ■

Proof of Theorem 2: Straightforward from Lemma 3, by assuming that the upper bound (31) is strictly lower than the lower bound (32). ■

Proof of Theorem 3: The proof is based on making that a lower bound for the left-hand-side of (19) to be greater than an upper bound for the right-hand-side of (19). Based in (32), a lower bound for the left-hand-side of (19) is given by

$$\begin{aligned}\max_{\substack{i \notin \mathcal{S}^{(k)} \\ i \leq p}} \left|\frac{\hat{x}^{(k)}_i}{1 - Z^{(k)}_{i,i}}\right| \geq& (1 - \tau^{(k)})|u_\ell| \\ &- \tau^{(k)}\left((p - q)|u_\ell| + \sum_{j=p+1}^{\gamma} |u_j|\right) \\ &- \epsilon^{(k)} \\ \geq& (1 - \tau^{(k)})|u_\ell| - \tau^{(k)}(p - q)|u_\ell| \\ &- \tau^{(k)}(\gamma - p)|u_{p+1}| - \epsilon^{(k)} \qquad (42)\end{aligned}$$

Similarly, based on (40), an upper bound for the right-hand-side of (19) is given by

$$\begin{aligned}\max_{\substack{j \notin \mathcal{S}^{(k)} \\ j > p}} \left|\frac{\hat{x}^{(k)}_j}{1 - Z^{(k)}_{j,j}}\right| \leq& (1 - \tau^{(k)})|u_{p+1}| + \tau^{(k)}(p - q)|u_\ell| \\ &+ \tau^{(k)}(\gamma - p)|u_{p+1}| + \epsilon^{(k)} \qquad (43)\end{aligned}$$

Then, the result is obtained by ensuring that (42) is greater than (43). ■

Proof of Theorem 4: To prove Theorem 4 we first derive an upper bound for $\tau^{(k)}$ as a function of $\tau^{(0)}$. Note that $\tau^{(k)}$ in (14) can be expressed as

$$\tau^{(k)} = \max_{\substack{j \notin \mathcal{S}^{(k)} \\ i \notin \mathcal{S}^{(k)} \\ i \neq j}} \left|\frac{Z^{(k)}_{i,j}}{1 - Z^{(k)}_{j,j}}\right|$$

From (11), the update of each element on $Z^{(k)}$ is given by

$$Z^{(k)}_{i,j} = Z^{(k-1)}_{i,j} + \frac{Z^{(k-1)}_{i,\ell}}{1 - Z^{(k-1)}_{\ell,\ell}} Z^{(k-1)}_{\ell,j}. \qquad (44)$$

Using algebra, we obtain the following relationship

$$\frac{Z^{(k)}_{i,j}}{1 - Z^{(k)}_{j,j}} = \frac{\left(\frac{Z^{(k-1)}_{i,j}}{1 - Z^{(k-1)}_{j,j}} + \frac{Z^{(k-1)}_{i,\ell}}{1 - Z^{(k-1)}_{\ell,\ell}} \frac{Z^{(k-1)}_{\ell,j}}{1 - Z^{(k-1)}_{j,j}}\right)}{1 - \frac{Z^{(k-1)}_{j,\ell}}{1 - Z^{(k-1)}_{\ell,\ell}} \frac{Z^{(k-1)}_{\ell,j}}{1 - Z^{(k-1)}_{j,j}}}. \qquad (45)$$

Suppose that $\tau^{(k-1)} < 1$, and using the definition of $\tau^{(k)}$ and $\tau^{(k-1)}$ in (45) we obtain that

$$\tau^{(k)} \leq \frac{(\tau^{(k-1)} + (\tau^{(k-1)})^2)}{1 - (\tau^{(k-1)})^2} = \frac{\tau^{(k-1)}}{1 - \tau^{(k-1)}} \qquad (46)$$

Equation (46) may be applied recursively to obtain that

$$\tau^{(k)} \leq \frac{\tau^{(0)}}{1 - k\tau^{(0)}} \qquad (47)$$

where it is assumed that $\tau^{(0)} < \frac{1}{k}$. This ensures that $\tau^{(q)} < 1$ for $q = 1, \ldots, k$. Equation (47) provides an upper bound of $\tau^{(k)}$ that is independent of the element selected by the algorithm in Table II. The next step is to find upper bounds for $\tau^{(k)}$, for $k = 0, \ldots, \gamma - 1$ that ensure the recovery of the sparsity pattern of the vector u. These upper bounds are obtained from Theorem 3 by considering that $u_{p+1} = 0$, $p = \gamma$, $q = k$ and $\epsilon^{(k)} = 0$ for all k. Under these conditions, (18) is given by

$$\tau^{(k)} < \frac{1}{1 + 2(\gamma - k)}. \tag{48}$$

The condition for exact sparse recovery is obtained by ensuring that the upper bound in (48) is larger than the bound in (47). Thus we obtain that

$$\tau^{(0)} < \frac{1}{1 + 2\gamma - k} \tag{49}$$

The final result in (20) follows from the evaluation of the tightest bound at $k = 0$. ■

References

[1] S.-Y. Rhee, J. Taylor, W. J. Fessel, D. Kaufman, W. Towner, P. Troia, P. Ruane, J. Hellinger, V. Shirvani, A. Zolopa *et al.*, "HIV-1 protease mutations and protease inhibitor cross-resistance," *Antimicrobial agents and chemotherapy*, vol. 54, no. 10, pp. 4253–4261, 2010.

[2] S.-Y. Rhee, J. Taylor, G. Wadhera, A. Ben-Hur, D. L. Brutlag, and R. W. Shafer, "Genotypic predictors of human immunodeficiency virus type 1 drug resistance," *Proceedings of the National Academy of Sciences*, vol. 103, no. 46, pp. 17 355–17 360, oct 2006.

[3] F. Clavel and A. J. Hance, "HIV drug resistance," *New England Journal of Medicine*, vol. 350, no. 10, pp. 1023–1035, 2004, pMID: 14999114.

[4] E. A. Hernandez-Vargas and R. H. Middleton, "Modeling the three stages in HIV infection," *Journal of Theoretical Biology*, vol. 320, pp. 33 – 40, 2013.

[5] E. A. Hernandez-Vargas, P. Colaneri, and R. H. Middleton, "Switching strategies to mitigate HIV mutation," *IEEE Transactions on Control Systems Technology*, vol. 22, no. 4, pp. 1623–1628, 2014.

[6] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[7] D. Needell and R. Vershynin, "Signal recovery from incomplete and inaccurate measurements via regularized orthogonal matching pursuit," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 310–316, April 2010.

[8] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, Feb 2006.

[9] E. J. Candès, Y. Plan *et al.*, "Near-ideal model selection by ℓ_1 minimization," *The Annals of Statistics*, vol. 37, no. 5A, pp. 2145–2177, 2009.

[10] R. Carvajal, G. Urrutia, and J. C. Agüero, "An optimization-based algorithm for model selection using an approximation of Akaike's information criterion," in *2016 Australian Control Conference (AuCC)*, Nov 2016, pp. 217–220.

[11] G. Urrutia, R. A. Delgado, R. Carvajal, D. Katselis, and J. C. Agüero, "Sparse logistic regression utilizing cardinality constraints and information criteria," in *2016 IEEE Conference on Control Applications (CCA)*, Sept 2016, pp. 798–803.

[12] C. Herzet, A. Drémeau, and C. Soussen, "Relaxed recovery conditions for OMP/OLS by exploiting both coherence and decay," *IEEE Transactions on Information Theory*, vol. 62, no. 1, pp. 459–470, 2016.

[13] Z. Zhang, Y. Xu, J. Yang, X. Li, and D. Zhang, "A survey of sparse representation: Algorithms and applications," *IEEE Access*, vol. 3, pp. 490–530, 2015.

[14] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[15] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM review*, vol. 43, no. 1, pp. 129–159, 2001.

[16] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *Information Theory, IEEE Transactions on*, vol. 50, no. 10, pp. 2231–2242, 2004.

[17] R. R. Hocking, "The analysis and selection of variables in linear regression," *Biometrics*, vol. 32, no. 1, pp. 1–49, 1976.

[18] Y. Pati, R. Rezaiifar, and P. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," in *Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on*, Nov 1993, pp. 40–44 vol.1.

[19] S. Chen, C. Cowan, and P. Grant, "Orthogonal least squares learning algorithm for radial basis function networks," *Neural Networks, IEEE Transactions on*, vol. 2, no. 2, pp. 302–309, Mar 1991.

[20] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE transactions on information theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[21] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Transactions on information theory*, vol. 52, no. 1, pp. 6–18, 2006.

[22] J. Zhang, S.-Y. Rhee, J. Taylor, and R. W. Shafer, "Comparison of the precision and sensitivity of the antivirogram and phenosense HIV drug susceptibility assays," *Journal of acquired immune deficiency syndromes (1999)*, vol. 38, no. 4, p. 439, 2005.

[23] S.-Y. Rhee, M. J. Gonzales, R. Kantor, B. J. Betts, J. Ravela, and R. W. Shafer, "Human immunodeficiency virus reverse transcriptase and protease sequence database," *Nucleic acids research*, vol. 31, no. 1, pp. 298–303, 2003.

[24] A. S. Bandeira, E. Dobriban, D. G. Mixon, and W. F. Sawin, "Certifying the restricted isometry property is hard," *IEEE Transactions on Information Theory*, vol. 59, no. 6, pp. 3448–3450, jun 2013.

[25] D. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Transactions on Information Theory*, vol. 47, no. 7, pp. 2845–2862, 2001.

[26] J. Duan, C. Soussen, D. Brie, J. Idier, and Y. P. Wang, "On LARS/homotopy equivalence conditions for over-determined lasso," *IEEE Signal Processing Letters*, vol. 19, no. 12, pp. 894–897, Dec 2012.

[27] R. A. Delgado, J. C. Agüero, G. C. Goodwin, and E. M. A. M. Mendes, "Application of rank-constrained optimization to nonlinear system identification," in *1st IFAC Conference on Modelling, Identification and Control of Nonlinear Systems (MICNON 2015)*, Saint-Petersburg, Russia, 2015.

[28] R. P. Aguilera, G. Urrutia, R. A. Delgado, D. Dolz, and J. C. Agüero, "Quadratic model predictive control including input cardinality constraints," *IEEE Transactions on Automatic Control*, vol. 62, no. 6, pp. 3068–3075, June 2017.

[29] R. A. Delgado, J. C. Agüero, and G. C. Goodwin, "A novel representation of rank constraints for real matrices," *Linear Algebra and its Applications*, vol. 496, pp. 452 – 462, 2016.

[30] B. Cassidy, V. Solo, and A. J. Seneviratne, "Grouped L0 least squares penalised magnetoencephalography," in *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*, May 2012, pp. 868–871.

[31] M. Nikolova, "Description of the minimizers of least squares regularized with ℓ_0-norm. Uniqueness of the global minimizer," *SIAM Journal on Imaging Sciences*, vol. 6, no. 2, pp. 904–937, jan 2013.

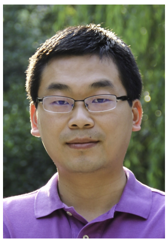
[32] A.-I. N. Tikhonov, *Solutions of Ill Posed Problems (Scripta series in mathematics)*. Vh Winston, 1977.

[33] World Health Organization, "Global health sector strategy on HIV 2016-2021. towards ending AIDS," 2016.

[34] D. L. Donoho and Y. Tsaig, "Fast solution of ℓ_1-norm minimization problems when the solution may be sparse," *IEEE Transactions on Information Theory*, vol. 54, no. 11, pp. 4789–4812, 2008.

[35] W. W. Hager, "Updating the inverse of a matrix," *SIAM Review*, vol. 31, no. 2, pp. 221–239, 1989.

Ramón A. Delgado received his professional title of Ingeniero Civil Electrónico and Master degree in Electronic Engineering from Universidad Técnica Federico Santa María, Chile, in 2009. He received a Ph.D degree in Electrical Engineering from the University of Newcastle, Australia, in 2014. He is currently a Research Academic at the University of Newcastle. In his current role, Dr Delgado is mainly focussed on the current collaboration project between the University of Newcastle and the Swedish company Ericsson AB. His research interests include system identification, control, signal processing and optimisation.



Zhiyong Chen received the B.E. degree from the University of Science and Technology of China, and the M.Phil. and Ph.D. degrees from the Chinese University of Hong Kong, in 2000, 2002, and 2005 respectively. He worked as a Research Associate at the University of Virginia during 2005-2006. He joined the University of Newcastle, Australia, in 2006 where he is currently an Associate Professor. He is also a Changjiang Chair Professor with Central South University, Changsha, China. His research interests include nonlinear systems and control, biological systems, and multi-agent systems. He is an Associate Editor of IEEE Transactions on Automatic Control and IEEE Transactions on Cybernetics.

Professor Richard H. Middleton completed his Ph.D. (1987) from the University of Newcastle, Australia. He was a Research Professor at the Hamilton Institute, The National University of Ireland, Maynooth from May 2007 till 2011 and is currently Professor at the University of Newcastle and Head of the School of Electrical Engineering and Computer Science. He has served as Program Chair (CDC 2006), CSS Vice President Membership Activities, and Vice President Conference Activities. In 2011, he was President of the IEEE Control Systems Society. He is a Fellow of IEEE and of IFAC, and his research interests include a broad range of Control Systems Theory and Applications, including Robotics, control of distributed systems and Systems Biology with applications to Parkinson's Disease and HIV Dynamics.