Dictionary Learning for Sparse Representations: A Pareto Curve Root Finding Approach

Mehrdad Yaghoobi [*] and Mike E. Davies

Institute for Digital Communications (IDCom),
the University of Edinburgh, EH9 3JL, UK
{m.yaghoobi-vaighan@ed.ac.uk,mike.davies}@ed.ac.uk
http://www.see.ed.ac.uk/research/IDCOM

Abstract. A new dictionary learning method for exact sparse representation is presented in this paper. As the dictionary learning methods often iteratively update the sparse coefficients and dictionary, when the approximation error is small or zero, algorithm convergence will be slow or non-existent. The proposed framework can be used in such a setting by gradually increasing the fidelity of the approximation. This technique has previously been used for the convex sparse representations. It has been extended here to the non-convex dictionary learning problem by allowing the dictionary be modified.

Key words: Sparse Inverse Problems, Dictionary Learning for Sparse Representations, Pareto Curves, Gradient Projection Method

1 Introduction

Inverse problems are the subject of different areas in science and engineering. We name computational tomography, seismology and radar as some examples. The objective is to recover the parameters which give the observed data by applying the forward operator. As the problem is often ill-posed, we need to assume a model for the parameters to resolve the recoverability ambiguity. The sparsity model, in which we assume few parameters are non-zero, can successfully model a variety of observed natural data. The forward operator is often assumed to be linear and finite dimensional. Hence the forward operator can be presented as a *fat* matrix, which is called a dictionary and each column is called an atom [1]. The inverse problem with this setting is an NP-hard problem, but many practical algorithms have been proposed to solve it approximately or exactly in some occasions [1, 2].

When the forward operator in this setting is not given, we can use the domain knowledge to select a *good* model. The term *good* means that such sparse

[*] This work is supported by EU FP7, FET-Open grant number 225913. MED acknowledges support of his position from the Scottish Funding Council and their support of the Joint Research Institute with the Heriot-Watt University as a component part of the Edinburgh Research Partnership.

parameters (coefficients) are possible to find for the observed data. The domain knowledge can be incorporated as a parametric model for the dictionary [3] or the dictionary can be adapted using a sample based dictionary learning method [4–8]. In the sample based dictionary learning, the algorithm starts with an initial guess for the dictionary and gradually changes it to provide sparser representations or less model mismatch. The learning process has thus two distinct steps, solving a sparse inverse problem with the current dictionary, then updating the dictionary to reduce model mismatch with fixed coefficients[1]. When the dictionary is normal, *i.e.* a dictionary with unit norm atoms, it is recoverable if the observed data uniquely represents the dictionary up to some column permutations and atom sign flips. The uniqueness condition has been explored in [9, 10]. Aharon *et al.* intuitively suggested the K-SVD dictionary learning algorithm as a candidate for the dictionary recovery. This method has been introduced for dictionary learning for sparse approximation, where there exists some model mismatch, *i.e.* approximation error. Although good results have been reported in [7, 8], the relation between exact recovery and K-SVD has not been shown. Recently a new framework for dictionary recovery has been introduced in [11, 12], which is based on dictionary learning for ℓ_1 exact sparse representations. Gribonval *et al.* [12] shows that the generative dictionary is a local minimum of the proposed (non-convex) optimization problem with high probability, when the coefficients and the dictionary follow some distributions. A difficulty with the proposed recovery framework is that most of the dictionary learning methods are not able to work in an exact sparse representation setting. An ℓ_1 exact dictionary learning [13] has been proposed recently. Unfortunately the stability and the practical performance of this method have not been explored. Here we present a new dictionary learning method which can be used for a high fidelity sparse representation. It is based on generalizing the Pareto Curve root finding technique for ℓ_1 sparse representation [14], to a dictionary learning framework. The new framework can also be used in dictionary learning for fixed levels of fidelity sparse *approximations.*

2 Sparse Representation using Pareto Curve Root Finding

Let $\mathbf{y} \in \mathbb{R}^m$, $\mathbf{x} \in \mathbb{R}^n$ and $f : \mathbb{R}^n \mapsto \mathbb{R}^m$ be respectively the observed data, the coefficient vector and the forward operator. As the forward operator lies in the finite dimensional space here, we represent it using a matrix $\mathbf{D} \in \mathbb{R}^{m \times n}$. The inverse problem can now be formulated as finding $\mathbf{x}$ such that $\mathbf{y} = \mathbf{D}\mathbf{x}$. As this problem often is ill-posed/underdetermined, *e.g* $m < n$, the solution is not unique. By assuming a sparsity model for the coefficients, the sparse inverse problem can be solved by finding the sparsest solution in $\{\mathbf{x} | \mathbf{y} = \mathbf{D}\mathbf{x}\}$. The convex envelope of the sparsity function, constrained to an ℓ_∞ ball, is the ℓ_1 norm, which is now the most popular regularization factor for sparse coding

[1] The K-SVD [7] is slightly different, as it also allows the coefficients to adapt in the dictionary update steps.

[2]. The sparse inverse problem can now be found by minimizing the following optimization problem,

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad s.t \quad \mathbf{y} = \mathbf{D}\mathbf{x}. \tag{1}$$

This optimization problem, which is called Basis Pursuit (BP), is convex and can be solved using the Linear Programing (LP) method [2]. Although LP is one the most powerful methods for *exactly* solving (1), we practically need to solve it up to few significant figures. The gradient based methods converge fast with this precision. The main issue with solving (1) is that the objective is not differentiable and the gradient projection method [15], which is an efficient method to solve convexly constrained continuously differentiable objectives, can not be applied. For the moment let $\mathbf{y} \approx \mathbf{D}\mathbf{x}$ and distance in the observation space be Euclidean, $d(\mathbf{y}, \tilde{\mathbf{y}}) = \|\mathbf{y} - \tilde{\mathbf{y}}\|_2$. A sparse approximation can be found using the following optimization problem, which is called LASSO [16],

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2^2 \quad s.t. \quad \|\mathbf{x}\|_1 \leq \tau, \tag{2}$$

where $\tau \in \mathbb{R}^+$ is the radius of the ℓ_1 ball. Let the solution of (2), for each τ, be called $\phi(\tau)$, *i.e.* $\phi(\tau) := \|\mathbf{y} - \mathbf{D}\mathbf{x}^*\|_2^2$, where $\mathbf{x}^*$ is the minimizer of (2). $\phi(\tau)$ is a non-increasing function and when $\mathbf{D}$ is full rank, the set $\mathcal{A} = \{\tau | \phi(\tau) = 0\}$ is non-empty. The *Pareto Curve* is generated by plotting $\phi(\tau)$ for $\tau \in \mathbb{R}^+$. It actually presents the optimal trade-off between the sparsity, here the ℓ_1 norm of the coefficients, and the ℓ_2 norm of the approximation error. Let τ^* be the smallest $\tau \in \mathcal{A}$. It is straight forward to show that any solution of (2) with τ^* is also a solution of (1). The good news is that (2) can efficiently be solved using the gradient projection method, for any given τ. We thus only need to find τ^*. Van den berg *et al.* showed in [14] that $\phi(\tau)$ is convex and differentiable, where $\phi(\tau) \neq 0$, and used a Newton's root finding method to iteratively update $\tau^{[n]}$, which is τ at the nth iteration, such that $\lim_{n \to \infty} \tau^{[n]} = \tau^*$. The Newton's method is guaranteed to converge with this setting, *i.e.* convex and differentiable function.

3 Dictionary Learning for ℓ_1 Exact Sparse Representation

In sample based dictionary learning, a set of training samples $\mathcal{Y} = \{\mathbf{y}_l\}_{l \in \mathcal{L}}$, where $|\mathcal{L}| = L$, is given. When L is large and the sparse signals are rich enough to uniquely define the dictionary, up to some permutations in columns of $\mathbf{D}$ and atom sign flips, Gribonval *et al.* [12] suggest to solve the following problem to recover the generative dictionary,

$$\min_{\mathbf{X}, \mathbf{D}} \|\mathbf{X}\|_1 \quad s.t. \quad \mathbf{D}\mathbf{X} = \mathbf{Y}, \mathbf{D} \in \mathcal{D}, \tag{3}$$

where $\mathbf{X}$, $\mathbf{Y} = [\mathbf{y}_l]_{l \in \mathcal{L}}$ are respectively coefficient and observation matrices, $\mathcal{D}$ is the dictionary admissible set and $\| \cdot \|_1 = \sum_{i,j} |\{\cdot\}_{i,j}|$ is an element-wise one norm in the matrix space. An admissible set $\mathcal{D}$, which constrains the amplitude, has been used to resolve the scale ambiguity, which prevents the dictionary and the coefficients to scale by $\alpha > 1$ and $1/\alpha$ respectively, to reduce the sparsity penalty $\|\cdot\|_1$, while preserving the admissibility $\mathbf{DX} = \mathbf{Y}$. A common choice for $\mathbf{D}$ is a fixed column ℓ_2 norm or a bounded column ℓ_2 norm to make a convex admissible set. A similar argument which was presented earlier, about the non-differentiability of the objective in (1), can be presented here. Let $\psi(\mathbf{X}, \mathbf{D}) = \|\mathbf{Y} - \mathbf{DX}\|_F^2$, where $\|\cdot\|_F$ is the Frobenius norm. A new formulation for the dictionary learning can be presented as,

$$\min_{\mathbf{X},\mathbf{D}} \psi(\mathbf{X}, \mathbf{D}) \quad s.t. \quad \mathbf{X} \in \mathcal{B}_1^\tau, \quad \mathbf{D} \in \mathcal{D}, \tag{4}$$

where $\mathcal{B}_1^\tau := \{\mathbf{X} | \|\mathbf{X}\|_1 \leq \tau\}$ is the ℓ_1 ball with radius τ and $\mathcal{D}$ is a convex admissible set. Let $\phi(\tau)$ similarly be the optimum value of (4) for each τ. Although $\phi(\tau)$ is non-increasing and $\mathcal{A}$ is not empty, it might now be non-convex.

The gradient projection can be used again, as the constraints are convex sets, to solve (4) for a fixed τ. The uniform convergence of this method can be shown if $\nabla\psi$ is locally Lipschitz continuous [15]. As $\mathcal{D}$ and $\mathcal{B}_1^\tau$, for a finite τ, are compact, it is straightforward to shows locally Lipschitz continuity of $\nabla\psi$. However in practice this results in a small gradient step, which slow down the convergence of the algorithm. Instead we use a block relaxation technique and keep fixed $\mathbf{X}$ or $\mathbf{D}$ while gradient projecting the other parameter. It allows us to choose a larger gradient step size at each step. The down side of this technique is that we only can show the *convergence of* $(\mathbf{X}, \mathbf{D})$ *to a set of accumulation points*, see for example [8, Appendices A and B]. The gradient steps in directions $\frac{\partial}{\partial \mathbf{X}}\psi$ and $\frac{\partial}{\partial \mathbf{D}}\psi$ are respectively smaller than $(\sigma_{max}(\mathbf{D}))^{-2}$ and $(\sigma_{max}(\mathbf{X}))^{-2}$, where σ_{max} operator finds the largest singular value.

Another difficulty with solving (4), with a fixed τ, is that the (global) minimum might not be found using a gradient projection method. This raises a big issue in the convergence proof of the overall algorithm, where the achievable *local* minimum of $\psi(\mathbf{X}, \mathbf{D})$ might increase in the next minimization step, after increasing τ. To resolve this issue, we initiate the gradient projection algorithm by the (local) minimum found using the previous τ. The algorithm is now guaranteed to reduce the objective after each gradient projection step [15]. As the objective is lower bounded, the stability of the algorithm is guaranteed[2].

When $\phi(\tau)$ is convex, as it has been shown to be in the ℓ_1 sparse representation, Newton's method finds the root. This is not true in the dictionary learning as $\phi(\tau)$ might not be convex. We can use the non-increasing feature of ϕ, by updating $(\mathbf{X}, \mathbf{D})$ as explained earlier, and find the root by applying a line search method. Although such an update scheme for τ may not be as efficient as using Newton's method, *i.e.* more updates needed to find τ^*, we practically found that the proposed line search method in Algorithm 1 converges fast. Note that when τ gets large enough, $\phi(\tau) \to 0$ and the gradient of objective in (4), with respect to each parameter, tends to zero. This is enough to show the convergence of the algorithm to some local minima $\mathbf{X}^*$, $\mathbf{D}^*$.

[2] The stability in a Lyapunov sense, which provides boundedness of the solutions (for $\mathbf{X}$, which might become infinitely large in general).

Algorithm 1 Pareto Curve root finding based Dictionary Learning (PCDL)

```
initialization: $0 < \delta \ll 1$, $\mathbf{D}_\tau = \mathcal{P}_{\mathcal{D}}\left([d_{i,j} = \mathcal{N}(1,0)]_{i,j}\right)$, $\mathbf{X}_\tau = \mathbf{D}_\tau^\dagger \mathbf{Y}$,
    $\tau = \|\mathbf{X}_\tau\|_1 / N$,$K$, $\mu = 1.5$
while $\left(\|\mathbf{D}_\tau \mathbf{X}_\tau - \mathbf{Y}\|_F^2 - \epsilon\right)^2 > .01\epsilon$ do
    $\mathbf{X}^{[0]} = \mathbf{X}_\tau$, $\mathbf{D}^{[0]} = \mathbf{D}_\tau$
    for $n = 1$ to $K$ do
        $\mathbf{A} = \mathbf{X}^{[n-1]} - \frac{2}{\sigma_{max}^2\left(\mathbf{D}^{[n-1]}\right)+\delta} \mathbf{D}^{[n-1]^T} \left(\mathbf{D}^{[n-1]} \mathbf{X}^{[n-1]} - \mathbf{Y}\right)$
        $\mathbf{X}^{[n]} = \mathcal{P}_{\mathcal{B}_1^\tau}(\mathbf{A})$
        $\mathbf{B} = \mathbf{D}^{[n-1]} - \frac{2}{\sigma_{max}^2\left(\mathbf{X}^{[n]}\right)+\delta} \left(\mathbf{D}^{[n-1]} \mathbf{X}^{[n]} - \mathbf{Y}\right) \mathbf{X}^{[n]}$
        $\mathbf{D}^{[n]} = \mathcal{P}_{\mathcal{D}}(\mathbf{B})$
    end for
    if $\|\mathbf{D}_\tau \mathbf{X}_\tau - \mathbf{Y}\|_F^2 < \epsilon$ then
        $\tau = \tau / \mu$
        $\mu = \mu^{1/3}$
    else
        $\mathbf{X}_\tau = \mathbf{X}^{[K]}$, $\mathbf{D}_\tau = \mathbf{D}^{[K]}$
    end if
    $\tau = \mu\tau$
end while
output: $\mathbf{D}_\tau$
```

Algorithm 1 presents a pseudocode for the Pareto Curve root finding based Dictionary Learning (PCDL) method. It initiates with a random dictionary after projecting onto the admissible set $\mathcal{D}$, $\mathcal{P}_{\mathcal{D}}(\cdot)$. The initial $\mathbf{X}_\tau$ was selected to be the minimum ℓ_2 norm inverse solution. The algorithm starts with a τ, which is a division of the ℓ_1 norm of the current solution, *i.e.* least square solution. The "for" loop includes K iterations of gradient-projection steps. $\mathcal{P}_{\mathcal{B}_1^\tau}$ is the projection onto the ℓ_1 ball, with radius τ. The "if" part is the line search for updating τ. For a given precision ϵ, if the approximation error is less than this precision, the algorithm steps back and choose a smaller scale factor μ. Otherwise it updates $(\mathbf{X}_\tau, \mathbf{D}_\tau)$. The algorithm stops when $\left(\|\mathbf{D}_\tau \mathbf{X}_\tau - \mathbf{Y}\|_F^2 - \epsilon\right)^2 \leq .01\epsilon$, which practically seems to be an acceptable accuracy.

4 Simulations

We demonstrate the dictionary recovery with a toy example in the first experiment. A normalized (normally distributed) random dictionary $\mathbf{D} \in \mathbb{R}^{20\times 40}$ and a set of $L = 1280$ training samples are generated, where the sparsity is changing from 3 to 7 in different experiments. The locations and magnitude of non-zero coefficients are respectively selected uniformly random and bounded in $[.2, 1]$ with a random sign pattern. By following the definition of the atom recovery from [7, 8], we call an atom is recovered if the Euclidean distance of the atom and one of the recovered atoms is not greater than $0.1\sqrt{2}$. The simulations were repeated 5 times, with new initial random dictionaries, for $K = 500$. The aver-

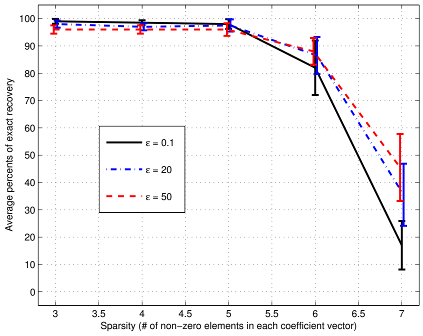

Fig. 1. Exact dictionary recovery with different approximation errors ϵ.

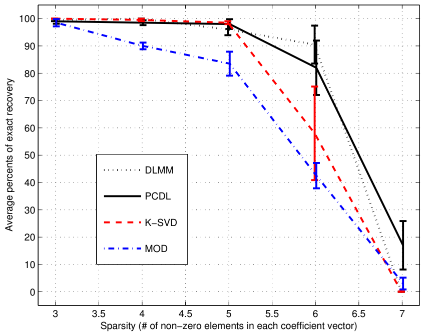

Fig. 2. Exact dictionary recovery using PCDL, MMDL, K-SVD and MOD methods.

age atom recovery percentages for three different $\epsilon = 0.1, 20$ and 50, which are roughly corresponding to $42, 19$ and $11dB$ signal to noise ratios, are plotted in Figure 1. The variance of the success is also shown by error bars. It shows that although sparse approximation with high fidelity improves the success rate for highly sparse coefficients, the noisy sparse approximation provides better recovery with less sparse coefficients. The success of dictionary recovery using sparse *approximations*, *i.e.* large ϵ, in a less sparse setting may be caused by the fact that small coefficients, which may actually be zero in the generated sparse coefficients, are now assumed to be noise. The algorithm thus learn a dictionary based on the coefficients which we are more confident to be non-zero. Figure 2 compares the proposed method, $\epsilon = 0.1$, with some other dictionary learning methods. The λ parameter, which is the Lagrange multiplier in the sparse approximation [2], is 0.4 in K-SVD [7], MOD [6] and DLMM [8]. An ℓ_1 sparse approximation with an extra de-biasing step, which is simply an orthogonal projection of the observed data onto the linear span of subdictionary indexed by non-zero coefficients, has been used in the sparse approximation steps of these methods. It demonstrates that PCDL performs almost the same as the best current methods for very sparse and it also shows superior performance for less sparse data. It deserves to mention that the computational complexity of PCDL is often higher than the standard fixed-sparsity dictionary learning methods, as it needs to iteratively reduce τ and learn a dictionary within new ℓ_1 ball. The total computational cost of the algorithm directly depends on the number of iterations in the outer loop, *i.e.* "while" loop in Algorithm 1. In practice we found that after 10 to 20 updates of τ, the algorithm converges to a solution in this example. We also observed that the inner loop, which includes K iterations of coefficients and dictionary updates, is faster than MMDL, for small λ's.

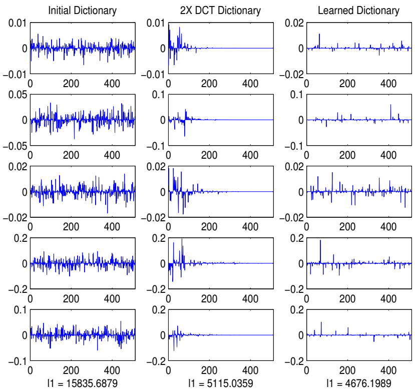

Fig. 3. Sparse representations of five different training data, using three different dictionaries.

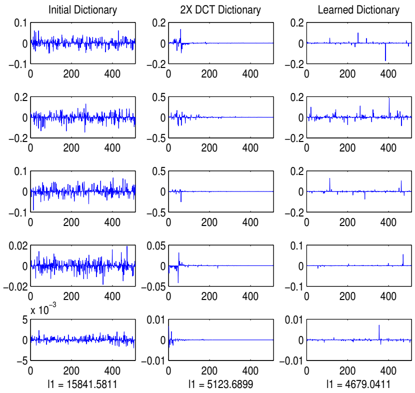

Fig. 4. Sparse representations of five different observed data, out of training set, using three different dictionaries.

In the second experiment we applied PCDL to the dictionary learning for audio data, which has been shown to have some sparse structures. We chose a 256 by 512 dictionary and randomly select 16384 audio samples, of length 256, from a long audio record from BBC radio 3, which plays classical music. The sparse representations of 5 random samples from training and some independent data are respectively shown in Figure 3 and 4. The initial random, a two times overcomplete DCT (oversampled frequency) and the learned dictionaries have been used to find the sparse codes in the left to right columns respectively. The total ℓ_1 of 16384 sparse codes are mentioned in the bottom line. The minimum ℓ_1 sparse representation, using the learned dictionary, provides good sparse codes, even though it may not be the optimum dictionary. The consistency of the learning is also demonstrated, as the learned dictionary works well for independently selected audio blocks. It is also demonstrated that the learned dictionary is superior to the two times overcomplete DCT in providing less ℓ_1, with the same ϵ.

5 Conclusions

We have introduced a new dictionary learning framework for sparse representation. It is based on Pareto Curve root finding which has previously been used for sparse representation. The new algorithm is guaranteed to be stable and we can also show the convergence to a set of fixed points. As the new framework needs to update τ using a line search method, a more efficient method may provide faster convergence using fewer updates of τ. We chose the current technique for updating τ as it provides a uniform reduction of a lower bounded objective.

The proposed algorithm can also be used in a dictionary learning for sparse approximation framework, by using an extra parameter ϵ, which measures the deviation from the exact representation subspace, *i.e.* $\mathbf{DX} = \mathbf{Y}$. This is particularly useful, when ϵ is small, as current dictionary learning methods often converge very slowly when using a small sparsity penalty, *i.e.* small λ.

References

1. Mallat, S., Zhang, Z.: Matching Pursuits with time frequency dictionaries. IEEE Trans. on Signal Processing, 41(12), 3397–3415 (1993)
2. Chen, S.S., Donoho, D.L., Saunders, M.A.: Atomic Decomposition by Basis Pursuit. SIAM Journal on Scientific Computing, 20(1), pp 33–61 (1998)
3. Yaghoobi, M., Daudet, L., Davies, M.: Parametric Dictionary Design for Sparse Coding. IEEE Trans. on Signal Processing, 57(12), 4800–4810 (2009)
4. Olshausen, B.A., Field, D.J.: Sparse coding with an overcomplete basis set: a strategy employed by V1? Vision Research , 37(23), 3311–3325 (1997)
5. Lewicki, M.S. and Sejnowski, T.J.: Learning Overcomplete Representations. Neural Comp. , 12(2), 337–365 (2000)
6. Engan, K., Aase, S.O., Hakon-Husoy, J.: Method of optimal directions for frame design. In: International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 2443–2446 (1999)
7. Aharon,M., Elad,E., Bruckstein, A.M.: K-SVD: an algorithm for designing of overcomplete dictionaries for sparse representation. IEEE Trans. on Signal Processing, 54(11), pp. 4311–4322 (2006)
8. Yaghoobi, M., Blumensath, T., Davies, M.: Dictionary Learning for Sparse Approximations with the Majorization Method. IEEE Trans. on Signal Processing, 57(6), pp. 2178–2191 (2009)
9. Georgiev, P., Theis, F., Cichocki, A.: Sparse component analysis and blind source separation of underdetermined mixtures. IEEE Trans. on Neural Networks, 16(4), pp. 992–996 (2005)
10. Aharon, M., Elad, M., Bruckstein, A.M.: On the uniqueness of overcomplete dictionaries and a practical way to retrieve them. Journal of Linear Algebra and Applications, 416, pp. 48–67 (2006)
11. Gribonval, R., Schnass, K.: Some Recovery Conditions for Basis Learning by L1-Minimization. In: International Symposium on Communications, Control and Signal Processing (ISCCSP),(2008)
12. Gribonval, R., Schnass, K.: Dictionary Identification: Sparse Matrix-Factorisation via ℓ_1 Minimisation. IEEE Trans. on Information Theory, 56(7), pp. 3523–3539 (2010)
13. Plumbley, M.D.: Dictionary Learning for l1-Exact Sparse Coding. In: International Conference on Independent Component Analysis and Signal Separation (ICA), pp. 406–413 (2007)
14. Van den Berg, E. and Friedlander, M. P.: Probing the Pareto frontier for basis pursuit solutions. SIAM Journal on Scientific Computing, 31(2), pp. 890–912 (2008)
15. Goldstein, A.A.: Convex programming in Hilbert space. Bulletin of the American Mathematical Society, 70(5), pp. 709–710 (1964)
16. Tibshirani, R.: Regression shrinkage and selection via the Lasso. Journal of Royal Statistical Society Series B, 58, pp. 267–288 (1996)