COLLABORATIVE REPRESENTATION, SPARSITY OR NONLINEARITY: WHAT IS KEY TO DICTIONARY BASED CLASSIFICATION?

Xu Chen and Peter J. Ramadge

Department of Electrical Engineering
Princeton University, Princeton, NJ, USA

ABSTRACT

Recent studies have suggested that the critical aspect of sparse representation-based classification (SRC) is collaborative representation, rather than sparsity. This has given rise to fast collaborative representation-based classification using 2-norm regularized least squares (CRC-RLS). This paper digs deeper into the difference between SRC and CRC-RLS. We show that linear coding schemes such as CRC-RLS share a common pairwise boundary class B. Moreover, the corresponding pairwise classifiers can be realized by quadratic SVMs. Using three datasets, we show empirically that collaborative representations are not always required, and that a quadratic SVM has superior generalization over CRC-RLS, with fast classification times. However, SRC exhibits the best prediction accuracy. This leads us to posit that the nonlinear coding of SRC is a key attribute.

Index Terms— Machine Learning, Sparse Representation, Collaborative Representation.

1. INTRODUCTION

Sparse representation with respect to an over-complete dictionary has been of recent interest in a broad range of classification applications. For example, Sparse Representation-based Classification (SRC), first introduced for robust face recognition [1], has since been used for speaker recognition [2], tumor classification [3], image classification [4] and music genre classification [5]. SRC computes a sparse coding for a test sample x with respect to a dictionary of labelled training samples. Based on this sparse coding, a class specific approximation $\tilde{x}_i$ to x is constructed for each class. Then x is assigned to the class of least approximation residual.

Recent work [6, 7, 8, 9] has questioned the advantage of sparsity in image classification. In [6] it is argued that it is collaborative representation, not sparsity, that plays the essential role in SRC. With this motivation, a new collaborative representation-based classification scheme using ℓ_2-regularized least squares (CRC-RLS) has been proposed. CRC-RLS has very competitive classification results in face recognition, with significantly less complexity than SRC. In [7, 8], this approach is further extended and several variations are developed. Similarly, [9] proposes a visual tracker using non-sparse linear representations, which admit efficient closed-form solutions without sacrificing accuracy.

We show that linear coding schemes such as CRC-RLS share a common pairwise boundary class B, and that these pairwise boundaries can be realized by quadratic SVMs. Using a synthetic dataset and two real datasets, we demonstrate empirically that quadratic SVMs can robustly learn better boundaries than CRC-RLS and that collaborative representations are not always required.

In §2 we review SRC and CRC-RLS. In §3 we study classification based on linear codings and use a simple synthetic example to show the limitations of this approach and to highlight alternatives. In §4 we give experimental results comparing SRC and CRC-RLS on the synthetic problem and on two standard datasets. We draw our conclusions in §5.

2. BACKGROUND

We first review the basic aspects of SRC [1] and CRC-RLS [6]. Let the columns of $D_i = [d_{i,1}, \ldots, d_{i,p_i}] \in \mathbb{R}^{n \times p_i}$ contain unit norm feature vectors (codewords or atoms) drawn from the i-th class. Form the joint dictionary $D = [D_1, D_2, ..., D_c] \in \mathbb{R}^{n \times p}$. SRC and CRC-RLS use this dictionary to perform multi-class classification as follows. One first solves a regularized least squares (RLS) problem to obtain a coding of unit norm $x \in \mathbb{R}^n$ with respect to D:

$$\arg\min_{w_i \in \mathbb{R}^{p_i}} {}^1\!/_2 \|x - \textstyle\sum_{i=1}^c D_i w_i\|_2^2 + \lambda \sum_{i=1}^c f(w_i), \quad (1)$$

where for SRC, $f(w_i) = \|w_i\|_1$, and for CRC-RLS, $f(w_i) = \|w_i\|_2^2$. Each w_i incurs a regularization cost $f(w_i)$, but the least squares cost results from a cooperative effort across the classes. This is collaborative representation.

The solution $\tilde{w}$ of (1) represents x as a linear combination of the columns of D. For SRC, $\tilde{w} = h(x)$ is sparse and is a nonlinear function of x, but solving (1) can be time consuming for large problems. Fast solvers, e.g., [10, 11, 12], have been proposed, and dictionary screening, e.g., [13, 14, 15, 16], can also help. In contrast, for ℓ_2 regularization, the solution of (1) is linear in x and given in closed form by $\tilde{w} = (D^T D + \lambda I)^{-1} D^T x$.

Next, the above codings are used to form class-specific approximations $\tilde{x}_i = D_i \tilde{w}_i$ to x, with residuals $x - \tilde{x}_i$.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

Here $\tilde{w}_i = E_i\tilde{w} \in \mathbb{R}^{p_i}$ extracts the entries in $\tilde{w}$ associated with class i. Finally, SRC selects the class of least residual: $\arg\min_i s_i(x) = \|x - D_i\tilde{w}_i\|_2^2$. CRC-RLS is similar, except for instance specific scaling of the residual [6]: $\arg\min_i s_i(x) = \|x - D_i\tilde{w}_i\|_2^2/\|\tilde{w}_i\|_2^2$.

3. COLLABORATION, SPARSITY, NONLINEARITY

As a contrast to collaborative representation, consider moving the sum in (1) within $\|...\|_2^2$, outside the squared norm. In the "uncollaborative" problem that results, each class independently seeks its own representation $\tilde{w}_i$ of x using its dictionary D_i. For ℓ_1 (resp. ℓ_2) regularization, we call this coding method *subspace-SRC* (S-SRC) (resp. *subspace-RLS* (S-RLS)). The codings of S-RLS are also linear in x.

We now consider what we lose, if anything, by moving from SRC to linear coding schemes such as CRC-RLS. We begin with classifiers that use linear codings $\tilde{w}_j = P_j x$ without residual scaling. The class specific approximations of x with respect to D are $\tilde{x}_j = D_j P_j x = L_j x$, with residual $r_j(x) = (I - L_j)x$. Let $s_j(x) = \|r_j(x)\|_2^2 = x^T Q_j x$, where $Q_j = (I - L_j)^T(I - L_j)$ is symmetric PSD (positive semidefinite). Then the classification of x is $\arg\min_j s_j(x) = x^T Q_j x$. Hence classification between classes j and k is determined by the sign of the quadratic form $d_{jk}(x) = x^T(Q_j - Q_k)x$, with x classified as class k if $d_{jk}(x) \geq 0$.

Write $Q_{jk} = Q_j - Q_k = V\Sigma V^T$ with V orthogonal and $\Sigma = \text{diag}(\boldsymbol{\sigma}) \in \mathbb{R}^{n\times n}$. Let $\mathcal{U}$ (resp. $\mathcal{U}^\perp$) denote the subspace spanned by the eigenvectors with $\sigma_i \geq 0$ (resp. $\sigma_i < 0$). A classification boundary divides $\mathbb{R}^n$ into classification regions for class k (containing $\mathcal{U}$) and class j (containing $\mathcal{U}^\perp$). Write $x = Vb$. Then $d_{jk}(x) = b^T V^T(V\Sigma V^T)Vb = \sum_{i=1}^n \sigma_i b_i^2$. So the boundary is determined by the quadratic: $\sum_i \sigma_i b_i^2 = 0$. The solutions of this equation are invariant under scaling and hence form a (double-sided) cone in $\mathbb{R}^n$. For unit norm x, the decision boundary is the intersection of this cone with the unit sphere. Hence, modulo V, the set of boundaries is:

$$B = \{b\colon \textstyle\sum_{j=1}^n \sigma_j b_j^2 = 0, \sum_{j=1}^n b_j^2 = 1\}. \qquad (2)$$

For $b \in B$, $b^{(2)} = [b_j^2]$ lies on a linear manifold in $\mathbb{R}^n$ of dimension $n-2$. The warping of this manifold resulting from the component-wise (positive) square root also has dimension $n-2$. The square root operator forms 2^n manifold sections depending on the sign given to each root. These sections join smoothly on the plane where the corresponding variable is zero. So the surface of the sphere is divided into 2 classification regions by $(n-2)$-dimensional, quadratic manifolds. In summary, linear coding results in pairwise classification using quadratic boundaries in B (modulo V). Fixing $\boldsymbol{\sigma}$ determines the boundary shape, and V "rotates" it to the desired location.

For c classes, the decision region for class k is the intersection of $c-1$ pairwise regions. But if we focus on the pairwise boundaries, we can make some interesting connections.

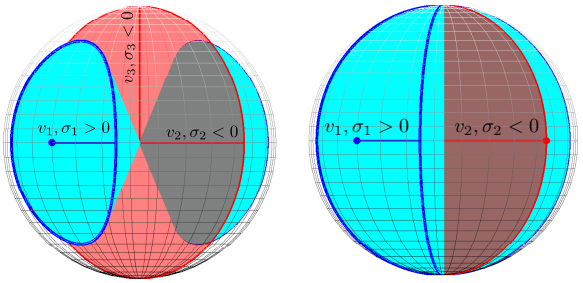


Fig. 1. Example 1. Decision regions of unscaled CRC-RLS for $n = 3$, $\sigma_1 > 0$, $\sigma_2 < 0$, and $\sigma_3 < 0$ (left) and $\sigma_3 = 0$ (right).

Write $d_{jk}(x) = x^T Q_{jk} x = q_{jk}^T \phi(x)$, where $q_{jk} \in \mathbb{R}^m$ is a weight vector and $\phi(x) = (x_1^2, \ldots, x_n^2, x_1x_2, \ldots x_{n-1}x_n)^T \in \mathbb{R}^m$, with $m = {}^{n(n+1)}/_2$. This shows that $d_{jk}(x)$ is a linear classifier in $\mathbb{R}^m$ under the data embedding ϕ. Moreover, we can learn q_{jk}, and hence Q_{jk}, using a SVM with quadratic kernel $(<x, y>)^2$. This quadratic SVM (Q-SVM) has VC dimension $m+1$, and classifies via $\text{sign}(s^T\phi(x) + \gamma)$. If we constrain $\gamma = 0$, it yields the required set of quadratic boundaries. Hence this boundary set has VC dimension m.

Example 1: For $n = 2$, assume that Q_{jk} has $\sigma_1 > 0$ and $\sigma_2 < 0$. By (2), the boundary consists of four points, with entries $b_1 = \pm\sqrt{-\sigma_2/(\sigma_1 - \sigma_2)}$ and $b_2 = \pm\sqrt{\sigma_1/(\sigma_1 - \sigma_2)}$, that divide the 1-sphere into two decision regions. The eigenvector for $\sigma_1 > 0$ is centered on the class k region, and for $\sigma_2 < 0$ on the class j region.

For $n = 3$, assume Q_{jk} has $\sigma_1 > 0$ and $\sigma_2, \sigma_3 < 0$. The positive eigenspace $\mathcal{U}$ (Fig. 1(left)) is centered on the class k decision region, and the negative eigenspace $\mathcal{U}^\perp$ (red tinted disk) on the class j decision region. The boundary has eight sections; four link to form the boundary curve in dark blue in the foreground of the figure. The other four form a matching curve in the opposite hemisphere. The boundary cone is shown as the cyan surface. The situation of one negative and two positive eigenvalues is similar. If one eigenvalue is zero, say $\sigma_1 > 0$, $\sigma_2 < 0$ and $\sigma_3 = 0$, the boundary joins at the poles (Fig. 1(right)) to form two boundary circles. □

Example 2: Consider data on the unit sphere in $\mathbb{R}^3$ with $x = \cos\theta\cos\alpha$, $y = \cos\theta\sin\alpha$, $z = \sin\theta$. For class 1, (α, θ) is uniformly distributed in $([-40°, -20°] \bigcup [20°, 40°]) \times [-15°, 15°]$, and for class 2, uniformly in $[-20°, 20°] \times [\beta - 15°, \beta + 15°]$. Here β controls the relative position of the classes. Training sets consisting of 400 samples/class are generated for $\beta = 0°$ and $15°$. These datasets are (almost surely) separable by B using two great circles with $\alpha^\star = \pm 20°$. This boundary is achieved by $\boldsymbol{\sigma} = (\sigma_1, -\sigma_2, 0)$ with $\sigma_1, \sigma_2 > 0$ and $\sigma_2/\sigma_1 = \tan^2\alpha^\star$, and $V = I$. For each dataset, we train S-RLS, unscaled CRC-RLS and Q-SVM using values of λ (or C for SVM) selected for each classifier and each value of β by cross-validation. The resulting quadratic decision boundaries and training data classifications are shown in Fig. 2. For S-RLS and unscaled CRC-RLS and both values of β, $d_{12}(x)$

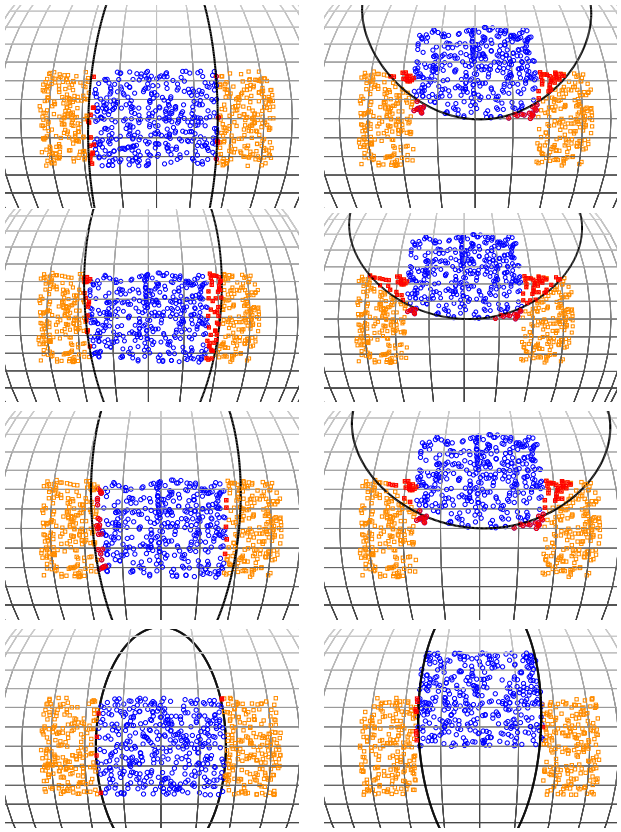


Fig. 2. Example 2. Boundaries and training point classifications from: S-RLS (top), unscaled CRC-RLS (second), CRC-RLS (third) and Q-SVM (bottom). KEY:- Left: $\beta = 0°$, Right: $\beta = 15°$, class 1: orange-□, class 2: blue-o; misclassified points: red. The decision boundary of unscaled CRC-RLS is shown on the CRC-RLS plot.

has one positive eigenvalue. The classifiers easily separate the classes when $\beta > 30°$ (not shown). As β decreases and class 2 moves between class 1, S-RLS and unscaled CRC-RLS struggle to learn a good boundary until near $\beta = 0°$, where the separating boundary is "discovered". In contrast, Q-SVM learns the separating boundary in both cases (fourth row). The third row of Fig. 2, shows the results for CRC-RLS and the boundary curve for unscaled CRC-RLS. There are local changes in classification, but the CRC-RLS decisions are largely determined by the linear coding. □

4. EXPERIMENTAL RESULTS

We now investigate the generalization performance of SRC, CRC-RLS, the subspace variants, and Q-SVM on the synthetic problem in Example 2 and two real datasets. In the following figures, error bars indicate ±standard error.

Fig. 3 shows the generalization performance on the synthetic dataset of the tested classifiers as class 2 slides from $\beta = 0°$ to $\beta = 30°$. Test accuracy is shown for training sets with $N = 20$ and $N = 200$ codewords/class. For each classifier, λ is selected for each value of β and N via leave-one-out

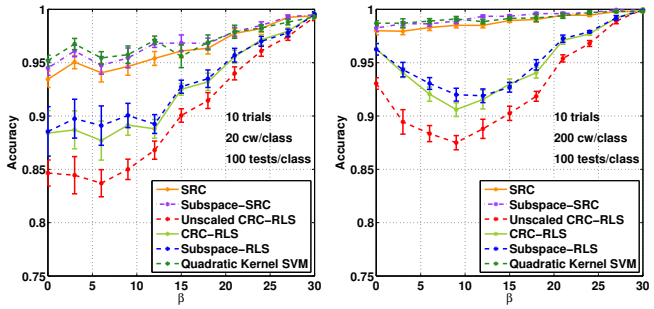
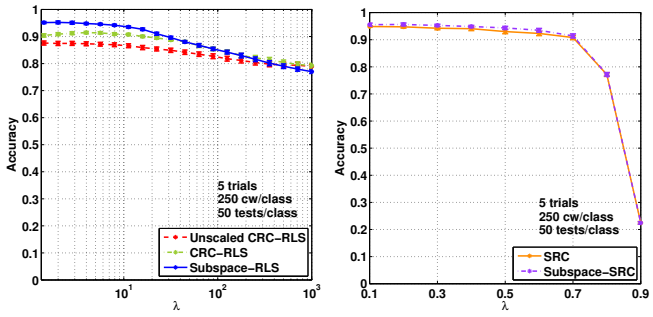


Fig. 3. Generalization results for Example 2. Test performance of six cross-validated classifers as class 2 slides from $\beta = 0°$ to 30° for $N = 20$ training samples/class (left) and $N = 200$ (right).

cross validation on the training set. Q-SVM was implemented and cross-validated using LIBSVM [17]. LIBSVM does not have an option to constrain $\gamma = 0$. So the tested Q-SVM has VC dimension $m + 1$. A comparison with the results obtained using SVM^{light} [18], with $\gamma = 0$, indicated equivalent performance.

Here are the salient points to note: 1) The accuracy of cross-validated CRC-RLS falls well below that of S-SRC, SRC and Q-SVM. Only when $\beta > 30°$ and the classes are linearly separable, can CRC-RLS achieve competitive accuracy. 2) Residual scaling in CRC-RLS has a significant effect, but not enough to change the overall trend of performance. 3) Surprisingly, the "uncollaborative" classifiers (S-RLS and S-SRC) have the same, or better, performance as the "collaborative" counterparts. So collaborative representation is not critical for this problem. 4) Q-SVM has similar accuracy to SRC and S-SRC. So if a strong learning method is used, then (as expected) a quadratic classifier suffices for accurate classification on these datasets. 5) For $N = 200$, CRC-RLS peaks in accuracy around $\beta = 0°$ where it has approximated the separating boundary (see Fig. 2). However, for $N = 20$, this peak disappears, indicating that on the smaller training set CRC-RLS failed to identify the boundary. In contrast, while the performance of Q-SVM also decreased as β approached 0, at all values of β it is competitive with SRC and S-SRC.

The **MNIST** dataset consists of 28×28 hand-written digit images [19]). From $60,000$ training images, we randomly sampled balanced sets of $2,500$ training images and 500 testing images and report average results over 5 random selections. Fig. 4 shows test accuracy plotted versus λ, for the tested classifiers. Multi-class Q-SVM is cross-validated over different parameters and hence is not shown in the plot. However it is included in the table below the plots, where we list the best accuracy and test time of each method. We place the average testing time of Q-SVM in quotes since testing was done using LIBSVM, while other testing times are based on MATLAB code. LIBSVM implements multi-class SVM using the one-vs-one method. Hsu et al. [20] indicate this has comparable performance with the one-vs-rest method. The salient points to note include: SRC is more accurate than

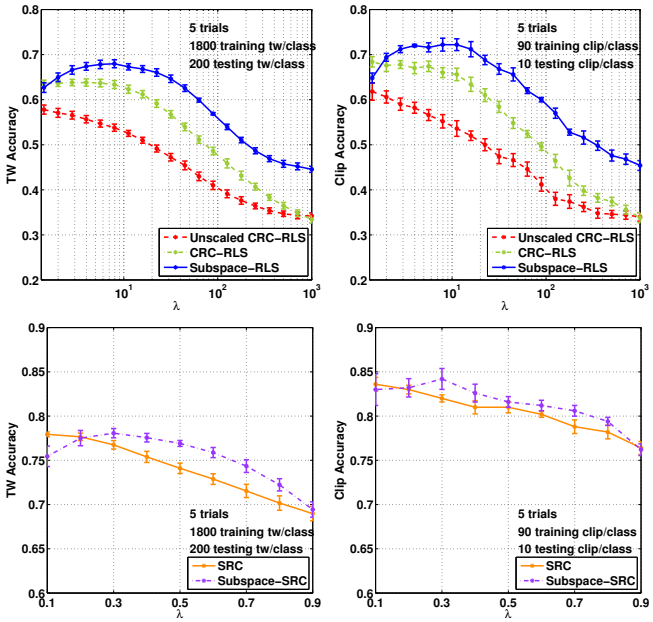


MNIST	SRC	S-SRC	CRC	S-RLS	Q-SVM
Acc (%)	**94.9**	**95.6**	91.4	95.2	94.0
Time (ms)	14	36	7	2	"4"

Fig. 4. Results for MNIST. Top: Classification accuracy vs λ for the tested methods. Bottom: Classification accuracy and average test time at the λ value where each classifier achieves its best accuracy.

CRC-RLS, but at the cost of a longer test time; the "uncollaborative" classifiers subspace-SRC and subspace-RLS again have a competitive accuracy with the collaborative counterparts; and multi-class Q-SVM outperforms CRC-RLS.

The **GTZAN** dataset consists of 100 music clips (30 sec, sampled at 22,050 Hz) for each of ten genres of music [21]. Clips are divided into 3-second, 50% overlap, texture windows (TW) with each TW represented by a 1st-order scattering vector $x \in \mathbb{R}^{199}$ [22]. We randomly select a dictionary of 18,000 TWs and generate test vectors from the remaining 2,000 TWs. The results (Fig. 5) show that for CRC-RLS residual scaling yields a significant improvement. However, SRC still has an average 15% accuracy advantage over CRC-RLS, both in TW and clip classification. Although not as accurate as SRC, multi-class Q-SVM also does well. S-SRC and S-RLS are again at least as good as the corresponding counterparts SRC and CRC-RLS.



Acc (%)	SRC	S-SRC	CRC	S-RLS	Q-SVM
Clip	**83.6**	**84.2**	68.4	72.2	80.8
TW	**77.9**	**78.1**	63.8	68.0	73.2

Time (s)	SRC	S-SRC	CRC	S-RLS	Q-SVM
Clip	1.23	2.89	0.21	0.08	"0.07"
TW	0.06	0.14	0.01	0.004	"0.003"

Fig. 5. Results for music genre classification. Top: Accuracy of the tested methods versus λ. Left: Texture-window accuracy; Right: Clip accuracy. Bottom: The classification accuracy and average test time at the λ values where each classifier achieves its best accuracy.

5. DISCUSSION AND CONCLUSION

We examined how linear coding methods such as CRC-RLS perform beyond the domain of face recognition. We have shown that linear coding classifiers of the SRC-form share a pairwise quadratic boundary class and that these boundaries can be realized by a SVM with kernel $K(x, y) = (<x, y>)^2$. Based on these insights, we performed three experiments leading to the following conclusions.

Multi-class Q-SVM is a viable alternative to CRC-RLS. Despite the separability of the synthetic problem by a quadratic boundary, S-RLS and CRC-RLS failed to learn a separating boundary over a large range of β. This results in the trough and peak in CRC-RLS accuracy in Fig. 3(right). This is exacerbated for the smaller training set (Fig. 3(left)). In contrast, Q-SVM had significantly better accuracy. Multi-class Q-SVM also outperformed CRC-RLS and S-RLS on the two real datasets while exhibiting competitive testing times. How this scales with c remains to be investigated.

Collaborative representation is not always necessary. For our datasets, well-tuned S-SRC (resp. S-RLS) was as accurate as SRC (resp. CRC-RLS). The collaborative representation of SRC did not improve accuracy, but it did save testing time compared with S-SRC. However, the performance of these classifiers may differ in robustness to tuning errors and to noise and other corruption in the test samples.

SRC & multi-class Q-SVM are the "right" benchmarks. SRC and the Q-SVM were on par for the synthetic dataset, and Q-SVM was competitive on the real datasets. The strong learning ability exhibited by Q-SVM and the small testing times, make it a better competitor to SRC. A one-vs-rest version of Q-SVM scales linearly with c, and deserves attention.

Sparsity induced nonlinearity is key to SRC. SRC (resp. S-SRC) exhibited higher accuracy than the linear coding method CRC-RLS (resp. S-RLS). Others have reported similar findings [8]. The most important difference between SRC, CRC-RLS and Q-SVM is that SRC uses a sparsity induced, nonlinear coding of test samples. We posit that rather than just sparsity or collaborative representation, it is this attribute that is SRC's key characteristic. It gives SRC the potential to "localize" a testing sample on a nonlinear manifold of labelled prior examples.

6. REFERENCES

[1] John Wright, Allen Y. Yang, Arvind Ganesh, S. Shankar Sastry, and Yi Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, February 2009.

[2] J.M.K. Kua, E. Ambikairajah, J. Epps, and R. Togneri, "Speaker verification using sparse representation classification," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, 2011, pp. 4548–4551.

[3] Chun-Hou Zheng, D. Zhang, To-Yee Ng, S. C K Shiu, and De-Shuang Huang, "Metasample-based sparse representation for tumor classification," *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, vol. 8, no. 5, pp. 1273–1282, 2011.

[4] J. Wright, Yi Ma, J. Mairal, G. Sapiro, T.S. Huang, and Shuicheng Yan, "Sparse representation for computer vision and pattern recognition," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1031–1044, 2010.

[5] Xu Chen and P. J. Ramadge, "Music genre classification using multiscale scattering and sparse representations," in *Proc. Annual Conference on Information Sciences and Systems*, 2013.

[6] D. Zhang, Meng Yang, and Xiangchu Feng, "Sparse representation or collaborative representation: Which helps face recognition?," *Computer Vision (ICCV), 2011 IEEE International Conference on*, pp. 471–478, 2011.

[7] Meng Yang, D. Zhang, D. Zhang, and Shenlong Wang, "Relaxed collaborative representation for pattern classification," *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 2224–2231, 2012.

[8] R. Timofte and L. Van Gool, "Weighted collaborative representation and classification of images," *Pattern Recognition (ICPR), 2012 21st International Conference on*, pp. 1606–1610, 2012.

[9] Xi Li, Chunhua Shen, Qinfeng Shi, A. Dick, and A. van den Hengel, "Non-sparse linear representations for visual tracking with online reservoir metric learning," *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012.

[10] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, 2009.

[11] D.M. Malioutov, M. Cetin, and A.S. Willsky, "Homotopy continuation for sparse signal representation," *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, vol. 5, 2005.

[12] Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Y. Ng, "Efficient sparse coding algorithms," *NIPS*, 2007.

[13] L. El Ghaoui, V. Viallon, and T. Rabbani., "Safe feature elimination in sparse supervised learning," *Pacific Journal of Optimization*, 2012.

[14] Z. J. Xiang and P. J. Ramadge, "Fast lasso screening tests based on correlations," *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2012.

[15] Yun Wang, Zhen James Xiang, and Peter J. Ramadge, "Tradeoffs in improved screening of lasso problems," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 3297–3301.

[16] J. Jie Wang, B. Lin, P. Gong, P. Wonka, and J. Ye, "Lasso screening rules via dual polytope projection," *arXiv:1211.3966v1 [cs.LG]*, Nov. 2012.

[17] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, 2011.

[18] Thorsten Joachims, "Advances in kernel methods," chapter Making large-scale support vector machine learning practical, pp. 169–184. MIT Press, Cambridge, MA, USA, 1999.

[19] Y. LeCun, C. Cortes, and C.C.J. Burges, "The mnist database of handwritten digits," 1998.

[20] Chih-Wei Hsu and Chih-Jen Lin, "A comparison of methods for multiclass support vector machines," *Neural Networks, IEEE Transactions on*, vol. 13, no. 2, pp. 415–425, 2002.

[21] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech and Audio Processing*, vol. 10, no. 5, 2002.

[22] J. Andén and S. Mallat, "Multiscale scattering for audio classification," *Proceedings of the ISMIR 2011 Conference*, 2011.