# THE USE OF ISOMETRIC TRANSFORMATIONS AND BAYESIAN ESTIMATION IN COMPRESSIVE SENSING FOR FMRI CLASSIFICATION

*Avishy Carmi*[1]*, Tara N. Sainath*[2]*, Pini Gurfil*[3]*, Dimitri Kanevsky*[2]*,*
*David Nahamoo*[2]*and Bhuvana Ramabhadran*[2]

[1] The Signal Processing Group, Department of Engineering, University of Cambridge, UK
[2]IBM T. J. Watson Research Center, Yorktown Heights, NY 10598
[3] The Faculty of Aerospace Engineering, Technion, Haifa 32000, Israel
[1]ac599@cam.ac.uk,[2]tnsainat@us.ibm.com, [3]pgurfil@technion.ac.il,
[2]{kanevsky, nahamoo, bhuvana}@us.ibm.com

## ABSTRACT

Compressive sensing (CS) is a popular technique used to reconstruct a signal from few training examples, a problem which arises in many machine learning applications. In this paper, we introduce a technique to guarantee that our data obeys certain isometric properties. In addition, we introduce a bayesian approach to compressive sensing, which we call ABCS, allowing us to obtain complete statistics for estimated parameters. We apply these ideas to fMRI classification and find that by isometrically transforming our data, significant improvements in classification accuracy can be achieved using the LASSO and Dantzig selector methods, two standard techniques used in CS. In addition, applying the ABCS method offers improvements in classification accuracy over both LASSO and Dantzig. Finally, we find that applying both the ABCS method together with isometric transformations, we are able to achieve an error rate of $0.0\%$.

***Index Terms*—** Compressive sensing, sparse representation, bayesian learning, image classification

## 1. INTRODUCTION

In recent years, compressive sensing (CS) ([1]) has become a popular technique used to limit the amount of data required to reconstruct signals, also known as sparse signal recovery. This technique has become popular in many practical machine learning applications, such as biomedical image classification and signal reconstruction, and well as in applications to prevent over-fitting and reduce storage capacities. Mathematically speaking, in a typical CS formulation, a sensing matrix $H$ is constructed consisting of possible examples of the signal, that is $H = [h_1; h_2 \ldots; h_n]$. To reconstruct a signal $y$ from $H$, a small number of samples from $H$ are found by solving the equation $y = H\beta$, where the vector $\beta$ selects a small number of examples from $H$. In the CS formulation, a regularization is imposed on $\beta$, typically an $l_1$ regularization, to ensure sparseness. In summary, the goal of CS is to solve the following problem, where $\| \beta \|_1 < \epsilon$ imposes an $l_1$ regularization on $\beta$.

$$y = H\beta \ \ \texttt{s.t.} \quad \| \beta \|_1 < \epsilon \ \ \texttt{for} \ \ \beta \tag{1}$$

State-of-the-art methods for sparse signal recovery commonly utilize convex relaxation methods, including LASSO [2] and the Dantzig selector [3]. It has been shown in [1] that the convex $l_1$ relaxation yields an exact solution to the recovery problem provided two conditions are met: 1) the signal is sufficiently sparse, and 2) the set of samples used for signal recovery obeys a Restricted isometry property (RIP) to facilitate accurate recovery of sparse signals. However, to date the LASSO and Dantzig selector techniques have been applied to many CS problems without guaranteeing that the set of samples obeys the RIP property ([4],[5]). In addition, while these convex relation methods have shown success in a variety of compressed sensing problems, there is a considerable amount of effort required to tune the sparseness constraint. Moreover, these methods only provide a point estimate for $\beta$, and can thus be considered suboptimal solutions. In this paper, we introduce two novel techniques to deal with the isometry and $\beta$ estimation problems in typical CS optimization techniques.

First, we introduce a novel method to transform $H$ to be isometric. Specifically, by randomly sampling samples from a binary distribution, we construct a matrix $P$ which has been shown to obey isometric properties [6]. Next, we find a transformation $T$ such that applying this transformation to $H$ produces $P$ (i.e., $P = TH$). This technique allows us to transform our data to an isometric space in which we will demonstrate that CS algorithms are more effective.

In addition, to address the $\beta$ estimation problems in CS, a non-convex optimization method known as Bayesian CS [7] has been introduced, which has the advantage of introducing a probabilistic framework to estimate the spareness parameters required for signal recovery. This technique limits the effort required to tune the sparseness constraint, and also provides complete statistics for the estimate of $\beta$. However, the approach introduced in [7] suffers from the fact that it is difficult to obtain a closed-form probability expression to estimate these parameters. In this paper, we introduce an approximation method into the Bayesian CS formulation which allows us to derive a closed form probability expression. Specifically, instead of utilizing a typical Laplacian sparseness-promoting prior (i.e., $l_1$ as used in [7] and given in Equation 1), we explore a semi-Gaussian type prior. Our motivation for this is twofold. First, [8] and [9] illustrate that a semi-gaussian prior gives more weight to good samples in $H$ compared to a Laplacian prior, will still promoting more sparseness than a gaussian prior. Secondly, incorporating a semi-gaussian prior into a Bayesian framework facilitates the derivation of a closed-form recursion for estimating sparseness parameters, which is difficult to do using Laplacian priors [7]. We will call this new technique Approximate Bayesian Compressive Sensing (ABCS).

We explore the benefits of the isometric transformation and ABCS techniques on fMRI classification [10]. First, we explore the classification accuracy of the LASSO and Dantzig selector methods,

with and without the isometric transformation, and find that using isometric transformation offers a $3-59\%$ relative improvement over classification accuracy (depending on the data set used). Second, we compare the behavior of the ABCS classifier to the Dantzig selector and LASSO classifiers, and find that the ABCS technique offers and improvement over the LASSO classifier. The ABCS technique also outperforms LASSO when isometric transformations are applied to the data, and in fact is able to offer a $0.0\%$ error rate.

The rest of this paper is organized as follows. Section 2 presents our novel formulation of isometric transformation for compressive sensing, which Section 3 discusses the ABCS formulation. Section 4 presents the experiments performed, followed by a discussion of the results in Section 5. Finally, Section 6 concludes the paper and discusses future work.

## 2. ISOMETRIC TRANSFORMATIONS

The theory of CS requires that in order for a sparse solution for $\beta$ [6], the sensing matrix $H$ should obey a so-called restricted isometry property (RIP) at a certain level. In detail, the RIP is defined as

$$(1-\delta_s) \parallel x \parallel_2^2 \leq \parallel Hx \parallel_2^2 \leq (1+\delta_s) \parallel x \parallel_2^2 \tag{2}$$

for some $\delta_s \in (0,1)$ and any $x$ that is $s$-sparse at most. In [11], we introduced the definition of a vector $x = [x_1 x_2 \dots x_d]$ to be $s$-block-sparse if its non-zero entries are concentrated in blocks of dimension $s$. In other words, $x$ is $s$-block sparse if $\#\{x_i \mid \ x_i \neq 0, \ i = 1, \dots, d\} << d$. Equation 2 infers that every subset of $H$ of dimension $m \times s$ acts as nearly orthonormal system. The RIP constant $\delta_s$ gives an indication of the actual proximity of any subset to orthogonality. In this section, given that our original matrix $H$ does not satisfy the RIP property, we present a solution to producing an RIP-satisfying data matrix out of the original matrix $H$.

The isometric transformation relies on the existence of some RIP-satisfying matrix of the same dimension as the original data set. Constructing such a matrix is generally a non-trivial task. Nevertheless, it is well known fact that some random matrices obey the RIP with high probability [6]. For example, it has been shown that a binary matrix is isometric if the entries are sampled according to:

$$\Pr\left(P_{ij} = \pm 1/\sqrt{m}\right) = 0.5 \tag{3}$$

Thus, given a se of Binary samples, an isomatric random matrix in constructed, which we will refer to as $P$. Here $P$ has the same dimension as $H$. Next, the columns of $H$ are reordered according to their measure of significance. In this work, the columns are reordered by computing a correlation coefficient between $y$ an the columns in $H$, described in more detail here [11]. As we will discuss in the next step, the columns of $H$ are split into smaller subsets and a transformation is applied to each of these smaller subsets to obtain an isometric matrix. The reordering of columns is done to ensure that when the columns of $H$ are split into smaller subsets, and a $\beta$ is obtained for each subset, there is a higher chance that this $\beta$ will be sparse.

Next, the matrices $H$ and $P$ are partitioned into $d$ smaller subsets, namely $H = [H_1, \dots, H_d]$ and $P = [P_1, \dots, P_d]$, where each subset is denoted by $H_i$ and $P_i$ with $i \in d$. For each $H_i$ and $P_i$, subset, we define a transformation matrix $\hat{T}_i = C_i D_i^T$ for all $i = 1, \dots, d$ where $C_i \Lambda_i D_i^T$ is the Singular Value Decomposition (SVD) of $H_i^{-1} P_i$. Here $\Lambda_i = \text{diag}(\lambda_i^1, \dots, \lambda_i^m)$, and $\lambda_i^1 \geq \lambda_i^2 \geq \dots \geq \lambda_i^m$. In [11], it can be shown that $H\hat{T}_i$ is an isometric matrix with the following property for some scalars $\alpha > 0$ and $\delta \in (0,1)$.

$$(1-\delta) \parallel x \parallel_2^2 \leq \parallel \alpha H\hat{T}x \parallel_2^2 \leq (1+\delta) \parallel x \parallel_2^2$$

Letting $\lambda_{\max} = \arg\max_{i \in [1,d]} \lambda_i^1$ and $\lambda_{\min} = \arg\min_{i \in [1,d]} \lambda_i^n$, the scaling factor $\alpha$ is approximated by:

$$\alpha \approx 2\left(\lambda_{\min}^{-1} + \lambda_{\max}^{-1}\right)^{-1}$$

and the RIP constant $\delta$ is given as

$$\delta = \frac{(1+\delta_m)\lambda_{\min}^{-1} - (1-\delta_m)\lambda_{\max}^{-1}}{(1+\delta_m)\lambda_{\min}^{-1} + (1-\delta_m)\lambda_{\max}^{-1}} < 1$$

where $\delta_m \in (0,1)$ is the RIP constant associated with $P$.

Finally, the transformed matrix $\bar{H}$ is defined as $\bar{H} = \alpha\left[H_1\hat{T}_1, \dots, H_d\hat{T}_d\right]$.

## 3. ORIGINAL ABCS DERIVATION

In this section, we formulate the ABCS solution. Ultimately, we would like to use CS to solve the following problem:

$$y = H\beta \ \ \texttt{s.t.} \quad \parallel \beta \parallel_1^2 < \epsilon \ \texttt{for} \ \beta \tag{4}$$

In ABCS we use a sparseness promoting semi-gaussian prior, denoted as $\parallel \beta \parallel_1^2 < \epsilon$, rather than the Laplacian prior (i.e., $\parallel \beta \parallel_1$) as given by Equation 1. In addition, $y$ is a frame of data from the test set such that $y \in \Re^m$ where $m$ is the dimension of the feature vector $y$. $H$ is a matrix of training examples and $H \in \Re^{m \times n}$ where $m << n$. We assume that $y$ satisfies a linear model as: $y = H\beta + \zeta$ where $\zeta \sim N(0, R)$. This allows us to represent $p(y|\beta)$ as a Gaussian distribution as:

$$p(y|\beta) \propto exp(-1/2(y - H\beta)^T R^{-1}(y - H\beta)) \tag{5}$$

Assuming $\beta$ is a random parameter with some prior $p(\beta)$ we can obtain the maximum a posteriori (MAP) estimate for $\beta$ given $y$ as follows: $\beta^* = \arg\max_\beta p(\beta|y) = \max_\beta p(y|\beta)p(\beta)$. In the ABCS formulation, we assume that $p(\beta)$ is actual the product of two prior constraints, namely a gaussian constraint $p_G(\beta)$ and a semi-gaussian constrain $p_{SG}(\beta)$ to enforce sparseness. Below, we present a two-step solution to solve the following problem in the ABCS framework.

$$\beta^* = \arg\max_\beta p(y|\beta)p_G(\beta)p_{SG}(\beta) \tag{6}$$

### 3.1. Step 1

In step 1, we solve for the $\beta$ which maximizes the following expression. Equation 7 is equivalent to solving the equation $y = H\beta$ without enforcing a sparseness constraint on $\beta$ [8].

$$\beta^* = \arg\max_\beta p(y|\beta)p_G(\beta) \tag{7}$$

We assume that $p_G(\beta)$ is a Gaussian, i.e., $p_G(\beta) = N(\beta|\beta_0, P_0)$. Here $\beta_0$ and $P_0$ are initialized statistical moments utilized in the algorithm. In [8], we show that the solution to Equation 7 has a closed form solution given by Equation 8.

$$\beta^* = \beta_1 = \left(I - P_0 H^T (HP_0 H^T + R)^{-1} H\right)\beta_0 + P_0 H^T (HP_0 H^T + R)^{-1} y \tag{8a}$$

Similarly, we can express the variance of $\beta_1$ as $P_1 = E\left[(\beta - \beta^1)(\beta - \beta^1)^T\right]$, given more explicitly by Equation 8b.

$$P_1 = (I - P_0 H^T (HP_0 H^T + R)^{-1} H)P_0 \tag{8b}$$

### 3.2. Step 2

Step 1 essentially solved for the pseudo-inverse of $y = H\beta$, of which there are many solutions. In this section, we impose an additional constraint that $\beta$ will have a sparseness-promoting semi-Gaussian prior, as given by Equation 9. Here $\sigma^2$ is a constant parameter which controls the degree of sparseness of $\beta$.

$$p_{SG}(\beta) = exp\left(-\frac{||\beta||_1^2}{2\sigma^2}\right) \tag{9}$$

Given the solutions to Step 1 in Equations 8, we can simply rewrite Equation 7 as another gaussian as $p'(\beta|y) = p(y|\beta)p_G(\beta) = N(\beta|\beta_1, P_1)$. Therefore, let us assume now that we would like to solve for the MAP estimate of $\beta$ given the constraint that it is semi-gaussian, in other words:

$$\beta^* = \arg\max_{\beta} p'(\beta|y)p_{SG}(\beta) \tag{10}$$

In order to represent $p_{SG}(\beta)$ as a Gaussian the same way that $p(y|\beta)$ in Equation 5 was represented, let us define $\beta^i$ to be the $i^{th}$ entry of the vector $\beta$. We introduce a matrix $\hat{H}$ of which the entries are set as $\hat{H}^i(\beta^i) = \text{sign}(\beta^i)$, for $i = 1, \ldots, n$. Here $\hat{H}^i(\beta^i) = +1$ for $\beta^i > 0$, $\hat{H}^i(\beta^i) = -1$ for $\beta^i < 0$, and $\hat{H}^i(\beta^i) = 0$ for $\beta^i = 0$. This matrix $\hat{H}$ is motivated from the fact that

$$\| \beta \|_1^2 = (\sum_i (|\beta^i|))^2 = (\sum_i (\hat{H}^i(\beta^i)\beta^i))^2 = (\hat{H}\beta)^2 \tag{11}$$

Substituting the expression for $\| \beta \|_1^2$ given in Equation 11 and assuming a that $y = 0$, we can rewrite Equation 9 as Equation 12. Notice that Equation 12 has the same form as Equation 5 with $H$ and $R$ now replaced by $\hat{H}$ and $\sigma$ respectively.

$$p_{SG}(\beta) = p(y = 0|\beta) = exp\left(\frac{-(0 - \hat{H}\beta)^2}{2\sigma^2}\right) \tag{12}$$

The only problem with using Equation 10 to solve for $\beta$ is the dependency of $\hat{H}$ on $\beta$ in Equation 8. Therefore, we make an assumption, by calculating $\hat{H}$ based on the sign of the previously estimated $\beta$. In other words $\hat{H}^i(\beta^i) \approx \hat{H}^i(\beta^i_{k-1})$. With this approximation we can use Equations 8a and 8b to solve Equation 12. However, because of this semi-gaussian approximation, we must estimate $\beta$ and $P$ iteratively. Equation 13 gives the recursive formula which solves Equation 10 at iteration $k$ for $k > 1$. Note that $p'(\beta|y) = N(\beta|\beta_{k-1}, P_{k-1})$, where for $k = 2$, $\beta$ and $P$ are computed in Step 1.

$$\beta_k = \beta_{k-1} - \frac{P_{k-1}\hat{H}^T}{\hat{H}P_{k-1}\hat{H}^T + \sigma^2}\hat{H}\beta_{k-1} \tag{13a}$$

$$P_k = \left[I - \frac{P_{k-1}\hat{H}^T}{\hat{H}P_{k-1}\hat{H}^T + \sigma^2}\right]P_{k-1} \tag{13b}$$

In [8], we show that for large $\sigma^2$ and large $k$, the estimate of $\beta$ and $P$ using the approximate semi-gaussian given in Equation 12 is bounded from the estimate of these parameters for the true semi-gaussian given in Equation 9 by $O(1/\sigma^2)$.

## 4. EXPERIMENTS

We analyze the benefit of isometric transformations and ABCS for fMRI classification. The fMRI data sets are those that were used in [10]. The data consists of a series of trials in which the subject is being shown either a picture $(+1)$ or a sentence $(-1)$. The brain activity is monitored over a time interval of 9 seconds during which a fMRI scan is performed every 1 second. We have used this set-up for producing two data sets for each subject. The first set, which we termed 'sliced', consists of the 1st scan in each trial whereas the second one involves the average of 6 fMRI scans (from 1 to 6). The resulting data sets consist of nearly 2000 features, and 40 relevant samples [10]. We use a 2-out cross-validation scheme for testing the underlying classifiers. This procedure involves 20 trials in which 2 samples (one of each class) are taken as a testing set while the remaining samples are used for training. The classifiers are applied using 12 data sets, these account for 6 sets (sliced and averaged for each of the 3 subjects) and their transformed versions.

Next to demonstrate the behavior of ABCS and isometric transformations in a multi-class classification problem, we conduct experiments on the fMRI data set described here [12]. This fMRI data set consists of 20 samples of a subject viewing 8 types of images which are labeled as 1 to 8 (i.e., total of 160 fMRI scans). The experimental setup uses 8-out cross-validation in which 8 samples, one of each label, are taken as a test set while the remaining samples are used for training the classifiers.

To reduce the number of 2000 features, a random field (RF) model is applied to select an optimal subset of these features. This method is described in more detail here [8]. In all tests a binary random matrix $P$ (see Section 2) was used for producing the transformed set. The realization of $P$ is chosen as the one that yields the best classification accuracy when applying the above procedure on a predetermined development set which in our case was composed out of 20 samples from the training set.

All classifiers were coded in Matlab's environment. The Dantzig selector implementation uses the built-in function 'linprog' that is based on a linear interior point solver. The LASSO classifier uses the MATLAB implementation of LARS algorithm [13] to solve the linear regression problem with the class label treated as a real-valued response variable; in order to obtain binary prediction on a test sample, we simply threshold the output of LASSO model (i.e., predict +1 if the output is positive and -1 otherwise).

First, we explore the classification accuracy of CS classifiers, namely Dantzig selector and LASSO, with and without isometric transformations. Please note that only for this implementation of LASSO, feature selection was not performed using the RF model. Instead the LASSO classifier approach runs on all features, and selects a desired number of variables (that is specified as an input parameter to LARS procedure) automatically. Second, we compare the classification of the ABCS method to the LASSO and Dantzig selector methods, without applying isometric transformations, for various number of optimal subset parameters selected using the RF model from the 2000 feature dimensional set. Finally, we explore the behavior of LASSO and ABCS techniques with isometric transformations, for various numbers of features. In our work, we choose all elements of $\beta_0$ to be 0 since $\beta$ is assumed to be sparse around 0 anyways. $P_0$ is chosen to be diagonal with elements having a large value of 100, reflecting the fact that the prior is uninformative.

## 5. RESULTS

### 5.1. Classification with Isometric Transformations

Tables 1 and 2 illustrate the classification results for the LASSO and Dantzig selector methods on the sliced and averaged data sets, with and without isometric transformations. The results for each data set are averaged across all three subjects. Notice that for both techniques, the isometric transformations offer improvements in accuracy, with relative improvements ranging from $3\%$ to $59\%$ depending on the data set. This demonstrates that transforming the data to be isometric provides for better signal recovery (and thus overall classification accuracy), an idea which was only theoretically proven in [6] but illustrated by application in Tables 1 and 2 .

| | Sliced Data set | | |
|---|---|---|---|
| Method | Original | **Transf.** | Rel. Imp. |
| RF-Dantzig | 0.52 | **0.83** | 0.59 |
| LASSO(100 vars, ttest) | 0.73 | **0.82** | .12 |

**Table 1**. Classification accuracy on Sliced Data Sets, averaged across all 3 subjects.

| | Averaged Data Set | | |
|---|---|---|---|
| Method | Original | **Transf.** | Rel. Imp. |
| RF-Dantzig | 0.76 | **0.90** | .18 |
| LASSO(100 vars, ttest) | 0.90 | **0.93** | .03 |

**Table 2**. Classification accuracy on Averaged Data Sets, averaged across all 3 subjects.

### 5.2. Classification with ABCS

To analyze the behavior of the ABCS method, Figure 1 compares the classification error rate for the LASSO and ABCS methods, with and without isometric transformations for different optimal number of features. Note that we have not included the Dantzig selector in this analysis, since Section 5.1 illustrated that LASSO outperformed the Dantzig selector with and without isometric transformations. First, notice that without isometric transformations the ABCS method outperforms the LASSO method when the number of features is above 40. This illustrates that by using ABCS to obtain better estimates of $\beta$, the classification accuracy is improved over LASSO.

The figure also illustrates that similar trends hold when isometric transformations are applied. Using these transformations, the ABCS method and LASSO method perform similarly when the number of features is less than 20. However, as the number of features grows, the ABCS technique offers better performance compared the LASSO method. In fact, the ABCS method is able to achieve $0.0\%$ error rate when using more than 80 features.

## 6. CONCLUSIONS AND FUTURE WORK

In this paper, we introduced a technique to guarantee that the sensing matrix $H$ was isometric. Applying this transformation to fMRI classification, we found that having isometrically transformed data offered improvements in classification accuracy for both the LASSO and Dantzig selector techniques. In addition, we developed an ABCS technique which utilizes a semi-gaussian prior to obtain complete statistics on the estimate of $\beta$. We found that this ABCS method outperformed the popular LASSO method, with and without isometric transformations. In fact, the ABCS technique with isometric transformations was able to achieve a $0.0\%$ error rate. In the future, we would like to perform better convergence assessments for

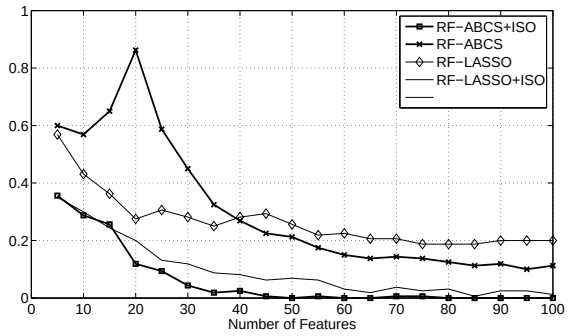


**Fig. 1**. Classification Errors for Different CS Methods

Step 2 of the ABCS technique, which will allow us to have a better idea of when to stop iterating Step 2, rather than choosing a large fixed number. In addition, we would like to explore applying the ABCS method to other machine learning tasks, where CS has been relatively unexplored.

## 7. REFERENCES


[1] D. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, pp. 1289–1306, 2006.

[2] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[3] E. Candes and T. Tao, "The dantzig selector: statistical estimation when p is much larger than n," *Annals of Statistics*, vol. 35, pp. 2313–2351, 2007.

[4] D. Gosh and A. M. Chinnaiyan, "Classification and Selection of Biomarkers in Genomic Data Using LASSO," *Journal of Biomedicine and Biotechnology*, vol. 2, pp. 147–154, 2005.

[5] B. A. Johnson, "Fast Restoration Dantzig Selection for Censored Data," Tech. Rep., Department of Biostatistics and Bioinformatics, Emory University, 2009.

[6] E. J. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, pp. 489–509, 2006.

[7] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Transactions on Signal Processing*, vol. 56, pp. 2346–2356, 2008.

[8] A. Carmi, P. Gurfil, D. Kanevsky, and B. Ramabhadran, "ABCS: Approximate Bayesian Compressed Sensing," Tech. Rep., Human Language Technologies, IBM, 2009.

[9] A. Carmi, P. Gurfil, and D. Kanevsky, "Methods for Sparse Signal Recovery Using Kalman Filtering with Embedded Pseudo-Measurement Norms and Quasi-Norms," *IEEE Transactions on Signal Processing*, vol. 58, no. 1, 2010.

[10] T. Mitchell, R. Hutchinson, R. Niculescu, F. Pereira, X. Wang, M. Just, and S. Newman, "Learning to Decode Cognitive States from Brain Images," *Machine Learning*, vol. 57, pp. 145–175, 2004.

[11] A. Carmi, I. Rish, G. Cecchi, D. Kanevsky, and B. Ramabhadran, "Isometry-enforcing Data Transformations for Improving Sparse Model Learning," Tech. Rep. RC 24801, Human Language Technologies, IBM, 2009.

[12] J. V. Haxby, M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, and P Pietrini, "Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex," *Science*, vol. 293, pp. 2425–2430, 2001.

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