RC24816 (W0906-115) June 25, 2009
Computer Science

IBM Research Report

ABCS: Approximate Bayesian Compressed Sensing

Avishy Carmy
The Signal Processing Group
Department of Engineering
University of Cambridge
UK

Pini Gurfil
The Faculty of Aerospace Engineering
Technion
Haifa 32000
Israel

Dimitri Kanevsky, Bhuvana Ramabhadran
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

ABCS: APPROXIMATE BAYESIAN COMPRESSED SENSING

Avishy Carmi[1], *Pini Gurfil*[2], *Dimitri Kanevsky*[3], *Bhuvana Rambahadran*[3]

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

ABSTRACT

In this work we present a new approximate Bayesian compressed sensing scheme. The new method is based on a unique type of sparseness-promoting prior, termed here semi-Gaussian owing to its Gaussian-like formulation. The semi-Gaussian prior facilitates the derivation of a closed-form recursion for solving the noisy compressed sensing problem. As part of this, the discrepancy between the exact and the approximate posterior pdf is shown to be of the order of a quantity that is computed online by the new scheme. In the second part of this work, a random field-based classifier utilizing the approximate Bayesian CS scheme is shown to attain a zero error rate when applied to fMRI classification.

1. INTRODUCTION

Recent studies have shown that sparse signals can be recovered accurately using less observations than what is considered necessary by the Nyquist/Shannon sampling principle; the emergent theory that brought this insight into being is known as compressed sensing (CS) [1–3]. The essence of the new theory builds upon a new data acquisition formalism, in which compression plays a fundamental role. From a signal processing standpoint, one can think about a procedure in which signal recovery and compression are carried out simultaneously, thereby reducing the amount of required observations. Sparse, and more generally, compressible signals arise naturally in many fields of science and engineering. A typical example is the reconstruction of images from under-sampled Fourier data as encountered in radiology, biomedical imaging and astronomy [4, 5]. Other applications consider model-reduction methods to enforce sparseness for preventing over-fitting and for reducing computational complexity and storage capacities. The reader is referred to the seminal work reported in [3] and [2] for an extensive overview of the CS theory.

The recovery of sparse signals is in general NP-hard [1, 6]. State-of-the-art methods for addressing this optimization problem commonly utilize convex relaxations, non-convex local optimization and greedy search mechanisms. Convex relaxations are used in various methods such as LASSO [7], the Dantzig selector [8], basis pursuit and basis pursuit de-noising [9], and least angle regression [10]. Non-convex optimization approaches include Bayesian methodologies such as the relevance vector machine, otherwise known as sparse Bayesian learning [11], as well as stochastic search algorithms that are mainly based on Markov chain Monte Carlo techniques [12–15]. Notable greedy search algorithms are the matching pursuit (MP) [16], the orthogonal MP [17], and the orthogonal least squares [18].

CS theory has drawn much attention to the convex relaxation methods. It has been shown 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 sensing matrix obeys the so-called restricted isometry property (RIP) at a certain level. A complementary result guarantees high accuracy when dealing with noisy observations, yielding recovery 'with overwhelming probability'. To put it informally, it is likely for the convex l_1 relaxation to yield an exact solution provided that the involved quantities, the sparseness degree s, and the sensing matrix dimensions $m \times n$ maintain relation of the type $s = \mathcal{O}(m/\log(n/m))$.

Recently, a Bayesian CS approach has been introduced in [19]. As opposed to the conventional non-Bayesian methods, the Bayesian CS has the advantage of providing the complete statistics of the estimate in the form of a posterior probability density function (pdf). Adopting this approach, however, suffers from the fact that rarely one can obtain a closed-form expression of the posterior and therefore approximation methods should be utilized.

In this work we present a new approximate Bayesian CS scheme. The new method is based on a unique type of sparseness-promoting prior, termed here semi-Gaussian owing to its Gaussian-like formulation. The semi-Gaussian prior facilitates the derivation of a closed-form recursion for solving the noisy compressed sensing problem. As part of this, the discrepancy between the exact and the approximate posterior pdf is shown to be of the order of a quantity that is computed online by the new scheme. In the second part of this work, a random field-based classifier utilizing the approximate Bayesian CS scheme is shown to attain a **zero error rate** when applied to fMRI classification.

2. A NEW COMPRESSED SENSING ALGORITHM

This section derives the approximate Bayesian CS (ABCS) method. The key idea behind the new algorithm is based on an approximate sparseness-promoting prior, which is a mixture of Gaussian and Laplace distributions. In what follows, we gradually develop this concept.

2.1. Bayesian Estimation

The Bayesian estimation methodology provides a convenient representation for dealing with complex observation models. In this work, however, we restrict ourselves to the conventional linear model used in CS theory,

$$y_k = H\beta + n_k \tag{1}$$

where y_k, $H \in \mathbb{R}^{m\times n}$ and n_k denote the kth $\mathbb{R}^m$-valued observation, a fixed sensing matrix, and the observation noise with a known pdf $p(n_k)$, respectively. The sought-after estimator of the random parameter (the signal) β is a $\mathbb{R}^n$-valued vector of which the prior pdf $p(\beta)$ is given. Following this, the complete statistics of β conditioned on the entire observation set $\mathscr{Y}_k := \{y_1, \ldots, y_k\}$ can be sequentially computed via the Bayesian recursion

$$p(\beta \mid \mathscr{Y}_k) = \frac{p(y_k \mid \beta)p(\beta \mid \mathscr{Y}_{k-1})}{\int p(y_k \mid \beta)p(\beta \mid \mathscr{Y}_{k-1})d\beta} \tag{2}$$

where the likelihood $p(y_k \mid \beta) = p_{n_k}(y_k - H\beta)$. Unfortunately, rarely can one obtain a closed-form analytic expression of the posterior pdf (2) and therefore approximation techniques are often utilized. One well-known example in which (2) does admit a closed-form solution is given by the following well-known theorem from estimation theory.

Theorem 1 (Gaussian pdf Update) *Assume that $p(\beta \mid \mathscr{Y}_{k-1})$ is a Gaussian pdf of which the first two statistical moments are given by $\hat{\beta}_{k-1} \in \mathbb{R}^n$ and $P_{k-1} \in \mathbb{R}^{n\times n}$, that is $p(\beta \mid \mathscr{Y}_{k-1}) = \mathcal{N}(\beta \mid \hat{\beta}_{k-1}, P_{k-1})$. Assume also that the observation y_k satisfies the linear model* (1) *where n_k is a $\mathbb{R}^m$-valued zero-mean*

Gaussian random variable $n_k \sim \mathcal{N}(0, R)$ that is statistically independent of β. Then the Bayesian recursion (2) *yields $p(\beta \mid \mathscr{Y}_k) = \mathcal{N}(\beta \mid \hat{\beta}_k, P_k)$ where*

$$\hat{\beta}_k = \hat{\beta}_{k-1} + P_{k-1}H^T \left(HP_{k-1}H^T + R\right)^{-1} \left[y_k - H\hat{\beta}_{k-1}\right] \tag{3a}$$

$$P_k = \left[I - P_{k-1}H^T \left(HP_{k-1}H^T + R\right)^{-1}\right] P_{k-1} \tag{3b}$$

The initial values of the above quantities are set according to the Gaussian prior $p(\beta) = \mathcal{N}(\beta \mid \hat{\beta}_0, P_0)$.

The proof is provided in the Appendix. Note that the quantity P_k in Theorem 1 is the estimation error covariance, i.e., $P_k = E\left[(\beta - \hat{\beta}_k)(\beta - \hat{\beta}_k)^T \mid \mathscr{Y}_k\right]$, where $\beta - \hat{\beta}_k$ is the estimation error of the unbiased estimator $\hat{\beta}_k$. In addition, under the restrictions of Theorem 1, $\hat{\beta}_k$ is the maximum *a posteriori* (MAP) estimator, i.e., $\hat{\beta}_k = \arg\max_\beta \log p(\beta \mid \mathscr{Y}_k) = \arg\min_\beta \sum_{i=1}^k \parallel y_i - H\beta \parallel_R^2 + \parallel \beta - \beta_0 \parallel_{P_0}^2$ where $\parallel a \parallel_R^2 := a^T R^{-1} a$.

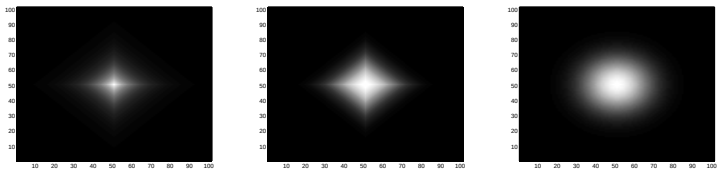
2.2. Sparseness-Promoting Semi-Gaussian Priors

Compressed sensing was embedded in the framework of Bayesian estimation by utilizing sparseness-promoting priors [19] such as the Laplace and Cauchy distributions. As opposed to the conventional CS methods, which provide a point estimate, the Bayesian approach yields the complete statistics $p(\beta \mid \mathscr{Y}_k)$ of which an exact expression is given by Theorem 1 for the linear Gaussian observation model. When resorting to non-Gaussian priors, however, the conditions of Theorem 1 are violated, which renders the recursion (3) inadequate. For instance, when using a Laplace prior, the sought-after estimator is the one that solves a problem of the form $\min_\beta \sum_{i=1}^k \parallel y_i - H\beta \parallel_R^2 + \lambda \parallel \beta \parallel_1$.

In this work, we consider a new type of prior, which facilitates the application of the closed-form recursion of Theorem 1. This sparseness-promoting prior is termed *semi-Gaussian*, and is given by

$$p(\beta) = c \exp\left(-\frac{1}{2}\frac{\parallel \beta \parallel_1^2}{\sigma^2}\right) \tag{4}$$

Comparing to the Laplace distribution, the semi-Gaussian distribution possesses greater concentrations in the vicinity of the origin. This is further illustrated in Fig. 1, in which the level maps are shown for Laplace, semi-Gaussian and Gaussian pdf's in the 2-dimensional case. The embedding of the prior (4) within the



(a) Laplace **(b) Semi-Gaussian** **(c) Gaussian**

Fig. 1: Laplace, semi-Gaussian and Gaussian pdf's in the 2-dimensional case.

Gaussian variant of the Bayesian recursion in Theorem 1 is not straightforward. This follows from the fact

that the restrictions under which Theorem 1 is derived involve a purely-Gaussian prior and a likelihood pdf that is based on a deterministic sensing matrix H,

$$p(y_k \mid \beta) \propto \exp\left(-\frac{1}{2}(y_k - H\beta)^T R^{-1}(y_k - H\beta)\right) \tag{5}$$

Theorem 1 provides an exact recursion for computing the Gaussian posterior based exclusively on the factors composing the above likelihood: the observation y_k, the sensing matrix H and the observation noise covariance R. This fact has motivated the following approach, which allows enforcing an approximate semi-Gaussian prior without changing the fundamental structure of the underlying update equations as obtained in Theorem 1.

2.3. Approximate Semi-Gaussian Prior

We introduce a state-dependent matrix $\hat{H} \in \mathbb{R}^{1\times n}$ of which the entries are set as $\hat{H}^i = \operatorname{sign}(\beta^i)$, $i = 1,\ldots,n$ (i.e., $\hat{H}^i = +1$ and $\hat{H}^i = -1$ for $\beta^i > 0$ and $\beta^i < 0$, respectively). Now, the semi-Gaussian prior can be expressed based on (5) while replacing H and R with $\hat{H}$ and σ, respectively, and assuming a fictitious observation $y = 0$, that is

$$p(\beta) = p(y = 0 \mid \beta, \hat{H}, \sigma) \propto \exp\left(-\frac{1}{2}\frac{(0 - \hat{H}\beta)^2}{\sigma^2}\right) \tag{6}$$

At this point, the only difficulty preventing us from using (3) for enforcing the semi-Gaussian prior (6) is the dependency of $\hat{H}$ upon β. We recall that Theorem 1 relies on a possibly varying – albeit deterministic – H, as opposed to the formulation in (6). This problem can be alleviated by letting

$$\hat{H}^i = \operatorname{sign}(\hat{\beta}_k^i), \quad i = 1,\ldots,n \tag{7}$$

i.e., by substituting the conditional mean instead of the actual β. This modification renders $\hat{H}$ a $\mathscr{Y}_k$-measurable quantity, as it depends on $\hat{\beta}_k$, which is a function of the entire observation set. This fact does not affect the expressions in Theorem 1 as the derivations are conditioned on $\mathscr{Y}_k$ (see Appendix). Applying this approximation facilitates the implementation of Theorem 1 based on the likelihood (6). Hence, an additional processing stage is obtained as

$$\hat{\beta}_{k+1} = \hat{\beta}_k - \frac{P_k\hat{H}^T}{\hat{H}P_k\hat{H}^T + \sigma^2}\hat{H}\hat{\beta}_k \tag{8a}$$

$$\hat{P}_{k+1} = \left[I - \frac{P_k\hat{H}^T}{\hat{H}P_k\hat{H}^T + \sigma^2}\right]P_k \tag{8b}$$

The above CS stage is implemented after the usual processing of the observations set $\mathscr{Y}_k$ (see (3)) where the initial covariance is taken as $P_0 \to \infty$.

At this point a natural question is raised concerning the validity of the approximation suggested above. The following theorem bounds the discrepancy between the exact posterior, which uses the semi-Gaussian prior (4), and the approximate posterior in terms of the estimation error covariance $\hat{P}_k$.

Theorem 2 *Denote $\hat{p}(\beta \mid \mathscr{Y}_k)$ the Gaussian posterior pdf obtained by using the approximate semi-Gaussian prior technique, and let $p(\beta \mid \mathscr{Y}_k)$ be the posterior pdf obtained by using the exact semi-Gaussian prior* (4)*. Then*

$$\mathrm{KL}\left(\hat{p}(\beta \mid \mathscr{Y}_k) \parallel p(\beta \mid \mathscr{Y}_k)\right) = \mathcal{O}\left(\sigma^{-2}\max\left\{\mathrm{Tr}(\hat{P}_k), \mathrm{Tr}(\hat{P}_k)^{1/2}\right\}\right) \tag{9}$$

where KL *and* Tr *denote the Kullback-Leibler divergence and the matrix trace operator, respectively.*

The proof is provided in the Appendix.

2.4. Discussion

The fundamental observation conveyed by Theorem 2 is that the approximation error of the new scheme in (3) and (8) is affected by both the prior variance σ^2 and the estimation error covariance $\hat{P}_k$. Consequently, regulating these factors is beneficial in getting close to the exact CS solution in the Bayesian sense. This can be attained by either increasing σ or having a sufficiently small $\hat{P}_k$. The former approach, however, might bring upon an adverse effect as the sparseness constraint is less restrictive in such cases.

A prominent advantage of the bound (9) is that it involves quantities that are either known (σ^2) or computed ($\hat{P}_k$). Whereas σ is a predetermined tuning parameter, the estimation error covariance $\hat{P}_k$, which is computed at every step, decreases rapidly as $k \rightarrow \infty$ (in the sense that $\hat{P}_{k+1} < \hat{P}_k$). This fact can be easily verified by recognizing that (3b) and (8b) translate into $P_k^{-1} = P_{k-1}^{-1} + H^T R^{-1} H$ and $\hat{P}_{k+1}^{-1} = P_k^{-1} + \sigma^{-2} \hat{H}^T \hat{H}$ owing to the matrix inversion lemma. This observation along with Theorem 2 further imply that it is advantageous to perform d consecutive updates of (8) using $d\sigma^2$, $d > 1$ rather than a single update with σ^2. This approach, motivated by the equivalence relation (see (6)) $\prod_{k=1}^{d} \exp\left(-\frac{1}{2d\sigma^2}(\hat{H}\beta)^2\right) = \exp\left(-\frac{1}{2\sigma^2}(\hat{H}\beta)^2\right)$ relieves the conservative bound (9), as $\hat{P}_k$ decreases after each update in which the prior variance is larger than σ^2.

The suggested ABCS algorithm is summarized below.

Algorithm 1 ABCS

1: Obtain $\hat{\beta}_k$ and P_k using (3).
2: *CS stage:* Let $\hat{P}_k = P_k$.
3: **for** $\tau = k+1, \ldots, k+d$ iterations **do**
4: Compute $\hat{\beta}_\tau$ and $\hat{P}_\tau$ using (7) and (8) with $d\sigma^2$ as the prior variance.
5: **end for**

It should be clarified that $\hat{P}_k$ does not represent the exact estimation error covariance, which is based on the semi-Gaussian prior (4). The only purpose of $\hat{P}_k$, stated by Theorem 2, is to indicate the proximity of the obtained solution (3), (8) to the exact Bayesian one. An upper bound on the exact estimation error covariance can be obtained based on the result in [2] (Theorem 4.1). Thus, assuming β is at most s-sparse and H satisfies the RIP at the level $\delta_{3s} + \delta_{4s} < 2$, we have

$$\mathrm{Tr}\; E\left[(\beta - \hat{\beta}_k)(\beta - \hat{\beta}_k)^T \mid \mathscr{Y}_k\right] \leq \left(c_1 \sqrt{\mathrm{Tr}(R)} + c_2 e\right)^2 \tag{10}$$

where e denotes the reconstruction error in the noiseless case. For reasonable values of δ_{4s}, the constants c_1 and c_2 in (10) are well behaved [2].

2.5. Simple Example

The estimation performance of the ABCS is demonstrated using a simple example similar to the one in [8]. The new algorithm is compared with the Dantzig Selector (DS), which is aimed at solving the CS problem

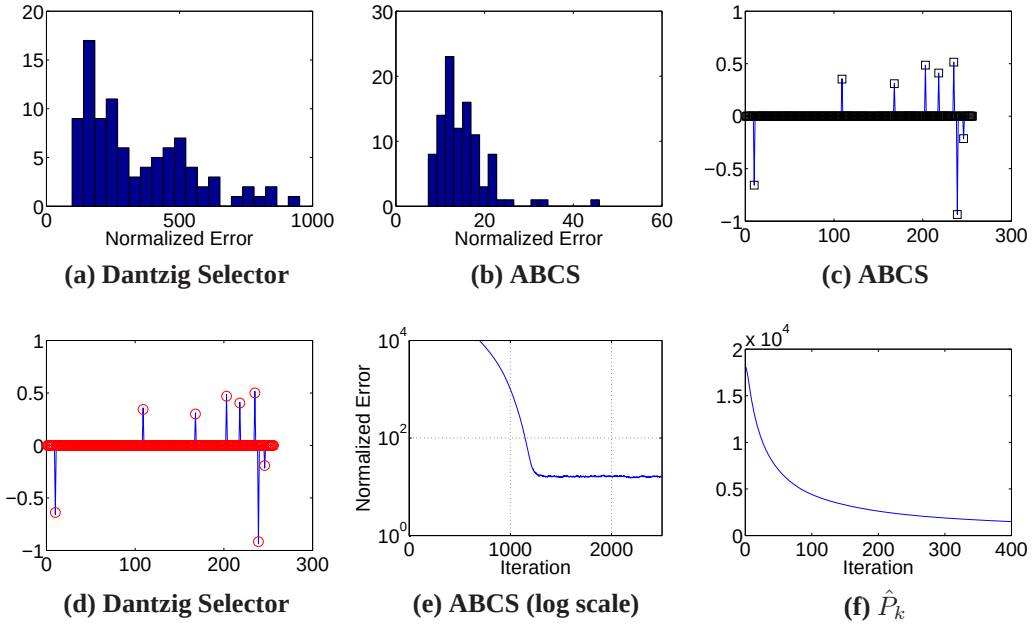
$$\min_{\beta} \parallel \beta \parallel_1 \text{ s.t. } \sum_{i=1}^{k} \parallel H^T(y_i - H\beta) \parallel_\infty \leq \epsilon \tag{11}$$

The sensing matrix $H \in \mathbb{R}^{72\times 256}$ consists of entries that are sampled according to $\mathcal{N}(0, 1/72)$. This type of matrix has been shown to satisfy the RIP with overwhelming probability for sufficiently sparse signals. The sparseness degree of the parameter vector β does not exceed 8.

The distributions of the estimation errors over 100 Monte Carlo runs of both the DS and the ABCS are depicted in Figs. 2d and 2b, respectively. In these figures, the histograms of the normalized errors defined in [8] as $\| \beta - \hat{\beta}_k \|_2^2 \, / \sum_{i=1}^{n} \min\left((\beta^i)^2, \mathrm{Tr}(R)\right)$ are shown. It can be clearly seen that the ABCS outperforms the DS in terms of estimation accuracy. While the averaged normalized error of the DS is around 300, the ABCS attains an average of approximately 16.

The actual (lines) and reconstructed (markers) signals of both the ABCS and DS in a typical run are shown in Figs. 2c and 2d, respectively. In these figures, the magnitudes of the 256 entries of β and $\hat{\beta}_k$ are shown along the abscissa.

The remaining figures, Figs. 2e, 2f, show the behavior of the normalized estimation error and of $\hat{P}_k$ with respect to the iteration τ of the ABCS method. Both these figures further demonstrate the previously discussed concepts.



(a) Dantzig Selector **(b) ABCS** **(c) ABCS**

(d) Dantzig Selector **(e) ABCS (log scale)** **(f)** $\hat{P}_k$

Fig. 2: Estimation performance of the ABCS and the Dantzig Selector.

3. LEARNING COMPRESSED RANDOM FIELD MODELS

In what follows, we demonstrate the application of the ABCS for learning sparse random field (RF) structures that are used for classifying high-dimensional data sets. The new classification method, termed here RF-ABCS, is tested and compared with an RF-based LASSO classifier and a naive Bayes (NB) classifier using multi-class (fMRI) data.

3.1. Classification Problem

Let $V = [v(1), \ldots, v(n)]$ be a feature space. Here we assume that the features $\{v(i)\}_{i=1}^{n}$ are vectors in $\mathbb{R}^m$ which in turn renders V a vector space. It is known that V is associated in some manner with a m-dimensional response vector Y of which the entries y_j, $j = 1, \ldots, m$ take values on some discrete space $\mathscr{Y}$. Now, given a new feature space (testing set) $V' = [v'(1), \ldots, v'(n)]$, $v'(i) \in \mathbb{R}^l$ we are interested in predicting the response vector $\hat{Y}$ associated with V' based on the pair (V, Y).

3.2. Sparse and Compressible Model Learning

A learning algorithm uses (V, Y) for constructing a model that encompasses significant relations between features. Usually the number of features n is much larger than m, the number of training samples (entries of $v(i)$). This fact has motivated the incorporation of various model selection techniques as part of the learning process. The resulting algorithm generally seeks to reduce the model complexity in terms of parameters, thereby promoting interpretable and robust structures.

The classification method derived in the ensuing learns a random field model individually for each value in $\mathscr{Y}$ (class). Following this approach, the predicted class is taken as the one of which the model best "explains" the new feature space in terms of likelihood.

3.2.1. Random Field Model

Let $G = (E^m, V^m)$ be a finite graph with a vertex set V^m and an edge set E^m. The sample space Ω consists of all possible assignments of the vertices in V^m. A random field on G is a probability distribution on Ω.

At this point we assume that our random field model obeys a Gibbs distribution. This, in turn, allows us to specify linear Gaussian connections of the form

$$V^m(i) = H(i)\beta(i) + \zeta(i), \quad \zeta(i) \sim \mathcal{N}(\mu_i, r_i^2 I) \tag{12}$$

where $H(i) = [V^m(j),\ j \neq i]$ is a matrix composed of the entire vertex set excluding the i-th one, and $\beta(i)$ is a parameter vector associated with the i-th vertex. An alternative formulation of (12) embeds a bias term within $\beta(i)$ and assumes a zero-mean noise, that is $H(i)$ is replaced by $[H(i) \quad \mathbf{1}]$ where $\mathbf{1}$ is a vector of which the entries are all 1's. Following this, we may write the conditional probabilities describing the connections as

$$p(V^m(i) \mid H(i), \beta(i), \sigma_i) \propto \exp\left(-\frac{1}{2\sigma_i^2} \parallel V^m(i) - [H(i) \quad \mathbf{1}]\beta(i)) \parallel_2^2\right) \tag{13}$$

3.2.2. Learning over the Feature Space

Let V^m be a set of n_f features from V. The random field structure associated with a given class θ in $\mathscr{Y}$ can be then learned by locally solving (12) for every feature [1] using ABCS, LASSO or any other CS method. The obtained parameters $\beta^\theta(i)$, $i = 1, \ldots, n_f$ associated with a class θ can be then used for approximating the corresponding noise variances r_i^θ

$$\left(r_i^\theta\right)^2 = (k_\theta - 1)^{-1} \sum_{j=1}^{k_\theta} \left[V_j^m(i) - [H_j(i) \quad \mathbf{1}]\beta^\theta(i)\right]^2 \tag{14}$$

where the subscript j denotes the j-th sample, and k_θ denotes the total number of samples for the class θ.

[1] Here, a feature consists of all samples that are associated with a specific class.

3.2.3. Classification Rule

Having the random field parameters for all classes θ, the predicted class of each and every sample in the new feature space V' is chosen as the one which maximizes the posterior probability $p\left((V')_j^m \mid \left\{\beta^\theta(i), r_i^\theta\right\}_{i=1}^{n_f}\right)$ where the subscript j denotes the j-th sample. In practice, the exact posterior may not be easy to compute. This, however, can be alleviated by computing the pseudo-likelihood over the entire network. An approximate solution is then given as

$$\hat{y}_j = \arg\max_{\theta \in \mathscr{Y}} \sum_{i=1}^{n_f} \log p\left((V')_j^m(i) \mid H(i), \beta^\theta(i), r_i^\theta\right) \tag{15}$$

where the conditionals are given in (13).

3.3. Multi-Class Example

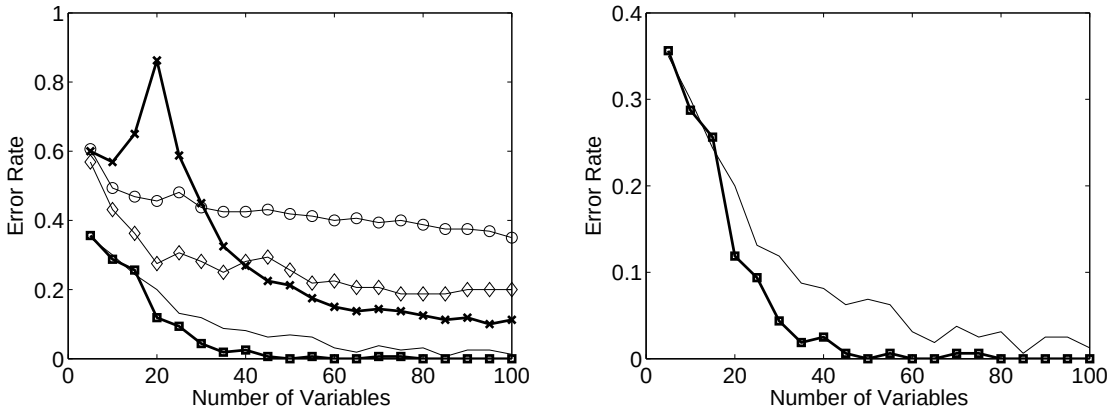
The RF-ABCS is applied for fMRI classification using the same data set of [20]. The performance of the new classifier is compared to both NB and RF-LASSO. The latter method uses a least angle regression (LARS) implementation of LASSO for solving (12) at each node. The fMRI data set consists of 20 samples 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.

In all tests we have used a limited number of features, which have been selected as those with the highest correlation with the response variable (i.e., the labels). The number of features used for each method varies between 5 to 100. Additionally, two of the methods, the RF-ABCS and the RF-LASSO, are tested using data sets on which an isometric transformation has been applied. The purpose of this procedure is to produce rotated RIP data sets based on the original ones. As mentioned previously, both c_1 and c_2 in the upper bound (10) are well-behaved owing to this property. The reader is referred to the appendix and [21] for further discussions and derivations pertaining to this transformation.

The prediction error rates of all methods are depicted versus the number of selected features (variables) in Fig. 3a. This figure shows a clear advantage of the RF-ABCS classifier when using more than 40 features. The error rates of the NB, RF-LASSO and RF-ABCS using 100 features are approximately **0.4, 0.2**, and **0.1**, respectively. The affect of the isometric transformation on the prediction accuracy of the RF-LASSO and RF-ABCS is demonstrated in this figure and in Fig. 3b, which provides a clearer picture of the corresponding error rates. With no exception, in this case as well, the RF-ABCS outperforms the RF-LASSO essentially attaining a **zero error rate** when using more than 80 features.

4. CONCLUSIONS

A new approximate Bayesian compressed sensing algorithm is derived based on a unique type of sparseness-promoting prior. The semi-Gaussian prior, which forms the core of the new method, yields a closed-form recursion for solving the noisy compressed sensing problem. The discrepancy between the exact and the approximate posterior pdf obtained by using the new algorithm is of the order of $\hat{P}_k$, a quantity which is computed online. In the last part of this work, a random field-based classifier utilizing the approximate Bayesian scheme is shown to attain a **zero error rate** when applied to fMRI classification.



(a) Classification Errors **(b) RF-ABCS and RF-LASSO + Isometric**

Fig. 3: Prediction error rates of the various classifiers based on 8-out cross-validation. Showing the NB (circles), RF-LASSO (diamonds), RF-ABCS (bold-crosses), RF-LASSO using the transformed data set (line), and RF-ABCS using the transformed data set (bold-squares).

Appendix

A. PROOF OF THEOREM 1

Proof We first notice that

$$p(y_k \mid \beta) \propto \exp\left(-\frac{1}{2}\left(y_k - H\beta\right)^T R^{-1}\left(y_k - H\beta\right)\right) \tag{16}$$

Recognizing that the denominator in (2) is independent of β while using the above likelihood immediately gives

$$p(\beta \mid \mathscr{Y}_k) \propto \exp\left(-\frac{1}{2}\left(y_k - H\beta\right)^T R^{-1}\left(y_k - H\beta\right) - \frac{1}{2}\left(\beta - \hat{\beta}_{k-1}\right)^T P_{k-1}^{-1}\left(\beta - \hat{\beta}_{k-1}\right)\right) \tag{17}$$

Now, the mean of $p(\beta \mid \mathscr{Y}_k)$ can be obtained by solving

$$\frac{\partial \log p(\beta \mid \mathscr{Y}_k)}{\partial \beta} = 0 \tag{18}$$

This in turn yields

$$\hat{\beta}_k = \left(P_{k-1}^{-1} + H^T R^{-1} H\right)^{-1}\left[P_{k-1}^{-1}\hat{\beta}_{k-1} + H^T R^{-1} y_k\right] \tag{19}$$

which translates into

$$\begin{aligned}\hat{\beta}_k &= \left(I - P_{k-1}H^T\left(HP_{k-1}H^T + R\right)^{-1}H\right)P_{k-1}\left[P_{k-1}^{-1}\hat{\beta}_{k-1} + H^T R^{-1} y_k\right] \\ &= \hat{\beta}_{k-1} + P_{k-1}H^T\left(HP_{k-1}H^T + R\right)^{-1}\left[-H\hat{\beta}_{k-1}\right] \\ &\quad + \left(P_{k-1}H^T R^{-1} - P_{k-1}H^T\left(HP_{k-1}H^T + R\right)^{-1}HP_{k-1}H^T R^{-1}\right) y_k \end{aligned} \tag{20}$$

by applying the matrix inversion lemma. Further elaborating the parentheses in the last term in (20) yields

$$\begin{aligned} &P_{k-1}H^T R^{-1} - P_{k-1}H^T\left(HP_{k-1}H^T + R\right)^{-1}HP_{k-1}H^T R^{-1} = \\ &\quad P_{k-1}H^T R^{-1} - P_{k-1}H^T R^{-1}\left(HP_{k-1}H^T R^{-1} + I\right)^{-1}HP_{k-1}H^T R^{-1} \\ &\quad = P_{k-1}H^T R^{-1}\left[I - \left(HP_{k-1}H^T R^{-1} + I\right)^{-1}HP_{k-1}H^T R^{-1}\right] \\ &\quad = P_{k-1}H^T R^{-1}\left[\left(HP_{k-1}H^T R^{-1} + I\right)^{-1}\left(HP_{k-1}H^T R^{-1} + I\right)\right. \\ &\qquad \left. - \left(HP_{k-1}H^T R^{-1} + I\right)^{-1}HP_{k-1}H^T R^{-1}\right] \\ &= P_{k-1}H^T R^{-1}\left(HP_{k-1}H^T R^{-1} + I\right)^{-1}\left[(HP_{k-1}H^T R^{-1} + I - (HP_{k-1}H^T R^{-1}\right] \\ &\quad = P_{k-1}H^T\left(HP_{k-1}H^T + R\right)^{-1} \end{aligned} \tag{21}$$

Finally, substituting (21) into (20) and collecting terms yields (3a).

The covariance of $p(\beta \mid \mathscr{Y}_k)$ is obtained as

$$P_k = E\left[(\beta - \hat{\beta}_k)(\beta - \hat{\beta}_k)^T \mid \mathscr{Y}_k\right] \tag{22}$$

with

$$\begin{aligned}\beta - \hat{\beta}_k &= \beta - \left(I - P_{k-1}H^T(HP_{k-1}H^T + R)^{-1}H\right)\hat{\beta}_{k-1} - P_{k-1}H^T(HP_{k-1}H^T + R)^{-1}y_k \\ &= \left(I - P_{k-1}H^T(HP_{k-1}H^T + R)^{-1}H\right)\left(\beta - \hat{\beta}_{k-1}\right) - P_{k-1}H^T(HP_{k-1}H^T + R)^{-1}n_k \quad (23)\end{aligned}$$

where the fact that $y_k = H\beta + n_k$ was used for obtaining the 2nd line in (23). Substituting (23) into (22) and after a few (albeit tedious) algebraic manipulations we get the simple recursion in (3b). *QED.*

B. PROOF OF THEOREM 2

Proof The exact and the approximate posterior pdf's are given by

$$p(\beta \mid \mathscr{Y}_k) = c\exp\left(-\frac{1}{2}\frac{\parallel \beta \parallel_1^2}{\sigma^2}\right)p'(\beta \mid \mathscr{Y}_k) \quad (24a)$$

$$\hat{p}(\beta \mid \mathscr{Y}_k) = \hat{c}\exp\left(-\frac{1}{2}\frac{(\hat{H}\beta)^2}{\sigma^2}\right)p'(\beta \mid \mathscr{Y}_k) \quad (24b)$$

where c and $\hat{c}$ denote the appropriate normalization constants, and $p'(\beta \mid \mathscr{Y}_k)$ is the Gaussian posterior without any sparseness promoting prior (i.e., with $P_0 \to \infty$). Now, explicitly writing the KL divergence between these two pdf's yields

$$\begin{aligned}&\mathrm{KL}\left(\hat{p}(\beta \mid \mathscr{Y}_k) \parallel p(\beta \mid \mathscr{Y}_k)\right) \\ &\quad = \int \hat{p}(\beta \mid \mathscr{Y}_k)\log\frac{\hat{p}(\beta \mid \mathscr{Y}_k)}{p(\beta \mid \mathscr{Y}_k)}d\beta = \log(\hat{c}/c) + \frac{1}{2\sigma^2}\int \hat{p}(\beta \mid \mathscr{Y}_k)\left[\parallel \beta \parallel_1^2 - (\hat{H}\beta)^2\right]d\beta \\ &\quad = \log(\hat{c}/c) + \frac{1}{2\sigma^2}\hat{E}\left[\parallel \beta \parallel_1^2 \mid \mathscr{Y}_k\right] - \frac{1}{2\sigma^2}\hat{E}\left[(\hat{H}\beta)^2 \mid \mathscr{Y}_k\right] \quad (25)\end{aligned}$$

where $\hat{E}[\cdot \mid \mathscr{Y}_k]$ denotes the expectation operator with respect to $\hat{p}(\beta \mid \mathscr{Y}_k)$. Applying the Jensen inequality while recalling that $\hat{H}\hat{\beta}_k = \parallel \hat{\beta}_k \parallel_1$, $\hat{\beta}_k = \hat{E}[\beta \mid \mathscr{Y}_k]$, gives

$$\begin{aligned}\mathrm{KL}\left(\hat{p}(\beta \mid \mathscr{Y}_k) \parallel p(\beta \mid \mathscr{Y}_k)\right) &\le \log(\hat{c}/c) + \frac{1}{2\sigma^2}\hat{E}\left[\parallel \beta \parallel_1^2 \mid \mathscr{Y}_k\right] - \frac{1}{2\sigma^2}\hat{E}\left[\hat{H}\beta \mid \mathscr{Y}_k\right]^2 = \\ &\log(\hat{c}/c) + \frac{1}{2\sigma^2}\hat{E}\left[\parallel \beta \parallel_1^2 \mid \mathscr{Y}_k\right] - \frac{1}{2\sigma^2}\parallel \hat{\beta}_k \parallel_1^2 \quad (26)\end{aligned}$$

Further letting $\delta\beta := \beta - \hat{\beta}_k$, Eq. (26) yields

$$\begin{aligned}\mathrm{KL}\left(\hat{p}(\beta \mid \mathscr{Y}_k) \parallel p(\beta \mid \mathscr{Y}_k)\right) &\le \log(\hat{c}/c) + \frac{1}{2\sigma^2}\hat{E}\left[\parallel \beta \parallel_1^2 \mid \mathscr{Y}_k\right] - \frac{1}{2\sigma^2}\parallel \hat{\beta}_k \parallel_1^2 \\ &\le \log(\hat{c}/c) + \frac{1}{2\sigma^2}\hat{E}\left[\left(\parallel \hat{\beta}_k \parallel_1 + \parallel \delta\beta \parallel_1\right)^2 \mid \mathscr{Y}_k\right] - \frac{1}{2\sigma^2}\parallel \hat{\beta}_k \parallel_1^2 \\ &= \log(\hat{c}/c) + \frac{1}{\sigma^2}\hat{E}\left[\parallel \delta\beta \parallel_1 \mid \mathscr{Y}_k\right]\parallel \hat{\beta}_k \parallel_1 + \frac{1}{2\sigma^2}\hat{E}\left[\parallel \delta\beta \parallel_1^2 \mid \mathscr{Y}_k\right] \quad (27)\end{aligned}$$

owing to the triangle inequality. Recalling that $\parallel \delta\beta \parallel_1 \leq \sqrt{n} \parallel \delta\beta \parallel_2$ we may then write

$$\begin{aligned}\mathrm{KL}\left(\hat{p}(\beta \mid \mathscr{Y}_k) \parallel p(\beta \mid \mathscr{Y}_k)\right) &\leq \log(\hat{c}/c) + \frac{\sqrt{n}}{\sigma^2}\hat{E}\left[\parallel \delta\beta \parallel_2 \mid \mathscr{Y}_k\right] \parallel \hat{\beta}_k \parallel_1 + \frac{n}{2\sigma^2}\hat{E}\left[\parallel \delta\beta \parallel_2^2 \mid \mathscr{Y}_k\right] \\ &\leq \log(\hat{c}/c) + \frac{\sqrt{n}}{\sigma^2}\hat{E}\left[\parallel \delta\beta \parallel_2^2 \mid \mathscr{Y}_k\right]^{1/2} \parallel \hat{\beta}_k \parallel_1 + \frac{n}{2\sigma^2}\hat{E}\left[\parallel \delta\beta \parallel_2^2 \mid \mathscr{Y}_k\right] \quad (28)\end{aligned}$$

where the 2nd line in (28) is due to the Jensen inequality. Recognizing that

$$\hat{E}\left[\parallel \delta\beta \parallel_2^2 \mid \mathscr{Y}_k\right] = \mathrm{Tr}\left(\hat{E}\left[(\beta - \hat{\beta}_k)(\beta - \hat{\beta}_k)^T \mid \mathscr{Y}_k\right]\right) = \mathrm{Tr}(\hat{P}_k) \quad (29)$$

Eq. (28) can be written as

$$\mathrm{KL}\left(\hat{p}(\beta \mid \mathscr{Y}_k) \parallel p(\beta \mid \mathscr{Y}_k)\right) \leq \log(\hat{c}/c) + \mathcal{O}\left(\sigma^{-2}\max\left\{\mathrm{Tr}(\hat{P}_k), \mathrm{Tr}(\hat{P}_k)^{1/2}\right\}\right) \quad (30)$$

At this point the theorem immediately follows based on the inequality $\log(\hat{c}/c) \leq 0$. In order to show this we first note that $\parallel \beta \parallel_1^2 \geq (\hat{H}\beta)^2$ owing to the fact that the left hand expression consists of summation of positive terms whereas the same terms on the right side of the equation have either positive or negative signs. This further implies

$$\begin{aligned}&-\parallel \beta \parallel_1^2 \leq -(\hat{H}\beta)^2 \Longrightarrow \exp\left(-\frac{1}{2}\frac{\parallel \beta \parallel_1^2}{\sigma^2}\right) \leq \exp\left(-\frac{1}{2}\frac{(\hat{H}\beta)^2}{\sigma^2}\right) \\ &\Longrightarrow \int \exp\left(-\frac{1}{2}\frac{\parallel \beta \parallel_1^2}{\sigma^2}\right) p'(\beta \mid \mathscr{Y}_k)d\beta \leq \int \exp\left(-\frac{1}{2}\frac{(\hat{H}\beta)^2}{\sigma^2}\right) p'(\beta \mid \mathscr{Y}_k)d\beta \\ &\Longrightarrow \left(\int \exp\left(-\frac{1}{2}\frac{\parallel \beta \parallel_1^2}{\sigma^2}\right) p'(\beta \mid \mathscr{Y}_k)d\beta\right)^{-1} \geq \left(\int \exp\left(-\frac{1}{2}\frac{(\hat{H}\beta)^2}{\sigma^2}\right) p'(\beta \mid \mathscr{Y}_k)d\beta\right)^{-1} \Longrightarrow c \geq \hat{c}\end{aligned} \quad (31)$$

QED.

C. ISOMETRIC DATA TRANSFORMATIONS

The theory of CS shows that the solutions of the noiseless convex problem

$$\min_{\hat{\beta}} \parallel \hat{\beta} \parallel_1 \text{ s.t. } Y = V\hat{\beta} \quad (32)$$

and the original NP-hard problem, in which the l_1 norm in (32) is substituted by the l_0 norm, coincides under the restriction that the sensing matrix V obeys a so-called restricted isometry property (RIP) at a certain level. In detail, the RIP is defined as

$$(1 - \delta_s) \parallel x \parallel_2^2 \leq \parallel Vx \parallel_2^2 \leq (1 + \delta_s) \parallel x \parallel_2^2 \quad (33)$$

for some $\delta_s \in (0, 1)$ and any x that is s-sparse at most. In other words, every subset of V of dimension $m \times s$ acts as nearly orthonormal system. The RIP constant δ_s gives an indication of the actual proximity of any subset to orthogonality. In the noisy case (i.e., similar to the model in (14)) the RIP constant sets an upper bound on the norm estimation error $\parallel \beta - \hat{\beta} \parallel_2$ where β is the actual sparse solution. The reader is referred to [1, 2] for an extensive discussion about the RIP and its role in CS.

C.1. Main Result

The classification method suggested in the Section 3.2 locally solves a regression problem for each feature in V^m. This is accomplished by applying a CS-based method or some other l_1 regularized technique using (14). Recalling that both c_1 and c_2 in (12) are well behaved whenever the sensing matrix satisfies the RIP, it is desired to have this property locally for every node in (14). However, this cannot be guaranteed for the original data set V. Bearing this in mind, we provide a detailed description of a technique for producing an RIP data matrix out of the original one while preserving distance ratios in the transformed space. Before proceeding, however, we introduce the notion of block-sparseness which is used in the ensuing.

Definition 1 *A vector $x \in \mathbb{R}^{dm}$ with $d, m \in \mathbb{N}$ is m-block-sparse if its non zero entries are concentrated in blocks of dimension m. That is, if*

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$$

where $x_i \in \mathbb{R}^m$, $i = 1, \ldots, d$, and

$$\#\{x_i \mid \ x_i \neq 0, \ i = 1, \ldots, d\} << d$$

then x is said to be m-block-sparse.

The following theorem is proved in [21].

Theorem 3 (Isometric Transformation) *Suppose that $H \in \mathbb{R}^{m \times dm}$ for some $d, m \in \mathbb{N}$, and let*

$$T = \operatorname{diag}(H_1^{-1} P_1, \ldots, H_d^{-1} P_d), \quad \operatorname{Ker}(T) = \emptyset \tag{34}$$

where $H_i \in \mathbb{R}^{m \times m}$ and $P_i \in \mathbb{R}^{m \times m}$, $i = 1, \ldots, d$ are the partitions of H and some RIP matrix $P \in \mathbb{R}^{m \times dm}$, respectively. Then there exists an orthogonal transformation $\hat{T} \in \mathbb{R}^{dm \times dm}$ and scalars $\alpha > 0$ and $\delta \in (0, 1)$ for which

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

for m-block-sparse x. In particular

$$\hat{T} = \operatorname{diag}(C_1 D_1^T, \ldots, C_d D_d^T)$$

where $C_i \Lambda_i D_i^T$, $\Lambda_i = \operatorname{diag}(\lambda_i^1, \ldots, \lambda_i^m)$, $\lambda_i^1 \geq \lambda_i^2 \geq \cdots \geq \lambda_i^m$ is the singular values decomposition (SVD) of $H_i^{-1} P_i$. 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 can be approximated by

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

and the RIP constant 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.

Proof Notice that by definition (34), $P = HT$ obeys the RIP

$$(1 - \delta_m) \parallel z \parallel_2^2 \leq \parallel HTz \parallel_2^2 \leq (1 + \delta_m) \parallel z \parallel_2^2 \tag{35}$$

for some m-sparse z. Now, let us set

$$z = D\Lambda^{-1}D^T x \tag{36}$$

where $C\Lambda D^T$ is the SVD of T. Notice that Λ^{-1} exists owing to $\mathrm{Ker}(T) = \emptyset$.

The transformation of the parameter space in (36) changes the sparsity basis of x. In general, we cannot expect that both vectors, the original and transformed, will have the same sparseness degree. However, the following subtle observation changes the whole picture. It can be easily verified that the block structure of T in (34) induces a similar structure of $D\Lambda^{-1}D^T$ in (36). This further implies that if x is m-block-sparse then so z.

The definition of z implies

$$\parallel z \parallel_2^2 = x^T D\Lambda^{-1}D^T D\Lambda^{-1}D^T x = \parallel \Lambda^{-1}D^T x \parallel_2^2 \tag{37}$$

Substituting (37) and the SVD of T into (35) yields

$$(1 - \delta_m) \parallel \Lambda^{-1}D^T x \parallel_2^2 \leq \parallel HCD^T x \parallel_2^2 \leq (1 + \delta_m) \parallel \Lambda^{-1}D^T x \parallel_2^2 \tag{38}$$

Without any loss of generality let us assume at this point that $\parallel x \parallel_2 = 1$. Hence,

$$\lambda_{\max}^{-1}(\Lambda) = \lambda_{\min}(\Lambda^{-1}D^T) \leq \parallel \Lambda^{-1}D^T x \parallel_2^2 \leq \lambda_{\max}(\Lambda^{-1}D^T) = \lambda_{\min}^{-1}(\Lambda) \tag{39}$$

Using the above in (38) while multiplying by some $\alpha > 0$ yields

$$\alpha^2(1 - \delta_m)\lambda_{\max}^{-1} \leq \parallel \alpha HCD^T x \parallel_2^2 \leq \alpha^2(1 + \delta_m)\lambda_{\min}^{-1} \tag{40}$$

Finally setting

$$\alpha^2(1 - \delta_m)\lambda_{\max}^{-1} = 1 - \delta \tag{41a}$$

$$\alpha^2(1 + \delta_m)\lambda_{\min}^{-1} = 1 + \delta \tag{41b}$$

and solving for α and δ yields the theorem. *QED.*

C.2. Block Sparseness Equivalence and Column Ordering

As can be easily recognized from (36) and the definition of T in (34), the isometric transformation assumes a block sparse form of the projected linear space. This fact raises a question. How can we impose the block sparse form on β, our original parameter space. A straight forward approach for alleviating this problem is to reorder the columns of V or more precisely of V^m so as to group significant features in blocks. This in turn increases the chance of having a block sparse solution to the CS problem

$$\min_{\beta^\theta(i)} \parallel \beta^\theta(i) \parallel_1 \quad \text{s.t.} \quad \parallel V^m(i) - H(i)\beta^\theta(i) \parallel_2 \leq \epsilon \tag{42}$$

which forms the heart of the random field classifier of Section 3.2.

Reordering of columns can be carried out using either a ranking method or a feature selection technique (e.g., correlation, t-test, LDA and Fisher linear discriminant). The columns will be then reordered according to their measure of significance. Notice, however, that the ordering of columns within the $m \times m$ blocks of V does not really affect the block sparseness degree of β.

C.3. Practical Implementation: Random Projections

The isometric transformation relies on the existence of some RIP 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 [1, 2].

Consider a matrix $P \in \mathbb{R}^{m\times n}$ of which the entries are independent identically distributed (iid) samples from $\mathcal{N}(0, m^{-1})$. Then if s, the maximal sparseness degree of the underlying parameter vector, satisfies

$$s = \mathcal{O}\left(m/\log(n/m)\right) \tag{43}$$

the matrix P obeys the RIP with probability exceeding $1 - \mathcal{O}\left(\exp(-\gamma n)\right)$ for some $\gamma > 0$ [2]. Similar result exists for a binary measurement matrix of which the entries are sampled according to

$$\Pr\left(P_{ij} = \pm 1/\sqrt{m}\right) = 0.5 \tag{44}$$

In the case of high dimensional feature space it seems that the random approach is the only one that can guarantee the RIP to some extent. Taking random P, however, imposes a conceptual problem. Thus, we can expect that there might be realizations of P that render the new data set less informative thereby deteriorating the classification accuracy. In order to avoid such instances we propose an additional stage in which a proper realization of P would be chosen by cross-validating over a transformed development set. This technique is demonstrated in the numerical study section in the ensuing.

C.4. Transductive Approach

In practice the isometric transformation can be applied to a feature space that is augmented by the (testing) data set V'

$$\bar{V} = \begin{bmatrix} V \\ V' \end{bmatrix} \in \mathbb{R}^{(m+l)\times n} \tag{45}$$

This technique, which can be thought of as a form of unsupervised learning, yields a transformation that depends on both the training and the unlabeled testing data sets. This approach, which is used in the numerical study part of this work, has shown to significantly improve the classification accuracy.

C.5. Summary

Given an arbitrary (augmented) data matrix $H \in \mathbb{R}^{(m+l)\times n}$ where $n = d(m + l)$ and $d, m, l \in \mathbb{N}$, the transformation is carried out as follows.

1. Generate a realization of an isometric (Gaussian or Binary) random matrix P of the same dimensions as H.
2. Reorder the columns of H according to some ranking or feature selection technique.
3. Partition $H = [H_1, \ldots, H_d]$, $P = [P_1, \ldots, P_d]$ where $H_i, P_i \in \mathbb{R}^{(m+l)\times(m+l)}$.
4. Compute the transformation $\hat{T}_i = C_i D_i^T$ for all $i = 1, \ldots, d$ where $C_i \Lambda_i D_i^T$ is the SVD of $H_i^{-1} P_i$.
5. Compute the scaling factor α.
6. The transformed data set is $\bar{H} = \alpha \left[H_1\hat{T}_1, \ldots, H_d\hat{T}_d\right]$.

D. REFERENCES

[1] 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.

[2] E. J. Candes, "Compressive sampling", Madrid, Spain, 2006, European Mathematical Society, Proceedings of the International Congress of Mathematicians.

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

[4] M. Lustig, D. Donoho, and J. M. Pauly, "Sparse mri: The application of compressed sensing for rapid mr imaging", *Magnetic Resonance in Medicine*, vol. 58, pp. 1182–1195, 2007.

[5] U. Gamper, P. Boesiger, and S. Kozerke, "Compressed sensing in dynamic mri", *Magnetic Resonance in Medicine*, vol. 59, pp. 365–373, 2008.

[6] R. Chartrand, "Exact reconstruction of sparse signals via nonconvex minimization", *IEEE Signal Processing Letters*, vol. 14, pp. 707–710, 2007.

[7] 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.

[8] 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.

[9] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit", *SIAM Journal of Scientific Computing*, vol. 20, no. 1, pp. 33 – 61, 1998.

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

[11] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine", *Journal of Machine Learning Research*, vol. 1, pp. 211 – 244, 2001.

[12] R. E. McCulloch and E. I. George, "Approaches for bayesian variable selection", *Statistica Sinica*, vol. 7, pp. 339 – 374, 1997.

[13] J. Geweke, *Bayesian Statistics 5*, chapter Variable selection and model comparison in regression, Oxford University Press, 1996.

[14] B. A. Olshausen and K. Millman, "Learning sparse codes with a mixtureof-gaussians prior", *Advances in Neural Information Processing Systems (NIPS)*, pp. 841 – 847, 2000.

[15] S. J. Godsil and P. j. Wolfe, "Bayesian modelling of time-frequency coefficients for audio signal enhancement", *Advances in Neural Information Processing Systems (NIPS)*, 2003.

[16] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries", *IEEE Transactions on Signal Processing*, vol. 4, pp. 3397 – 3415, 1993.

[17] Y. C. Pati, R. Rezifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition", 27th Asilomar Conf. on Signals, Systems and Comput., 1993.

[18] S. Chen, S. A. Billings, and W. Luo, "Orthogonal least squares methods and their application to non-linear system identification", *International Journal of Control*, vol. 50, pp. 1873 – 1896, 1989.

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

[20] 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.

[21] A. Carmi, I. Rish, Cecchi G., D. Kanevsky, and B. Ramabhadran, "Isometry-enforcing data transformations for improving sparse model learning", Tech. Rep. RC 24801, Human Language Technologies, IBM, 2009.