Journal of Neuroscience Methods 276 (2017) 46–55

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

journal homepage: www.elsevier.com/locate/jneumeth

A robust sparse-modeling framework for estimating schizophrenia biomarkers from fMRI

Keith Dillon[a,b,*], Vince Calhoun[c,d], Yu-Ping Wang[a,b]

[a] Department of Biomedical Engineering, Tulane University, New Orleans, LA, USA
[b] Department of Global Biostatistics and Data Science, Tulane University, New Orleans, LA, USA
[c] The Mind Research Network & LBERI, Albuquerque, NM, USA
[d] Department of Electrical Engineering, University of New Mexico, New Mexico, USA

HIGHLIGHTS

- We consider robust components as constant over all possible unknown mechanisms.
- We derive a method to incorporate a preference for sparsity in the mechanism.
- Improvement in robustness is demonstrated with simulation.
- Application to fMRI demonstrates superior accuracy in classifying schizophrenia.

ARTICLE INFO

Article history:
Received 19 September 2016
Received in revised form 9 November 2016
Accepted 10 November 2016
Available online 17 November 2016

Keywords:
Schizophrenia
Functional MRI
Sparsity
PCA
Optimization

ABSTRACT

Background: Our goal is to identify the brain regions most relevant to mental illness using neuroimaging. State of the art machine learning methods commonly suffer from repeatability difficulties in this application, particularly when using large and heterogeneous populations for samples.
New method: We revisit both dimensionality reduction and sparse modeling, and recast them in a common optimization-based framework. This allows us to combine the benefits of both types of methods in an approach which we call unambiguous components. We use this to estimate the image component with a constrained variability, which is best correlated with the unknown disease mechanism.
Results: We apply the method to the estimation of neuroimaging biomarkers for schizophrenia, using task fMRI data from a large multi-site study. The proposed approach yields an improvement in both robustness of the estimate and classification accuracy.
Comparison with existing methods: We find that unambiguous components incorporate roughly two thirds of the same brain regions as sparsity-based methods LASSO and elastic net, while roughly one third of the selected regions differ. Further, unambiguous components achieve superior classification accuracy in differentiating cases from controls.
Conclusions: Unambiguous components provide a robust way to estimate important regions of imaging data.

1. Introduction

In this paper our goal is to find the most relevant brain regions given labeled neuroimaging data; the ultimate goal is to use those results to understand disease mechanisms, as well as to provide biomarkers to help diagnose (i.e., classify) patients as having disease or not. There is a significant need for techniques which can robustly extract information in such a problem. Neuroimaging, particularly functional neuroimaging, has provided a wealth of intriguing information regarding brain function, but has yet to show clear value to psychiatric diagnosis (Krystal and State, 2014). Despite this, impressive results have been achieved with machine learning techniques such as support vector machines, which demonstrate high classification accuracies (Orr et al., 2012). Reproducibility problems persist however (Buck, 2015), with an apparent trend towards poorer performance for larger studies (Schnack and Kahn, 2016).

The identification of meaningful components of the data is a key benefit of many feature selection techniques (Guyon and Elisseeff,

* Corresponding author at: Department of Biomedical Engineering, Tulane University, New Orleans, LA, USA.
E-mail address: kdillon1@tulane.edu (K. Dillon).

http://dx.doi.org/10.1016/j.jneumeth.2016.11.005
0165-0270/

2003), in addition to providing improvements in performance of subsequent classification stages (Chu et al., 2012). In a typical neuroimaging study there may be tens or hundreds of subjects, each with an image consisting of up to hundreds of thousands of voxels, resulting in an extremely underdetermined problem. A popular category of feature selection approaches is regularized regression techniques such as LASSO (Tibshirani, 1996) and related methods employing sparse models (Cao et al., 2014; Lin et al., 2014). Such supervised techniques impose task-specific information (the data labels), with a penalty term to incorporate prior knowledge. In the case of LASSO, the prior knowledge amounts to a presumption of sparsity on the relationship between image data and labels, i.e., that the underlying biological mechanism involves a limited number of the imaged voxels. Unfortunately, if the problem is both very underdetermined and very noisy, then the regularized solution may not be a particularly superior choice; many solutions may potentially be of similar or even equal probability to each other. For example in the underdetermined case, the LASSO solution may not be unique for highly-structured datasets (Tibshirani, 2013; Zhang et al., 2014). Along these lines, we simply may not have sufficient confidence in the validity of our prior knowledge formulation to presume that the most regular solution is preferable to those even moderately as regular.

From a different direction, dimensionality reduction techniques (Lemm et al., 2011) offer a more robust approach to feature selection in neuroimaging data. An example is principal component analysis (PCA) (Dunteman, 1989), which finds basis vectors for the space containing the data variation. This set of basis vectors can be viewed as robust in the sense that they are common to all solutions of any linear regression based on the data. In statistics a closely-related concept is estimable functions (Milliken and Johnson, 2009). We will refer to components with such a property as unambiguous, and examine this more formally in the next section. Of course such components only describe the data itself, not necessarily the aspects of the data pertinent to our application, such as for finding information most related to a disease phenotype. A common approach is to utilize PCA and related factoring methods in a supervised fashion by choosing a subset of factors which best correlate with the labels. Supervised factoring techniques such as the "Supervised PCA" of Barshan et al. (2011), and related methods, can be viewed as a more sophisticated version of this technique, finding a transformation of the data such that the correlation with the labels is maximized. However these techniques are not able to incorporate prior knowledge, such as sparsity of the mechanism, into this transformation. Techniques have been developed which do incorporate sparsity into unsupervised factoring techniques (e.g., sparse PCA of Zou et al., 2006) in a heuristic sense, though this differs from presuming sparsity of the underlying mechanism; the presumption of sparsity is applied to the structure of the component itself rather than to the unknown mechanism. Hybrid methods have been proposed which perform PCA following a pre-screening step which picks a subset of variables over a correlation threshold (Bair et al., 2006) or in known pathways (Ma and Dai, 2011). However we would prefer to incorporate multi-variable relationships in the screening component.

In this paper, we develop an approach which combines the benefits of both regularized estimates and dimensionality reduction by simultaneously enforcing unambiguity and prior knowledge in calculating components. We start by reviewing dimensionality reduction from the perspective of unambiguous components. Then we review related regularization methods and show how they motivate our approach to incorporate prior knowledge into unambiguous components. By maximizing the correlation with the mechanism, we calculate components which identify the most important regions in the data. We use a simulation to show how this component performs robustly in the face of inaccurate prior knowledge, by demonstrating that the correlation still remains controlled as the prior knowledge is relaxed. Finally, we show a successful application to biomarker identification where we identify features of fMRI data which relate to schizophrenia more accurately than other methods which utilize sparsity as prior knowledge.

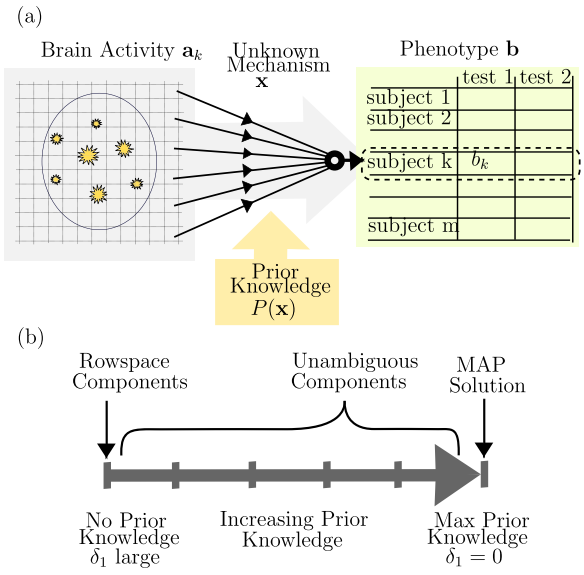

Fig. 1. (a) Mathematical model $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{x}$ describes the mechanism that relates brain activity to phenotype (psychiatric assessments). The contrast map for a single subject, $\mathbf{a}_k$, provides the kth row of $\mathbf{A}$. As there are still many unknown biological variables, the problem is underdetermined and $\mathbf{x}$ cannot be found uniquely; instead we must settle for a probable result such as a maximum likelihood solution which utilizes prior knowledge, or an estimable component of $\mathbf{x}$. (b) Continuum between rowspace components and most probable solution, based on increasing confidence in the prior knowledge, which we control by the relaxation parameter δ_1.

2. Materials and methods

We will consider the linear model $\mathbf{Ax} = \mathbf{b} + \mathbf{n}$ where $\mathbf{A}$ is a $m \times n$ data matrix with $n > m$, containing samples as rows, and variables as columns; $\mathbf{b}$ is the phenotype encoded into a vector of labels such as case or control; the solution $\mathbf{x}$ is the unknown model parameters that relate $\mathbf{A}$ to $\mathbf{b}$; and $\mathbf{n}$ is a noise vector about which we have only statistical information. We will also assume the means have been removed from $\mathbf{b}$ and the columns of $\mathbf{A}$ to simplify the presentation. The rows of $\mathbf{A}$ are provided by the contrast images from individual study subjects, so a predictor $\mathbf{x}$ selects a weighted combination of voxels (i.e., columns of $\mathbf{A}$) which relates the imagery to the case–control status. By examining the weightings in this combination we hope to learn more about the spatial distribution of causes or effects of the disease, which we will term the "mechanism" in this paper. The model is depicted in Fig. 1(a), where we depict the true solution $\mathbf{x}$ as the mechanism whereby brain activity relates to the measured phenotypes. Of course there are far more unknown variables than samples, hence our linear system is underdetermined and there will be many possible $\mathbf{x}$ which solve the system. One way to address this problem is to impose prior knowledge about the biological mechanism, such as a preference for sparser $\mathbf{x}$, and select the solution which best fulfills this preference. We will review this approach in a later section. Another approach is to restrict our analysis to components of the solution which may be more easily estimated, such as via dimensionality reduction; an intuitive example of this approach is to group voxels into low-resolution regions. These alternatives are depicted in Fig. 1(b), as extremes on a continuum of possible methods, where the goal of this paper is to find intermediate information which utilizes the benefits of both extremes.

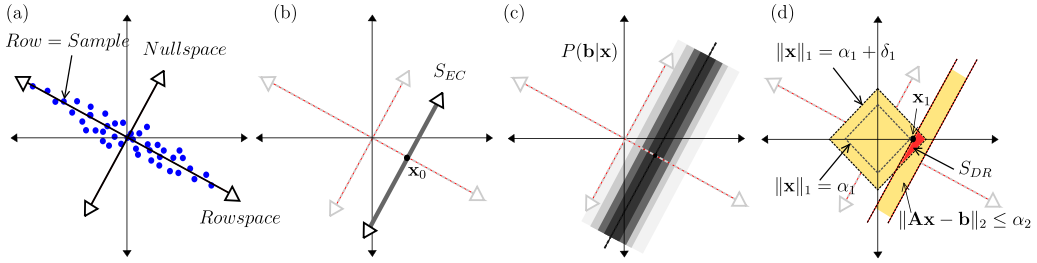

Fig. 2. Two-dimensional example: (a) The rowspace of **A** is the direction of data diversity, while the nullspace is perpendicular directions, which lack diversity. (b)The solution set to $\mathbf{Ax} = \mathbf{b}$ is the affine space S_{EC}, which is parallel to the nullspace; the ambiguity due to lack of data diversity results in ambiguity of possible solutions forming S_{EC}. (c) The distribution of **x** resulting from the likelihood $P(\mathbf{b}|\mathbf{x})$, the high probability region is nearest S_{EC}. (d) The set S_{DR} formed by the intersection of a bound on $P(\mathbf{b}|\mathbf{x})$ enforced by $\|\mathbf{Ax} - \mathbf{b}\|_2 \leq \alpha_2$, and a bound on the prior distribution $P(\mathbf{x})$ enforced by $\|\mathbf{x}\|_1 \leq \alpha_1$; the LASSO solution $\mathbf{x}_1$ is based on a trade-off where the combination of bounds must be as tight as possible.

2.1. Dimensionality reduction

To see how dimensionality reduction can apply to the estimation of mechanism, consider the Singular Value Decomposition (SVD) of $\mathbf{A} = \mathbf{USV}^T$, where **S** is a diagonal matrix of singular values σ_i, and **U** and **V** contain the left and right singular vectors $\mathbf{u}_i$ and $\mathbf{v}_i$. In Principal Components Analysis (PCA) the focus is on this expansion itself, however our focus here is on the mechanism which we model by **x**. If we plug $\mathbf{USV}^T$ into $\mathbf{Ax} = \mathbf{b}$ and apply the transformation $\mathbf{U}^T$ to both sides, we get a diagonalized system with decoupled equations of the form $\sigma_i \mathbf{v}_i^T \mathbf{x} = \mathbf{u}_i^T \mathbf{b}$. This equation tell us that, while we cannot calculate the true **x** in the underdetermined case, we can calculate components of **x** corresponding to nonzero singular values. One way to view this property is by considering that these components are constant for all possible **x** given our linear system. In other words, $\sigma_i \mathbf{v}_i^T \mathbf{x}$ calculates the same value for any solution in the set $\{\mathbf{x} \mid \mathbf{Ax} = \mathbf{b}\}$. In terms of our application, this means while we cannot identify the true mechanism **x**, we can extract reliable components of it. For example, we might be able to coarsely identify large regions containing important activity, without being able to pinpoint particular voxels within those regions. Of course the SVD selects vectors based on orthogonal directions of ranked data variation, which may not be best suited to provide information about the disease mechanism. We will address this later by optimizing the choice of component, but first we will formally consider this property of constant components so that we may extend it later to incorporate prior knowledge.

2.2. Unambiguous components without prior knowledge

In this section we will introduce the idea of unambiguous components in the noise-free case without prior knowledge. The concept is demonstrated geometrically in Fig. 2(a), where the useful information in **A** forms the matrix rowspace; this contains the dimensions over which we have a diversity of data which we can compare to the phenotype **b**. Dimensions perpendicular to the rowspace form the nullspace, directions over which our data does not vary. For example if our dataset was composed of subjects with the same age, then we cannot perform a regression to see how disease risk relates to age. In terms of components of the data, if **c** is a loading vector and $\mathbf{a}_i$ is a sample (row of **A**), then $\mathbf{c}^T \mathbf{a}_i$ is potentially useful information when **c** is in the rowspace, and useless information (in fact always zero) when **c** is in the nullspace. An equivalent perspective, depicted in Fig. 2(b), is the geometry of the solution set to the regression model. This is the affine set S_{EC}, composed of an offset (the least-length solution $\mathbf{x}_0$) plus a vector from the nullspace.

$$S_{EC} = \{\mathbf{x} \mid \mathbf{Ax} = \mathbf{b}\}. \tag{1}$$

The subscript "*EC*" refers to **x** being only equality-constrained (i.e., we have imposed no prior knowledge yet). The rowspace vectors yield estimable components of **x** given the data, while nullspace vectors give the dimensions of ambiguity we have about **x**. We will formulate this relationship rigorously next.

Recall that the rowspace of **A** is defined as the set of all possible linear combinations of rows, i.e., $\mathcal{R}(\mathbf{A}^T) = \{\mathbf{c} \mid \mathbf{A}^T\mathbf{y} = \mathbf{c} \ \ \forall \ \ \mathbf{y} \in \mathbb{R}^m\}$. We can equivalently view this as a set whose members take on a constant value over S_{EC}. In other words, they form the set of unambiguous components for S_{EC}. We state this simple but important fact in Theorem 1.

Theorem 1. *The following two statements are equivalent:*

1. $\mathbf{c} \in \mathcal{R}(\mathbf{A}^T)$.
2. $\mathbf{c}^T\mathbf{x} = \mu$ *for all* $\mathbf{x} \in S_{EC}$, *where* μ *is a constant.*

Proof. It is straightforward to note that, as $\mathbf{A}^T\mathbf{y} = \mathbf{c}$, we have $\mathbf{c}^T\mathbf{x} = \mathbf{y}^T\mathbf{Ax} = \mathbf{y}^T\mathbf{b} \equiv \mu$. However it will be useful to our subsequent extension to note the fact that the rowspace is the orthogonal complement to the nullspace, and can be also written as $\mathcal{R}(\mathbf{A}^T) = \{\mathbf{c} \mid \mathbf{c}^T\mathbf{z} = 0 \ \ \forall \ \ \mathbf{z} \in \mathcal{N}(\mathbf{A})\}$. Equivalently, any vector **z** in the nullspace may be described as a difference between solutions, i.e., $\mathbf{z} = \mathbf{x}_1 - \mathbf{x}_2$, where $\mathbf{x}_1, \mathbf{x}_2 \in S_{EC}$. Therefore the rowspace may also be written as $\mathcal{R}(\mathbf{A}^T) = \{\mathbf{c} \mid \mathbf{c}^T\mathbf{x}_1 = \mathbf{c}^T\mathbf{x}_2 \ \ \forall \ \ \mathbf{x}_1, \mathbf{x}_2 \in S_{EC}\}$, which explicitly states the equivalence in Theorem 1. □

The principal components of **A** corresponding to the *r* nonzero singular values must be in the rowspace. Again, given a singular-value decomposition $\mathbf{A} = \mathbf{USV}^T$, we have $\mathbf{VSU}^T\mathbf{y} = \mathbf{c}$ for $\mathbf{c} \in \mathcal{R}(\mathbf{A}^T)$, and hence $\mathbf{V}^T\mathbf{c} = \mathbf{SU}^T\mathbf{y} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i$, where the $\mathbf{u}_i$ for $i \in \{1, \ldots, r\}$ form a basis for the rowspace. So principal components are unambiguous components.

Later we will also consider how to optimize the choice of component such that it also computes a score which is maximally useful for an application (such as to identify important image regions for classification of disease). But first we will extend the concept of unambiguous components to incorporate prior knowledge. Our focus on components of the mechanism **x** rather than of the data $\mathbf{a}_i$ allows us to incorporate prior knowledge which applies to **x** into the method, such as the presumption of a sparse biological mechanism. In effect, we will utilize prior knowledge to further restrict the variation in **x** and potentially increase the estimable components of **x**.

2.3. LASSO and Elastic Net

In this section we will review closely-related estimation methods which utilize prior knowledge, to provide an intuitive basis for our mathematical formulation. The Maximum A Posteriori (MAP) estimate $\mathbf{x}_{MAP}$ is the solution which maximizes the

posterior probability $P(\mathbf{x}|\mathbf{b})$ over $\mathbf{x}$. Using Bayes' theorem, we can form the equivalent problem, $\text{argmax}_{\mathbf{x}}\ P(\mathbf{b}|\mathbf{x})P(\mathbf{x})$, which utilizes the likelihood $P(\mathbf{b}|\mathbf{x})$ (essentially the distribution for $\mathbf{n}$), and the prior probability distribution $P(\mathbf{x})$. It is the form of this prior distribution which we refer to as the prior knowledge. A common approach is to take the log of this objective to get a penalized regression problem. With a Gaussian distribution for $\mathbf{n}$ and a Laplace distribution for $\mathbf{x}$, taking the log yields the most well-known version of LASSO, $\min_{\mathbf{x}}\|\mathbf{Ax}-\mathbf{b}\|_2^2+\lambda\|\mathbf{x}\|_1$. LASSO was originally proposed in the form, $\min_{\mathbf{x}}\|\mathbf{Ax}-\mathbf{b}\|_2^2$ *subject to* $\|\mathbf{x}\|_1\le\alpha_1$, with the other version sometimes described as the Lagrangian form. It can be shown that for any λ, a α_1 exists such that these optimization problems have the same minimizer, $\mathbf{x}_1$. We can similarly form the feasibility problem (an optimization problem which stops when any feasible point is found, hence we simply use a zero for the objective as it is irrelevant (Boyd and Vandenberghe, 2004)),

$$\begin{aligned}\mathbf{x}_1 = \quad & \underset{\mathbf{x}}{\text{argmin}} \quad 0 \\ & \|\mathbf{x}\|_1 \le \alpha_1 \\ & \|\mathbf{Ax}-\mathbf{b}\|_2 \le \alpha_2\end{aligned} \tag{2}$$

In this paper we will focus on sets of the general form of this feasible set we will denote as S_{DR}. In terms of the statistical distributions, the constraints on the norms amount to constraints on the probabilities, as in

$$S_{DR} = \left\{\mathbf{x} \;\middle|\; \|\mathbf{x}\|_1 \le \alpha_1, \quad \|\mathbf{Ax}-\mathbf{b}\|_2 \le \alpha_2\right\}. \tag{3}$$

$$S_{DR} = \left\{\mathbf{x} \;\middle|\; P(\mathbf{x}) \ge P_{\min}, \quad P(\mathbf{Ax}-\mathbf{b}) \ge P_{\min}\right\}. \tag{4}$$

The subscript *DR* refers to the combination of denoising and regularization constraints. It can also be shown that $\mathbf{x}_1$ which solves the previous versions, will provide a member of this set. Therefore, we may write the LASSO problem as

$$\begin{aligned}\mathbf{x}_1 = \quad & \underset{\mathbf{x}}{\text{argmin}} \quad 0 \\ & \mathbf{x} \in S_{DR}\end{aligned} \tag{5}$$

The elastic net (Zou and Hastie, 2005) can be viewed as a variation on LASSO, where (under the Bayesian framework) the Laplace prior distribution for the prior is replaced by a product of Laplace and Gaussian distributions. Taking the log of the posterior distribution leads to the well-known problem $\min_{\mathbf{x}}\|\mathbf{Ax}-\mathbf{b}\|_2^2+\lambda_1\|\mathbf{x}\|_1+\lambda_2\|\mathbf{x}\|_2^2$. As with LASSO, we may achieve the same solution using different optimization problems. We wish to use the same constraints as earlier, so we form the related problem,

$$\begin{aligned}\mathbf{x}_2 = \quad & \underset{\mathbf{x}}{\text{argmin}} \quad \|\mathbf{x}\|_2^2 \\ & \|\mathbf{x}\|_1 \le \alpha_1 \quad = \quad \underset{\mathbf{x}}{\text{argmin}} \quad \|\mathbf{x}\|_2^2 \\ & \|\mathbf{Ax}-\mathbf{b}\|_2 \le \alpha_2 \qquad\quad \mathbf{x} \in S_{DR}.\end{aligned} \tag{6}$$

Recall that LASSO seeks any single solution within S_{DR}. In that case, α_1 and α_2 are chosen as small as possible so the set is (hopefully) a singleton. Elastic net, on the other hand, seeks the least length solution in S_{DR}, where we may choose α_1 and α_2 more loosely, to trade-off desirable properties of this solution. These are depicted in Fig. 2 (d), where the LASSO solution, $\mathbf{x}_1$ is formed by tightening the bounds on the two sets (one representing the prior and one representing the noise) to minimal intersection. The elastic net solution, by contrast, relaxes the constraints on the two sets forming the larger set S_{DR}, and selection of the least length solution in this intersection. Next we will extend the unambiguous component idea to utilize S_{DR}, thereby incorporating prior knowledge into the framework.

Finally, note that the norms $\|\cdot\|_1$ and $\|\cdot\|_2$ may be replaced with other choices of norms representing other forms of prior knowledge, such as the ℓ_∞-norm, which imposes hard limits on the unknowns or the error, and which might result from box constraints (De Angelis et al., 1997) or a minimax regression (Radhakrishna Rao and Toutenburg, 2013). More complex forms of penalties are used in the elastic net (Zou and Hastie, 2005), and methods utilizing mixed norms (Kowalski, 2009) such as group LASSO (Yuan and Lin, 2006). The methods and theoretical results in this paper are generally the same as long as the choice fulfills the mathematical properties of a norm (though this is not a strict requirement).

2.4. Unambiguous components subject to prior knowledge

Earlier, we introduced unambiguous components as components of the unknown solution, which computed a constant score over all possible solutions. Now we will extend this concept to replace the set of possible solutions (S_{EC}) with a set of highly-probable solutions, for which we will use S_{DR}. We will also extend the requirement that the score be constant to a requirement that their variation is limited within a fixed bound. So in effect, given a pool of solutions, which we may view as the most likely hypotheses for the underlying mechanism, we will seek components which compute approximately-constant scores for every likely mechanism.

We incorporate prior knowledge as well as bounds on the score by relaxing the rowspace with the following generalization,

$$R_S(\epsilon) = \left\{\mathbf{c} \;\middle|\; |\mathbf{c}^T\mathbf{x}_1 - \mathbf{c}^T\mathbf{x}_2| \le \epsilon \quad \forall \quad \mathbf{x}_1, \mathbf{x}_2 \in S_{DR}\right\} \tag{7}$$

$$R_S(\epsilon) = \left\{\mathbf{c} \;\middle|\; \mathbf{c}^T\mathbf{x}_1 - \mathbf{c}^T\mathbf{x}_2 \le \epsilon \quad \forall \quad \mathbf{x}_1, \mathbf{x}_2 \in S_{DR}\right\}. \tag{8}$$

Here we have relaxed the ambiguity to potentially allow a positive bound ϵ, as well as replaced S_{EC} with S_{DR}. Note that we were able to remove the absolute value in Eq. (8) because membership in $R_S(\epsilon)$ requires the inequality hold for all $\mathbf{x}_1, \mathbf{x}_2 \in S_{DR}$, hence if it holds for $\mathbf{x}_1', \mathbf{x}_2' \in S_{DR}$, it must also hold for $\mathbf{x}_1'', \mathbf{x}_2''$ where $\mathbf{x}_1'' = \mathbf{x}_2'$ and $\mathbf{x}_2'' = \mathbf{x}_1'$. To highlight the view of this set as a generalization of the rowspace, we state the following simple generalization of Theorem 1, without proof,

Theorem 2. *The following two statements are equivalent:*

1. $\mathbf{c} \in R_S(\epsilon)$.
2. $\mu - \frac{1}{2}\epsilon \le \mathbf{c}^T\mathbf{x} \le \mu + \frac{1}{2}\epsilon$ *for all* $\mathbf{x} \in S_{DR}$, *where* μ *is a constant.*

Note that while this formulation appears quite simple, it will generally not be possible to describe $R_S(\epsilon)$ in a closed form or analytically determine whether a vector is a member of $R_S(\epsilon)$. However we may use convex optimization theory to formulate $R_S(\epsilon)$ as a system of inequalities. We defined $R_S(\epsilon)$ via the test $\mathbf{c}^T\mathbf{x}_1 - \mathbf{c}^T\mathbf{x}_2 \le \epsilon$, which may also be written as $\mathbf{c}^T(\mathbf{x}_1 - \mathbf{x}_2) \le \epsilon$. We may use optimization to directly perform this test as follows,

$$\begin{aligned}p = \quad & \max_{\mathbf{x},\mathbf{x}'} \quad \mathbf{c}^T(\mathbf{x}-\mathbf{x}') \\ & \mathbf{x} \in S_{DR} \\ & \mathbf{x}' \in S_{DR}.\end{aligned} \tag{9}$$

If the optimal value $p \le \epsilon$, then we were unable to find any pair of solutions to demonstrate an ambiguity in the component score greater than ϵ, and hence $\mathbf{c} \in R_S(\epsilon)$. Similar methods have been proposed elsewhere to use optimization to test boundedness (Dillon and Fainman, 2013) and uniqueness (Dillon and Fainman, 2016; Tibshirani, 2013) of solutions to systems with particular forms of

prior knowledge; in those cases the test was performed over individual variables, whereas we will use a test for a general component $\mathbf{c}$ here. S_{DR} is a convex set so this is a straightforward convex optimization problem for either maximization or minimization, and so we can be assured of finding a global optima with an efficient algorithm (Boyd and Vandenberghe, 2004). If we replace S_{DR} with S_{EC} we have the equality-constrained linear program (Gill et al., 1991),

$$\begin{aligned} p = \max_{\mathbf{x},\mathbf{x}'} \ & \mathbf{c}^T(\mathbf{x}-\mathbf{x}') \\ & \mathbf{A}\mathbf{x} = \mathbf{b} \\ & \mathbf{A}\mathbf{x}' = \mathbf{b}. \end{aligned} \tag{10}$$

Unless $\mathbf{c}$ is in the rowspace of $\mathbf{A}$, this optimization problem will be unbounded. So the optimality condition for Eq. (10) is the now-familiar requirement that a solution can be found to $\mathbf{A}^T\mathbf{y} = \mathbf{c}$.

It is not useful to relax the ambiguity in the equality-constrained case (there, p is either zero or infinite), but for S_{DR} we can form general conditions for $\mathbf{c} \in R_S(\epsilon)$ incorporating a relaxed upper limit $p \leq \epsilon$. The analogous set of conditions achieved using S_{DR} instead of S_{EC}, and allowing a non-zero ϵ, is the following, as we will show with Theorem 3,

$$\begin{aligned} \mathbf{b}^T(\mathbf{y}+\mathbf{y}') + \alpha_1(\lambda_1+\lambda_1') + \alpha_2(\lambda_2+\lambda_2') \leq \epsilon \\ \|\mathbf{A}^T\mathbf{y} - \mathbf{c}\|_\infty \leq \lambda_1 \\ \|\mathbf{A}^T\mathbf{y}' + \mathbf{c}\|_\infty \leq \lambda_1' \\ \|\mathbf{y}\|_2 \leq \lambda_2 \\ \|\mathbf{y}'\|_2 \leq \lambda_2'. \end{aligned} \tag{11}$$

This is the generalization of the rowspace condition $\mathbf{A}^T\mathbf{y} = \mathbf{c}$ to $R_S(\epsilon)$. We show this with the following theorem,

Theorem 3. *If there exists a* $\mathbf{y}$, $\mathbf{y}'$, λ_1, λ_1', λ_2 *and* λ_2' *such that* $\mathbf{c}$ *is a solution to Eq. (11), then* $\mathbf{c} \in R_S(\epsilon)$.

Proof. We can test the limits of $\mathbf{c}^T\mathbf{x}$ with the optimization problem of Eq. (9) with $S = S_{DR}$,

$$\begin{aligned} p = \max_{\mathbf{x},\mathbf{x}'} \ & \mathbf{c}^T(\mathbf{x}-\mathbf{x}') \\ & \|\mathbf{x}\|_1 \leq \alpha_1 \\ & \|\mathbf{x}'\|_1 \leq \alpha_1 \\ & \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2 \leq \alpha_2 \\ & \|\mathbf{A}\mathbf{x}' - \mathbf{b}\|_2 \leq \alpha_2. \end{aligned} \tag{12}$$

By forming the dual (Boyd and Vandenberghe, 2004) of the optimization problem in Eq. (12), we can get an upper bound on the optimal. The dual optimization problem is

$$\begin{aligned} d = \min_{\substack{\mathbf{y},\mathbf{y}', \\ \lambda_1,\lambda_1',\lambda,\lambda'}} \ & \left\{\mathbf{b}^T(\mathbf{y}+\mathbf{y}') + \alpha_1(\lambda_1+\lambda_1') + \alpha_2(\lambda+\lambda')\right\} \\ & \|\mathbf{A}^T\mathbf{y} - \mathbf{c}\|_1^* \leq \lambda_1 \\ & \|\mathbf{A}^T\mathbf{y}' + \mathbf{c}\|_1^* \leq \lambda_1' \\ & \|\mathbf{y}\|_2^* \leq \lambda_2 \\ & \|\mathbf{y}'\|_2^* \leq \lambda_2', \end{aligned} \tag{13}$$

where $\|\cdot\|^*$ is the dual norm of $\|\cdot\|$. For example the dual norm for the ℓ_1-norm is the ℓ_∞-norm (we provide this general result to allow for other choices of norms if desired). The weak duality condition (Boyd and Vandenberghe, 2004), which always holds, tells us that $p \leq d$. By constraining the objective of Eq. (13) to be bounded by ϵ, which would mean $\mathbf{c} \in R_S(\epsilon)$, we get the conditions of Eqs. (11). □

Note that strong duality holds for convex optimization problems as well as many non-convex problems under a broad set of conditions, the most well-known being Slater's condition (Boyd and Vandenberghe, 2004). When strong duality holds, the converse of Theorem 3 holds as well, and so our conditions hold for all elements of $R_S(\epsilon)$. We will presume this is the case. Next we will consider the choice of useful components which fulfill the conditions of Eqs. (11).

2.5. Optimizing components

The motivation and strategy of our approach are summarized in Fig. 3. For a hypothetical example, we give several potential solutions to $\mathbf{A}\mathbf{x} = \mathbf{b}$ which we denote as $\mathbf{x}_{(i)}$, where $\mathbf{A}\mathbf{x}_{(i)} = \mathbf{b}$ for all i. Recall each solution is a potential mechanism. For example, potential solution $\mathbf{x}_{(1)}$ suggests the mechanism is based on activity at voxel k_2, while potential solution $\mathbf{x}_{(4)}$ suggests the mechanism is based on activity at voxel k_1. Choosing any single solution such as the MAP solution is risky as it may lead us to the wrong voxel if incorrect. We can analyze the risk of a choice of solution by looking at its correlation with other high-probability solutions. If we treat the choice of solution as an alternative candidate for a component, we may equivalently evaluate this risk by examining the marginal distribution of the score computed by this component. In Fig. 3 we depict multiple options for components; $\mathbf{c}_{(1)}$ is the MAP solution, while $\mathbf{c}_{(2)}$ calculates the an average over a large region. Fig. 3 also gives the distributions for the scores computed by these components (i.e., the range of correlations with potential solutions), where we see that these two result in broad distributions. Such broad distributions imply a wide possible variation in correlation with the different possible solutions, meaning these components are not very unambiguous, and hence are poor representatives of the set of high-probability solutions. Our strategy is depicted in Fig. 3 by the distribution for $P(\mathbf{c}^{*T}\mathbf{x})$, where we set desired limits on the spread of the correlations, and use optimization to choose an optimal $\mathbf{c}^*$ which fulfills these requirements. We can implement this strategy, therefore, using the conditions we derived in the previous section, recalling the relationship between the constraint sets and the distributions of Eq. (4).

Further, in order to choose the unambiguous component which is most useful for identifying properties of the mechanism, not only do we seek a component which is unambiguous, we also wish for the correlation with potential solutions to be maximized. To motivate this approach, consider the following robust strategy. Suppose instead of seeking an unambiguous component, we simply sought a component which had high correlations with all possible $\mathbf{x} \in S_{DR}$ using the following minimax strategy,

$$\mathbf{x}^* = \underset{\|\mathbf{c}\|=1}{arg\max} \ \min_{\mathbf{x} \in S_{DR}} \mathbf{c}^T\mathbf{x} \tag{14}$$

Here we seek the $\mathbf{c}$ for which the worst-possible correlation with high-probability solutions $\mathbf{x} \in S_{DR}$ is maximized. If we exchange the maximization and minimization to get $\min_{\mathbf{x} \in S_{DR}} \max_{\|\mathbf{c}\|=1} \mathbf{c}^T\mathbf{x}$, then we can analytically solve the inner maximization to get $\min_{\mathbf{x} \in S_{DR}} \|\mathbf{x}\|_2$, which is a variation on the elastic net of Eq. (6). So the elastic net can be viewed as a robust strategy to choose a vector $\mathbf{c}$ which maximizes the worst-case correlation with possible mechanisms, defined by S_{DR}.

For an unambiguous component, the correlations $\mathbf{c}^T\mathbf{x}$ will be approximately equal for all $\mathbf{x} \in S_{DR}$ by design. Therefore, we can simply choose an unambiguous component with maximum

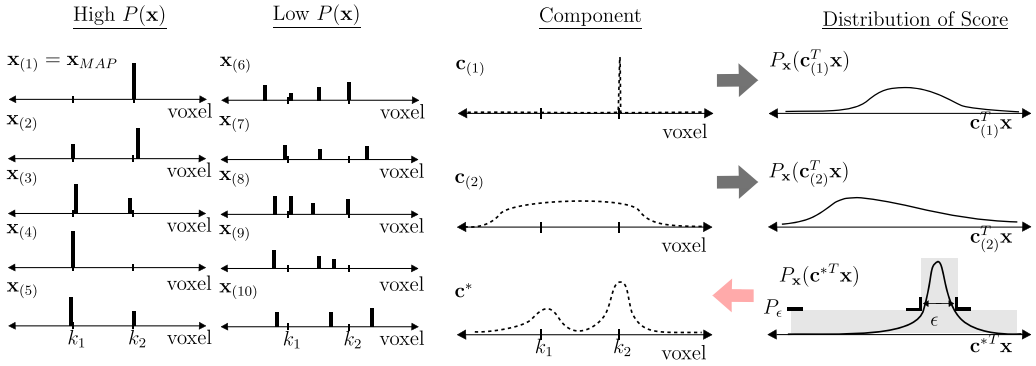

Fig. 3. Example giving several solutions $\mathbf{x}$ with higher and lower $P(\mathbf{x})$, candidate components, and the marginal probability distribution of feature values. The broader the spread of the values for a given feature, the more ambiguous the feature. The design problem of calculating $\mathbf{c}$ based on constraints ϵ and P_ϵ (which is based on α_1 and α_2), amounts to finding a component who's value has a controlled statistical spread.

correlation using the following optimization (which we will call the UMAX problem, for Unambiguous Maximum correlation),

$$\begin{aligned} \mathbf{c}_{UMAX} = \ & \underset{\mathbf{c},\mathbf{x}}{\operatorname{argmax}} \ \mathbf{c}^T\hat{\mathbf{x}} \\ & \mathbf{c} \in R_S \\ & \|\mathbf{c}\| \leq 1. \end{aligned} \tag{15}$$

where $\hat{\mathbf{x}}$ is any solution in S_{DR}, such as the LASSO solution $\mathbf{x}_1$. We also relaxed the unit length constraint to an inequality constraint $\|\mathbf{c}\| \leq 1$, making the optimization convex. In the extreme case where S_{DR} is small and only contains close approximations to $\mathbf{x}_1$, the optimal $\mathbf{c}$ will converge to $\mathbf{x}_1$.

Optimization also provides an opportunity to impose additional properties on the component vector $\mathbf{c}$. In particular, we may independently impose sparsity of our component (not to be confused with the sparsity prior which applies to the mechanism). In Eq. (16), we trade off a degree of sparsity on $\mathbf{c}$ and correlation with the mechanism, based on a choice of regularization parameter η.

$$\begin{aligned} \mathbf{c}_\eta = \ & \underset{\substack{\mathbf{c},\mathbf{y},\mathbf{y}',\\ \lambda_1,\lambda_1',\lambda_2,\lambda_2'}}{\operatorname{arg\,max}} \ \left\{\mathbf{c}^T\mathbf{x}_0 - \eta\|\mathbf{c}\|_1\right\} \\ & \mathbf{b}^T(\mathbf{y}+\mathbf{y}') + \alpha_1(\lambda_1+\lambda_1') + \alpha_2(\lambda_2+\lambda_2') \leq \epsilon \\ & \|\mathbf{A}^T\mathbf{y} - \mathbf{c}\|_\infty \leq \lambda_1 \\ & \|\mathbf{A}^T\mathbf{y}' + \mathbf{c}\|_\infty \leq \lambda_1' \\ & \|\mathbf{y}\|_2 \leq \lambda_2 \\ & \|\mathbf{y}'\|_2 \leq \lambda_2' \\ & \|\mathbf{c}\|_2 \leq 1. \end{aligned} \tag{16}$$

For comparison, the following provides a similar optimization without the imposition of prior knowledge,

$$\begin{aligned} \mathbf{c}_{EC}^* = \ & \underset{\mathbf{c},\mathbf{y}}{\operatorname{argmax}} \ \left\{\mathbf{c}^T\mathbf{x}_0 - \eta\|\mathbf{c}\|_1\right\} \\ & \mathbf{A}^T\mathbf{y} = \mathbf{c} \\ & \|\mathbf{c}\|_2 \leq 1. \end{aligned} \tag{17}$$

This component is simply the best choice according to our heuristic which can be found within the rowspace. Techniques which select combinations of principal components such as sparse PCA or supervised PCA, essentially form a variation of Eq. (17).

3. Results

In this section we will demonstrate the method, first with simulations which illustrate the ability to control the variance of the component, then with real data where we compare the method to LASSO and elastic net for feature selection.

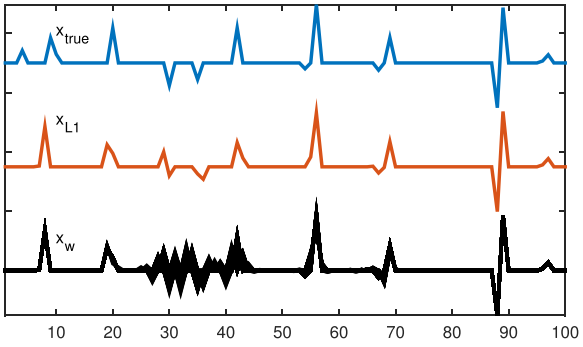

Fig. 4. True $\mathbf{x}$, Basis Pursuit solution ($\mathbf{x}_{L1}$), and various other randomly-found solutions ($\mathbf{x}_w$) plotted on top of each other, where $\|\mathbf{x}_{L1}\|_1 = \|\mathbf{x}_w\|_1$.

3.1. Simulation

First, we performed a simulation which demonstrated the robustness of the unambiguous component (i.e., the limited spread of its correlation with all other potential mechanisms). For this example, the minimum ℓ_1-norm (Basis Pursuit, Chen et al., 2001) solution is itself not unique. To explore this, we use this solution set as S_{DR}, so $\alpha_2 = 0$, and α_1 computed from the Basis Pursuit solution, $\alpha_1 = \min_{\mathbf{x}} \|\mathbf{x}\|_1$ *subject to* $\mathbf{Ax} = \mathbf{b}$. The matrix $\mathbf{A}$ is 20×100 with binary random elements $A_{ij} \in \{-1, +1\}$. Such highly-structured matrices often fail to yield unique solutions for ℓ_1-norm based techniques (Dillon and Wang, 2016). The true solution $\mathbf{x}_{true}$ is a sparse vector, plotted in Fig. 4, and $\mathbf{b} = \mathbf{Ax}_{true}$. Fig. 4 also gives the Basis Pursuit solution, and a variety of other solutions that have the same ℓ_1-norm (computed randomly), to demonstrate the non-uniqueness of the Basis Pursuit solution for this system.

In Fig. 5 we give the components computed via Eq. (16) with $\eta = 0.1$ ($\mathbf{c}_{0.1}$) and with $\eta = 0$ ($\mathbf{c}_{max}$). We used $\epsilon = 0.01$ to allow for finite numerical precision. We also computed the analogous maximum component in the rowspace, $\mathbf{c}_{EC}$, using $\eta = 0$ in Eq. (17).

To demonstrate the robust behavior of the components, we used the optimization approach of Eq. (12) to measure the range of values each component $\mathbf{c}^T\mathbf{x}$ could take. We repeated this optimization over

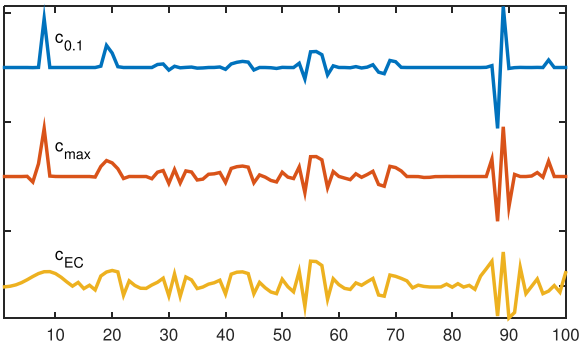

Fig. 5. Different candidates for computing feature.

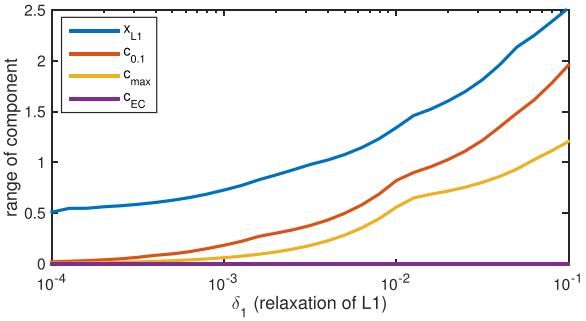

Fig. 6. $p(\delta_1)$ computed via Eq. (12) for component values $\mathbf{c}^T\mathbf{x}$ over a relaxed version of S_{DR} based on using $\alpha_1 + \delta_1$, legend entries are arranged in same order as plot traces, from top to bottom.

a range of relaxations of the prior knowledge, where we replaced α_1 with $\alpha_1 + \delta_1$ as in the following,

$$\begin{aligned} p(\delta_1) = \quad & \max_{\mathbf{x},\mathbf{x}'} \ \mathbf{c}^T(\mathbf{x}-\mathbf{x}') \\ & \|\mathbf{x}\|_1 \le \alpha_1 + \delta_1 \\ & \|\mathbf{x}'\|_1 \le \alpha_1 + \delta_1 \\ & \|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2 \le \alpha_2 \\ & \|\mathbf{A}\mathbf{x}'-\mathbf{b}\|_2 \le \alpha_2. \end{aligned} \tag{18}$$

So when $\delta_1 = 0$, we have a test of the correlations with all possible solutions that have the same ℓ_1-norm as the Basis Pursuit solution. In Fig. 6 we see that the unambiguous components we computed achieve $p \le 0.01$ for $\delta_1 \to 0$, as they were designed to, and as does the rowspace solution $\mathbf{c}_{EC}$. Once this prior knowledge is relaxed by increasing δ_1, however, $\mathbf{c}_{max}$ and $\mathbf{c}_{0.1}$ cease to yield unique values for $\mathbf{c}^T\mathbf{x}$. For comparison we also demonstrate the correlation with the Basis Pursuit solution itself, $\mathbf{x}_{L1}$. We see that it always results in a higher range p for the ambiguity, which demonstrates that $\mathbf{x}_{L1}$ is a poorer choice for a component, as its correlation varies significantly more over the possible predictors.

Next we repeated the example but with more relaxed choices for S_{DR} in the component design, achieved by choosing larger values for α_1. We also used different values for ϵ, the constraint on allowable variation in correlations. The results for three different combinations are given in Fig. 7 for three different combinations of ϵ and α_1. For each example, we chose a new pair of both α_1' and ϵ, where α_1' is relaxed by a given amount versus the Basis Pursuit optimal α_1 used in the previous example. We again plotted the performance of these components over increasingly-relaxed versions of S_{DR}, where we see the features are bounded by the chosen amount of relaxation within the limit given by the chosen ϵ. Further we see that the choice of ϵ and α_1 can control the robustness of the feature beyond the constraint set, as the range of correlations

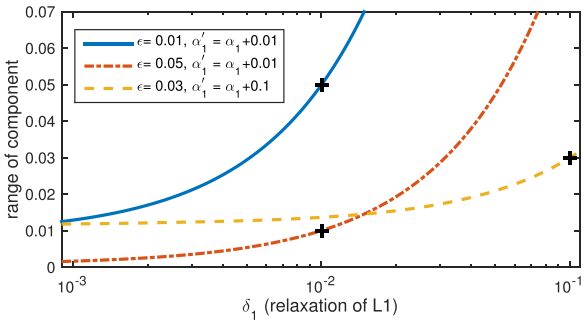

Fig. 7. $p(\delta_1)$ computed via Eq. (12) for component values $\mathbf{c}^T\mathbf{x}$ over a relaxed set using three different combinations of δ_1 and ϵ.

smoothly increase with a curve which depends on the conditions used. This suggests that by carefully designing S_{DR}, we can calculate components which robustly compute the information we seek.

3.2. Biomarker identification from real fMRI data

Next we will use the calculated components for finding regions which most accurately classify diseases. We consider a dataset consisting of functional MRI images for a number of subjects which are labeled as cases or controls. We will first use cross-validation to optimize a LASSO regression estimate, in order to find a MAP estimate with the best possible accuracy in differentiating cases versus controls, and this will set the standard we wish to improve on. Then with this as a starting point, we will relax the prior knowledge a controlled amount and use cross-validation to test for improvement in terms of accuracy with unambiguous components. Accuracy is directly computed using the sign of the component score $\mathbf{c}^T\mathbf{a}_i$ versus the true class, where $\mathbf{a}_i$ is an image in the test set. The goal is to take advantage of the behavior we saw in the simulations as the prior is relaxed, to improve robustness while at the same time maximizing the ability to discriminate cases versus controls.

We used the data from a study comparing psychiatric patients to controls during an auditory sensorimotor task, conducted by The Mind Clinical Imaging Consortium (MCIC) (Gollub et al., 2013). The study included 208 participants, 92 of whom were diagnosed with schizophrenia, schizophreniform or schizoaffective disorder; the remaining 116 were healthy controls. The fMRI data were preprocessed using the statistical parametric mapping (SPM) software (Penny et al., 2011); contrast images associated with the auditory stimuli form the data samples we use. Further details of the data collection and preprocessing can be found in our previous study (Chen et al., 2012).

Fig. 8 provides projections of the components or predictors calculated with several different methods. For PCA, Logistic LASSO, and Elastic Net, we used Matlab (Grace, 1992). The sparse principal component was calculated using the SpaSM toolbox (Sjstrand et al., 2012), with a setting of 100 nonzero voxels. The regularization parameters for all methods were chosen using 10-fold cross-validation. For UMAX (the proposed method) we started from the norms of the LASSO solution, then performed a line search for the optimal choice of relaxation, and picked the result with highest accuracy.

Both the first principal component and the sparse principal component took largest values in the occipital lobe, unlike the supervised methods, suggesting the signal variability was highest there, but not significantly related to the phenotype. The univariate correlation was highest in the precentral and postcentral gyrus, as noted in Chen et al. (2012). This was also evident in the LASSO, Logistic elastic, and UMAX results. The logistic elastic and UMAX

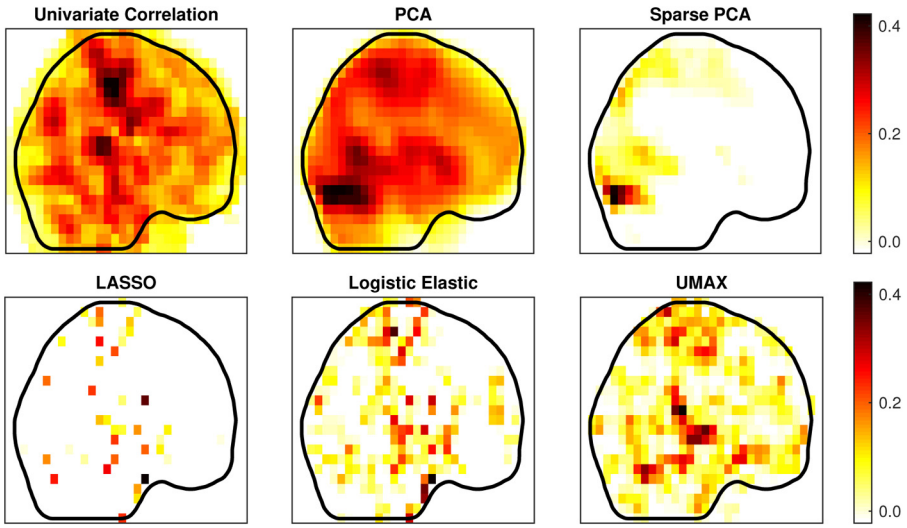

Fig. 8. Projections of components calculated using: Pearson correlation, PCA, Sparse PCA, LASSO, Logistic Elastic Net, and UMAX (proposed method).

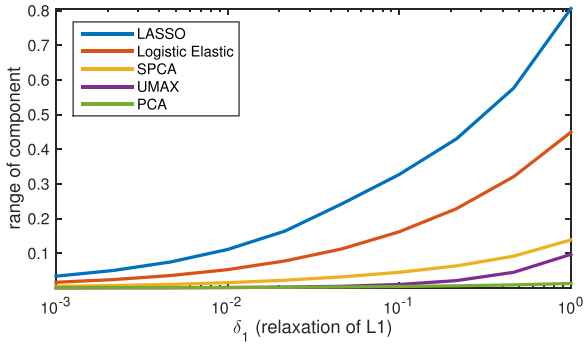

Fig. 9. Component ranges for different elements versus sets of L1-regularized solutions; legend entries are arranged in same order as plot traces, from top to bottom.

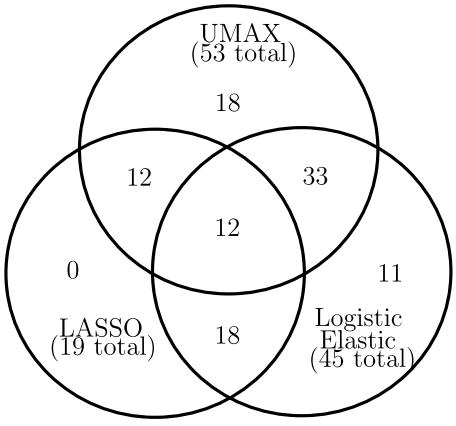

Fig. 10. Intersections and differences in ROI containing the nonzero signal for the different components; the elastic net result is essentially a superset of LASSO, while the UMAX overlaps with roughly two thirds of each.

Table 1
Methods and selectivity of different components compared to disease status.

Method	Accuracy	Correlation
PCA	48.6 %	−0.08
Sparse PCA	54.3 %	0.03
LASSO	62.1 %	0.23
Elastic	62.5 %	0.26
Logistic elastic	65.4 %	0.31
UMAX	69.2 %	0.35

result also included more of the neighboring sensory and motor association cortex, and UMAX in particular included a much higher weight to regions of the prefrontal cortex. In Fig. 9 we give the results of testing these components over relaxations of S_{DR}, suggesting how robust they are to the actual degree of sparsity of the true predictor (measured by relaxations of the ℓ_1-norm). Again we see the controlled ambiguity of the UMAX method. Interestingly, we see that the sparse PCA component had higher ambiguity.

To compare the components' value as biomarkers or for inspiring further research into particular regions, we considered their accuracy at finding regions relevant to the disease. Hence we tested the other features' classification performance directly by comparing the sign of the feature score $\mathbf{c}^T\mathbf{a}_i$ to the sign of b_i, where $\mathbf{a}_i$ is a sample in our test set, $\mathbf{c}$ is a feature tested, and $b_i \in \{+1, -1\}$ is the phenotype for the sample. In Table 1 we give the best accuracy (defined as the fraction of total test samples which were correctly classified, as used to determine the parameter values) and Pearson correlation achieved by each of the six features in 10-fold cross validation. As expected, we find that the PCA and sparse PCA components are not relevant to the disease. The LASSO solution performed fairly poorly and extensions such as elastic net (which allow a denser solution) and Logistic Elastic Net, yielded minor improvements. Relaxation using the UMAX achieved a more significant improvement, both in terms of accuracy and correlation. We note that the overall accuracies achieved here are in line with meta-analyses (Schnack and Kahn, 2016), which generally suggest accuracies of 65–70 percent for samples of this size.

Next we identified the regions of interest (ROI) containing the signal for each of the components, using the Automated Anatomical Labeling (AAL) parcellation (Tzourio-Mazoyer et al., 2002). In Fig. 10 we give the number of common and different ROI for the three methods. We see that the elastic net essentially identifies a superset of the LASSO selections, while the proposed method only agrees with about two-thirds of each. Projections of the common and different ROI (for a comparison of UMAX versus Logistic Elastic Net) are provided in Fig. 11, and the names of ROI's are provided in the appendix.

4. Discussion

In this paper we provided a robust supervised framework for utilizing prior knowledge to find pertinent components of

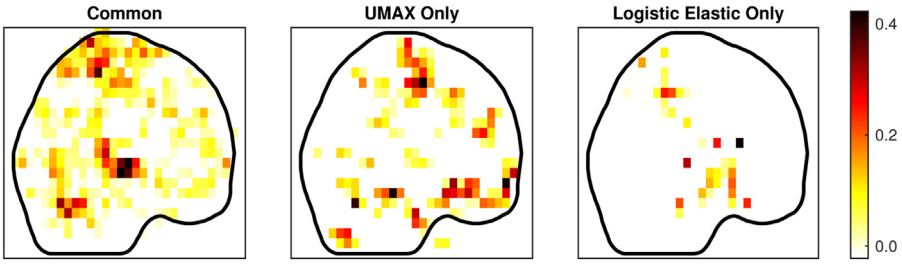

Fig. 11. Projections of components contained in ROI's common to Logistic Elastic Net and UMAX estimate, and projections of those in only one result but not other.

biological mechanisms. We demonstrated promising preliminary results using the technique for neuroimaging data, where we found that the choice of a maximum correlation component provided better accuracy as a classifier of case versus control. The use of the ℓ_1-norm assumes a Laplacian prior, and therefore common significance testing approaches, which presume Gaussian statistics, are not applicable. Hence we used the best accuracy seen in cross-validation as the metric to determine performance. We found a limited degree of relaxation of the prior provides improvement, as the simulations suggested we might, particularly given the fact that the prior appeared to be of limited effectiveness. As the scope of this paper is the introduction of a broadly-applicable method, where many possible forms of prior may be used, we did not delve more deeply into the more specific issues of ℓ_1-regression and associated statistical issues. In future work we intend to focus on both the development of more specialized priors, as well as associated significance testing approaches.

Potential difficulties of the method relate to the computational complexity, both in terms of algorithmic complexity, as well as the number of parameters to be chosen. We can view the latter (indeed all variables except **c**) as internal variables chosen by the algorithm itself, for example via cross-validation as in the previous section. The advantages of the method result from its rigorous formulation as a combination of dimensionality reduction with prior knowledge. We saw that this provides significant advantage for extremely noisy and underdetermined problems such as classification with neuroimaging data.

Acknowledgements

The authors wish to thank the NIH (R01 GM109068, R01 MH104680, R01 MH107354) and NSF (1539067) for their partial support.

Appendix: ROI Listing.

Table 2
ROI common to UMAX and Logistic Elastic Net.

Postcentral_R	Temporal_Sup_R	Postcentral_L
Frontal_Mid_L	Temporal_Sup_L	Precentral_L
Cingulum_Mid_L	Temporal_Mid_L	Paracentral_Lobule_L
Cuneus_L	Precuneus_L	Precuneus_R
Cerebelum_6_L	Frontal_Sup_R	Parietal_Sup_R
Calcarine_L	Cingulum_Post_L	Cerebelum_6_R
Frontal_Sup_Medial_L	Supp_Motor_Area_R	Cerebelum_Crus1_L
Frontal_Inf_Tri_R	Frontal_Med_Orb_R	Temporal_Pole_Sup_L
Cerebelum_4_5_R	Fusiform_R	Lingual_L
Occipital_Mid_L	Paracentral_Lobule_R	Temporal_Mid_R
Lingual_R	Rolandic_Oper_L	Caudate_R

Table 3
ROI in either UMAX or Logistic Elastic Net which is not common to other.

UMAX only	Logistic elastic only
Frontal_Mid_R	Parietal_Inf_L
Precentral_R	Olfactory_L
Temporal_Inf_R	Insula_R
Frontal_Med_Orb_L	Caudate_L
Cerebelum_Crus2_L	Parietal_Sup_L
ParaHippocampal_R	Putamen_L
Supp_Motor_Area_L	Thalamus_R
Cingulum_Ant_L	Cerebelum_3_R
Olfactory_R	SupraMarginal_R
Parietal_Inf_R	Rolandic_Oper_R
Frontal_Sup_Orb_L	Temporal_Pole_Sup_R
Fusiform_L	Hippocampus_L
Frontal_Sup_Orb_R	
Cerebelum_4_5_L	
SupraMarginal_L	
Calcarine_R	
Frontal_Inf_Orb_L	
Cingulum_Mid_R	
Rectus_R	
Insula_L	

References

Bair, E., Hastie, T., Paul, D., Tibshirani, R., 2006. Prediction by supervised principal components. J. Am. Stat. Assoc. 101 (473), 119–137.
Barshan, E., Ghodsi, A., Azimifar, Z., Jahromi, M.Z., 2011. Supervised principal component analysis: visualization, classification and regression on subspaces and submanifolds. Pattern Recogn. 44 (7), 1357–1371.
Boyd, S.P., Vandenberghe, L., 2004, March. Convex Optimization. Cambridge University Press.
Buck, S., 2015. Solving reproducibility. Science 348 (6242), 1403.
Cao, H., Duan, J., Lin, D., Shugart, Y.Y., Calhoun, V., Wang, Y.-P., 2014. Sparse representation based biomarker selection for schizophrenia with integrated analysis of fMRI and SNPs. NeuroImage 102 (Pt1 (November)), 220–228.
Chen, S.S., Donoho, D.L., Saunders, M.A., 2001. Atomic decomposition by basis pursuit. SIAM Rev. 43 (January (1)), 129–159.
Chen, J., Calhoun, V.D., Pearlson, G.D., Ehrlich, S., Turner, J.A., Ho, B.-C., Wassink, T.H., Michael, A.M., Liu, J., 2012. Multifaceted genomic risk for brain function in schizophrenia. NeuroImage 61 (4), 866–875.
Chu, C., Hsu, A.-L., Chou, K.-H., Bandettini, P., Lin, C., 2012. Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. NeuroImage 60 (1), 59–70.
De Angelis, P.L., Pardalos, P.M., Toraldo, G., 1997. Quadratic programming with box constraints. In: Bomze, I.M., Csendes, T., Horst, R., Pardalos, P.M. (Eds.), Developments in Global Optimization, Number 18 in Nonconvex Optimization and Its Applications. Springer, US, pp. 73–93.
Dillon, K., Fainman, Y., 2013. Bounding pixels in computational imaging. Appl. Opt. 52 (April (10)), D55–D63.
Dillon, K., Fainman, Y., 2016, April. Element-wise uniqueness, prior knowledge, and data-dependent resolution. Signal Image Video Process., 1–8.
Dillon, K., Wang, Y.-P., 2016, June. Imposing uniqueness to achieve sparsity. Signal Process. 123, 1–8.
Dunteman, G.H., 1989, May. Principal Components Analysis. SAGE.
Gill, P.E., Murray, W., Wright, M.H., 1991. Numerical Linear Algebra and Optimization. Addison-Wesley Pub. Co., Advanced Book Program.
Gollub, R.L., Shoemaker, J.M., King, M.D., White, T., Ehrlich, S., Sponheim, S.R., Clark, V.P., Turner, J.A., Mueller, B.A., Magnotta, V., OLeary, D., Ho, B.C., Brauns, S., Manoach, D.S., Seidman, L., Bustillo, J.R., Lauriello, J., Bockholt, J., Lim, K.O., Bruce, R., Rosen, S., Schulz, C., Calhoun, V.D., Andreasen, N.C., 2013. The MCIC collection: a shared repository of multi-modal, multi-site brain image data

from a clinical investigation of schizophrenia. Neuroinformatics 11 (3), 367–388.
Grace, Andrew, 1992. MATLAB Optimization Toolbox. The MathWorks Inc., Natick, USA.
Guyon, I., Elisseeff, A., 2003, March. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182.
Kowalski, M., 2009. Sparse regression using mixed norms. Appl. Comput. Harm. Anal. 27 (November (3)), 303–324.
Krystal, J.H., State, M.W., 2014. Psychiatric disorders: diagnosis to therapy. Cell 157 (1), 201–214.
Lemm, S., Blankertz, B., Dickhaus, T., Mller, K.-R., 2011. Introduction to machine learning for brain imaging. NeuroImage 56 (2), 387–399.
Lin, D., Cao, H., Calhoun, V.D., Wang, Y.-P., 2014, November. Sparse models for correlative and integrative analysis of imaging and genetic data. J. Neurosci. Methods 237, 69–78.
Ma, S., Dai, Y., 2011. Principal component analysis based methods in bioinformatics studies. Brief. Bioinform. 12 (6), 714–722, November.
Milliken, G.A., Johnson, D.E., 2009, March. Analysis of Messy Data Volume 1: Designed Experiments, Second ed. CRC Press.
Orr, G., Pettersson-Yeo, W., Marquand, A.F., Sartori, G., Mechelli, A., 2012. Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. Neurosci. Biobehav. Rev. 36 (4), 1140–1152.
Penny, W.D., Friston, K.J., Ashburner, J.T., Kiebel, S.J., Nichols, T.E., 2011, April. Statistical Parametric Mapping: The Analysis of Functional Brain Images: The Analysis of Functional Brain Images. Academic Press.
Radhakrishna Rao, C., Toutenburg, H., 2013, June. Linear Models: Least Squares and Alternatives. Springer Science & Business Media.
Schnack, H.G., Kahn, R.S., 2016. Detecting Neuroimaging Biomarkers for Psychiatric Disorders: Sample Size Matters. Neuroimaging and Stimulation., pp. 50.
SpaSM, http://www2.imm.dtu.dk/projects/spasm/, last accessed November 18, 2016.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B (Methodol.) 26, 7–288.
Tibshirani, R.J., 2013. The lasso problem and uniqueness. Electron. J. Stat. 7, 1456–1490.
Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002 January. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. NeuroImage 15 (1), 273–289.
Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. J. R. Stat. Soc.: Ser. B (Stat. Methodol.) 68 (February (1)), 49–67.
Zhang, H., Yin, W., Cheng, L., 2014. Necessary and sufficient conditions of solution uniqueness in 1-norm minimization. J. Optim. Theory Appl. 164 (1), 109–122.
Zou, H., Hastie, T., 2005. Regularization and variable selection via the Elastic Net. J. R. Stat. Soc. Ser. B 67, 301–320.
Zou, H., Hastie, T., Tibshirani, R., 2006. Sparse principal component analysis. J. Comput. Graph. Stat. 15 (2), 265–286.