Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

A Sparse Combined Regression-Classification Formulation for Learning a Physiological Alternative to Clinical Post-Traumatic Stress Disorder Scores

Sarah M. Brown
Northeastern University
Charles Stark Draper Laboratory
brown@ece.neu.edu

Andrea Webb and **Rami S. Mangoubi**
Charles Stark Draper Laboratory
Cambridge, MA
awebb,mougoubi@draper.com

Jennifer G. Dy
Northeastern University
Boston, MA
jdy@ece.neu.edu

Abstract

Current diagnostic methods for mental pathologies, including Post-Traumatic Stress Disorder (PTSD), involve a clinician-coded interview, which can be subjective. Heart rate and skin conductance, as well as other peripheral physiology measures, have previously shown utility in predicting binary diagnostic decisions. The binary decision problem is easier, but misses important information on the severity of the patients condition. This work utilizes a novel experimental set-up that exploits virtual reality videos and peripheral physiology for PTSD diagnosis. In pursuit of an automated physiology-based objective diagnostic method, we propose a learning formulation that integrates the description of the experimental data and expert knowledge on desirable properties of a physiological diagnostic score. From a list of desired criteria, we derive a new cost function that combines regression and classification while learning the salient features for predicting physiological score. The physiological score produced by Sparse Combined Regression-Classification (SCRC) is assessed with respect to three sets of criteria chosen to reflect design goals for an objective, physiological PTSD score: parsimony and context of selected features, diagnostic score validity, and learning generalizability. For these criteria, we demonstrate that Sparse Combined Regression-Classification performs better than more generic learning approaches.

Introduction

Prevalence rates of Post-Traumatic Stress Disorder (PTSD) in veterans returning from Iraq varies from $4 - 18\%$ (Richardson, Frueh, and Acierno 2010) and costs related to treatment for PTSD and depression in this population is estimated to be $923 million (Kilmer et al. 2011). Like other mental pathologies, PTSD is currently defined by behavioral symptoms as interpreted by a clinician (American Psychiatric Association 2013). For PTSD, a structured clinical interview is the current gold standard diagnostic tool. To diagnose, a clinician codes the patient's responses to questions that each address one of three categories of symptoms, counts the number of expressed symptoms per category and combines in a predetermined fashion to produce a clinical score, then compares the score to a threshold to reach a final diagnosis (healthy or PTSD). The clinical interview provides a score based on patient self-reports after a time consuming and expensive process, so a more objective, easier to administer diagnostic tool is sought. Although definitive physiological bases have yet to be characterized well enough to build a test, prior work has demonstrated that measurements of peripheral physiology can be used to predict diagnostic class assigned by the clinician (Orr, Metzger, and Pitman 2002; Webb et al. 2013; Pitman et al. 1987). We seek to advance this line of work by producing a score which provides a finer degree of granularity, necessary for treatment planning and monitoring.

The ultimate goal is a parsimonious but generalizable feature set and a function for synthesizing them into a single score that can assist in the diagnostic process. The problem at hand is somewhat circular; the gold standard for diagnosis is the clinical score, but the motivation for an alternate score is the weaknesses in that score. This requires a careful construction of the problem for learning and creates a more urgent need for a collaboration between an AI researcher and the domain experts – black box application of machine learning tools would provide an inadequate result for this important societal challenge.

Technically, this goal resembles two well studied machine learning tasks: feature selection and regression. As we understand the context of the problem however, we find that subtleties in the data available for learning and additional contextual problems motivate a novel formulation of the problem. For a well defined task and data that meets underlying distribution assumptions, an algorithm's relative performance can be reliably distilled to a single quantity. For a real world problem, including research questions in other disciplines, this is more challenging. Carefully constructing performance measures suited to the unique application area is an imperative step toward enabling the insights provided by the solution to have an impact in the application area (Wagstaff 2012).

To advance the state of the art in PTSD diagnosis, our solution should be a parsimonious, generalizable, and diagnostically valid model. We will refer to these as *solution desiderata*. Next we describe the specific experimental protocol we will use to develop a physiological PTSD score. We use the context of this experiment, expert knowledge about current diagnostic procedures, and the solution desiderata to formulate an application-specific technical problem in

the form of a list of *learning desiderata*. We propose a sparse combined regression and classification formulation that meets the *learning desiderata* and position it within related technical literature. Finally, we assess the suitability of the learning formulation through computational experiments using a set of performance measures that combine the long term objective to improve diagnostic procedures, and the specific goals of the current stage of the project (proof of concept) through the *solution desiderata*.

The contributions of this paper are: with respect to the application, (1) a candidate physiological scoring function as an alternative to the more expensive and time-consuming clinician interview-based gold standard clinical score; and with respect to machine learning, (2) a Sparse Combined Regression-Classification (SCRC) formulation that takes into account the desired properties of a physiological score as provided by a domain-expert; and (3) evaluation measures tailored to this application.

Problem Definition

While mental pathologies are presently defined by behavioral changes, we rely on the assumption that these changes are the result of a change in the way a person's brain processes various events. Current diagnostic methods require the clinician to ask the patient to introspectively assess past behavior as an indirect measure of this change in brain processing. We propose peripheral physiological signals as a more objective, though still indirect measure of this change. Experimentally, we expose the subjects to non-idiographic virtual reality videos thematically reminiscent of their trauma and record physiology in order to develop a proof of concept advance in the science. The underlying mechanisms of PTSD are not well understood so we limit our investigation to discriminative approaches.

In contrast to the current gold standard for diagnosis, (the Clinician Administered PTSD Scale (CAPS) (Blake et al. 1995)) which is computed from a clinician-coded structured clinical interview, we propose a diagnostic score objectively computed from measured signals. Specifically, the current task is to learn a scoring function for computing a physiological PTSD diagnostic score that is in agreement with current clinical understanding. First we present a detailed description of the data to define the problem with respect to the application and then a technical specification of the problem we can use to develop and assess the solution.

PTSD Study and Pre-processing

For this work, we use data from a pilot-scaled study examining physiological response of PSTD by presenting the subjects with virtual reality videos (Webb et al. 2013). In particular, male veteran subjects were shown two non-idiographic videos generated with the Virtual Iraq software originally designed for treatment applications[1]. One video was designed to emulate a foot patrol in a city setting, and the other was a humvee driving scenario. Each contained five increasingly intense events (i.e., helicopter flying over to insurgent firing a weapon) spaced approximately every 45 seconds.

All procedures were IRB approved and all subjects provided informed consent. Prior to the experimental protocol, a clinician-administered interview was conducted to determine the clinical score for each subject. The clinician codes subject responses into binary scores for each item, counts to produce sub-scores by category (predetermined groups of items) and combines them following a standard procedure to compute a total score (integer valued). During the experimental protocol, the BIOPAC system was used to measure physiological signatures from each subject, using only the two most expressive BIOPAC channels as determined by prior work: Electrocardiogram (ECG) and Galvanic Skin Response (GSR) (skin conductance) (Carson et al. 2000; Orr et al. 1998; Goldfinger, Amdur, and Liberzon 1998; Orr et al. 2000; Pole 2007).

The Inter-Beat-Interval (IBI) signal was computed from the ECG recording using automated peak-detection. The IBI and GSR signals were segmented into 20 second windows, called response curves, beginning at each video event. Eleven (11) features commonly used to study various psychological concepts, including peak amplitude, standard deviation, and area to full recovery were extracted from each of the ten (10) response curves (for a total of 110 features per subject) (Kircher and Raskin 1988; Webb et al. 2013). Fifty-seven ($D = 57$) features remained for use in learning a score after preliminary class-wise analyses (Webb et al. 2013). These are concatenated into a single feature vector for each subject.

To allow meaningful comparisons across subjects and physiological channels, the within-subject mean was subtracted from each feature and features were normalized by the within-subject standard deviation to produce the final feature vector, $\mathbf{x}_i \in \mathbb{R}^D$, for each subject $i \in 1 \ldots, N$. After removing subjects with missing data for any reason $N = 38$ subjects (22 trauma, 16 PTSD) remained for learning and assessment as described in the rest of this paper. The matrix constructed by concatenating data from all subjects is $\mathbf{X} \in \mathbb{R}^{D \times N}$ and the vector of all clinical scores is $\mathbf{y} \in \mathbb{R}^N$.

Learning Desiderata

Given the context afforded by expert knowledge of the problem, the experimental description above is an insufficient technical specification. To solve, we must also quantify the desired properties of the new score and the limitations of the measurements drawn from expert knowledge. We distill insight from expert consultation and data-set examination into the following four learning desiderata, including both desired properties for a physiological score, y_i^p, and limitations on how to learn from the clinical score, y_i^c (for subject i).

1. **Linearity**: Linear with respect to physiological features: $y_i^p = f(\mathbf{x}_i) = \mathbf{x}_i^T \beta$.
2. **Sparsity**: Dependent on only a small subset of the physiological features. Several $\beta_f = 0$.
3. **Severity**: Preserve ranking provided by clinical scores: $y_i^p = g(y_i^c)$, with g nondecreasing $y_i > y_j \rightarrow y_i^p > y_j^p$.

[1]The authors thank Albert Skip Rizzo at the Institute for Creative Technologies for providing the Virtual Iraq software used for the creation of the virtual reality videos.

4. **Ambiguity**: Identical clinical scores do not indicate identical status. Zero scores are especially non-specific, it only indicates that these subjects present no symptoms. $y_i^c = 0 \rightarrow y_i^p < \epsilon$ for some ϵ near 0.

This list is motivated by the solution desiderata and will serve as a framework for developing a learning algorithm. We expand upon each item by highlighting the key insights summarized and relation to the solution desiderata expressed in the introduction.

Linearity Prior work has suggested that linear models using a standard set of features works well for the binary diagnosis problem (Webb et al. 2013). As an application, as encapsulated by our solution desideratum of parsimony, it is important to provide a solution that is human interpretable, so that users will trust it (Giraud-Carrier 1998; Rudin and Wagstaff 2014); linear models fulfill this objective. Further, linear models are computationally attractive as they rely on fewer parameters than more complex choices.

Sparsity A sparse coefficient vector yields a score that is dependent on only a small number of features. This supports the solution desideratum for parsimony. Sparsity or feature selection is necessary because although after preprocessing we only have 57 features, this is large compared to only 38 subjects. Using all of the candidate features to learn the function parameters would over-fit to this experimental data thus contradicting our solution desideratum of generalizability.

Severity The gold standard clinical diagnostic tool, the CAPS score, may be interpreted as a ranking of the severity of subjects' condition because it does not have physically interpretable units, it is derived from a coded interview. Maintaining this ranking is essential to the solution desideratum of diagnostic validity. The distinction between the technical specification of the linearity and severity criteria is an important subtlety of the requirements. The learned function should be linear in the features, but it does not need to be linearly related to the clinical score.

Ambiguity Subjects with a zero clinical score experienced trauma, but present with no symptoms. The fact that they have the same CAPS score does not necessarily indicate the same underlying health status. The CAPS score was designed to diagnose unhealthy subjects, so it is nonspecific in this realm. Our physiological score need not lump these subjects tightly together, only ensure they are not scored near the subjects presenting with symptoms. This learning desideratum formalizes the distinction between achieving the solution desideratum of diagnostic validity and solving a prediction problem.

Combined Sparse Regression-Classification

General supervised learning frameworks posit that for data $(\mathbf{x}, y)$, where $\mathbf{x} \in \mathbb{R}^D$ is input data sample and y is the target output, there exists a function $y = f(\mathbf{x})$. In this setting, the goal is to use training samples of $(\mathbf{x}, y)$ pairs to uncover f or its parameters, so that for new observations of x the appropriate y can be predicted. Alternatively we interpret the description in the prior section to define data $(\mathbf{x}, y^c)$, a description of the relationship $y^p = g(y^c)$ and some properties of the form of f, where $y^p = f(\mathbf{x})$. The end goal is to compute $y^p = f(\mathbf{x})$ for a new observation of $\mathbf{x}$. However, learning f is less straight forward in this case. In other words, our target task output is the physiological score, y^p. However, our training data contains pairs of $(\mathbf{x}, y^c)$. The information on $y^p = g(y^c)$ and the form of f are represented in the list of learning desiderata.

Formulation

We propose that a unifying cost function combining regression loss, classification loss, and sparse regularization will meet all of the learning desiderata. Further, we assert that loss should be defined differently per subject, based on the information available. In (1a) we define our SCRC formulation using ℓ_2 regression loss, hinge classification loss, and an ℓ_1 regularization term. To denote the subjects with zero and nonzero clinical scores, we will use subscripts Z and $\bar{Z}$ respectively, $X = [X_Z X_{\bar{Z}}]$ up to subject permutation. For example $\mathbf{X}_{\bar{Z}}$ is the matrix of all of the features for the subjects with nonzero clinical scores and $y^c_{\bar{Z}}$ is the corresponding clinical score vector. We use $i \in Z$ to denote the set of indices corresponding to subjects with a zero clinical score, such that $Z \cup \bar{Z} = 1, \ldots, N$.

$$\min_{\beta} \ \gamma_1 \|\mathbf{X}_{\bar{Z}}^T \beta - \mathbf{y}_{\bar{Z}}^c\|_2^2 + \gamma_2 \sum_{i=1}^{N} \mathcal{L}_H(y_i^d, \mathbf{x_i}^T \beta) + \lambda \|\beta\|_1 \tag{1a}$$

$$\mathcal{L}_H(y_i^d, \mathbf{x_i}^T \beta) = \max(0, 1 - y_i^d \mathbf{x_i}^T \beta) \tag{1b}$$

$$y_i^d = \begin{cases} d & y_i^c > 0 \\ -d & y_i^c = 0 \end{cases} \tag{1c}$$

where $\mathcal{L}_H$, defined in (1b), is called hinge loss and y_i^d is a scaled ($d = 65$, middle of y^c range) binary classification label for each subject having a zero clinical score or not as defined in (1c). We use γ_1 and γ_2 to weigh those two terms based on the number of subjects used in each and find that solutions are not dependent on this choice.

Linearity After learning the β that satisfies the optimization problem in (1a), we compute a physiological score, y_i^p, that is linear with respect to the features, $\mathbf{x}_i$, $y_i^p = \mathbf{x}_i^T \beta$.

Sparsity To learn the subset of features that is most useful, we use a sparse regularization term: ℓ_1 in (1a). The λ controls the degree of sparsity in β, with λ inversely proportional to the number of features retained.

Severity As a function of y_i^c, this objective function is piecewise linear, which is nondecreasing; thus, meeting our severity-preserving criterion. The squared loss term is linear, hence in that range of y_i^c we expect $y_i^p = y_i^c, i \in \bar{Z}$, which preserves the ranking for those subjects. The hinge loss term uses a saturated version of y_i^c.

Ambiguity The hinge loss term in (1a) models the ambiguity in the clinical scores of zero by optimizing such

that $y_i^p < y_j^p, \forall i \in Z, j \in \bar{Z}$. For subjects without symptoms ($i \in \bar{Z}$), the physiological score only needs to be below a threshold, since these subjects are not included in the squared loss term which has specific target values.

Optimization

We solve the optimization problem using the Alternating Direction Method of Multipliers (ADMM) which solves unconstrained optimization problems by considering each term to operate on a local variable and constraining the solution to force consensus (Boyd 2010). The ADMM iterations for (1a) are shown in (2) with iteration indexed by superscript t. The first update is regularized least squares, or ridge regression, the second resembles a Support Vector Machine (SVM), and the third is a shrinkage operator.

$$\beta_1^{t+1} = \left(\mathbf{X}\mathbf{X}^T + \rho I\right)^{-1}\left(\mathbf{X}^T\mathbf{y} + \rho(\beta^t - \mathbf{u}_1^t)\right) \tag{2a}$$

$$\beta_2^{t+1} = \underset{\beta_2}{\operatorname{amin}} \sum_{i=1}^{N} \mathcal{L}_H\left(y_i^d, \mathbf{x_i}^T\beta_2\right) + \frac{\rho}{2}\|\beta_2 + \beta^t + \mathbf{u}_2^t\|_2^2 \tag{2b}$$

$$\beta^{t+1} = \mathrm{S}_{\frac{\lambda}{\rho}}(.5(\sum_i \beta_i^{t+1} - \sum_i u_i^k)) \tag{2c}$$

$$\mathbf{u}_{1,2}^{t+1} = \mathbf{u}_{1,2}^t + \beta_{1,2}^{t+1} + \beta^{t+1} \tag{2d}$$

where ρ is the augmented Lagrangian variable and controls how much the difference between solutions regularizes the next iteration thus only influencing convergence rate, not the final solution. The $\mathrm{S}_\kappa(a)$ is the soft threshold function. We use a stochastic gradient descent for the hinge loss term. ADMM convergences under nonrestrictive conditions: (1) the function in each term must be closed, proper and convex, and (2) the Lagrangian must have a saddle point (Boyd 2010). Squared loss, hinge loss and ℓ_1 are each closed, proper and convex and their independent solutions imply a saddle point in the ADMM form of the objective.

Related Work

There are key distinctions between SCRC and related technical solutions. First we note that combining the two loss functions into a single optimization problem is different from executing two learned models in sequence. Because the features for each task, regression and classification, are different, a subject that is near the decision boundary in the binary decision feature space can have a score based on the regression features that indicates a severe condition. If the binary decision boundary moved just a small amount, that subject would then be marked healthy- this sensitivity is undesirable. We avoid this by learning the features and weights that satisfy both objectives simultaneously.

Linear Regression and LASSO

Linear regression is a standard approach to learning coefficients for a linear combination of the input variable to predict the output variable. The Least Absolute Shrinkage and Selection Operator (LASSO) method adds a ℓ_1 regularization term to the squared loss term of linear regression (Tibshirani 1996). This modification produces a sparse result, thus performing feature selection integrated into the regression problem.

$$\min_\beta \|X^T\beta - y\|_2^2 + \lambda\|\beta\|_1 \tag{3}$$

Relative to LASSO the SCRC incorporates the additional knowledge that we have about the zero clinical scores by excluding them from the squared loss term and treating them as a classification problem.

Combined Regression and Ranking

Combined regression and ranking is a tunably (via α) weighted sum of two loss functions and a ℓ_2 regularization for model complexity tradeoff, shown in (4) (Sculley 2010). Any loss function can be used for either the regression term or the ranking term; they differ in how the data are used. A regression loss function is computed for each sample directly between the prediction $f(\beta, \mathbf{x_i})$ and the measured label, y_i. A ranking loss term is computed for each pair of samples from a prediction made on signed distance in feature space, $f(\beta, \mathbf{x}_i - \mathbf{x}_j)$, and a binary representation of their ranking $t(y_i, y_j)$.

$$\min_\beta\ (1-\alpha) \sum_{i,j \in \mathrm{P}(N)} \mathcal{L}_{\text{rank}}(t(y_i, y_j), f(\beta, \mathbf{x_i} - x_j)) + \lambda\|\beta\|_2^2 + \alpha \sum_{i=1}^{N} \mathcal{L}_{\text{reg}}(y_i, f(\beta, \mathbf{x_i})) \tag{4}$$

where $\mathrm{P}(N)$ is all of the unique i, j pairs for N samples. The SCRC is nearly a sparse variation of (4) using least squares loss for regression and hinge loss for ranking due to the similarity between a ranking loss and classification. However, we use additional knowledge about the context of the problem to ignore the regression loss for some and reduce the pairwise ranking loss to a subject-wise classification loss. This insight eliminates the need for a computational trick to speed up computation, since we are not comparing every pairwise relationship.

SVM and Rank SVM

A linear SVM uses hinge loss to find the projection that provides the maximum margin between the two classes of data so that a hyperplane can be applied as a decision boundary (Cortes and Vapnik 1995). RankSVM solves the problem of ranking a series of objects by considering the set of all pairwise comparisons and optimizing to preserve pairwise rankings problem that can thus be solved with an SVM (Joachims 2002). The optimization problem for RankSVM is:

$$\min_\beta \sum_{i,j \in \mathrm{P}(N)} \left(1 - (y_i - y_j)\beta^T(\mathbf{x}_i - \mathbf{x}_j)\right)_+ + \|\beta\|_2^2 \tag{5}$$

where $\mathrm{P}(N)$ is again all of the unique i, j pairs for N samples. The first term is hinge loss, $\mathcal{L}_{\mathrm{H}}((y_i - y_j), \beta^T(\mathbf{x}_i - \mathbf{x}_j))$ as in our formulation. SVM solutions are sparse, but with respect to the samples, not the features as in SCRC. Further, we add the additional condition that we also have a regression loss term.

Computational Experiments

A comprehensive, context sensitive evaluation is important, as the application is central to the contribution of this paper. Here we assess the solution provided by our SCRC formulation with respect to the *solution desiderata* defined in the introduction: parsimony and context, diagnostic validity, and generalizability. First we describe the general environment used for testing, then define how we assess and discuss the results relevant to each desideratum.

We present results comparing the SCRC only to LASSO as defined in (3). It is the closest approach in computation and provided outputs. The combined regression and ranking formulation with sparsity suffers similar challenges to LASSO as a key novelty is in the division of subjects. Both problems were solved using ADMM, the LASSO code is from supplemental materials for (Boyd 2010).

Results presented use $\rho = 1.0$, based on guidance in (Boyd 2010) and confirmation of solution insensitivity to this ADMM parameter. The data matrices $X_{\bar{Z}}$ and X_Z are augmented with a column of ones to provide a linear offset term, as is typical in linear modeling. We fit the models for 100 values of λ spaced uniformly in log-space in a range selected dependent on the infinity norm of the training data. We run tests for both $K = 7$ fold (folds balanced for diagnostic class) and leave-one-out cross validation.

Diagnostic Score Validity

To assess clinical validity, we compute average performance in the $K = 7$ fold cross validation results. As the primary measure of fit, we introduce $\mathrm{MSE_{NZ}}$ (6c): normal mean-square error for the subjects with a non-zero clinical score ($i \in \bar{Z}$) and one-sided error for the subjects with a zero clinical score ($i \in Z$). For subjects with a clinical score of zero, any negative physiological score is considered zero error. Per the severity desiridatum, ranking is important, so we compare the solutions on the Spearman correlation coefficient, ρ_S (6d). We test the physiological score in its ability to return a diagnostic classification that matches the clinical diagnosis ($d_i^c = y_i^c > \theta^c$) by comparing the learned score y_i^p to a range of thresholds (θ^i) to produce a physiological diagnosis and receiver-operating curve, from which we compute an area under the curve (AUC).

$$y_i^p = \beta^T \mathbf{x_i} \tag{6a}$$

$$A = \bar{Z} \cup \{i; y_i^p > 0\} \tag{6b}$$

$$\mathrm{MSE_{NZ}}(\beta) = \frac{1}{N} \sum_{i \in A} (y_i^p - y_i^c)^2 \tag{6c}$$

$$\rho_S = \mathrm{Lin}\left(\mathrm{Rank}\left(\mathbf{y}^p\right), \mathrm{Rank}\left(\mathbf{y}^c\right)\right) \tag{6d}$$

We denote Pearson linear correlation with $\mathrm{Lin}(\cdot)$ and define Rank as ascending with the position average assigned to ties (i.e.: $\mathrm{Rank}\left([10, 37, 25, 40, 25]\right) = [1, 4, 2.5, 5, 2.5]$). Figure 1 shows that SCRC outperforms (lower $\mathrm{MSE_{NZ}}$, higher AUC) LASSO for small values of λ. The smallest values of λ for which lasso is better in $\mathrm{MSE_{NZ}}$, it is worse in AUC, so in this range, *although the average fit is better, LASSO makes mistakes in more crucial areas, near the diagnostic threshold.* The performance of SCRC is nearly constant for small values of λ, suggesting a more robust learning scheme and more generalizable solution.

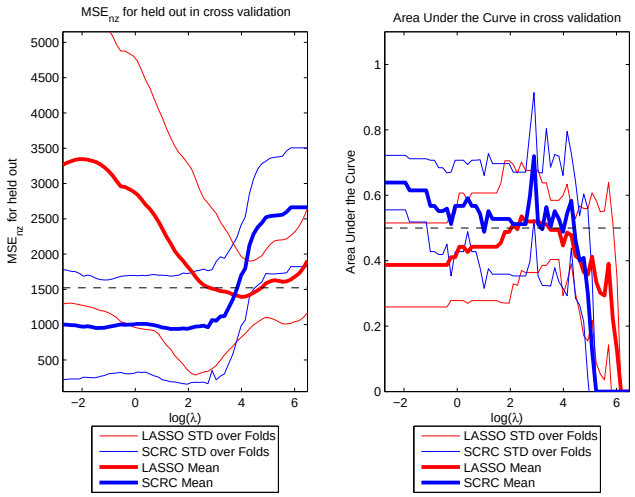

Figure 1: The model fit, as defined in (6c). *Left:* $\mathrm{MSE_{NZ}}$ and *Right* mean AUC, both versus $\log(\lambda)$. Red traces are the LASSO solution, blue are the SCRC. The bold traces are means and finer traces are +/- one standard deviation. The dashed line shows MSE (left) and AUC (right) for a naive prediction (sample mean/chance).

Learning Generalizabilty

We propose stability-related metrics to quantify generalizabilty based on numerous results linking notions of cross validation stability to the generalization power and statistical consistency of learning algorithms (Bousquet and Elisseeff 2002; Mukherjee et al. 2006; Lange et al. 2002). The intuition is that swapping out one sample is a small perturbation at the input of the optimization problem, and that therefore the changes at the output, the resulting β, should be small as well, if the model matches the data. We use $\mathbf{B}(\lambda) \in \mathbb{R}^{D \times N}$ to denote the N solutions provided for each λ and use $k \in 1, \ldots, K$ to index these solutions. We use f to index elements of β, corresponding to individual features.

We introduce the concept of feature persistence to measure feature-selection stability of SCRC across the whole set of Leave-One-Out (LOO) solutions as an alternative to pairwise subset similarity measures as in (Yu, Ding, and Loscalzo 2008; Kalousis, Prados, and Hilario 2005; Kuncheva 2007). The Feature Persistence Rate (FPR) measures how often a feature, f is *active*, as the percentage of folds, k, for which $\mathbf{B}_{f,k}(\lambda)$ is nonzero, for a given λ as shown in (7a). A feature, f is persistent at level $\alpha \in [0, \ldots, 1]$ for a given value of λ if $\mathrm{FPR}(f, \lambda) > \alpha$. Every feature would have a FPR of either one or zero (always *on* or always *off*) in an ideal solution.

$$\mathrm{FPR}(f, \lambda) = \frac{\mathrm{count}_k(|\mathbf{B}_{f,k}(\lambda)| > 0)}{N} \tag{7a}$$

$$\bar{\beta}_f(\lambda, \alpha) = \begin{cases} \frac{1}{K} \sum_k \mathbf{B}_{f,k}(\lambda) & \mathrm{FPR}(f, \lambda)) > \alpha \\ 0 & - \end{cases} \tag{7b}$$

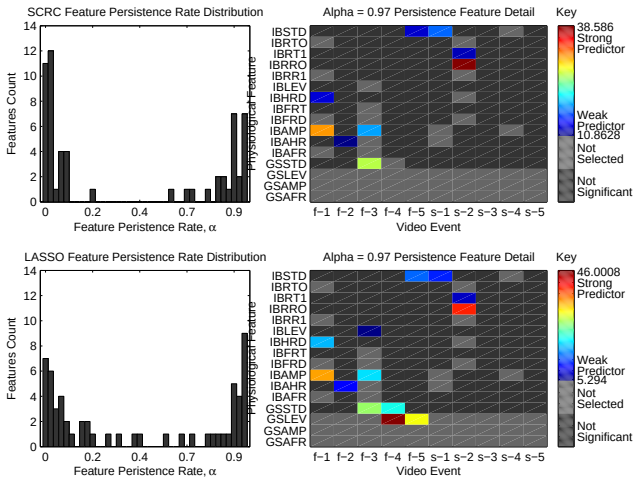

Figure 2: Feature Persistence $\log(\lambda) = 1.92, \alpha = 0.97$. *Left:* Histogram counts of features for each FPR. *Right:* (in color) Mean feature weights for the $\alpha = 0.97$ persistent features, dark gray were excluded in preprocessing, the light gray were excluded by sparsity; columns represent video events and the rows physiological features.

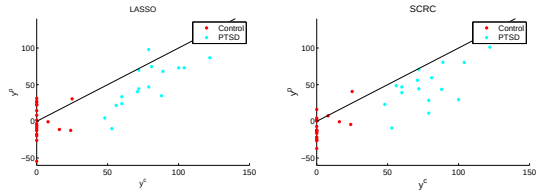

Figure 3: Qualitative illustration of model fit for $\alpha = 0.97$ persistent features for $\log(\lambda) = 1.92$, these use the features as shown in Figure 2. Each point is a subject, represented by a (y_i^c, y_i^p) pair, using the mean weights for $\alpha = 0.97$ persistent features to compute y^p.

In Figure 2 we see that in the left column, the distribution of FPRs from SCRC is more polarized than that of LASSO, a more persistent solution. We present an overall solution computed using $\bar{\beta}(\log(\lambda) = 1.92, \alpha = .97)$ to compute the physiological score for each subject as shown in Figure 3). This figure qualitatively illustrates the difference in the solutions provided by LASSO and SCRC. This shows that the performance difference is largely in the subjects with a zero score, SCRC keeps these subjects' scores below zero, confirming *that the unique learning paradigm we present achieves the desired effect.* We present the diagnostic validity measures on this solution in Table 1, which shows that with fewer features the overall solution for SCRC outperforms on all metrics.

Parsimony and Context

Finally we consider the context of the solution by examining the structure of the returned coefficient vector, β. To achieve parsimony, β should be sufficiently sparse to avoid over-fitting. We quantitatively assess this through the number of retained features. To agree with other literature, based on expert insight, the nonzero elements of β should correspond to features from both measurement channels (IBI and GSR) and multiple video events.

Metric	LASSO	SCRC
ρ_S	0.70	0.83
MSE_{NZ}	646.0284	628.0155
AUC	0.85	0.89
Features Count	13	9

Table 1: Diagnostic validity measures for the solution shown in the right subfigures of Figure 2 and all of Figure 3.

The right column of Figure 2 illustrates an averaged, persistent at level $\alpha = 0.97$ solution instead of choosing a single fold. We note that features kept are from both the GSR and IBI signals. Time is not modeled explicitly but the learned score can depend on changes in feature values over the course of time, because each feature value is computed for a variety of times (video events). The solutions from the two models are similar, but as we saw above, SCRC uses slightly different weights and fewer features, which is *more parsimonious, supporting our final solution desideratum.*

Discussion and Conclusions

In this work we present a learning paradigm to support development of a physiologically based PTSD diagnostic score. Our method uses the available measurements from an experiment where virtual reality videos were used to evoke a physiological response measured through IBI and GSR. The solution unifies regression and classification loss functions with a sparse regularization term. As a directly competitive method does not exist, we compare the SCRC to a computationally similar, but application-naive approach, LASSO. Our method provides more parsimonious, diagnostically valid, and generalizable results than the alternative.

Experimentally, these results demonstrate merit for an expanded subject enrollment and further collaboration with clinically focused researchers. Additionally, conducting another, longer study would allow for inclusion of multiple clinical interviewers to assess inter-rater reliability and agreement of each with these physiologically based scores. This work stems from a proof-of-concept scaled exploration into using non-idiographic virtual reality videos, like those previously used in treatment. Individual items or groups of items of the CAPS, which correspond to symptoms and categories of symptoms, is also a candidate area for future exploration.

Future technical extensions of the work can explore analytical relationships of the heuristically derived performance metrics, and the automated selection of the regularization parameter, λ. The method is presented as derived from an empirical risk minimization perspective, but a probabilistic interpretation may provide added insight.

Acknowledgments This work was supported by a National Science Foundation Graduate Research Fellowship (NSF DGE-0946746) and Charles Stark Draper Laboratory.

References

American Psychiatric Association. 2013. *The Diagnostic and Statistical Manual of Mental Disorders: DSM 5*. bookpointUS.

Blake, D. D.; Weathers, F. W.; Nagy, L. M.; Kaloupek, D. G.; Gusman, F. D.; Charney, D. S.; and Keane, T. M. 1995. The development of a clinician-administered PTSD scale. *J. Trauma. Stress* 8(1):75–90.

Bousquet, O., and Elisseeff, A. 2002. Stability and generalization. *J. Mach. Learn. Res.* 2:499–526.

Boyd, S. 2010. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Found. Trends Mach. Learn.* 3(1):1–122.

Carson, M. A.; Paulus, L. A.; Lasko, N. B.; Metzger, L. J.; Wolfe, J.; Orr, S. P.; and Pitman, R. K. 2000. Psychophysiologic assessment of posttraumatic stress disorder in Vietnam nurse veterans who witnessed injury or death. *J. Consult. Clin. Psychol.* 68(5):890.

Cortes, C., and Vapnik, V. 1995. Support-vector networks. *Mach. Learn.* 297:273–297.

Giraud-Carrier, C. 1998. Beyond predictive accuracy: what? In *Proc. ECML-98 Work. Upgrad. Learn. to Meta-Level Model Sel. Data Transform.*, 78–85.

Goldfinger, D. A.; Amdur, R. L.; and Liberzon, I. 1998. Psychophysiologic responses to the Rorschach in PTSD patients, noncombat and combat controls. *Depress. Anxiety* 8(3):112–120.

Joachims, T. 2002. Optimizing search engines using clickthrough data. *Proc. eighth ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '02* 133.

Kalousis, A.; Prados, J.; and Hilario, M. 2005. Stability of feature selection algorithms. In *Data Mining, Fifth IEEE Int. Conf.*, 8—-pp. IEEE.

Kilmer, B.; Eibner, C.; Ringel, J. S.; and Pacula, R. L. 2011. Invisible wounds, visible savings? Using microsimulation to estimate the costs and savings associated with providing evidence-based treatment for PTSD and depression to veterans of Operation Enduring Freedom and Operation Iraqi Freedom. *Psychol. Trauma Theory, Res. Pract. Policy* 3(2):201.

Kircher, J. C., and Raskin, D. C. 1988. Human versus computerized evaluations of polygraph data in a laboratory setting. *J. Appl. Psychol.* 73(2):291–302.

Kuncheva, L. 2007. A stability index for feature selection. *Artif. Intell. Appl.* 390–395.

Lange, T.; Braun, M. L. M.; Roth, V.; and Buhmann, J. M. 2002. Stability-based model selection. In *Adv. Neural Inf. Process. Syst.*, 617–624.

Mukherjee, S.; Niyogi, P.; Poggio, T.; and Rifkin, R. 2006. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Adv. Comput. Math.* 25(1-3):161–193.

Orr, S. P.; Meyerhoff, J. L.; Edwards, J. V.; and Pitman, R. K. 1998. Heart rate and blood pressure resting levels and responses to generic stressors in Vietnam veterans with posttraumatic stress disorder. *J. Trauma. Stress* 11(1):155–164.

Orr, S. P.; Metzger, L. J.; Lasko, N. B.; Macklin, M. L.; Peri, T.; and Pitman, R. K. 2000. De novo conditioning in trauma-exposed individuals with and without posttraumatic stress disorder. *J. Abnorm. Psychol.* 109(2):290.

Orr, S. P.; Metzger, L. J.; and Pitman, R. K. 2002. Psychophysiology of post-traumatic stress disorder. *Psychiatr. Clin. North Am.* 25(2):271–93.

Pitman, R. K.; Orr, S. P.; Forgue, D. F.; de Jong, J. B.; and Claiborn, J. M. 1987. Psychophysiologic assessment of posttraumatic stress disorder imagery in Vietnam combat veterans. *Arch. Gen. Psychiatry* 44(11):970–5.

Pole, N. 2007. The psychophysiology of posttraumatic stress disorder: a meta-analysis. *Psychol. Bull.* 133(5):725.

Richardson, L. K.; Frueh, B. C.; and Acierno, R. 2010. Prevalence estimates of combat-related post-traumatic stress disorder: critical review. *Aust. N. Z. J. Psychiatry* 44(1):4–19.

Rudin, C., and Wagstaff, K. L. 2014. Machine learning for science and society. *Mach. Learn.* 95(1):1–9.

Sculley, D. 2010. Combined regression and ranking. *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '10* 979.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Methodol.* 58(1):266–288.

Wagstaff, K. L. 2012. Machine Learning that Matters. In *Proc. 29th Int. Conf. Mach. Learn.*, 529–536.

Webb, A. K.; Vincent, A. L.; Jin, A.; and Pollack, M. H. 2013. Wearable sensors can assist in PTSD diagnosis. *2013 IEEE Int. Conf. Body Sens. Networks* 1–6.

Yu, L.; Ding, C.; and Loscalzo, S. 2008. Stable feature selection via dense feature groups. *Proceeding 14th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD 08* 803.