Published in final edited form as:

Emot Rev. 2014 April ; 6(2): 160–174. doi:10.1177/1754073913512519.

Advancing emotion theory with multivariate pattern classification

Philip A. Kragel and Kevin S. LaBar

Duke University

Abstract

Characterizing how activity in the central and autonomic nervous systems corresponds to distinct emotional states is one of the central goals of affective neuroscience. Despite the ease with which individuals label their own experiences, identifying specific autonomic and neural markers of emotions remains a challenge. Here we explore how multivariate pattern classification approaches offer an advantageous framework for identifying emotion specific biomarkers and for testing predictions of theoretical models of emotion. Based on initial studies using multivariate pattern classification, we suggest that central and autonomic nervous system activity can be reliably decoded into distinct emotional states. Finally, we consider future directions in applying pattern classification to understand the nature of emotion in the nervous system.

Keywords

emotion specificity; multivariate pattern classification; autonomic nervous system; central nervous system; model comparison

Emotion plays a prominent role in human experience. It changes the way we see the world, remember past events, interact with others, and make decisions. Emotional experiences produce feelings ranging from tranquility to full-fledged rage. Despite the ease with which these experiences are categorized and labeled, emotional feelings are often characterized as private, subjective, and notoriously difficult to quantify. As stated by Fehr and Russell, “Everyone knows what an emotion is, until asked to give a definition” (1984, p. 464). The scientific investigation of emotion has utilized behavioral observation, experiential self-report, psychophysiological monitoring, and functional neuroimaging to understand the nature of emotions.

Currently, there is considerable debate as to whether the categorical structure of experienced emotions is respected by the brain and body, stems from cognitive appraisals, or is a construction of the mind. One key issue at hand concerns whether the autonomic responses, central nervous system activity, and subjective feelings that occur during an emotional episode invariantly reflect distinct emotions or map on to more basic psychological

processes from which emotions are constructed. Although reviews and meta-analyses of univariate studies have not consistently supported emotion specific patterning within these response systems (e.g. Barrett, 2006a), recent proposals have suggested that multivariate analysis of autonomic and neuroimaging data may be better suited to identifying biomarkers for specific emotions compared to univariate approaches (Friedman, 2010; Hamann, 2012). It is possible that statistical limitations of univariate approaches have led to the premature dismissal of emotions as having distinct neural signatures. The goal of the present paper is to consider the promises and potential pitfalls of multivariate pattern classification and to review how these multivariate tools have been applied so far to test the structure of emotion in both the central and autonomic nervous systems.

Drawing formal inferences with multivariate pattern classification

The crux of the problem in defining distinct emotions is that, while many changes occur throughout the body during events given an emotional label, are they consistent or specific enough to indicate the occurrence of naturally separable phenomena? For example, research has shown that fearful stimuli often elicit increases in heart rate, sympathetic nervous system activity, and neural activation within the amygdala. If fear is categorically represented in the nervous system, then it should be possible to reliably infer its occurrence given such a response profile. This type of reverse inference is notoriously difficult to make accurately if the response profile is not highly selective, as any individual measure may change with the occurrence of any number of stimuli or psychological processes (Poldrack, 2006). For instance, the frequency of insula activation in neuroimaging studies has been reported to be almost one in five studies (Nelson et al., 2010; Yarkoni, Poldrack, Nichols, Van Essen, & Wager, 2011), making it difficult to make meaningful inferences based on the presence its activation alone. Within the domain of emotion, meta-analyses have suggested that no individual region responds both consistently and specifically to distinct emotions (Barrett & Wager, 2006; Lindquist & Barrett, 2012), which has been taken as disconfirming evidence that they are represented distinctly in the brain (Barrett, 2006a).

Utilizing machine learning algorithms for assessing the discriminability and specificity of multivariate patterns (Jain, Duin, & Mao, 2000) is a promising new approach for formally inferring mental states. This method, commonly called multivariate pattern classification (MVPC; Haynes & Rees, 2006; Norman, Polyn, Detre, & Haxby, 2006) was first employed using functional magnetic resonance imaging (fMRI) by Haxby and colleagues (2001) to characterize the functional response of the ventral temporal cortex during the perception of objects. This work aimed to advance a debate centered on the organization of neural systems involved in visual perception. While one prevalent model predicted the representation of objects was encoded broadly throughout ventral temporal cortex in a distributed fashion, alternative theories posed a modular functional organization of the region based on the category of object, including regions selectively responsive to faces (Kanwisher, McDermott, & Chun, 1997). By constructing a model to predict the class of perceived objects from multivariate samples of imaging data, pattern classification demonstrated that overlapping distributions of neural activation encoded stimulus categories. Beyond its characterization of object selective cortex, this study highlighted that individual voxels may exhibit increased activation to many stimuli and concurrently belong to a larger pattern of

activation which is functionally specific to a mental percept. In this way, the method can go beyond mean signal intensity changes averaged over voxels in univariate designs, offering a new means of drawing inferences.

Since its introduction to human neuroimaging, pattern classification has been used to infer mental states previously considered beyond the scope of fMRI. Some examples include the perception of complex scenes (Kay, Naselaris, Prenger, & Gallant, 2008), color (Brouwer & Heeger, 2009), the orientation of lines (Kamitani & Tong, 2005), the direction of motion (Kamitani & Tong, 2006), the source and content of spoken sound (Formisano, De Martino, Bonte, & Goebel, 2008), the contents of working memory (Harrison & Tong, 2009), and more broadly the occurrence of different cognitive (or subjective) processes (Poldrack, Halchenko, & Hanson, 2009). Because of its increased sensitivity and ability to quantify the specificity of subjective mental states, MVPC overcomes the shortcomings of univariate methods that complicate studying the nature of emotion in the nervous system. Thus, adopting the multivariate framework allows researchers to formally test the extent to which emotional states are discretely represented and whether such states are organized along dimensional constructs.

Decoding neural representations of emotion with MVPC

Before reviewing how MVPC studies may inform models of emotion, here we outline the steps involved in pattern classification, as there are multiple decisions to be made in the analysis pipeline for MVPC that constrain what information can be utilized in the formulation of a classification model. These steps include how ground truth is defined for different data points, how features are selected for use in classification, what algorithm is used to distinguish between classes, how data are split between training and testing, and how to quantify performance. The selection of these analysis parameters bears particular importance when assessing how distinct emotions are represented in the nervous system (Fig 1).

Establishing ground truth

As pattern classification is a supervised learning technique (Duda, Hart, & Stork, 2001), knowledge about each instance of data used to construct a classifier is required in order to assign class membership. In the case of classifying emotional states, different theories offer a variety of possibilities. Theories positing a small number of categorically separate emotions (on the basis of overt behavior, or specific neural circuitry) can be directly translated into the labeling of multiple classes. Although cognitive appraisal and constructionist theories do not consider emotions to be purely categorical in nature, they do propose that distinct emotions emerge from more basic dimensions or unique patterns of appraisal (Russell, 1980; Scherer, 1984). Another alternative is to consider regularities in states produced by reinforcement contingencies (Rolls, 1990) or action tendencies (Frijda, 1987) to guide the definition of emotional states. In this sense, using categorical labels is appropriate for comparing different theories, although alternative labeling based on dimensions of valence, arousal, or components of cognitive appraisal may improve

classification performance if responses cluster along constructs beyond basic emotion categories.

Feature selection

Determining which combination of response variables (referred to as ‘features’) to use in pattern classification is critical because selecting too many or the wrong ones can result in poor performance. Due to the high dimensionality of fMRI datasets (in the tens of thousands for typical datasets), reducing the number of uninformative inputs is important because the predictive power of a classifier decreases as a function of the number of features for a fixed number of instances (Hughes, 1968). While this is less of a concern when studying relatively few signals measured from the autonomic nervous system, its importance increases when confronted with thousands of measures of brain activity in the context of neuroimaging experiments. Common methods for feature selection in neuroimaging include the use of *a priori* regions of interest, spherical or cuboid searchlights centered around individual voxels (Kriegeskorte, Goebel, & Bandettini, 2006), masks of thresholded univariate statistics (e.g., an F map from an ANOVA), independent or principal component analysis, or a restricted field of view at high resolution. These different approaches extract information at different spatial scales and are accordingly useful for testing alternative hypothesis about the representation of emotional states in the brain. For instance, examining local patterning with searchlights may be more appropriate for identifying categorical representation of emotion in subcortical circuits as postulated by some categorical emotion theories (Ekman, 1992; J. Panksepp, 1982). Additionally, the ability of dimensionality reduction methods to extract functionally coherent networks associated with basic psychological processes is well matched to test the constructionist hypothesis that emotions emerge from the interaction of large scale networks (Kober et al., 2008; Lindquist & Barrett, 2012). Thus, different methods for reducing the dimensionality of fMRI data may reveal complementary ways in which emotion is represented in the brain.

Algorithm selection

The ability of a classifier to create an accurate mapping between input data and the ground truth class labels is dependent upon the learning algorithm used. While the varieties in learning algorithms are vast, they can broadly be designated as either linear or nonlinear (i.e. a curved boundary) based on the shape of the decision surface used to classify data. Linear methods (e.g. linear discriminant analysis (LDA) or support vector machines (SVM) with linear kernels) are more frequently used with fMRI as they have been shown to perform similar to, if not better than, nonlinear approaches (Misaki, Kim, Bandettini, & Kriegeskorte, 2010) while being straightforward to interpret. Although nonlinear algorithms are more complex, they are capable of solving a broader range of classification problems and mapping higher order representations (e.g. Hanson, Matsuka, & Haxby, 2004). Considering different organizations of affective space, there may be instances where nonlinear approaches are advantageous for decoding distinct emotional states. In particular, if the hypothesized dimensionality of the data is smaller than the number of invariant emotional states, as is the case in most constructionist and appraisal theories, then mapping data to a higher dimensional space with a nonlinear algorithm may be advantageous.

Generalization testing

In order to determine if the decision rule learned by a classifier generalizes beyond the data with which it was constructed, classification must be tested on an independent set of data (Kriegeskorte, Simmons, Bellgowan, & Baker, 2009). Typically, generalization involves the partitioning of a few trials within a run, different runs of data acquisition, or independent subjects for use in testing (beyond those used for training the classification model) to produce an unbiased estimate of accuracy. The way in which data are split into training and testing sets provides an opportunity to examine different claims of various emotion theories. For example, a fundamental aspect of some categorical emotion theories is the universality of emotional expressions and the conservation of their underlying neural mechanisms (Ekman & Cordaro, 2011). Between subject classification is one way of providing evidence for this claim (although negative results would be difficult to interpret as poor generalization may result from differences in fine scale neural anatomy, temporal dynamics, or other factors unrelated to representational content). Alternatively, testing on independent trials or runs may prove more fruitful if emotional states are hypothesized to be idiosyncratic to the individual or particular context as emphasized by appraisal and constructionist theories (Barrett, 2006b; Scherer, 1984) because these testing schemes relax assumptions of response invariance.

Quantifying performance

A classifier is generally considered to exhibit learning if it is capable of classifying independent data at levels beyond chance. While this criterion is important when testing performance, it does little to differentiate between alternative emotion theories. For example, two classifiers could exhibit the same levels of overall accuracy, sensitivity, and specificity, yet have uniquely different error distributions. Testing how the structure of errors corresponds to predictions of different emotion theories is particularly relevant in decoding emotional states, because distinct emotions have been hypothesized by different theories to be completely orthogonal (Ekman, 1992; Tomkins, 1962), clustered within a multidimensional space (Arnold, 1960; Scherer, 1984; Smith & Ellsworth, 1985), or arranged in a circumplex about dimensions of valence and arousal (Barrett, 2006b; Russell, 1980). Identifying the structure of information used in classification provides a link between neural data and conceptual and computational theories of emotion. In this way, multivariate classification can reveal how different constructs are instantiated in the brain or periphery instead of focusing on whether certain emotional states are biologically basic.

Model comparison

Given that psychological models propose that distinct emotional states emerge through different processes, the way in which a classifier performs can be used to test predictions of alternative theoretical models. While both categorical emotion theories and appraisal models assume some degree of consistency in the neural mechanisms underlying emotional states, constructionist models claim that distinct emotions do not have invariant neural representations. Altering the generalization testing in a classification pipeline is one means of testing this hypothesis. For instance, a classifier trained during the experience of core disgust and tested on moral disgust occurring from the violation of social norms could reveal

which response components are shared between the two emotional states. Additionally, appraisal models and constructionist models propose that different factors organize distinct emotions, and as such the similarity of neural activation patterns should reflect this structure. If there is increased similarity for emotional states that are more closely related along dimensions of valence and arousal (e.g., fear and anger), then evidence would support a constructionist model of emotion, whereas if they are related along appraisal dimensions such as novelty, valence, agency, or goals, then appraisal models would be favored. Generally, the similarity of emotional states reflected in neural activation patterns, quantified with classification errors, can be related to the similarity proposed by competing models (for an example using models of perception see Kriegeskorte, Mur, & Bandettini, 2008).

While different models of emotion are often characterized as competing, classification results supporting one model are not necessarily evidence against alternative models. It is possible that different branches of the nervous system, separate brain regions, or neural activation at diverse spatial scales, better conform to alternative conceptions of emotion. Moreover, given separate bodies of behavioral work supporting both dimensional and categorical aspects of emotion, it may not be advantageous to search for any single ‘best’ model. Investigating in what neural systems or under which contexts dimensional or categorical constructs are observed may provide more leverage for advancing theories of emotion. Thus, depending on the structure of information revealed by classification, MVPC is capable of discerning how distinct emotions are reflected in activity across multiple levels of the neuraxis – either as constructions, patterns of appraisal, or independent categories.

Limitations

While multivariate classification approaches are theoretically promising in terms of identifying emotion specific patterns of neural activity, there are several caveats to their use which limit the inferences that can be drawn from significant results. Importantly, classifiers will utilize any information which is capable of differentiating the cases to be labeled. In the context of classifying patterns of neural activity into emotional states it is possible that results are driven by factors unrelated to emotion per se, such as motor behavior, task induced artifact (e.g., movement-related or psychophysiological confounds across conditions), perceptual or semantic differences in stimuli. When the goal of a study is to identify the patterns of BOLD response which best represent a single construct, adequately accounting for potential confounds is necessary – especially given the increased sensitivity of the method relative to conventional regression models (Todd, Nystrom, & Cohen, 2013). Both carefully designed experiments and appropriate inferential models are thus critical in using MVPC to inform theories of emotion.

Classification of emotional responding in the central nervous system

Relative to the large body of univariate work examining the relationship between fMRI activation and distinct emotions, a small number of studies are beginning to use MVPC to map neural activation onto specific emotional states. Here, the extent to which emotional states, as defined by either dimensional or categorical models (i.e. classifying subjective valence/arousal or distinct emotions such as fear and anger), are represented by specific

patterns of neural activation is reviewed. Because emotional perception and experience engage different processes and are subserved by separate neural systems (Wager et al., 2008), the studies reviewed are organized based on the process most likely engaged during functional imaging (Table 1).

Decoding emotional perception

The majority of studies using MVPC to decode emotions have investigated the perception of emotional stimuli. The goal of these studies is to use blood-oxygen level dependent (BOLD) activation patterns to infer the perceived emotional content in a stimulus, as determined by subjective evaluative judgments. This approach differs from predicting the physical similarity of stimuli, which may not necessarily correspond to the perceptual state of participants (for a review and discussion, see Tong & Pratte, 2012). Stimulus sets are often comprised of facial, vocal, and gestural expressions due to their ability to elicit discrete emotional responses and convey unique information critical for motivating behavior and social communication (Adolphs, 2002).

Studies examining the perception of emotion in facial expressions are beginning to reveal how discrete emotion categories are represented neurally. One investigation (Said, Moore, Engell, Todorov, & Haxby, 2010) tested whether multivariate patterns of activation in regions that commonly respond to facial expressions, i.e., frontal operculum, along with anterior and posterior aspects of the superior temporal sulcus (STS), could be rated by an independent group of subjects. These results demonstrate that fine scale patterns of activation within STS contain movement related information capable of specifying the category of perceived facial affect. Most importantly, univariate analysis of the same activation patterns yielded a nonspecific mapping to the perception of emotion. Thus, the information carried within the STS depends on the underlying spatial distribution of neural activation in a fashion univariate analysis cannot detect. Therefore, the functional specificity of the STS may be improperly assessed if only tested with a univariate approach.

Another study (Pessoa & Padmala, 2007) examined emotional perception by showing participants faces with happy, neutral, and fearful facial expressions at near subliminal durations (33 or 67 ms) followed by neutral face masks and subsequently testing if a fearful face was perceived. The aim of the study was to determine if patterns of fMRI activation could predict the response of a participant, thus representing a mental state of perceived fear. Combinations of voxels from multiple *a priori* regions of interest (found to discriminate behavioral choice when assessed with a univariate analysis; Pessoa & Padmala, 2005) predicted the choice with increasing accuracy as more regions were included in the analysis. Further, the information contributed when incorporating multiple regions was found to add synergistically, such that the total decision-related information was greater when two regions were considered jointly compared to the sum of the information when each was considered individually. Information contributed by the amygdala in particular was found to most significantly improve the performance of classifiers. These findings indicate that regions implicated in different aspects of affective processing (e.g. representing the stimulus, or modulating autonomic responses) form a distributed representation of an emotional percept, and that conjoint analysis of independent information can more accurately characterize

emotional states. While studies examining emotional perception using facial expressions have found discriminable patterns of BOLD response, the reported studies used posed rather than naturalistic stimuli. Understanding whether response patterns to naturally occurring facial expressions are similar to those for staged or prototypical expressions remains an open area of research.

Studies investigating emotion perception of human vocalizations have similarly established categorical specificity of fMRI activation patterns. One such study (Ethofer, Van De Ville, Scherer, & Vuilleumier, 2009) demonstrated that spatially distinct patterns of activation within voice sensitive regions of auditory cortex, established with an independent localizer, could discriminate between pseudowords spoken with emotional prosody of anger, sadness, neutral, relief, and joy. While voice sensitive cortex was consistently activated across all stimuli in the experiment, no differences in the average response amplitude were observed between categories in a univariate analysis. Furthermore, errors committed by classifiers were most frequent for stimuli which shared similar arousal, but not valence. The authors interpreted this relationship to be driven by alterations in the fundamental frequency of vocalizations, which may be captured in the spatial variability of activation patterns. Other work aiming to decode the emotional prosody of speech from local patterns of fMRI activation (Kotz, Kalberlah, Bahlmann, Friederici, & Haynes, 2012) yielded similar results. Using the searchlight approach, these authors found that a distributed network of regions discriminated among the perception of angry, happy, neutral, sad and surprised vocal expressions. In addition to superior temporal gyrus, this predominantly right-lateralized network included inferior frontal operculum, inferior frontal gyrus, anterior superior temporal sulcus, and middle frontal gyrus. While univariate studies have implicated these regions in a number of aspects of vocal processing (e.g., speaker identity, fundamental frequency, or semantic content), future work is needed to confirm that they contribute distinct information to a distributed representation of perceived emotion.

In sum, MVPC studies of emotional perception have demonstrated that information carried in fine-grain structure across multiple cortical sites is capable of signifying affective information in the environment in distinct ways. Findings that information reflected in local patterns (e.g. Ethofer et al., 2009; Said et al., 2010), but not mean activation levels at the centimeter scale, is consistent with a growing body of evidence from multivariate decoding of perception. Together, these findings suggest that the perceptual qualities of a stimulus that confer emotional significance are reflected in patterns of BOLD activation within a number of cortical regions. Further, classification utilizing voxels spanning multiple distributed regions showed that different types of information could be combined synergistically to represent a perceptual state. By considering multivariate information at various scales, MVPC studies have revealed patterning specific to the perception of distinct emotional stimuli, a difficult goal to achieve with conventional univariate fMRI (e.g. Winston, O'Doherty, & Dolan, 2003) given its relatively limited functional specificity and spatial resolution.

Decoding emotional experience

In addition to examining emotional percepts, MVPC has been used to infer the subjective experience of emotion on the basis of functional activation patterns. Although the distinction between emotional perception and experience is often blurred, as the perception of emotional stimuli can lead to the experience of emotional feelings, here studies that attempted to decode subjective states on the basis of introspective self-report are reviewed. Work examining the neural underpinnings of emotional feelings using univariate approaches (Damasio et al., 2000) has qualitatively characterized the similarities and differences between distributed patterns of activation during emotional episodes. While comparisons of this nature are capable of grossly characterizing which brain regions are engaged during the experience of a specific emotional feeling, they lack the quantitative precision necessary to map patterns of neural activation to specific components of an emotional episode. For instance, Damasio and colleagues (2000) examined response patterns within subcortical structures (e.g. hypothalamus, brainstem nuclei) and cortical regions (insula, anterior cingulate, SII), which collectively exhibited altered response profiles during the experience of different emotional feelings. Some regions were selectively engaged during one emotion relative to others (for instance, happiness and sadness produced activation in different portions of the insula - among many other regions), but the extent to which these differences were specific to a particular emotion was not assessed with the univariate approach employed.

The application of MVPC can precisely characterize how patterns of neural activation signify distinct emotional experiences by quantifying where in the brain information specific to a particular emotion is represented, and the extent to which neural activation patterns specifically and accurately reflect a given emotional state. Initial MVPC studies decoding the experience of emotion have often utilized dimensional models, primarily focusing on the classification of positive and negative episodes. One such study (Baucom, Wedell, Wang, Blitzer, & Shinkareva, 2012) implemented an implicit emotion induction where participants incidentally viewed blocked presentations of images that had been previously validated to elicit positive and negative affect at high and low levels of arousal. Using classifiers constructed to decode the experience of valence and or arousal, whole brain activation patterns revealed a common neural representation of emotional experience sparsely distributed throughout the brain. Further, classification accuracies were above chance levels when cross-validation was performed within and between subjects. Additional multivariate analyses utilizing multidimensional scaling of the most stable 400 voxels identified from pattern classification revealed that the most variance in fMRI activation could be explained along dimensions of valence and arousal. While findings from this work demonstrate the feasibility of inferring emotional states from fMRI activation patterns, one main issue limits the implications of the study. Although the stimuli used were previously shown to produce emotional responses differing in valence and arousal, no on-line reports of emotional experience were conducted. As such, it remains unclear whether information related to the perceptual processing of stimuli or the affective experience of participants was informing the classification of fMRI data.

Additional work examining the neural representation of positive and negative experience has focused on the subjective pleasantness of thermal stimulation (Rolls, Grabenhorst, & Franco, 2009). In this work, the subjective pleasantness of thermal stimulation to the hand was successfully inferred from local patterns of fMRI activation in frontal cortex. Consistent with results from the whole brain classification conducted by Baucom and colleagues (2012; see Fig 2), information was redundantly pooled across local voxels such that classification performance increased sub-linearly as the number of voxels used increased, even when adding voxels from different regions. This finding suggests that the information encoded by each voxel was not independent, but rather was redundantly carried through multiple voxels. Taken together, these initial studies suggest widely distributed and highly redundant patterns of BOLD response contain valence-related information, possibly implicating a common neuromodulatory source, such as dopamine, given its role in appetitive motivation and its broad projections to prefrontal cortex (Ashby, Isen, & Turken, 1999). A regional modulation of neural activity due to transient dopamine release (Phillips, Ahn, & Howland, 2003) is consistent with the location and redundancy of information observed. Given the causal role of neurotransmitter systems in generating affective states proposed by neurobiological models of emotion (Jaak Panksepp, 1998), understanding the link between these systems and patterns of neural activation is a critical area for future research.

In addition to decoding the experience of emotion along affective dimensions, categorical emotion models have been used to guide MVPC. In one such study (Sitaram et al., 2011), pattern classifiers implementing support vector machine algorithms to classify patterns of whole brain activity were shown to accurately discriminate among the experience of happiness, sadness, and disgust during imagery in real time for individual participants. Off-line analysis of classifiers demonstrated that a number of cortical and subcortical regions each contributed to classification performance. In this analysis, regions of interest from prior meta-analyses (Frith & Frith, 2006; Ochsner, 2004; Ochsner & Gross, 2005; Phan, Wager, Taylor, & Liberzon, 2002), including cortical (middle frontal gyrus, superior frontal gyrus, and superior temporal gyrus) and subcortical (amygdala, caudate, putamen, and thalamus) regions, could be used to decode the experience of disgust versus happiness. These results are striking in comparison to univariate meta-analytic work (Lindquist, Wager, Kober, Bliss-Moreau, & Barrett, 2012; Phan et al., 2002) in which several of these regions, the medial prefrontal cortex in particular, exhibited little to no specificity for distinct emotions when treated as homogeneous functional units. One important limitation of this study, however, was the relatively small number of emotions sampled. Because often only a single positive and negative emotion pair was decoded (i.e. disgust and happiness), information used during classification could either reflect differences in basic emotion categories or affective dimensions such as valence.

Other work using imagery to induce distinct emotional states by Kassam and colleagues (2013) identified the most stable voxels from whole brain acquisitions of fMRI data to classify states of anger, disgust, envy, fear, happiness, lust, pride, sadness, and shame. Participants were presented with two cue words for each emotion which remained in view for 9 seconds during while imagery was performed. The authors demonstrated the localization, generalizability, and underlying structure of emotion-related information. Multi-voxel patterns drawn from the most stable 240 voxels commonly included a

distributed set of areas spanning frontal, temporal, parietal, occipital, and subcortical regions. Using Gaussian Naïve Bayes classifiers, the nine states were classified with 84% accuracy when generalization was performed within subjects and 71% accuracy for between subject classification. Furthermore, when generalizing to the perception of images, disgusting and neutral stimuli were classified within individuals at 91% accuracy. Investigating the occurrence classification errors revealed similarities between positive states of happiness and pride and isolated states of lust as being the least similar to the other elicited states. Consistent with these findings, a factor analysis conducted on neural activation patterns revealed constructs interpreted as valence, arousal, sociality, and lust to underlie the variability in patterns. While comprehensive in its characterization of emotion patterning, it is unclear to what extent the classification results were dependent on semantic processing. Given the repeated presentation of related words (e.g. repeated presentations of 'afraid' and 'frightened' were discriminated from those of 'gloomy' and 'sad') and the presence of a left lateralized prefrontal network implicated in semantic processing (Martin & Chao, 2001), it is possible that the observed results are partly due to the classification of semantic information.

In sum, the few studies that have conducted multivariate classification of emotional experience suggest that regions conventionally considered to be nonspecific in a univariate sense are capable of specifying the experience of a distinct emotion at a multivariate level. While this line of investigation is in its infancy, it shows promise for moving beyond limitations of univariate, locationist approaches to analyzing brain function. Consider, for instance, activation in medial prefrontal cortex (mPFC) to emotional stimuli. Meta-analyses of fMRI activation show that this region is commonly engaged during multiple emotional states, implicating the mPFC as a functional unit engaged by a broad array of emotional content (Phan et al., 2002). On the other hand, evidence from MVPC reviewed above suggests that at finer spatial scales, activity within mPFC is capable of discriminating between emotional states. As this region of medial frontal cortex has been hypothesized to play many roles, including emotional appraisal (Kalisch, Wiech, Critchley, & Dolan, 2006), attribution (Ochsner et al., 2004) and regulation (Goldin, McRae, Ramel, & Gross, 2008), it is possible that these processes are differentially engaged during the experience of different emotions. Such varied recruitment may be evident in the structure of spatial activation patterns that predict subjective emotional experience. Identifying which processes contribute to distinct emotional representations in mPFC remains an open area of research.

In addition to mPFC, activation patterns within the amygdala could consistently be used to decode the subjective experience of emotion along dimensions of valence and arousal (Baucom et al., 2012; Sitaram et al., 2011), converging with a large body of univariate work. The amygdala has been implicated in numerous affective processes (Davis & Whalen, 2001); more specifically it has been found to functionally covary with the subjective experience of arousal (Anderson et al., 2003; Colibazzi et al., 2010; Liberzon, Phan, Decker, & Taylor, 2003; Phan et al., 2003) and valence (Anders, Lotze, Erb, Grodd, & Birbaumer, 2004; Gerber et al., 2008; Posner et al., 2009). One possibility is that pattern classifiers read out information from distinct populations of neurons specific to positive and negative valence (Paton, Belova, Morrison, & Salzman, 2006) to predict the ongoing affective state. Given the potential of MVPC to utilize information beyond the conventional resolution of

fMRI by capitalizing on changes in the distribution of neurons from voxel to voxel (Freeman, Brouwer, Heeger, & Merriam, 2011; Gardner, 2010; Kamitani & Tong, 2005; but see Op de Beeck, 2010), these findings show promise in advancing our understanding of the functional architecture of the amygdala at increasingly fine resolution. Future work is required to determine whether there are neural populations within the amygdala that specifically code for valence and arousal, and whether distinct neural populations within the amygdala are essential for experiences commonly labeled as fear.

Classification of emotional responding in the autonomic nervous system

In addition to work investigating emotion specificity of central nervous system activation, a growing body of evidence is supporting the ability of MVPC to extract emotion specific patterns from multivariate sampling of the autonomic nervous system (ANS). While the emotions induced, stimuli used, and general experimental protocol are similar between studies examining central and autonomic patterning, the response variables being classified are inherently quite different. Psychophysiological studies typically measure both the sympathetic and parasympathetic branches of the ANS, including measures of cardiovascular, electrodermal, respiratory, thermoregulatory, and gastric activity. Here we review psychophysiological studies whose primary goal is to use pattern classification to predict the emotional state of participants (Table 2). In particular we focus on the extent to which information carried in the activity of the ANS is capable of discriminating distinct emotional states and how this information is organized (e.g., as discrete emotion categories or along dimensional constructs such as valence or arousal).

Early work applying pattern classification methods to test the emotional specificity of autonomic responding performed by Schwartz and colleagues (1981) sought to differentiate states of happiness, sadness, fear, anger, relaxation, and a control condition using cardiovascular measures. Entering measures of heart rate and systolic/diastolic blood pressure into a classifier utilizing stepwise discriminant analysis, linear combinations of cardiovascular response were found to significantly differentiate the six conditions with an accuracy of 42.6%, while chance was approximately 16.6%. Of the five discriminant functions produced to differentiate between conditions, two accounted for 96% of the total explained variance. Notably, the first discriminant function (the linear combination of heart rate and blood pressure changes which best differentiated the six conditions) corresponded to broad cardiovascular activation, positively weighting all three measures. This function generally corresponded to arousal, wherein discriminant weights increased along a spectrum from relaxation to anger. The second discriminant function less clearly mapped onto a psychological construct, having positive weights for both fear and sadness, and negative weights for happiness and a neutral control condition. While this study sampled a single branch of the ANS and accordingly was limited to three inputs for use in classification, it clearly demonstrates that differences in cardiovascular arousal are capable of partially differentiating distinct emotional states.

Subsequent studies examining response patterning during emotional imagery have further established the specificity of autonomic responding by incorporating a number of additional physiological measures beyond cardiovascular activity. In one such study (Sinha & Parsons,

1996), the extent to which fear, anger, action, and neutral states induced with imagery could be differentiated using measures of electrodermal, cardiovascular, thermal, and facial muscle activity was tested using discriminant function analysis. While this work did not exclusively test the specificity of autonomic responses (as classification was conducted using both autonomic and somatic responses), the largest reported differences between the induction of fear and anger were cardiovascular changes – suggesting that autonomic responding likely contributed considerably to classification. Classification of peripheral responses from two separate sessions yielded an internal accuracy of 84% and a cross-validated accuracy of 66.5% when generalizing from one session to the other (in both cases chance was approximately 33.3%). Although little can be inferred about the structure of affective information because only two negative emotional states were decoded, this work further established the predictive capacity of autonomic responses.

Work by Rainville and colleagues (2006) additionally examined the specificity of autonomic responses, specifically cardiovascular and respiratory changes, during states of fear, anger, sadness, and happiness induced with imagery. To more precisely characterize the contribution of sympathetic and parasympathetic branches of the ANS, the authors extracted multiple temporal and spectral measures of cardiac and respiratory activity from the raw electrocardiogram and respiratory data. This feature extraction produced a set of 18 variables which were subsequently decomposed into a smaller number of dimensions using principal component analysis, producing five components explaining 91% of the variance in the original data. Loadings of the five components showed that the principal component analysis separated variance related to parasympathetic activity coupled with respiration, parasympathetic activity independent of respiration, and sympathetic activity (as well as respiration amplitude and variability). Pattern classification utilizing discriminant analysis on component loadings yielded an accuracy of 65.3% using resubstitution and 49.0% with leave-one-out cross-validation, with chance accuracy equal to 25%. By using more innovative feature extraction and reduction methods, this work suggests that multiple independent mechanisms underlie autonomic changes capable of discriminating emotional states in a specific manner.

In addition to imagery, instrumental music or cinematic film clips are commonly used to induce emotional states in research investigating autonomic patterning, due to the specificity of experiences produced and the availability of validated stimuli. One such study by Christie and Friedman (2004) used film clips to induce distinct emotional states of amusement, contentment, anger, fear, sadness, disgust, and neutral during concurrent acquisition of cardiovascular and electrodermal responses. Feature extraction from raw measures yielded 6 variables for classification: heart period, mean successive difference of heart period, tonic skin conductance level, systolic blood pressure, diastolic blood pressure, and mean arterial pressure. With the exception of disgust, all conditions could be predicted above chance levels from these variables with an average correct classification rate of 37.4% (chance being 14.3%). This absolute level of accuracy is somewhat lower than that observed in studies classifying a smaller number of emotional states (e.g. Kreibig, Wilhelm, Roth, & Gross, 2007), due to both lower chance rates and increased difficulty of the classification problem. A separate discriminant function analysis excluding response patterns for disgust, as it was not discriminated above chance levels in pattern classification, revealed that self-

report variables differentiated emotional states along a structure organized by valence and activation (66.11% and 14.31% of explained variance), whereas autonomic variables better mapped on to dimensions of activation and action tendency (58.29% and 14.4% of explained variance). Discriminant weights on the primary autonomic factor revealed that high activation was accompanied by high values of skin conductance and negative values for mean successive difference in heart period, suggestive of sympathetic activation and decreased vagal influences. This work suggests that the information content of autonomic signals predicts emotional states in a manner distinct from the self-report of emotional experience.

The predominance of broad physiological activation in the ANS is consistent with other work decoding cardiac and respiratory responses into the experience of happiness, sadness, serenity, agitation, and neutral states elicited with instrumental music (Nyklicek, Thayer, & van Doornen, 1997). In this study, discriminant analysis capable of classifying five emotional states with an average accuracy of 46.5% produced two functions interpreted as arousal and valence (62.5% and 10.0% of explained variance, respectively). The first discriminant function differentiated happiness and agitation from sadness and serenity. This function was weighted negatively for respiration rate and weighted positively for inhale time, exhale time, inter-beat interval, and respiratory sinus arrhythmia measures. Although the second discriminant function was considered uninterpretable as it did not meaningfully correspond to any psychological construct, the third discriminant function mapped on to valence, differentiating happiness and serenity from sadness and agitation. The physiological weights of the function were negatively associated with inter-beat interval, diastolic blood pressure, and left ventricular ejection time. Contrary to other studies reviewed here, this work identified valence and arousal as the predominant factors in differentiating emotional states using autonomic measures – a structure commonly found in self-reported affect. It is possible, however, that the selection of stimuli drove these differences as the emotional states examined all varied either on dimensions of valence or arousal.

Studies investigating autonomic patterning across both film and instrumental music have examined how well classification models can find solutions that are capable of differentiating emotional states induced by stimuli with vastly different perceptual properties. Stephens and colleagues (2010) examined the extent to which measures of heart rate variability, peripheral vascular changes, systolic time intervals, respiratory changes, and electrodermal activity could be used to predict the induction of contentment, amusement, surprise, fear, anger, sadness and a neutral state from across both music and film. Discriminant analysis using ANS measures could correctly classify emotional state with an accuracy of 44.6% given chance rates of 14.3%. While accuracy rates were above chance for both music and film inductions, the weights of discriminant functions were not examined, possibly missing the role of constructs such as valence or arousal in the differentiation of autonomic responses.

Using a modified version of the Stephens et al. (2010) paradigm, Kragel and LaBar (2013) classified the same seven emotional states using electrodermal, cardiovascular, respiratory, and gastric activity with a support vector machine algorithm using a radial basis function kernel. An observed classification accuracy of 58.0% confirmed the extent to which

autonomic patterns varied between emotional states. In order to characterize the structure of information used in classification, the distribution of errors produced from classifying autonomic responses was correlated with the distance between ratings of experienced emotion using dimensional ratings of valence and arousal and categorical items separately (Fig 3). This analysis revealed that the structure of information used in classifying autonomic responses better conformed to a categorical rather than dimensional configuration of affective space (i.e. a multidimensional space where each dimension is defined based on the rating on a self-report item). More specifically, errors in classifying autonomic measures better correlated with the similarity of categorical items of emotional experience compared to dimensional items. This finding contradicts earlier studies identifying dimensional constructs within autonomic patterns (e.g. Nyklicek et al., 1997; Schwartz et al., 1981), though the incorporation of additional autonomic measures and more complex algorithms may have moved the classification solution away from information that varies continuously across multiple emotions.

In sum, this body of work using MVPC shows there is information in patterns of ANS activity that can be mapped to distinct emotions, although the true nature of this information is far from clear. There is mixed evidence concerning the extent to which emotion specific patterns are organized along a dimension of arousal, as only 3 of the 8 studies reviewed identified such a relationship. Only a single study identified discriminatory patterns which corresponded to the experience of valence, which only accounted for 10% of variance in autonomic responses (Nyklicek et al., 1997). Further, Kragel and LaBar (2013) directly compared the structure of information in autonomic patterning to that of subjective emotional experience, and found that classifiers did not produce errors corresponding to constructs of valence and arousal. Given these differences, it is unlikely that information contained in the ANS is represented equivalently as emotional states are experienced (either as valence and arousal or discrete categories), but that information is organized in an alternative low dimensional embedding, or configuration, of affective space. One possibility is that ANS activity follows along such dimensions within regions of affective space, for example sympathetic activation tracking experienced arousal, but interactions between sympathetic and parasympathetic branches produce nonlinear regions (see Berntson, Cacioppo, Quigley, & Fabro, 1994), which complicate a direct mapping.

Future directions and concluding remarks

Studies decoding emotional states have demonstrated that patterns of central and autonomic nervous system activity carry emotion-specific information, yet much remains to be understood about how this information is integrated into a distinct, coherent emotional experience. The representational format of information remains to be precisely characterized along psychological models of emotion. In addition, the extent to which emotion-specific activation patterns are essential to the perception, experience or expression of emotion remains to be determined. Altering the focus from asking if distinct emotional states are represented in nervous system activity to how the structure of information in neural activity maps on to unique emotional states allows theoretical and conceptual models of emotion to be effectively compared, addressing longstanding debates and advancing the field.

Despite the successful application of MVPC to decode emotional states from nervous system activity, the true nature of information driving classification remains elusive. The majority of studies reviewed primarily tested the hypothesis that nervous system activity could discriminate between distinct emotional states, rather than examining if the structure of observed patterns better conformed to one emotional model over another (but see Kragel & LaBar, 2013). One avenue for future studies is to complement the use of classification with other multivariate approaches such as representational similarity analysis (Kriegeskorte et al., 2008). This approach provides a method for making comparisons between theoretical and computational models, self-report, nervous system activity, and experienced affect. By characterizing the similarity of response patterns in an equivalent manner across different systems, representational similarity analysis offers a more formal framework for testing the presence of emotional constructs than exploratory dimension reduction techniques (e.g. using principal component analysis or multidimensional scaling). Future research should complement multivariate approaches with formal model comparisons to examine how well neural systems map on to aspects of psychological theories.

Additionally, it remains unclear whether specific neural systems are essential for the emergence of distinct emotional states. Presumably, if the information driving classification stems from the activation of emotion specific neural populations, then manipulations capable of disrupting their function, such as pharmacological manipulations or transcranial magnetic stimulation, should alter multivariate activation patterns and emotional responses in a consistent manner. Tests of this nature are critical because successful MVPC within a region does not guarantee neural activity within the region is specific to the proposed function (Bartels, Logothetis, & Moutoussis, 2008). It is possible that modulatory influences from presynaptic inputs or local processing in a given region, rather than function specific neural firing, are reflected in fMRI activation. For this reason, it is critical that future work is capable of both modulating the profiles of neural activation within a region and causally altering behavior to test whether multivariate patterns are essential to a particular mental state. However, if the systems involved are too broad or redundantly encode information-- as the present evidence suggests-- it may be challenging to completely disrupt or alter the emotional process involved.

Although much remains to be understood about the nature of emotional responses in the nervous system, a growing body of work has established MVPC as a tenable method for comparing different models and theories of emotion. Moving from univariate approaches to predictive multivariate models allows researchers to infer the mental or emotional state of individuals in a formal statistical framework. This advance stands to increase the synergy between affective neuroscience research and psychological theories of emotion, with the promise of yielding a better understanding of what defines an emotion.

Acknowledgments

The authors would like to thank R. McKell Carter for insightful comments and suggestions on a previous version of this manuscript. This work was supported in part by NIH grants R01 DA027802 and R21 MH098149.

References

1. Adolphs R. Recognizing emotion from facial expressions: psychological and neurological mechanisms. *Behav Cogn Neurosci Rev.* 2002; 1(1):21–62. [PubMed: 17715585]
2. Anders S, Lotze M, Erb M, Grodd W, Birbaumer N. Brain activity underlying emotional valence and arousal: a response-related fMRI study. *Hum Brain Mapp.* 2004; 23(4):200–209. DOI: 10.1002/hbm.20048 [PubMed: 15449355]
3. Anderson AK, Christoff K, Stappen I, Panitz D, Ghahremani DG, Glover G, ... Sobel N. Dissociated neural representations of intensity and valence in human olfaction. *Nat Neurosci.* 2003; 6(2):196–202. DOI: 10.1038/nn1001 [PubMed: 12536208]
4. Arnold, MB. Emotion and personality. New York: Columbia University Press; 1960.
5. Ashby FG, Isen AM, Turken AU. A neuropsychological theory of positive affect and its influence on cognition. *Psychological Review.* 1999; 106(3):529–550. [PubMed: 10467897]
6. Barrett LF. Are Emotions Natural Kinds? Perspectives on Psychological Science. 2006a; 1(1):28–58. DOI: 10.1111/j.1745-6916.2006.00003.x [PubMed: 26151184]
7. Barrett LF. Solving the emotion paradox: categorization and the experience of emotion. *Pers Soc Psychol Rev.* 2006b; 10(1):20–46. DOI: 10.1207/s15327957pspr1001_2 [PubMed: 16430327]
8. Barrett LF, Wager TD. The structure of emotion - Evidence from neuroimaging studies. *Current Directions in Psychological Science.* 2006; 15(2):79–83. DOI: 10.1111/j.0963-7214.2006.00411.x
9. Bartels A, Logothetis NK, Moutoussis K. fMRI and its interpretations: an illustration on directional selectivity in area V5/MT. *Trends Neurosci.* 2008; 31(9):444–453. DOI: 10.1016/j.tins.2008.06.004 [PubMed: 18676033]
10. Baucom LB, Wedell DH, Wang J, Blitzer DN, Shinkareva SV. Decoding the neural representation of affective states. *Neuroimage.* 2012; 59(1):718–727. S1053-8119(11)00814-7 [pii]. DOI: 10.1016/j.neuroimage.2011.07.037 [PubMed: 21801839]
11. Berntson GG, Cacioppo JT, Quigley KS, Fabro VT. Autonomic space and psychophysiological response. *Psychophysiology.* 1994; 31(1):44–61. [PubMed: 8146254]
12. Brouwer GJ, Heeger DJ. Decoding and reconstructing color from responses in human visual cortex. *J Neurosci.* 2009; 29(44):13992–14003. 29/44/13992 [pii]. DOI: 10.1523/JNEUROSCI.3577-09.2009 [PubMed: 19890009]
13. Christie IC, Friedman BH. Autonomic specificity of discrete emotion and dimensions of affective space: a multivariate approach. *International Journal of Psychophysiology.* 2004; 51(2):143–153. DOI: 10.1016/j.ijpsycho.2003.08.002 [PubMed: 14693364]
14. Colibazzi T, Posner J, Wang Z, Gorman D, Gerber A, Yu S, ... Peterson BS. Neural systems subserving valence and arousal during the experience of induced emotions. *Emotion.* 2010; 10(3):377–389. 2010-09991-007 [pii]. DOI: 10.1037/a0018484 [PubMed: 20515226]
15. Damasio AR, Grabowski TJ, Bechara A, Damasio H, Ponto LL, Parvizi J, Hichwa RD. Subcortical and cortical brain activity during the feeling of self-generated emotions. *Nat Neurosci.* 2000; 3(10):1049–1056. DOI: 10.1038/79871 [PubMed: 11017179]
16. Davis M, Whalen PJ. The amygdala: vigilance and emotion. *Mol Psychiatry.* 2001; 6(1):13–34. [PubMed: 11244481]
17. Duda, RO.; Hart, PE.; Stork, DG. Pattern classification. 2. New York: Wiley; 2001.
18. Ekman P. An Argument for Basic Emotions. *Cognition & Emotion.* 1992; 6(3–4):169–200.
19. Ekman P, Cordaro D. What is Meant by Calling Emotions Basic. *Emotion Review.* 2011; 3(4):364–370. DOI: 10.1177/1754073911410740
20. Ethofer T, Van De Ville D, Scherer K, Vuilleumier P. Decoding of emotional information in voice-sensitive cortices. *Curr Biol.* 2009; 19(12):1028–1033. S0960-9822(09)01053-7 [pii]. DOI: 10.1016/j.cub.2009.04.054 [PubMed: 19446457]
21. Fehr B, Russell JA. Concept of emotion viewed from a prototype perspective. *Journal of Experimental Psychology: General.* 1984; 113(3):464–486. DOI: 10.1037/0096-3445.113.3.464
22. Formisano E, De Martino F, Bonte M, Goebel R. “Who” is saying “what”? Brain-based decoding of human voice and speech. *Science.* 2008; 322(5903):970–973. 322/5903/970 [pii]. DOI: 10.1126/science.1164318 [PubMed: 18988858]

23. Freeman J, Brouwer GJ, Heeger DJ, Merriam EP. Orientation Decoding Depends on Maps, Not Columns. *Journal of Neuroscience*. 2011; 31(13):4792–4804. DOI: 10.1523/Jneurosci.5160-10.2011 [PubMed: 21451017]
24. Friedman BH. Feelings and the body: The Jamesian perspective on autonomic specificity of emotion. *Biological Psychology*. 2010; 84(3):383–393. DOI: 10.1016/j.biopsycho.2009.10.006 [PubMed: 19879320]
25. Frijda NH. Emotion, cognitive structure, and action tendency. *Cognition & Emotion*. 1987; 1(2): 115–143. DOI: 10.1080/02699938708408043
26. Frith CD, Frith U. The neural basis of mentalizing. *Neuron*. 2006; 50(4):531–534. S0896-6273(06)00344-8 [pii]. DOI: 10.1016/j.neuron.2006.05.001 [PubMed: 16701204]
27. Gardner JL. Is cortical vasculature functionally organized? *Neuroimage*. 2010; 49(3):1953–1956. S1053-8119(09)00768-X [pii]. DOI: 10.1016/j.neuroimage.2009.07.004 [PubMed: 19596071]
28. Gerber AJ, Posner J, Gorman D, Colibazzi T, Yu S, Wang Z, ... Peterson BS. An affective circumplex model of neural systems subserving valence, arousal, and cognitive overlay during the appraisal of emotional faces. *Neuropsychologia*. 2008; 46(8):2129–2139. S0028-3932(08)00091-2 [pii]. DOI: 10.1016/j.neuropsychologia.2008.02.032 [PubMed: 18440572]
29. Goldin PR, McRae K, Ramel W, Gross JJ. The neural bases of emotion regulation: reappraisal and suppression of negative emotion. *Biol Psychiatry*. 2008; 63(6):577–586. S0006-3223(07)00592-6 [pii]. DOI: 10.1016/j.biopsycho.2007.05.031 [PubMed: 17888411]
30. Hamann S. Mapping discrete and dimensional emotions onto the brain: controversies and consensus. *Trends Cogn Sci*. 2012; 16(9):458–466. S1364-6613(12)00172-6 [pii]. DOI: 10.1016/j.tics.2012.07.006 [PubMed: 22890089]
31. Hanson SJ, Matsuka T, Haxby JV. Combinatorial codes in ventral temporal lobe for object recognition: Haxby (2001) revisited: is there a “face” area? *Neuroimage*. 2004; 23(1):156–166. S105381190400299X [pii]. DOI: 10.1016/j.neuroimage.2004.05.020 [PubMed: 15325362]
32. Harrison SA, Tong F. Decoding reveals the contents of visual working memory in early visual areas. *Nature*. 2009; 458(7238):632–635. nature07832 [pii]. DOI: 10.1038/nature07832 [PubMed: 19225460]
33. Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*. 2001; 293(5539):2425–2430. 293/5539/2425 [pii]. DOI: 10.1126/science.1063736 [PubMed: 11577229]
34. Haynes JD, Rees G. Decoding mental states from brain activity in humans. *Nat Rev Neurosci*. 2006; 7(7):523–534. nrn1931 [pii]. DOI: 10.1038/nrn1931 [PubMed: 16791142]
35. Hughes G. On the mean accuracy of statistical pattern recognizers. *Information Theory, IEEE Transactions on*. 1968; 14(1):55–63. citeulike-article-id:4122657. DOI: 10.1109/TIT.1968.1054102
36. Jain AK, Duin RPW, Mao JC. Statistical pattern recognition: A review. *Ieee Transactions on Pattern Analysis and Machine Intelligence*. 2000; 22(1):4–37.
37. Kalisch R, Wiech K, Critchley HD, Dolan RJ. Levels of appraisal: a medial prefrontal role in high-level appraisal of emotional material. *Neuroimage*. 2006; 30(4):1458–1466. S1053-8119(05)02470-5 [pii]. DOI: 10.1016/j.neuroimage.2005.11.011 [PubMed: 16388969]
38. Kamitani Y, Tong F. Decoding the visual and subjective contents of the human brain. *Nat Neurosci*. 2005; 8(5):679–685. nn1444 [pii]. DOI: 10.1038/nn1444 [PubMed: 15852014]
39. Kamitani Y, Tong F. Decoding seen and attended motion directions from activity in the human visual cortex. *Curr Biol*. 2006; 16(11):1096–1102. S0960-9822(06)01464-3 [pii]. DOI: 10.1016/j.cub.2006.04.003 [PubMed: 16753563]
40. Kanwisher N, McDermott J, Chun MM. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J Neurosci*. 1997; 17(11):4302–4311. [PubMed: 9151747]
41. Kassam KS, Markey AR, Cherkassky VL, Loewenstein G, Just MA. Identifying Emotions on the Basis of Neural Activation. *PLoS One*. 2013; 8(6):e66032. PONED-12-32413 [pii]. doi: 10.1371/journal.pone.0066032 [PubMed: 23840392]
42. Kay KN, Naselaris T, Prenger RJ, Gallant JL. Identifying natural images from human brain activity. *Nature*. 2008; 452(7185):352–355. nature06713 [pii]. DOI: 10.1038/nature06713 [PubMed: 18322462]

43. Kober H, Barrett LF, Joseph J, Bliss-Moreau E, Lindquist K, Wager TD. Functional grouping and cortical-subcortical interactions in emotion: a meta-analysis of neuroimaging studies. *Neuroimage*. 2008; 42(2):998–1031. S1053-8119(08)00294-2 [pii]. DOI: 10.1016/j.neuroimage.2008.03.059 [PubMed: 18579414]
44. Kolodyazhnyi V, Kreibig SD, Gross JJ, Roth WT, Wilhelm FH. An affective computing approach to physiological emotion specificity: toward subject-independent and stimulus-independent classification of film-induced emotions. *Psychophysiology*. 2011; 48(7):908–922. DOI: 10.1111/j.1469-8986.2010.01170.x [PubMed: 21261632]
45. Kotz SA, Kalberlah C, Bahlmann J, Friederici AD, Haynes JD. Predicting vocal emotion expressions from the human brain. *Hum Brain Mapp*. 2012; doi: 10.1002/hbm.22041
46. Kragel PA, LaBar KS. Multivariate Pattern Classification Reveals Autonomic and Experiential Representations of Discrete Emotions. *Emotion*. 2013; 2013-10201-001 [pii]. doi: 10.1037/a0031820
47. Kreibig SD, Wilhelm FH, Roth WT, Gross JJ. Cardiovascular, electrodermal, and respiratory response patterns to fear- and sadness-inducing films. *Psychophysiology*. 2007; 44(5):787–806. DOI: 10.1111/j.1469-8986.2007.00550.x [PubMed: 17598878]
48. Kriegeskorte N, Goebel R, Bandettini P. Information-based functional brain mapping. *Proc Natl Acad Sci U S A*. 2006; 103(10):3863–3868. 0600244103 [pii]. DOI: 10.1073/pnas.0600244103 [PubMed: 16537458]
49. Kriegeskorte N, Mur M, Bandettini P. Representational similarity analysis - connecting the branches of systems neuroscience. *Front Syst Neurosci*. 2008; 2:4.doi: 10.3389/neuro.06.004.2008 [PubMed: 19104670]
50. Kriegeskorte N, Simmons WK, Bellgowan PS, Baker CI. Circular analysis in systems neuroscience: the dangers of double dipping. *Nat Neurosci*. 2009; 12(5):535–540. nn.2303 [pii]. DOI: 10.1038/nn.2303 [PubMed: 19396166]
51. Liberzon I, Phan KL, Decker LR, Taylor SF. Extended amygdala and emotional salience: a PET activation study of positive and negative affect. *Neuropsychopharmacology*. 2003; 28(4):726–733. 1300113 [pii]. DOI: 10.1038/sj.npp.1300113 [PubMed: 12655318]
52. Lindquist KA, Barrett LF. A functional architecture of the human brain: emerging insights from the science of emotion. *Trends Cogn Sci*. 2012; 16(11):533–540. S1364-6613(12)00221-5 [pii]. DOI: 10.1016/j.tics.2012.09.005 [PubMed: 23036719]
53. Lindquist KA, Wager TD, Kober H, Bliss-Moreau E, Barrett LF. The brain basis of emotion: a meta-analytic review. *Behavioral and Brain Sciences*. 2012; 35(3):121–143. S0140525X11000446 [pii]. DOI: 10.1017/S0140525X11000446 [PubMed: 22617651]
54. Martin A, Chao LL. Semantic memory and the brain: structure and processes. *Current Opinion in Neurobiology*. 2001; 11(2):194–201. S0959-4388(00)00196-3 [pii]. [PubMed: 11301239]
55. Misaki M, Kim Y, Bandettini PA, Kriegeskorte N. Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. *Neuroimage*. 2010; 53(1):103–118. S1053-8119(10)00783-4 [pii]. DOI: 10.1016/j.neuroimage.2010.05.051 [PubMed: 20580933]
56. Nelson SM, Dosenbach NU, Cohen AL, Wheeler ME, Schlaggar BL, Petersen SE. Role of the anterior insula in task-level control and focal attention. *Brain Struct Funct*. 2010; 214(5–6):669–680. DOI: 10.1007/s00429-010-0260-2 [PubMed: 20512372]
57. Norman KA, Polyn SM, Detre GJ, Haxby JV. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn Sci*. 2006; 10(9):424–430. S1364-6613(06)00184-7 [pii]. DOI: 10.1016/j.tics.2006.07.005 [PubMed: 16899397]
58. Nyklicek I, Thayer JF, van Doornen LJP. Cardiorespiratory differentiation of musically-induced emotions. *Journal of Psychophysiology*. 1997; 11(4):304–321.
59. Ochsner KN. Current directions in social cognitive neuroscience. *Curr Opin Neurobiol*. 2004; 14(2):254–258. S095943880400042X [pii]. DOI: 10.1016/j.conb.2004.03.011 [PubMed: 15082333]
60. Ochsner KN, Gross JJ. The cognitive control of emotion. *Trends Cogn Sci*. 2005; 9(5):242–249. S1364-6613(05)00090-2 [pii]. DOI: 10.1016/j.tics.2005.03.010 [PubMed: 15866151]
61. Ochsner KN, Knierim K, Ludlow DH, Hanelin J, Ramachandran T, Glover G, Mackey SC. Reflecting upon feelings: an fMRI study of neural systems supporting the attribution of emotion to

- self and other. *J Cogn Neurosci*. 2004; 16(10):1746–1772. DOI: 10.1162/0898929042947829 [PubMed: 15701226]
62. Op de Beeck HP. Against hyperacuity in brain reading: spatial smoothing does not hurt multivariate fMRI analyses? *Neuroimage*. 2010; 49(3):1943–1948. S1053-8119(09)00202-X [pii]. DOI: 10.1016/j.neuroimage.2009.02.047 [PubMed: 19285144]
 63. Panksepp J. Toward a General Psycho-Biological Theory of Emotions. *Behavioral and Brain Sciences*. 1982; 5(3):407–422.
 64. Panksepp, J. *Affective neuroscience : the foundations of human and animal emotions*. New York: Oxford University Press; 1998.
 65. Paton JJ, Belova MA, Morrison SE, Salzman CD. The primate amygdala represents the positive and negative value of visual stimuli during learning. *Nature*. 2006; 439(7078):865–870. nature04490 [pii]. DOI: 10.1038/nature04490 [PubMed: 16482160]
 66. Pessoa L, Padmala S. Quantitative prediction of perceptual decisions during near-threshold fear detection. *Proc Natl Acad Sci U S A*. 2005; 102(15):5612–5617. 0500566102 [pii]. DOI: 10.1073/pnas.0500566102 [PubMed: 15800041]
 67. Pessoa L, Padmala S. Decoding near-threshold perception of fear from distributed single-trial brain activation. *Cereb Cortex*. 2007; 17(3):691–701. bhk020 [pii]. DOI: 10.1093/cercor/bhk020 [PubMed: 16627856]
 68. Phan KL, Taylor SF, Welsh RC, Decker LR, Noll DC, Nichols TE, ... Liberzon I. Activation of the medial prefrontal cortex and extended amygdala by individual ratings of emotional arousal: a fMRI study. *Biol Psychiatry*. 2003; 53(3):211–215. S0006322302014853 [pii]. [PubMed: 12559653]
 69. Phan KL, Wager T, Taylor SF, Liberzon I. Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI. *Neuroimage*. 2002; 16(2):331–348. DOI: 10.1006/nimg.2002.1087 [PubMed: 12030820]
 70. Phillips AG, Ahn S, Howland JG. Amygdalar control of the mesocorticolimbic dopamine system: parallel pathways to motivated behavior. *Neurosci Biobehav Rev*. 2003; 27(6):543–554. S0149763403001015 [pii]. [PubMed: 14599435]
 71. Poldrack RA. Can cognitive processes be inferred from neuroimaging data? *Trends Cogn Sci*. 2006; 10(2):59–63. S1364-6613(05)00336-0 [pii]. DOI: 10.1016/j.tics.2005.12.004 [PubMed: 16406760]
 72. Poldrack RA, Halchenko YO, Hanson SJ. Decoding the large-scale structure of brain function by classifying mental States across individuals. *Psychol Sci*. 2009; 20(11):1364–1372. PSCI2460 [pii]. DOI: 10.1111/j.1467-9280.2009.02460.x [PubMed: 19883493]
 73. Posner J, Russell JA, Gerber A, Gorman D, Colibazzi T, Yu S, ... Peterson BS. The neurophysiological bases of emotion: An fMRI study of the affective circumplex using emotion-denoting words. *Hum Brain Mapp*. 2009; 30(3):883–895. DOI: 10.1002/hbm.20553 [PubMed: 18344175]
 74. Rainville P, Bechara A, Naqvi N, Damasio AR. Basic emotions are associated with distinct patterns of cardiorespiratory activity. *International Journal of Psychophysiology*. 2006; 61(1):5–18. DOI: 10.1016/j.ijpsycho.2005.10.024 [PubMed: 16439033]
 75. Rolls ET. A Theory of Emotion, and Its Application to Understanding the Neural Basis of Emotion. *Cognition & Emotion*. 1990; 4(3):161–190. DOI: 10.1080/02699939008410795
 76. Rolls ET, Grabenhorst F, Franco L. Prediction of subjective affective state from brain activations. *J Neurophysiol*. 2009; 101(3):1294–1308. 91049.2008 [pii]. DOI: 10.1152/jn.91049.2008 [PubMed: 19109452]
 77. Russell JA. A circumplex model of affect. *Journal of Personality and Social Psychology*. 1980; 39(6):1161–1178. DOI: 10.1037/h0077714
 78. Said CP, Moore CD, Engell AD, Todorov A, Haxby JV. Distributed representations of dynamic facial expressions in the superior temporal sulcus. *J Vis*. 2010; 10(5):11. 10.5.11 [pii]. doi: 10.1167/10.5.11 [PubMed: 20616141]
 79. Scherer, KR. On the nature and function of emotion: a component process approach. In: Scherer, KR.; Ekman, P., editors. *Approaches to emotion*. Hillsdale, N.J: L. Erlbaum Associates; 1984. p. 293-317.

80. Schwartz GE, Weinberger DA, Singer JA. Cardiovascular Differentiation of Happiness, Sadness, Anger, and Fear Following Imagery and Exercise. *Psychosomatic Medicine*. 1981; 43(4):343–364. [PubMed: 7280162]
81. Sinha R, Parsons OA. Multivariate response patterning of fear and anger. *Cognition & Emotion*. 1996; 10(2):173–198. DOI: 10.1080/026999396380321
82. Sitaram R, Lee S, Ruiz S, Rana M, Veit R, Birbaumer N. Real-time support vector classification and feedback of multiple emotional brain states. *Neuroimage*. 2011; 56(2):753–765. S1053-8119(10)01074-8 [pii]. DOI: 10.1016/j.neuroimage.2010.08.007 [PubMed: 20692351]
83. Smith CA, Ellsworth PC. Patterns of cognitive appraisal in emotion. *J Pers Soc Psychol*. 1985; 48(4):813–838. [PubMed: 3886875]
84. Stephens CL, Christie IC, Friedman BH. Autonomic specificity of basic emotions: Evidence from pattern classification and cluster analysis. *Biological Psychology*. 2010; 84(3):463–473. DOI: 10.1016/j.biopsycho.2010.03.014 [PubMed: 20338217]
85. Todd MT, Nystrom LE, Cohen JD. Confounds in multivariate pattern analysis: Theory and rule representation case study. *Neuroimage*. 2013; 77:157–165. DOI: 10.1016/j.neuroimage.2013.03.039 [PubMed: 23558095]
86. Tomkins, SS. *Affect, imagery, consciousness*. New York: Springer Pub. Co; 1962.
87. Tong F, Pratte MS. Decoding patterns of human brain activity. *Annu Rev Psychol*. 2012; 63:483–509. DOI: 10.1146/annurev-psych-120710-100412 [PubMed: 21943172]
88. Wager, T.; Barrett, LF.; Bliss-Moreau, E.; Lindquist, K.; Duncan, S.; Kober, H.; ... Mize, J. The Neuroimaging of Emotion. In: Lewis, M.; Haviland-Jones, JM.; Barrett, LF., editors. *Handbook of emotions*. 3. New York: Guilford Press; 2008. p. 249-267.
89. Winston JS, O'Doherty J, Dolan RJ. Common and distinct neural responses during direct and incidental processing of multiple facial emotions. *Neuroimage*. 2003; 20(1):84–97. S1053811903003033 [pii]. [PubMed: 14527572]
90. Yarkoni T, Poldrack RA, Nichols TE, Van Essen DC, Wager TD. Large-scale automated synthesis of human functional neuroimaging data. *Nat Methods*. 2011; 8(8):665–670. nmeth.1635 [pii]. DOI: 10.1038/nmeth.1635 [PubMed: 21706013]

Figure 1. Different approaches for testing the functional organization of emotion in the brain with MVPC

The flexibility of analysis parameters in pattern classification permits posing different experimental questions about the neural basis of emotion under the same framework (left column). By adjusting the method of feature selection employed (middle column), information contained in the activity of large scale networks, specific regions, or searchlights throughout the brain focus analysis onto different spatial scales. Adopting different learning algorithms and classification schema (right column) permits testing of diverse hypotheses.

Figure 2. Redundant encoding of subjective valence in BOLD response patterns

Classification accuracy obtained using an increasing number of voxels in discriminating positive and negative valence. (A) Solid lines depict gains in information and accuracy (left panels), which increase asymptotically in a nonlinear fashion when voxels from mPFC (right panel) are added to classification¹. (B) Box-plot (left) illustrates that including additional voxels in a whole brain feature search improves the classification accuracy of valence in a sub-linear fashion².

1. Adapted from “Prediction of subjective affective state from brain activations,” by E. T. Rolls, F. Grabenhorst, & L. Franco, 2009, *J Neurophysiol*, 101 p. 1299. Copyright 2009 by the American Physiological Society. 2. Adapted from “Decoding the neural representation of affective states,” by L. B. Baucom, D. H. Wedell, J. Wang, D. N. Blitzer, & S. V. Shinkareva, 2012, *NeuroImage*, 59 ps. 723–724. Copyright 2011 by Elsevier Inc. Adapted with permission.

Figure 3. Relating classification errors to experienced affect

Scatterplots depicting the relationship between the number of errors made in pattern classification and the subjective experience of emotion for 21 pairwise combinations of seven distinct emotional states (Kragel & LaBar, 2013). The number of errors in classifying autonomic response patterns decreases with Euclidean distance in an affective space created using the experience of categorical items (e.g. ‘fear’, ‘anger’, ‘sadness’; white circles in right panel) but not in a space constructed using dimensional items of valence and arousal (left panel). Adapted from “Multivariate pattern classification reveals autonomic and experiential representations of discrete emotions,” by P. A. Kragel & K. S. LaBar, *Emotion*, 13, *in press*. Copyright 2013 by the American Psychological Association. Adapted with permission.

Table 1

Studies decoding patterns of BOLD fMRI response into distinct emotions.

Said et al. 2010	Categorical	Anger, Disgust, Fear, Sadness, Happiness, Surprise	Perception	Visual	Multiple <i>a priori</i> regions of interest	Logistic regression (linear)	Run
Pessoa and Padmala 2007	Categorical	Fear, No Fear	Perception	Visual	Multiple <i>a priori</i> regions of interest	SVM (linear and nonlinear)	Trial
Ethofer et al. 2009	Categorical	Anger, Sadness, Joy, Relief, Neutral	Perception	Auditory	Single <i>a priori</i> region of interest	SVM (linear)	Trial
Kotz et al. 2012	Categorical	Anger, Sadness, Happiness, Surprise, Neutral	Perception	Auditory	Whole brain searchlight	SVM (linear)	Run
Baucom et al. 2012	Dimensional	P, N or LA, HA or HAN, LAN, LAP, HAP	Experience	Visual	Whole brain	Logistic regression (linear)	Subject, run
Rolls et al. 2009	Dimensional	P, N	Experience	Thermal	Multiple <i>a priori</i> regions of interest	Probability estimation, multilayer perceptron, SVM (linear)	Trial
Sitaram et al. 2011	Categorical	Happiness, Disgust or Sadness, Disgust, Happiness	Experience	Imagery	Whole brain and multiple <i>a priori</i> regions of interest	SVM (linear)	Subject, run
Kassam et al. 2013	Categorical	Anger, Disgust, Envy, Fear, Happiness, Lust, Pride, Sadness, Shame	Experience	Imagery	Whole brain	Gaussian Naïve Bayes (linear)	Subject, trial

Note. P = positive, N = negative, LA = low arousal, HA = high arousal, HAN = high arousal negative, LAN = low arousal negative, LAP = low arousal positive, HAP = high arousal positive, SVM = support vector machine

Table 2
Studies decoding patterns of autonomic nervous system activity into distinct emotional states

Study	Class Labels	Stimulus Modality	ANS Measures	Classification Algorithm	Classification Accuracy (Chance; <i>I</i>)
Schwartz et al., 1981	Happiness, Sadness, Anger, Fear, Relaxation, Neutral	Imagery	Cardiovascular	LDA	42.6% (16.6%; .312)
Sinha and Parsons, 1996	Fear, Anger, Neutral	Imagery	Cardiovascular, electrodermal, and thermal	LDA	66.5% (33.3%; .498)
Rainville et al., 2006	Fear, Anger, Sadness, Happiness	Imagery	Cardiovascular and respiratory	LDA	49.0% (25.0%; .320)
Christie and Friedman, 2004	Amusement, Anger, Contentment, Disgust, Fear, Sadness, Neutral	Audiovisual	Cardiovascular and electrodermal	LDA	37.4% (14.3%; .270)
Kreibig et al., 2007	Fear, Sadness, Neutral	Audiovisual	Cardiovascular, electrodermal, and respiratory	LDA *	69.0% (33.3%; .535)
Nyklíček et al., 1997	Happiness, Serenity, Sadness, Agitation	Auditory	Cardiovascular and respiratory	LDA	46.5% (20%; .331)
Stephens et al., 2010	Amusement, Contentment, Surprise, Fear, Anger, Sadness, Neutral	Audiovisual	Cardiovascular, electrodermal, and respiratory	LDA	44.6% (14.3%; .354)
Kragel and LaBar, 2013	Amusement, Contentment, Surprise, Fear, Anger, Sadness, Neutral	Audiovisual	Cardiovascular, electrodermal, gastric, and respiratory	SVM (linear and nonlinear)	58.0% (14.3%; .510)

Note. LDA = linear discriminant analysis, SVM = support vector machine, *I* = proportional reduction in error rate.

* Kolodyazhni and colleagues (2011) reanalyzed this dataset using quadratic discriminant analysis, radial basis function neural networks, multilayer perceptrons, and k-nearest neighbors clustering – with nonlinear algorithms showing improvements over linear methods