Evaluating the effects of regularization and cross-validation parameters on the performance of SVM-based decoding of EEG data

Guanghui Zhang[1,2,3] , Xinran Wang[1,2], and Steven J. Luck[3]

[1]Institute of Psychological and Brain Sciences, Liaoning Normal University, Dalian, Liaoning, China

[2]Key Laboratory of Brain and Cognitive Neuroscience, Liaoning Province, Dalian, China

[3]Center for Mind & Brain, University of California-Davis, Davis, CA, USA

Correspondence

Guanghui Zhang, Institute of Psychological and Brain Sciences, Liaoning Normal University, No. 850, Huanghe Road, Shahekou District, Dalian 116029, China. Email: zhang.guanghui@foxmail.com

Funding Information

This study was made possible by grants R01MH087450 and R01EY033329 from the National Institutes of Health to Steven J. Luck; This study was also supported by Liaoning Normal University High-level Scientific Research Achievements Cultivation Project (No. 25GDL004), and the Natural Science Foundation of Liaoning Province (No. 2025-BS-0780) to Guanghui Zhang.

Abstract

Regularization has been extensively used in multivariate pattern classification (MVPA; decoding) of EEG data to mitigate the risk of overfitting. N-fold cross-validation is also used to mitigate this risk, and it is often combined with averaging across trials to improve the signal-to-noise ratio. However, the impact of different regularization and cross-validation parameters on decoding performance remains unclear. This study aimed to evaluate the effects of variations in the support vector machine (SVM) regularization parameter (C) and the number of crossfolds (and the number of trials per average) on the performance of SVM-based decoding analyses. To achieve this, we examined the decoding performance in relatively simple binary classification tasks from seven commonly used event-related potential paradigms (N170, mismatch negativity, N2pc, P3b, N400, lateralized readiness potential, and error-related negativity). Additionally, we evaluated the decoding performance in more challenging multiclass tasks, including decoding face identity, facial expression, stimulus location, and stimulus orientation. The results revealed that both decoding accuracy and effect size were highest when the regularization strength was equal to or greater than 1. Furthermore, using between 3 and 5 folds with at least 10 trials per average yielded optimal decoding performance in most cases. Researchers applying SVM-based decoding to datasets similar to those examined here—in terms of population, recording systems, class numbers, and paradigms—might benefit from using the parameters that we found to be optimal here.

Keywords: EEG; ERP; MVPA; regularization; cross-validation; pseudotrial

1. Introduction

In recent decades, multivariate pattern classification (MVPA or decoding) has gained growing interest in electroencephalogram (EEG) and event-related potential (ERP) research. This is due to the ability of MVPA to detect subtle differences between experimental conditions that traditional univariate approaches often fail to identify (Carrasco et al., 2024; Hebart & Baker, 2018; Peelen & Downing, 2023). For example, Carrasco et al. (2024) demonstrated that MVPA produced larger effect sizes across a wide range of ERP components compared to traditional univariate analyses. In addition, decoding analyses can handle noisier data more effectively because the machine learning algorithm can focus on the most informative channels (Ashton et al., 2022; Carlson et al., 2019; Grootswagers et al., 2017). Here we will focus on classification via support vector machines (SVMs), which outperform other classification methods in many EEG datasets (Trammel et al., 2023).

It is essential to deal with the problem of overfitting in MVPA so that the classification is based on consistent differences between classes rather than noise. In SVMs, one strategy for this is *regularization*, which helps control model complexity by penalizing large weights in the decision function. This reduces the risk of overfitting, especially when the number of features (e.g., EEG channels) is large relative to the number of trials (Hastie et al., 2009; Peng et al., 2021; Pereira et al., 2009). In many implementations of the SVM algorithm (e.g., Matlab's fitcsvm function), the regularization strength is governed by the *box constraint parameter* (C). This parameter balances the trade-off between minimizing error in the training set and maintaining a simpler decision boundary that is more likely to generalize to new cases. A value of 1 gives equal weight to these two goals. A smaller C produces greater regularization, allowing more

misclassifications in the training data in exchange for better generalization. By contrast, a larger C emphasizes correct classification of the training examples, potentially at the cost of overfitting.

A second essential technique to address overfitting in EEG/ERP decoding is *N-fold cross-validation* (Lemm et al., 2011; Valente et al., 2021; Varoquaux, 2018; Varoquaux et al., 2017) . This is often combined with averaging, which improves decoding performance by increasing the signal-to-noise ratio of the data (Grootswagers et al., 2017; Petit et al., 2020; Williams et al., 2020, 2023). Specifically, after excluding trials with artifacts or other problems, the remaining trials are divided into *N* non-overlapping subsets, and the trials within each subset are averaged together to create pseudotrials. In every round of cross-validation, one pseudotrial from each class is used for testing, while the remaining pseudotrials are used for training. This process is repeated *N* times, with each pseudotrial serving as the test set once.

For example, one of the datasets analyzed by Carrasco et al. (2024) was an N170 experiment in which the two classes of interest were photographs of faces and photographs of cars, with 80 trials for each class (prior to the rejection of trials with errors or artifacts). A 5-fold leave-one-out cross-validation procedure was used, which created 5 averaged ERPs of approximately 16 trials for each class. In the first round of cross-validation, the first averaged ERP from each class was used as the test set, while ERPs 2–5 from each class served as the training set. In the second round, ERP 2 from each class was used as the test set and ERPs 1, 3, 4, and 5 were used as the training sets. This cycled three more times, using ERPs 3, 4, and 5 from each class as the test sets and the other ERPs as the training sets.

In this procedure, the number of crossfolds (*N*) is equivalent to the number of averaged ERPs per class, and the number of trials per ERP (*T*) is the number of total trials per class divided by N (and then rounded down to the nearest integer). The total number of trials per class

is therefore N*T. Increasing the value of N will tend to increase decoding performance, all else being equal, because increasing N increases the number of training cases per class. However, increasing N reduces T and therefore decreases the signal-to-noise ratio of each training and test case, and reducing T will tend to decrease decoding performance. A central goal of the present study was to determine what balance of N and T leads to the best decoding performance. Note that we did not consider any other cross-validation approaches (e.g., single-trial analyses), and we assume that the number of trials available for decoding is held constant such that N and T are reciprocally related to each other (as would be true in most real experiments, in which the total number of trials is fixed). We use the phrase *cross-validation parameters* to refer to the combination of N and T.

In current practice, N and T vary depending on the research topic and population, and they may even vary across studies within the same laboratory. The number of crossfolds (N) typically ranges from 2 to 25 or more, with the number of trials per average (T) typically varying between around 3 and 80 (Bae & Luck, 2018; Bayet et al., 2020; Bo et al., 2022; Chaisaen et al., 2020; Hong et al., 2020; Kikumoto & Mayr, 2018; Y. Li et al., 2022; Mares et al., 2020; Sanchez et al., 2020; Sarrett & Toscano, 2024; Turoman et al., 2024; Zhang, Carrasco, et al., 2024). A few previous studies have systematically varied the cross-validation parameters to assess their impact on decoding performance using real or simulated EEG datasets. For example, using real EEG data from a single study of auditory speech sounds, Sarrett et al.(2024) demonstrated that decoding accuracy improved as T increased and N decreased when they varied T from approximately 10 to 95 and N from 3 to 25 (Sarrett & Toscano, 2024). Furthermore, a study using simulated EEG data suggested that decoding performance could be optimized by selecting

N such that T is approximately 5-10% of the total number of available trials per class (Scrivener et al., 2023).

1.1 The goal of the current study

Because these previous studies of the cross-validation parameters each focused on a single dataset (Sarrett & Toscano, 2024; Scrivener et al., 2023), it is difficult to know whether their conclusions would generalize to datasets employing different experimental paradigms, electrode densities, and numbers of stimulus classes. The goal of the present study was to provide a more systematic and generalizable assessment of the impact of the cross-validation parameters and the regularization strength on SVM-based EEG/ERP decoding performance. We accomplished this by using a wide range of EEG/ERP experiments with diverse sample sizes, a wide range of numbers of stimulus classes, and different electrode densities.

We analyzed the data from multiple publicly available EEG/ERP datasets. One of these datasets was the ERP CORE (Compendium of Open Resources and Experiments; Kappenman et al., 2021), which includes six common ERP paradigms that can be used to elicit seven distinct ERP components: N170, mismatch negativity (MMN), P3b, N400, error-related negativity (ERN), N2pc, and lateralized readiness potential (LRP). For these experiments, binary decoding was applied to the primary classes (e.g., faces versus cars for N170 the experiment, standard versus deviant auditory stimuli for the MMN experiment). These datasets consist of recordings from 32 electrode sites with a moderate density across the scalp.

We also analyzed three additional datasets that involved more challenging multiclass decoding tasks, where the differences among classes were difficult to detect using conventional univariate analyses. In one of the paradigms, the stimulus set consisted of 16 different photographs of faces, combining 4 face identities with 4 emotional expressions (Bae, 2021). We

decoded face identity while disregarding facial expression and facial expression while disregarding face identity. For this dataset, recordings were obtained from 64 electrodes, providing greater electrode density than in the other datasets

In the other two paradigms, participants were required to perceive and remember one of 16 different orientations, and data were recorded from 32 electrodes (Bae & Luck, 2018). In one of these experiments, there were also 16 different stimulus locations. For these two paradigms, we decoded the data in the frequency domain as well as in the time domain (as in the original report of these datasets).

2. Methods

The ERP CORE dataset was downloaded from https://osf.io/thsqg/, and the other datasets were downloaded from https://osf.io/tgzew/. The data analysis procedures were implemented with MATLAB 2021b (MathWorks Inc), using EEGLAB Toolbox v2023.0 (Delorme & Makeig, 2004) combined with ERPLAB Toolbox v11.00 (Lopez-Calderon & Luck, 2014). The processing scripts and resulting data files are available at https://osf.io/un6pq.

The following sections provide a brief overview of the stimuli, tasks, recording methods, and preprocessing techniques for each dataset. Comprehensive descriptions are available in the original papers that detail these datasets, which we referred to as *ERP CORE* (Kappenman et al., 2021), *Orientations* (Bae & Luck, 2018), and *Faces* (Bae, 2021) datasets. All the studies were approved by the UC Davis Institutional Review Board.

2.1 Participants

All the participants were highly cooperative neurotypical college students from the UC-Davis community. The number of participants with usable data for the ERP CORE, Orientations, and Faces experiments is shown in Table 1.

2.2 Paradigms

Each of the ERP CORE paradigms lasted approximately 10 minutes, with each participant completing all six tasks in a single session (Kappenman et al., 2021). For the N170 face perception task (Figure 1a), participants were instructed to categorize each stimulus as a "texture" (scrambled faces or scrambled cars) or an "object" (faces or cars). Here we analyzed data only from face and car stimuli. As shown in Figure 1b, a passive auditory oddball task was used to elicit the mismatch negativity (MMN). During the task, participants heard a task-irrelevant sequence of standard (80 dB) and deviant (70 dB) auditory tones while viewing a silent video. The P3b component (Figure 1c) was elicited using an active visual oddball task, in which participants made one response to target letters and a different response to non-target letters. In each block, one of five letters (A, B, C, D, E; p = .2 each) was designated the target and the remaining letters served as non-targets. The N400 component (Figure 1d) was elicited using a word-pair judgment task. Each trial consisted of a prime word and a target word, and participants were required to determine whether the target word was semantically related to the prime word (p = .5 for related and unrelated). The lateralized readiness potential (LRP) and error-related negativity (ERN; Figure 1e) were elicited via an Eriksen flankers task, in which participants responded to the direction of a central arrowhead surrounded by congruent or incongruent arrowheads by pressing corresponding buttons with their left or right hand. The N2pc component (Figure 1f) was elicited during a visual search task, in which 24 colored squares were presented on each trial participants reported the location (top or bottom) of a gap in

the one square drawn in a designated target color (pink or blue). In all tasks, participants were instructed to maintain fixation on a central point. In all but the MMN paradigm, participants pressed one of two buttons on each trial to report the category of the current-trial stimulus.

In the Orientations-1 paradigm (Bae & Luck, 2018; Figure 1g), a teardrop shape (the *sample* stimulus) appeared at the center of the display for 200 ms at the start of each trial. The teardrop could take one of 16 possible orientations, spaced 22.5° apart. After a 1300-ms blank delay, participants adjusted the orientation of a *test* teardrop using a mouse to match the remembered orientation of the sample teardrop. In the Orientations-2 paradigm (Figure 1h), the sample teardrop was again presented in one of 16 random orientations, but its location varied across 16 possible positions around an invisible circle. The orientation and location were independent of each other. After the delay period, participants reproduced the teardrop's orientation with a test teardrop in a random location, making the location task-irrelevant.

The Faces dataset comes from a single experiment (Figure 1i) (Bae, 2021). On each trial, a sample face was first presented at the center of the display for 500 ms, followed by a 1000-ms blank interval. The face was randomly chosen from a set of 16 combinations of four identities (ID 1, ID 2, ID 3, and ID 4) and four emotional expressions (neutral, fearful, happy, and angry). On 89% of trials, the delay was followed by the next face stimulus. On the remaining 11% of trials, a probe display with four faces appeared. Participants were required to select the face in the probe that matched the immediately preceding sample face in either identity or expression. Identity and expression probes were presented randomly, requiring participants to remember both aspects of each face.

2.3 EEG recording and preprocessing

For the ERP CORE paradigms, the EEG was recorded using a Biosemi ActiveTwo system (Biosemi B.V.) at a sampling rate of 1024 Hz. During acquisition, a fifth-order sinc filter with a half-power cutoff at 204.8 Hz was applied to prevent aliasing. The dataset contains EEG signals from 30 scalp electrodes: FP1/2, Fz, F3/4/7/8, FCz, FC3/4, Cz, C3/4/5/6, CPz, Pz, P3/4/7/8, PO3/4/7/8/9/10, and Oz/1/2. The dataset also contains signals from two electrooculogram (EOG) electrodes placed lateral to the eyes and one EOG electrode below the right eye.

For the Faces and Orientations paradigms, the EEG was recorded using a Brain Products actiCHamp system (Brain Products GmbH) with a sampling rate of 500 Hz. To avoid aliasing, a cascaded integrator-comb filter was applied during recording, with a half-power cutoff at 130 Hz. These datasets included the same EOG signals as in the ERP CORE dataset, as well as EEG electrodes placed at the left and right mastoids.. For both Orientations paradigms, EEG was collected from 27 EEG sites: FP1/2, Fz, F3/4/7/8, Cz, C3/4, Pz, P3/4/5/6/7/8/9/10, POz, PO3/4/7/8, and Oz/1/2. The Faces dataset includes 59 EEG sites: FP1/2, AFz/3/4/7/8, Fz/1/2/3/4/5/6/7/8, FCz/1/2/3/4/5/6, Cz/1/2/3/4/5/6, T7/8, CPz/1/2/3/4/5/6, TP7/8, Pz/1/2/3/4/5/6/7/9/10, POz/3/4/7/8, Oz/1/2.

The key preprocessing steps for the ERP CORE paradigms are provided below, and more detailed descriptions can be found in the original paper (Kappenman et al., 2021). The continuous EEG signal was downsampled to 256 Hz and referenced to the average of the P9 and P10 electrodes (located near the left and right mastoids). However, in the N170 paradigm, the average of all scalp sites served as the reference. A noncausal Butterworth bandpass filter (half amplitude cutoff = 0.1 – 30 Hz , roll-off = 12 dB/oct) was applied. Note that subsequent research has demonstrated that different filter settings should be used for each paradigm to optimize

statistical power for univariate analyses (Zhang, Garrett, & Luck et al., 2024a, 2024b), but we held the filter settings constant across paradigms here to avoid introducing a confound in comparing the different paradigms. To correct for blinks, independent component analysis (ICA; Jung, Makeig, Humphries, et al., 2000; Jung, Makeig, Westerfield, et al., 2000) was employed. Instead of the original ICA procedure of ERP CORE paradigms, we used an optimized ICA procedure that has been shown to be more effective (Dimigen, 2020; Klug & Gramann, 2021; Luck, 2022; Zhang, Garrett, Simmons, et al., 2024). In this procedure, the ICA decomposition is performed on a copy of the dataset that has been heavily filtered (1-30 Hz) and then the weights are transformed back to the original dataset prior to performing the correction. After the correction, the datasets were segmented into epochs and baseline-corrected using the time windows shown in Table 1. Channels identified as problematic were interpolated using ERPLAB's spherical spline interpolation method. The decoding analysis included all trials except those with behavioral errors, because we have shown that the rejection of trials containing artifacts has little impact on decoding performance across a broad range of EEG and ERP paradigms (Zhang & Luck, 2025) .

The remaining preprocessing steps for the Faces and Orientations paradigms mirrored those of the ERP CORE, with the exception that the data were downsampled to 250 Hz and referenced to the average of the mastoids. In addition, no trials were excluded because of behavioral errors, which were very rare. For further details, refer to Bae (2021) and Bae and Luck (2018), and consult Table 1 for epoch timing.

2.4 SVM-based decoding methods

We performed decoding multiple times for each dataset, each time using one of several different regularization or cross-validation parameters. We performed binary decoding for the ERP CORE

paradigms. In the N170, MMN, P3b, and N400 paradigms, we decoded which of two stimulus classes was presented (e.g., targets versus nontargets in the P3b paradigm). In the N2pc paradigm, we decoded whether the target was on the left or right side of the display. In the flankers paradigm, we decoded whether the participant responded with the left hand or the right hand for the LRP analysis (for correct responses), and we decoded whether the response was correct or incorrect for the ERN analysis (disregarding whether the left or right hand was used to respond).

For the Faces dataset, we performed two independent 4-class decoding runs for each set of regularization and cross-validation parameters. In the first run, we decoded which of the four face identities was present, disregarding the emotional expression of the face. In the second run, we decoded which of the four emotional expressions was displayed, disregarding the identity of the face.

When we manipulated the regularization parameter for the Orientations experiments, we classified which of the 16 orientations was presented, disregarding stimulus location. For Orientations-2, we also decoded which of the 16 stimulus locations was used, disregarding orientation. We modified the approach somewhat to test the cross-validation parameters in the Orientations experiments. There were only 40 trials per orientation or per location, which made it difficult to examine a broad set of T and N values. To assess the effects of the cross-validation parameters, we therefore collapsed the 16 orientations or the 16 locations into 4 classes (upper left, upper right, lower left, and lower right) and decoded which of these 4 orientation classes or location classes was present. This yielded 120 trials per class.

We decoded the data separately at each time point for each participant in each experiment, using the set of EEG channels as the feature set for decoding. For binary decoding, we used the MATLAB function *fitcsvm()* to train the SVM classifier. For multiclass decoding, we used the

MATLAB function *fitcecoc()* to implement the error-correcting output codes approach. After training, we used the MATLAB function *predict()* to compute decoding accuracy for the test data.

In our leave-one-out cross-validation approach, we first split the available trials for a given participant into N subsets or folds for each stimulus class, resulting in T trials per subset where T is equal to the total number of available trials per class divided by N (Zhang et al., 2025). We then averaged the T trials for each subset to create N averaged ERPs for each class. For the data at a specific time point, we used N-1 of the averaged ERPs from each class to train the SVM decoder, and we then tested the decoder with the remaining averaged ERP from each class.

We repeated this procedure N times, with each set of averaged ERPs used as the test set one time and as a training set N-1 times. The total number of tests was therefore NK for each repetition of the procedure, where K is the number of classes being decoded. We then iterated this whole process J times, resulting in a proportion correct over a total of JKN tests. The key factor in controlling the precision of a proportion value is the total number of tests. For each paradigm, we chose a number of iterations (J) that yielded 5000 tests. This held the precision of the proportion correct variable constant even though the number of classes (K) varied across paradigms.

To examine the effect of the regularization strength (the box constraint parameter C), we performed the decoding procedure for a given dataset once for each of several values of C (0.001, 0.01, 0.1, 1, 10, 100, and 1000). For these analyses, we used an N of 3 for the N-fold cross-validation procedure, with the maximum number of trials per average (T) that was possible for that N in a given dataset (see the supplementary materials for evidence that variations in C

produce similar effects whether N is 20 or 3). By holding N constant, we could ensure that any differences in decoding performance across regularization parameters reflected the effect of regularization rather than the effects of the cross-validation parameters.

To test the effects of the N and T cross-validation parameters, we held the regularization strength constant at a balanced level (C = 1), and we repeated the decoding procedure for each dataset once with each of several values of N and T. Specifically, we examined N = 2, 3, 4, 5, 6, 10, 20, and 40 crossfolds for each paradigm (except that we could not test 40 crossfolds for the P3b paradigm because the total number of trials for one condition was smaller than 40), with T set to the highest value possible given that value of N for a given dataset.

Bae and Luck (2018) found that location and orientation in the Orientations datasets could be decoded from the alpha-band EEG power as well as from the ERP amplitude (Bae & Luck, 2018). We therefore decoded each of these datasets using both alpha-band power and ERP amplitude (separately). The alpha-band decoding was identical to the ERP decoding, except that we applied a bandpass filter from 8-12 Hz (two-way least-squares finite impulse response filter) and passed the filtered data through the Hilbert transform to extract the phase-independent power in this frequency range.

2.5 Time-window averaging and effect size calculation

For the sake of simplicity, decoding accuracy was averaged within a time window that was defined a priori for each paradigm (see Table 1). For the ERP CORE experiments, we used the same time windows that were used for the original univariate analyses of each experiment. For the other paradigms, we used both a perceptual time window (100–500 ms) and a working memory time window (500–1000 ms), as in previous methodology studies using these datasets (Bae & Luck, 2018).

Although researchers often try to maximize decoding accuracy, it is also important to consider statistical power, which is influenced by both the mean decoding accuracy across participants and variability across participants. We therefore also examined the Cohen's d_z metric of effect size, which is directly related to statistical power and reflects both mean decoding accuracy and participant variability. We calculated Cohen's d_z for each analysis as the mean across participants divided by the standard deviation across participants (after first subtracting the chance level for a given analysis from each participant's decoding accuracy). Bootstrapping with 10,000 iterations was employed to estimate the standard error of this metric of effect size.

3 Results

3.1 Binary decoding cases

This section presents the decoding results for the ERP CORE paradigms, each of which involved a binary classification. Note that the differences between experimental conditions for these ERP components could be easily detected using conventional univariate analyses (see Kappenman et al., 2021).

The top row of Figure 2 shows how mean decoding accuracy and effect size varied across different regularization settings in the ERP CORE experiments. For most ERP components, both metrics showed very little change as a function of the regularization strength. However, for the N2pc and LRP components, decoding accuracy and effect size decreased considerably when the regularization strength was less than 0.1.

The top row of Figure 3 shows how mean decoding accuracy varied as a function of the number of crossfolds (N) and the number of trials per average (T) for the ERP CORE paradigms.

Decoding accuracy improved as T increased from 2 to at least 18 trials for each ERP component. We observed a peak and then decline in decoding accuracy as T increased for the LRP and N2pc components. The other components appeared to be nearing a peak at the maximum value of T, but we did not have enough trials to determine conclusively whether decoding accuracy would eventually start falling. In all cases, the highest decoding accuracy was achieved when N was between 2 and 5. This suggests that having clean data (i.e., many trials per average) is more important for maximizing decoding accuracy than having a large number of training cases (i.e., a large number of crossfolds).

The bottom row of Figure 3 shows the corresponding effect sizes, which take into account both the mean decoding accuracy and the standard deviation across participants. The impact of variations in N and T was not as pronounced for effect size as it was for mean decoding accuracy. In all cases, the effect size increased at least slightly as T increased, then reached a peak and began to decline as T increased further. The optimal T varied quite a bit across components, with the largest effect size occurring between 7 and 12 trials per average for the P3b component and between 34 and 50 trials per average for the MMN. In all cases, the effect size was maximized when N was between 3 and 10. However, variations in N and T had only a minimal impact on the effect size except for the N400, ERN, and LRP components. Note that, whereas the mean decoding accuracy was maximal with only 2 crossfolds in several cases, the maximum effect size required at least 3 crossfolds in all cases.

3.2 Multiclass decoding cases

We also examined more challenging decoding cases in which the number of classes was greater (4 or 16) and the differences between classes would be difficult to detect using conventional univariate analyses. Both a perceptual time window (100–500 ms) and a working memory time

window (500–1000 ms) were analyzed for these paradigms, and we decoded alpha-band power in addition to ERP amplitude in the Orientations paradigms.

The effects of varying the regularization strength for the Orientation and Face datasets are shown in the middle and bottom rows of Figure 2. Decoding accuracy (left column) showed only slight variations as the regularization strength increased. Effect size (right column) exhibited more noticeable changes in some cases, dropping when the regularization strength was less than 1.

The effect of the cross-validation parameters on decoding accuracy is shown in the top row of Figure 4 for the 100-500 ms time window and in the top row of Figure 5 for the 500-1000 ms time window. Decoding accuracy increased as T increased up to 20-60 trials per average (corresponding to N = 2-5 crossfolds) and then began to decline. As observed for the ERP CORE paradigms, variations in N and T had a smaller impact on the effect size than on the mean decoding accuracy (bottom row in Figures 4 and 5), and the peak effect size was observed with a smaller T and larger N for effect size than for mean decoding accuracy. Specifically, the maximum effect size occurred with a T of 5–32 trials per average and an N of 5-20 crossfolds for most cases.

4 Discussion

This study investigated how regularization and cross-validation parameters influence EEG/ERP decoding performance. To our knowledge, this is the first study to systematically explore these effects across a wide range of EEG/ERP paradigms involving different electrode configurations, sample sizes, and experimental conditions. Using the ERP CORE, Orientations-1, Orientations-2, and Face datasets, we found that regularization strengths of less than C = 1 negatively impacted

mean decoding accuracy and the corresponding effect size compared to a strength of C = 1, with little or no impact of increasing the regularization strength above 1. The analyses of the cross-validation parameters indicated that keeping the number of crossfolds low (N = 2-5) and the number of trials per average correspondingly high (typically T = 20-50) yielded the highest mean decoding accuracy. By contrast, the effect size (which takes into account both the mean and the standard deviation across participants) benefited from more crossfolds (typically N = 3-10) even though this meant fewer trials per average (typically T = 5-30). Note, however, that the effects of variations in N and T on the effect size were modest in most cases, especially when N and T were not extreme.

4.1 Effect of regularization strength

The present findings provide quantitative evidence about the optimal regularization strength, whereas previous work has often relied on default or heuristic settings without systematic evaluation (Blankertz et al., 2011; Grootswagers et al., 2017). Regularization in SVM controls the trade-off between minimizing classification errors during training and improving generalization to the test data by maximizing the margin(Cortes & Vapnik, 1995) , with values of C less than 1 favoring generalization. Interestingly, we found that decoding accuracy for the test set (and the corresponding effect size) declined when C was substantially less than 1 (when the regularization was very strong, with very wide margins). This suggests that high levels of regularization ($C \ll 1$) did not allow sufficiently good fitting of the training data, which then led to poor results for the test data. However, we never saw a substantial benefit of having C exceed 1, so we recommend using a value of 1, which provides an equal balance between good fitting of the training data and good generalization to the test data. However, this recommendation may not generalize to datasets that are quite different from those examined here. For example, they may

not generalize to data obtained with small numbers of electrodes, noisier recording setups (e.g., dry electrodes, G.-L. Li et al., 2020) , or very different participant populations (e.g., infants, Ashton et al., 2022; Turk-Browne & Aslin, 2024).

4.2 Effect of the cross-validation parameters

Increasing the number of crossfolds may be beneficial for decoding by increasing the number of training cases, but it may be detrimental by decreasing the number of trials per average and therefore the signal-to-noise ratio of both the training data and the test data. Consistent with these competing factors, we found that decoding was typically best when extreme values of T were avoided and the number of crossfolds/trials was moderate to low. However, the optimal balance was somewhat different depending on whether we examined mean decoding accuracy or the effect size. Mean decoding accuracy typically benefited by maximizing the number of trials per average, as long as the number of crossfolds did not get too small (e.g., N = 2). By contrast, effect size benefited by increasing the number of crossfolds somewhat (e.g., N = 5-10) at the cost of fewer trials per average. Maximizing the decoding accuracy is primarily important in engineering applications, whereas maximizing the effect size (and therefore statistical power) is usually more important in scientific studies (see Hebart & Baker, 2018). Most scientific studies should therefore avoid low numbers of crossfolds. Note, however, that the impact of the number of trials/crossfolds on effect size was relatively modest in most cases. As a result, researchers should be able to obtain good effect sizes as long as they avoid extreme numbers of trials or crossfolds. In almost all cases examined here, researchers should be able to achieve near maximum performance for both decoding accuracy and effect size by using N = 3–5 crossfolds with at least T = 10 trials per average.

The decoding results were fairly consistent across a wide range of experimental paradigms, which varied in the sample size, the electrode densities, the nature of the experimental conditions, and the number of l classes. Consequently, the conclusion that decoding performance benefits from using N = 3-5 and T >= 10 will likely generalize to a broad range of datasets. Note that, because we were focused on typical cases in which the total number of trials is fixed and N and T are therefore reciprocally related, we did not examine the effects of variations in T for a fixed value of N or variations in N for a fixed value of T. However, Scrivener et al. (2023) demonstrated with simulated EEG data that increasing T improved decoding accuracy when N was held constant Presumably, increasing N while holding T constant would also increase decoding accuracy by increasing the amount of training data.

4.3 Generalizability

It is important to consider whether the present conclusions would be expected to generalize to other datasets and other decoding methods. All of the datasets examined here were collected from similar adult populations (i.e., highly cooperative college students), using similar recording systems and analyzed through comparable processing pipelines. Therefore, the results might not generalize to different other types of datasets collected from different populations, such as infants, children, and people with brain disorders (Marsicano et al., 2024; Ng et al., 2022; Turoman et al., 2024). Additionally, the patterns shown in this study might also not generalize to datasets recorded from different recording systems (e.g., dry electrodes) (Habibzadeh Tonekabony Shad et al., 2020; G.-L. Li et al., 2020), mobile EEG systems (Niso et al., 2023; Park et al., 2015), and different recording environments (e.g., hospitals) (Lei et al., 2024).

Additionally, although some previous studies have shown that single-trial classification can yield similar results to the averaged-trial approach (Williams et al., 2020) , it is unknown

whether the present conclusions about regularization will generalize to single-trial classification. This question was beyond the scope of the current study.

It is also unknown whether the present results will generalize to other relatively simple classification algorithms, such as the random forest approach (Breiman, 2001) and linear discriminant analysis (Fisher, 1936), and to more advanced algorithms, such as Bayesian classification, convolutional neural networks, recurrent neural networks, and deep belief networks (Al-Saegh et al., 2021; Craik et al., 2019; Gong et al., 2022; Hosseini et al., 2021). These approaches may benefit from a different balance between training accuracy and test accuracy (i.e., regularization strength) or a different balance between the number of classes and the number of trials per class. This is an issue that deserves study, especially in cases where these other approaches are superior to SVMs. However, SVMs are very widely used for EEG/ERP decoding at present, so the present conclusions will be valuable to many researchers.

CRediT authorship contribution statement

Guanghui Zhang: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Methodology, Investigation, Conceptualization.

Xinran Wang: Visualization, Validation, Supervision, Software, Resources, Methodology.

Steven J. Luck: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

None.

Acknowledgments

We thank the members of the Luck Lab for their valuable contributions to the community and thoughtful insights, including but not limited to: Carlos Daniel Carrasco, John Kiat, David Garrett, Kurt Winsler, and Brett Bahle.

Reference

Al-Saegh, A., Dawwd, S. A., & Abdul-Jabbar, J. M. (2021). Deep learning for motor imagery EEG-based classification: A review. *Biomedical Signal Processing and Control*, *63*, 102172. https://doi.org/10.1016/j.bspc.2020.102172

Ashton, K., Zinszer, B. D., Cichy, R. M., Nelson, C. A., Aslin, R. N., & Bayet, L. (2022). Time-resolved multivariate pattern analysis of infant EEG data: A practical tutorial. *Developmental Cognitive Neuroscience*, *54*, 101094. https://doi.org/10.1016/j.dcn.2022.101094

Bae, G.-Y. (2021). The time course of face representations during perception and working memory maintenance. *Cerebral Cortex Communications*, *2*(1), tgaa093. https://doi.org/10.1093/texcom/tgaa093

Bae, G.-Y., & Luck, S. J. (2018). Dissociable decoding of spatial attention and working memory from EEG oscillations and sustained potentials. *The Journal of Neuroscience*, *38*(2), 409–422. https://doi.org/10.1523/JNEUROSCI.2860-17.2017

Bayet, L., Zinszer, B. D., Reilly, E., Cataldo, J. K., Pruitt, Z., Cichy, R. M., Nelson, C. A., & Aslin, R. N. (2020). Temporal dynamics of visual representations in the infant brain. *Developmental Cognitive Neuroscience*, *45*, 100860. https://doi.org/10.1016/j.dcn.2020.100860

Blankertz, B., Lemm, S., Treder, M., Haufe, S., & Müller, K.-R. (2011). Single-trial analysis and classification of ERP components—A tutorial. *NeuroImage*, *56*(2), 814–825. https://doi.org/10.1016/j.neuroimage.2010.06.048

Bo, K., Cui, L., Yin, S., Hu, Z., Hong, X., Kim, S., Keil, A., & Ding, M. (2022). Decoding the temporal dynamics of affective scene processing. *NeuroImage*, *261*, 119532. https://doi.org/10.1016/j.neuroimage.2022.119532

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Carlson, T. A., Grootswagers, T., & Robinson, A. K. (2019). *An introduction to time-resolved decoding analysis for M/EEG* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1905.04820

Carrasco, C. D., Bahle, B., Simmons, A. M., & Luck, S. J. (2024). Using multivariate pattern analysis to increase effect sizes for event-related potential analyses. *Psychophysiology*, *61*(7), e14570. https://doi.org/10.1111/psyp.14570

Chaisaen, R., Autthasan, P., Mingchinda, N., Leelaarporn, P., Kunaseth, N., Tammajarung, S., Manoonpong, P., Mukhopadhyay, S. C., & Wilaiprasitporn, T. (2020). Decoding EEG rhythms during action observation, motor imagery, and execution for standing and sitting. *IEEE Sensors Journal*, *20*(22), 13776–13786. https://doi.org/10.1109/JSEN.2020.3005968

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Craik, A., He, Y., & Contreras-Vidal, J. L. (2019). Deep learning for electroencephalogram (EEG) classification tasks: A review. *Journal of Neural Engineering*, *16*(3), 031001. https://doi.org/10.1088/1741-2552/ab0ab5

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Dimigen, O. (2020). Optimizing the ICA-based removal of ocular EEG artifacts from free viewing experiments. *NeuroImage*, *207*, 116117. https://doi.org/10.1016/j.neuroimage.2019.116117

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*(2), 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

Gong, S., Xing, K., Cichocki, A., & Li, J. (2022). Deep learning in EEG: Advance of the last ten-year critical period. *IEEE Transactions on Cognitive and Developmental Systems*, *14*(2), 348–365. https://doi.org/10.1109/TCDS.2021.3079712

Grootswagers, T., Wardle, S. G., & Carlson, T. A. (2017). Decoding dynamic brain patterns from evoked responses: A tutorial on multivariate pattern analysis applied to time series neuroimaging data. *Journal of Cognitive Neuroscience*, *29*(4), 677–697. https://doi.org/10.1162/jocn_a_01068

Habibzadeh Tonekabony Shad, E., Molinas, M., & Ytterdal, T. (2020). Impedance and noise of passive and active dry EEG electrodes: A review. *IEEE Sensors Journal*, *20*(24), 14565–14577. https://doi.org/10.1109/JSEN.2020.3012394

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer New York. https://doi.org/10.1007/978-0-387-84858-7

Hebart, M. N., & Baker, C. I. (2018). Deconstructing multivariate decoding for the study of brain function. *NeuroImage*, *180*, 4–18. https://doi.org/10.1016/j.neuroimage.2017.08.005

Hong, X., Bo, K., Meyyappan, S., Tong, S., & Ding, M. (2020). Decoding attention control and selection in visual spatial attention. *Human Brain Mapping*, *41*(14), 3900–3921. https://doi.org/10.1002/hbm.25094

Hosseini, M.-P., Hosseini, A., & Ahi, K. (2021). A review on machine learning for EEG signal processing in bioengineering. *IEEE Reviews in Biomedical Engineering*, *14*, 204–218. https://doi.org/10.1109/RBME.2020.2969915

Jung, T., Makeig, S., Humphries, C., Lee, T., McKeown, M. J., Iragui, V., & Sejnowski, T. J. (2000). Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*, *37*(2), 163–178. https://doi.org/10.1111/1469-8986.3720163

Jung, T.-P., Makeig, S., Westerfield, M., Townsend, J., Courchesne, E., & Sejnowski, T. J. (2000). Removal of eye activity artifacts from visual event-related potentials in normal and clinical subjects. *Clinical Neurophysiology*, *111*(10), 1745–1758. https://doi.org/10.1016/S1388-2457(00)00386-2

Kappenman, E. S., Farrens, J. L., Zhang, W., Stewart, A. X., & Luck, S. J. (2021). ERP CORE: An open resource for human event-related potential research. *NeuroImage*, *225*, 117465. https://doi.org/10.1016/j.neuroimage.2020.117465

Kikumoto, A., & Mayr, U. (2018). Decoding hierarchical control of sequential behavior in oscillatory EEG activity. *eLife*, *7*, e38550. https://doi.org/10.7554/eLife.38550

Klug, M., & Gramann, K. (2021). Identifying key factors for improving ICA-based decomposition of EEG data in mobile and stationary experiments. *European Journal of Neuroscience*, *54*(12), 8406–8420. https://doi.org/10.1111/ejn.14992

Lei, M., Chen, W., Wu, J., Zhang, Y., & Lv, Y. (2024). Neurophysiological measures in hospitality and tourism: Review, critique, and research agenda. *Journal of Hospitality & Tourism Research*, *48*(1), 3–31. https://doi.org/10.1177/10963480221091117

Lemm, S., Blankertz, B., Dickhaus, T., & Müller, K.-R. (2011). Introduction to machine learning for brain imaging. *NeuroImage*, *56*(2), 387–399. https://doi.org/10.1016/j.neuroimage.2010.11.004

Li, G.-L., Wu, J.-T., Xia, Y.-H., He, Q.-G., & Jin, H.-G. (2020). Review of semi-dry electrodes for EEG recording. *Journal of Neural Engineering*, *17*(5), 051004. https://doi.org/10.1088/1741-2552/abbd50

Li, Y., Zhang, M., Liu, S., & Luo, W. (2022). EEG decoding of multidimensional information from emotional faces. *NeuroImage*, *258*, 119374. https://doi.org/10.1016/j.neuroimage.2022.119374

Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: An open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00213

Luck, S. J. (2022). *Applied event-related potential data analysis*. LibreTexts. https://doi.org/10.18115/D5QG92

Mares, I., Ewing, L., Farran, E. K., Smith, F. W., & Smith, M. L. (2020). Developmental changes in the processing of faces as revealed by EEG decoding. *NeuroImage*, *211*, 116660. https://doi.org/10.1016/j.neuroimage.2020.116660

Marsicano, G., Bertini, C., & Ronconi, L. (2024). Decoding cognition in neurodevelopmental, psychiatric and neurological conditions with multivariate pattern analysis of EEG data. *Neuroscience & Biobehavioral Reviews*, *164*, 105795. https://doi.org/10.1016/j.neubiorev.2024.105795

Ng, B., Reh, R. K., & Mostafavi, S. (2022). A practical guide to applying machine learning to infant EEG data. *Developmental Cognitive Neuroscience*, *54*, 101096. https://doi.org/10.1016/j.dcn.2022.101096

Niso, G., Romero, E., Moreau, J. T., Araujo, A., & Krol, L. R. (2023). Wireless EEG: A survey of systems and studies. *NeuroImage*, *269*, 119774. https://doi.org/10.1016/j.neuroimage.2022.119774

Park, J. L., Fairweather, M. M., & Donaldson, D. I. (2015). Making the case for mobile cognition: EEG and sports performance. *Neuroscience & Biobehavioral Reviews*, *52*, 117–130. https://doi.org/10.1016/j.neubiorev.2015.02.014

Peelen, M. V., & Downing, P. E. (2023). Testing cognitive theories with multivariate pattern analysis of neuroimaging data. *Nature Human Behaviour*, *7*(9), 1430–1441. https://doi.org/10.1038/s41562-023-01680-z

Peng, S., Xu, R., Yi, X., Hu, X., Liu, L., & Liu, L. (2021). Early screening of children with autism spectrum disorder based on electroencephalogram signal feature selection with L1-norm regularization. *Frontiers in Human Neuroscience*, *15*, 656578. https://doi.org/10.3389/fnhum.2021.656578

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: A tutorial overview. *NeuroImage*, *45*(1, Supplement 1), S199–S209. https://doi.org/10.1016/j.neuroimage.2008.11.007

Petit, S., Badcock, N. A., Grootswagers, T., & Woolgar, A. (2020). Unconstrained multivariate EEG decoding can help detect lexical-semantic processing in individual children. *Scientific Reports*, *10*(1), 10849. https://doi.org/10.1038/s41598-020-67407-6

Sanchez, G., Hartmann, T., Fuscà, M., Demarchi, G., & Weisz, N. (2020). Decoding across sensory modalities reveals common supramodal signatures of conscious perception. *Proceedings of the National Academy of Sciences*, *117*(13), 7437–7446. https://doi.org/10.1073/pnas.1912584117

Sarrett, M. E., & Toscano, J. C. (2024). Decoding speech sounds from neurophysiological data: Practical considerations and theoretical implications. *Psychophysiology*, *61*(4), e14475. https://doi.org/10.1111/psyp.14475

Scrivener, C. L., Grootswagers, T., & Woolgar, A. (2023). *Optimising analysis choices for multivariate decoding: Creating pseudotrials using trial averaging and resampling*. Neuroscience. https://doi.org/10.1101/2023.10.04.560678

Trammel, T., Khodayari, N., Luck, S. J., Traxler, M. J., & Swaab, T. Y. (2023). Decoding semantic relatedness and prediction from EEG: A classification method comparison. *NeuroImage*, *277*, 120268. https://doi.org/10.1016/j.neuroimage.2023.120268

Turk-Browne, N. B., & Aslin, R. N. (2024). Infant neuroscience: How to measure brain activity in the youngest minds. *Trends in Neurosciences*, *47*(5), 338–354. https://doi.org/10.1016/j.tins.2024.02.003

Turoman, N., Fiave, P. A., Zahnd, C., deBettencourt, M. T., & Vergauwe, E. (2024). Decoding the content of working memory in school-aged children. *Cortex*, *171*, 136–152. https://doi.org/10.1016/j.cortex.2023.10.019

Valente, G., Castellanos, A. L., Hausfeld, L., De Martino, F., & Formisano, E. (2021). Cross-validation and permutations in MVPA: Validity of permutation strategies and power of cross-validation schemes. *NeuroImage*, *238*, 118145. https://doi.org/10.1016/j.neuroimage.2021.118145

Varoquaux, G. (2018). Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage*, *180*, 68–77. https://doi.org/10.1016/j.neuroimage.2017.06.061

Varoquaux, G., Raamana, P. R., Engemann, D. A., Hoyos-Idrobo, A., Schwartz, Y., & Thirion, B. (2017). Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *NeuroImage*, *145*, 166–179. https://doi.org/10.1016/j.neuroimage.2016.10.038

Williams, J. M., Samal, A., & Johnson, M. R. (2023). *Trial averaging for deep EEG classification*. Neuroscience. https://doi.org/10.1101/2023.02.09.527905

Williams, J. M., Samal, A., Rao, P. K., & Johnson, M. R. (2020). Paired trial classification: A novel deep learning technique for MVPA. *Frontiers in Neuroscience*, *14*, 417. https://doi.org/10.3389/fnins.2020.00417

Zhang, G., Carrasco, C. D., Winsler, K., Bahle, B., Cong, F., & Luck, S. J. (2024). Assessing the effectiveness of spatial PCA on SVM-based decoding of EEG data. *NeuroImage*, *293*, 120625. https://doi.org/10.1016/j.neuroimage.2024.120625

Zhang, G., Garrett, D. R., & Luck, S. J. (2024a). Optimal filters for ERP research I: a general approach for selecting filter settings. *Psychophysiology*, *61*(6), e14531. https://doi.org/10.1111/psyp.14531

Zhang, G., Garrett, D. R., & Luck, S. J. (2024b). Optimal filters for ERP research II: recommended settings for seven common ERP components. *Psychophysiology*, *61*(6), e14530. https://doi.org/10.1111/psyp.14530

Zhang, G., Garrett, D. R., Simmons, A. M., Kiat, J. E., & Luck, S. J. (2024). Evaluating the effectiveness of artifact correction and rejection in event-related potential research. *Psychophysiology*, *61*(5), e14511. https://doi.org/10.1111/psyp.14511

Zhang, G., & Luck, S. J. (2025). Assessing the impact of artifact correction and artifact rejection on the performance of SVM- and LDA-based decoding of EEG signals. *NeuroImage*, *316*, 121304. https://doi.org/10.1016/j.neuroimage.2025.121304

Zhang, G., Xin, Y., Song, L., Wang, X., Chen, L., He, W., & Luo, W. (2025). Temporal dynamics of perceptual integrity and semantic congruency during color-word processing: An ERP and decoding study. *NeuroImage*, *321*, 121486. https://doi.org/10.1016/j.neuroimage.2025.121486

Table 1. Epoch window, baseline period, electrode site, number of classes, and measurement window used for each experiment. Note that we collapsed the 16 stimuli into 4 classes for the Orientations-1 and Orientation-2 experiments in our analyses of the effects of the cross-validation parameters.

Experiment	Epoch (ms)	Baseline Period (ms)	Measurement window (ms)	# of Subjects	# of classes (M)	# of channels
N170	-200 to 800	-200 to 0	110 to 150	39	2 (faces vs. cars)	28
MMN	-200 to 800	-200 to 0	125 to 225	40	2 (deviants vs. standards)	28
P3b	-200 to 800	-200 to 0	300 to 600	38	2 (targets vs. nontargets)	28
N400	-200 to 800	-200 to 0	300 to 500	39	2 (related vs. unrelated)	28
ERN	-600 to 400	-400 to -200	0 to 100	22	2 (correct vs. error)	28
LRP	-800 to 200	-800 to -600	-100 to 0	40	2 (left hand vs. right hand)	28
N2pc	-200 to 800	-200 to 0	200 to 275	40	2 (left target vs. right target)	28
Orientations-1	-500 to 1500	-500 to 0	100 to 500 or 500 to 1000	16	4 or 16 (orientation classes)	27
Orientations- 2	-500 to 1500	-500 to 0	100 to 500 or 500 to 1000	16	4 or 16 (orientation or location classes)	27
Faces	-500 to 1500	-500 to 0	100 to 500 or 500 to 1000	22	4 (face identities or emotional expressions)	59

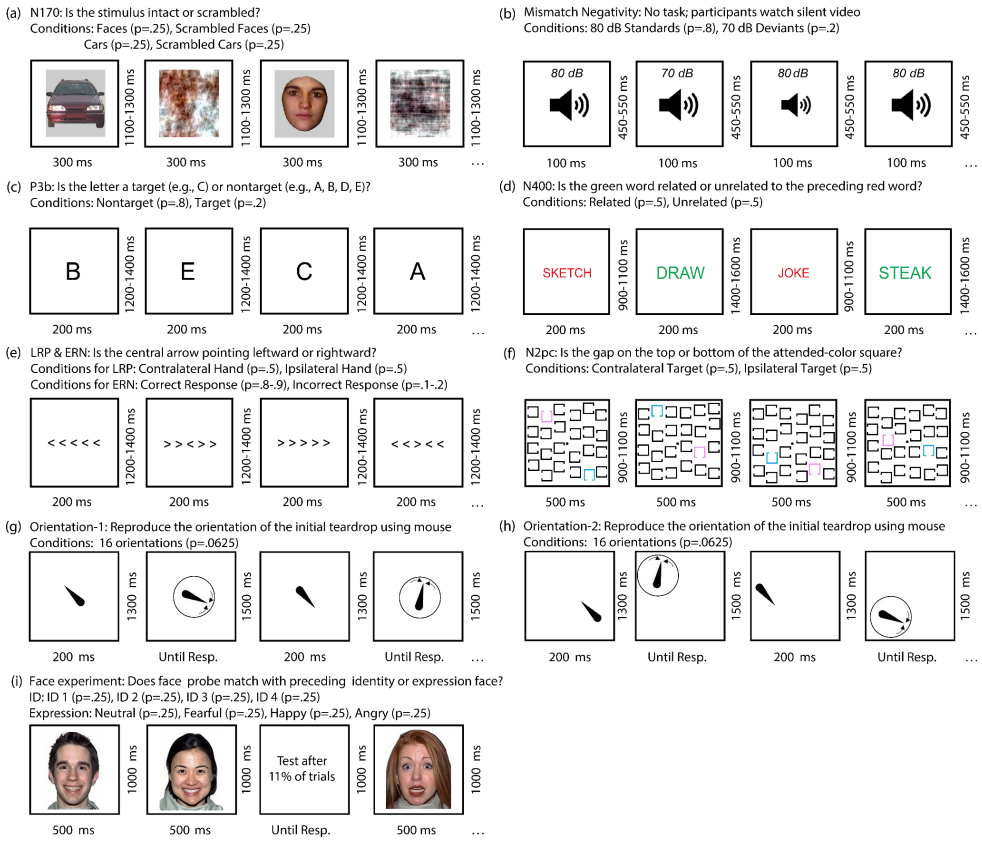

Figure 1. Example trials from the experimental paradigms. (a) N170 Paradigm: Participants identified whether a given stimulus was intact (faces or cars) or scrambled (scrambled faces or cars). This study focused exclusively on the face and car trials. (b) Mismatch Negativity Paradigm: Participants watched a silent video while task-irrelevant standard tones (80 dB, probability = .8) and deviant tones (70 dB, probability = .2) were presented. (c) P3b Paradigm: The letters A, B, C, D, and E were presented in random order, each with a probability of .2. One letter was designated as the target for each trial block, and participants identified whether a given stimulus was the target or a nontarget. (d) N400 Paradigm: Participants determined whether a green target word was semantically related or unrelated to a preceding red prime word. (e) Flankers Task: This task elicited the lateralized readiness potential (LRP) and error-related negativity (ERN). Participants reported the direction of the central arrow (leftward or rightward) while ignoring the surrounding arrows. (f) N2pc Paradigm: For each trial block, either pink or blue was designated as the target color. Participants reported whether the gap on the target-colored square was at the top or bottom of the square. (g) Orientations-1 Paradigm: Participants memorized the orientation of an initial teardrop, retained it over a 1300-ms delay period, and then used a mouse to adjust the orientation of a test teardrop to match the original. (h) Orientations-2 Paradigm: Identical to Orientations-1 except the teardrop appeared randomly at one of 16 different locations. Note that the orientations or locations located in the same quadrant are combined to one new orientation or location. (i) Face Paradigm: In each trial, one of 16 faces was presented, representing a combination of four unique identities and four distinct emotional expressions. On 11% of the trials, participants were asked to recall the most recent face shown.

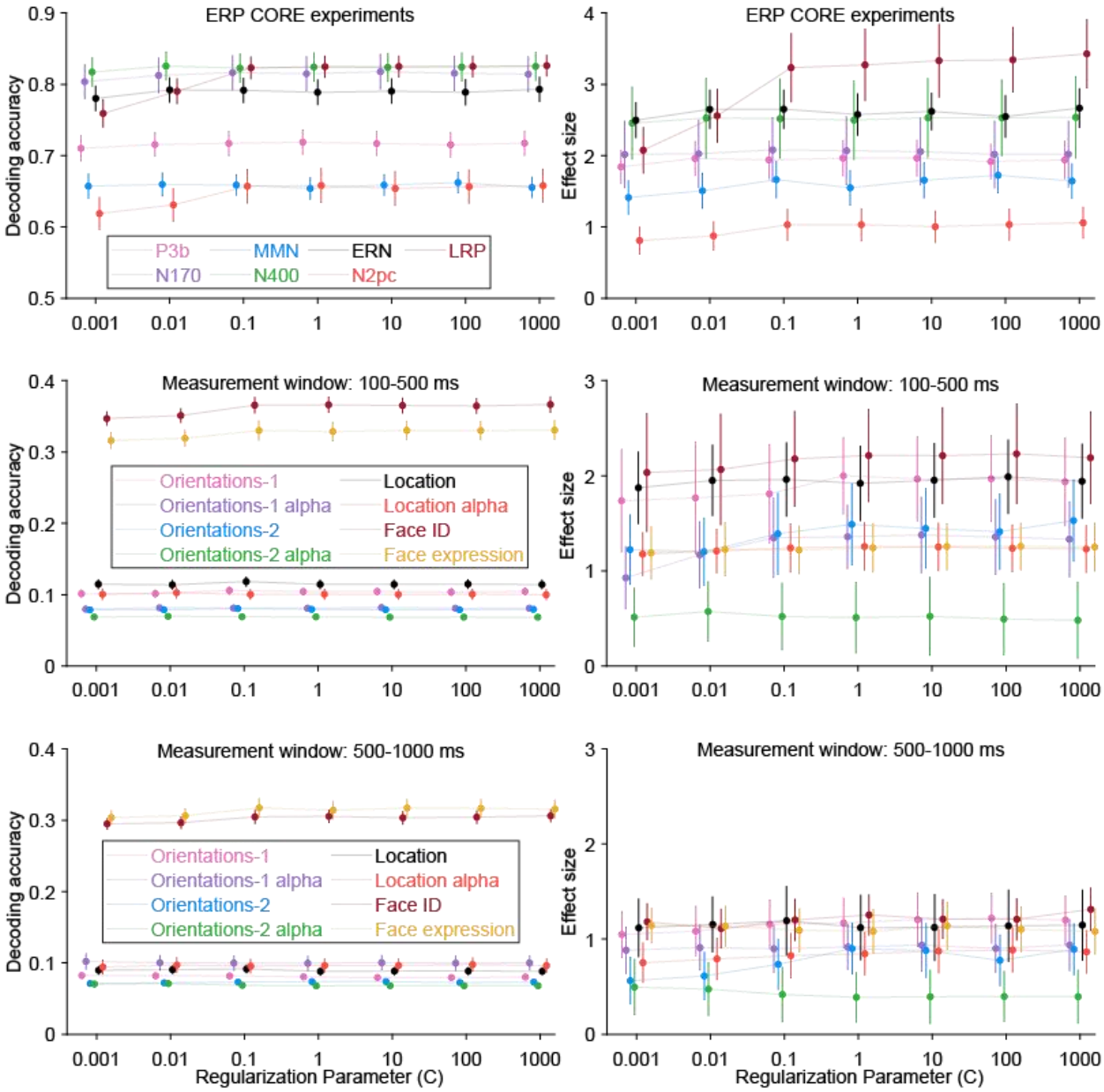

Figure 2. Decoding accuracy (left column) and effect size (right column) for different regularization strengths, separately for each dataset. Error bars show the standard error for each value.

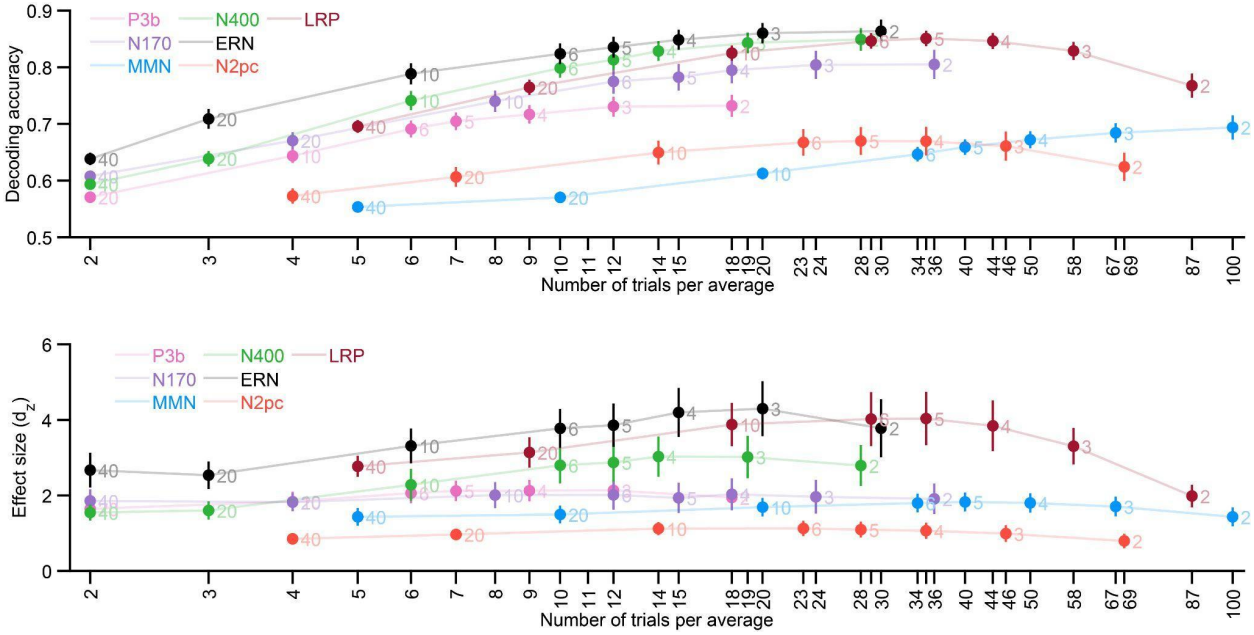

Figure 3. Decoding accuracy (top row) and effect size (bottom row) for different numbers of crossfolds for the ERP CORE paradigms, averaged across the time-window shown in Table 1. The X axis labels indicate the number of trials per average, and the number next to each data point indicates the corresponding number of crossfolds. Error bars show the standard error for each value.

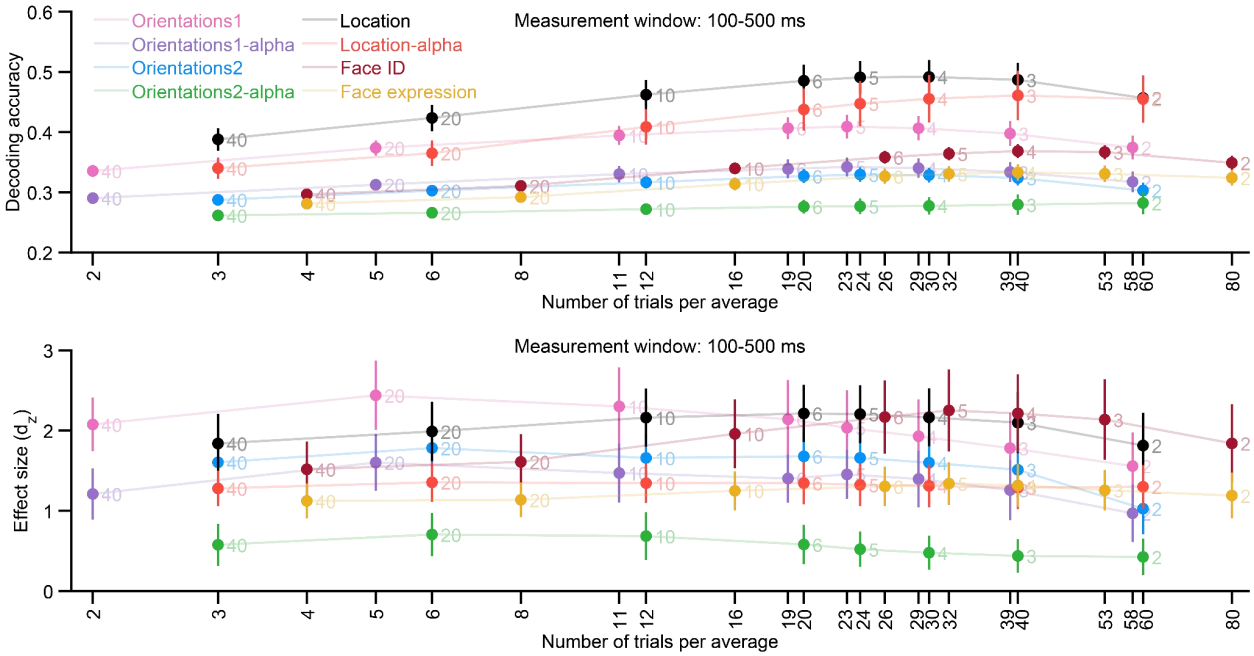

Figure 4. Decoding accuracy (top row) and effect size (bottom row) for different numbers of crossfolds when decoding orientation in the Orientations-1 & 2 datasets when decoding location of the Orientations-2 dataset, and when decoding face identity or facial expression in the Faces dataset. Decoding accuracy was averaged across the perceptual time-window of 100-500 ms. The X axis labels indicate the number of trials per average, and the number next to each data point indicates the corresponding number of crossfolds. Error bars show the standard error for each value.

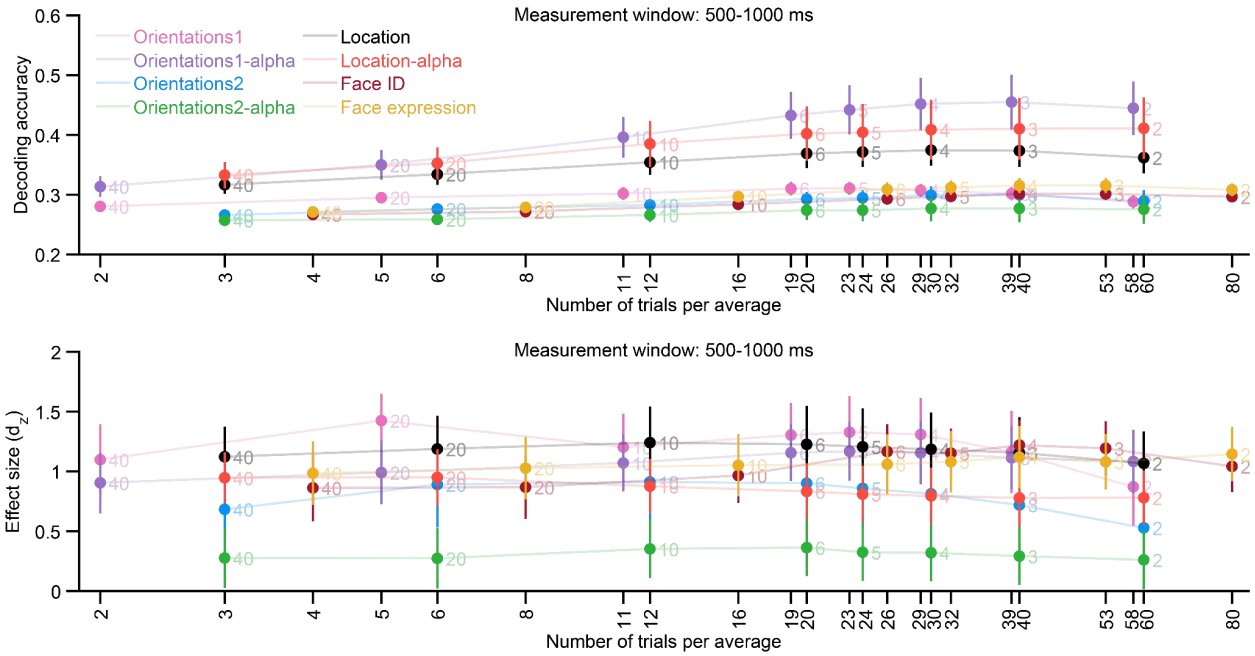

Figure 5. Decoding accuracy (top row) and effect size (bottom row) for different numbers of crossfolds when decoding orientation in the Orientations-1 & 2 datasets when decoding location of the Orientations-2 dataset, and when decoding face identity or facial expression in the Faces dataset. Decoding accuracy was averaged across the working memory time-window of 500-1000 ms. The X axis labels indicate the number of trials per average, and the number next to each data point indicates the corresponding number of crossfolds. Error bars show the standard error for each value.

Supplementary Materials for Evaluating the effects of regularization and cross-validation parameters on the performance of SVM-based decoding of EEG data

Guanghui Zhang[1,2,3] , Xinran Wang[1,2], and Steven J. Luck[3]

[1]Institute of Psychological and Brain Sciences, Liaoning Normal University, Dalian, Liaoning, China
[2]Key Laboratory of Brain and Cognitive Neuroscience, Liaoning Province, Dalian, China
[3]Center for Mind & Brain, University of California-Davis, Davis, CA, USA

Contents

Figure S1. Effects of the cross-validation parameters on decoding performance of ERP CORE datasets……35
Figure S2. Effects of the cross-validation parameters on decoding performance of face dataset……366
Figure S3. Effects of the cross-validation parameters on decoding performance of Orientations1 dataset……37
Figure S4. Effects of the cross-validation parameters on decoding performance of Orientations2 dataset……38

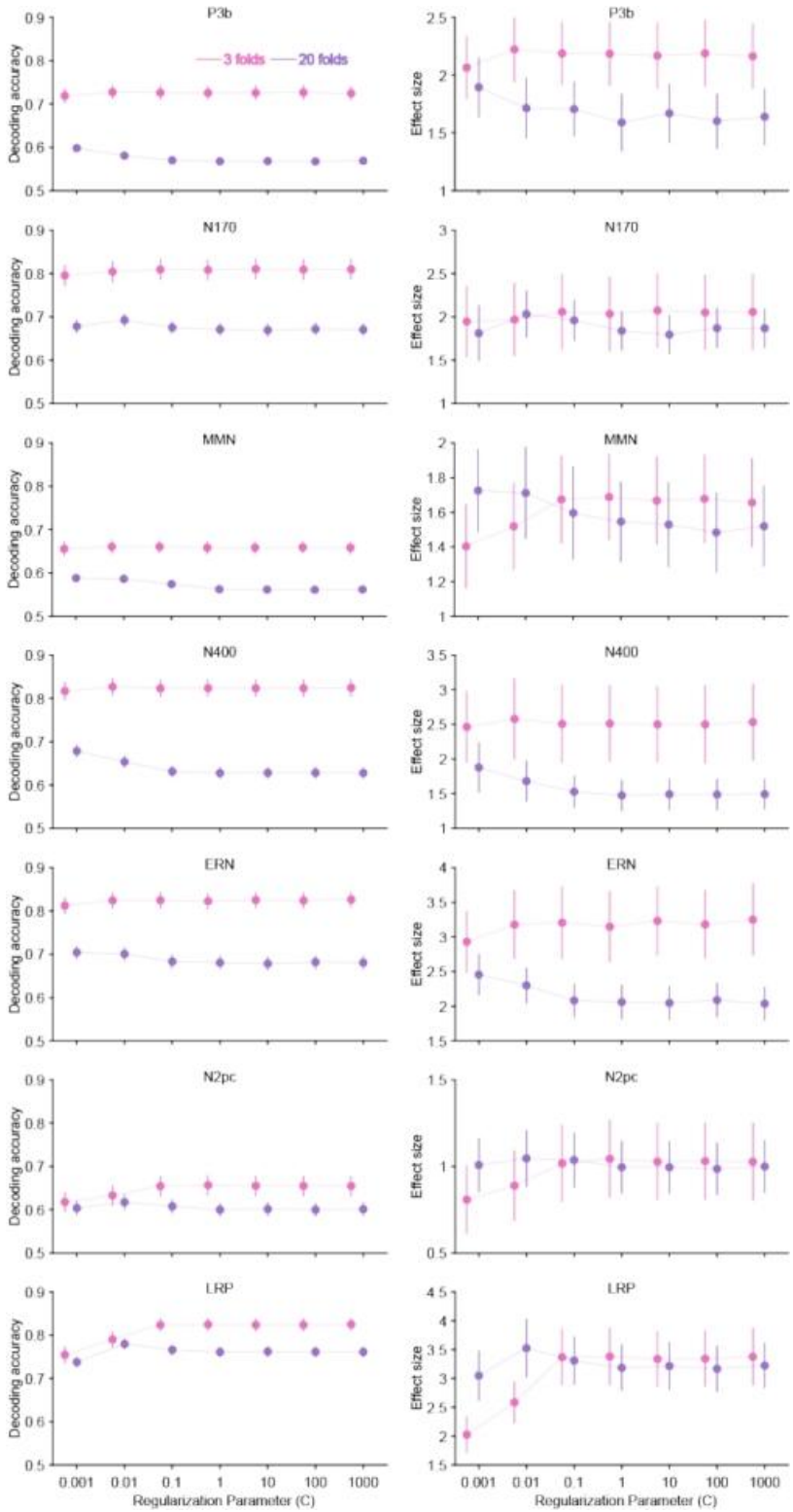

Figure S1. Decoding accuracy (left column) and effect size (right column) for different regularization strengths for ERP CORE dataset for different crossfold numbers (3 vs. 20). Error bars show the standard error for each value.

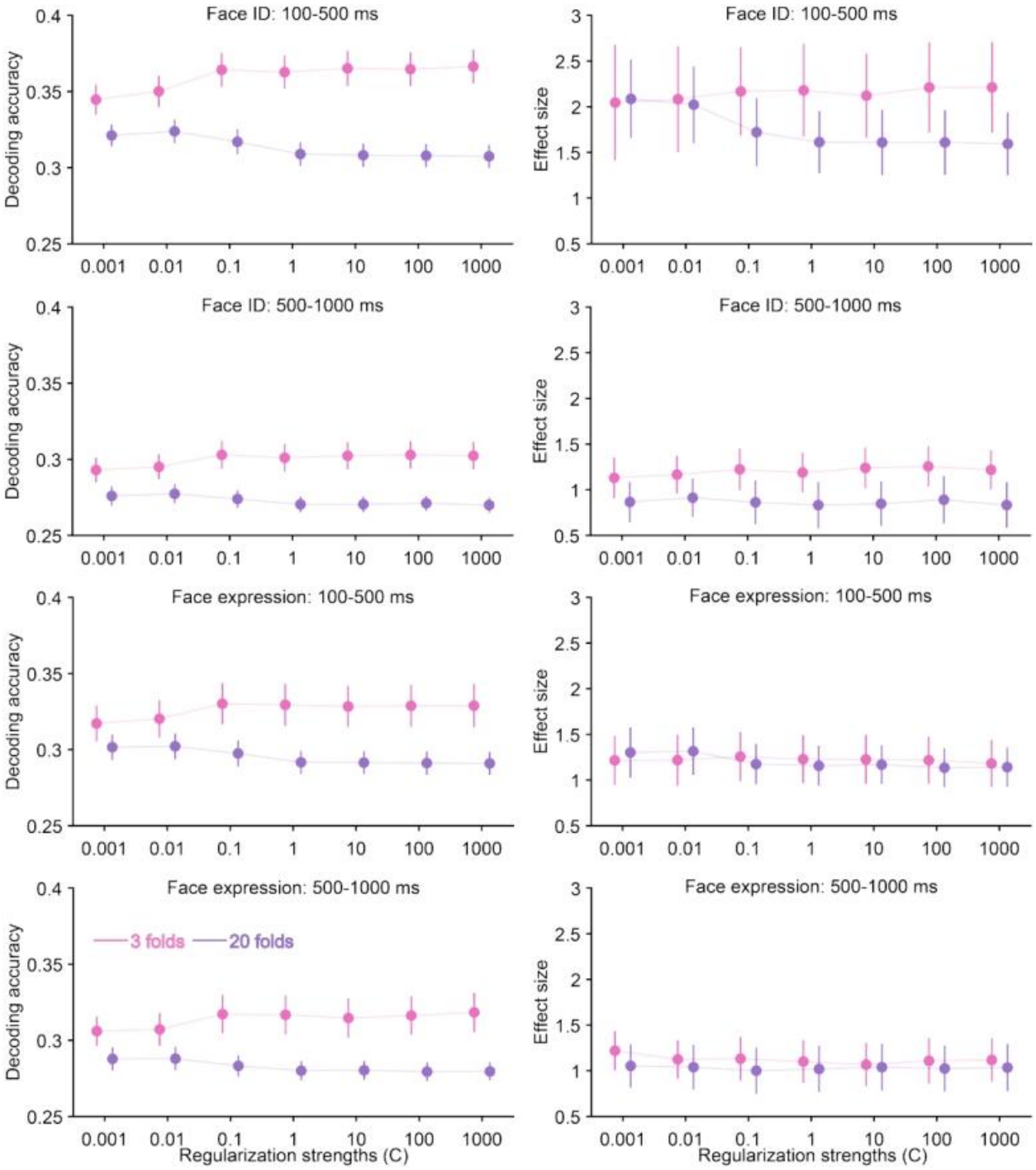

Figure S2. Decoding accuracy (left column) and effect size (right column) for different regularization strengths for Face dataset for different crossfold numbers (3 vs. 20). Error bars show the standard error for each value.

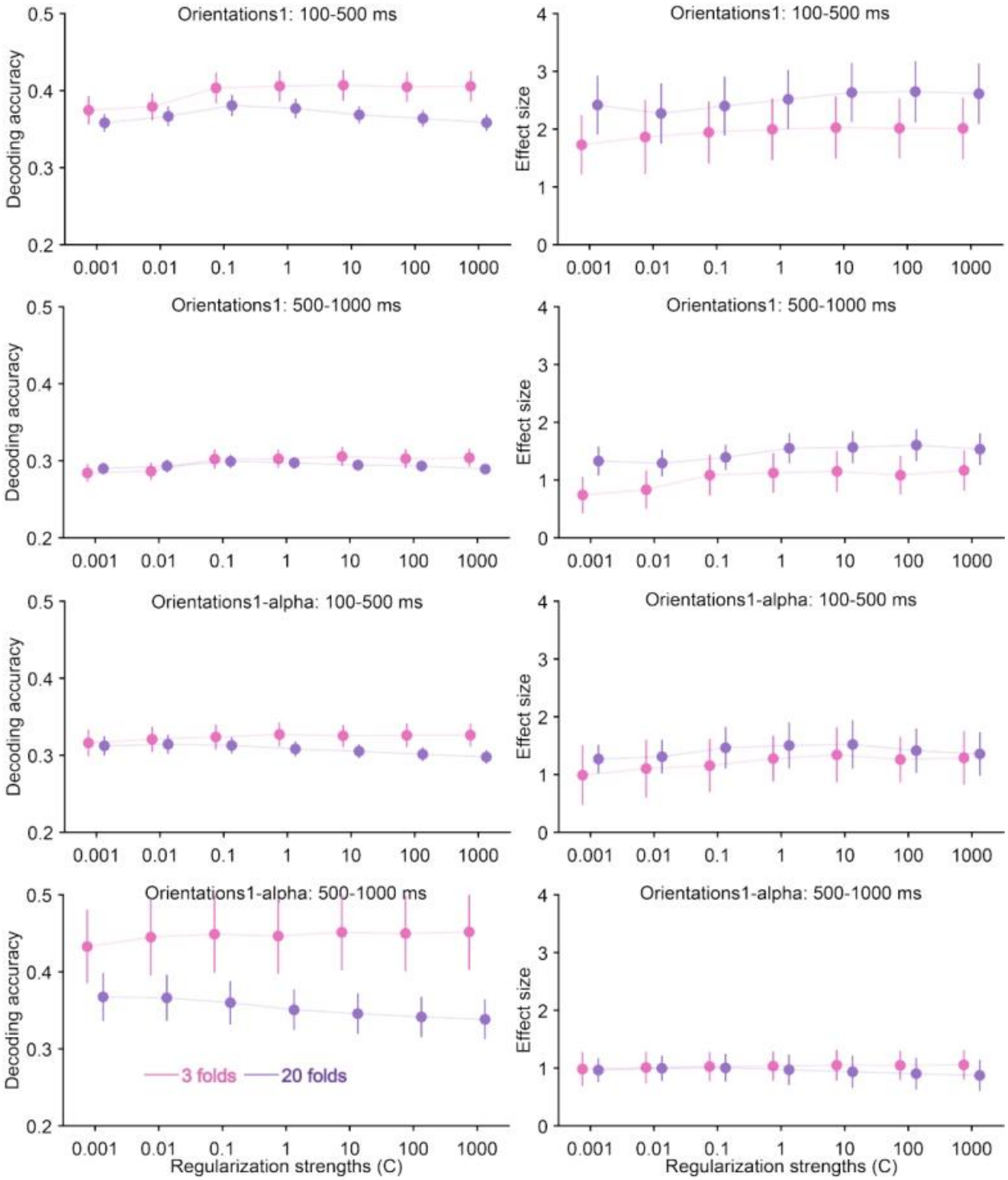

Figure S3. Decoding accuracy (left column) and effect size (right column) for different regularization strengths for Orientations1 dataset for different crossfold numbers (3 vs. 20). Error bars show the standard error for each value.

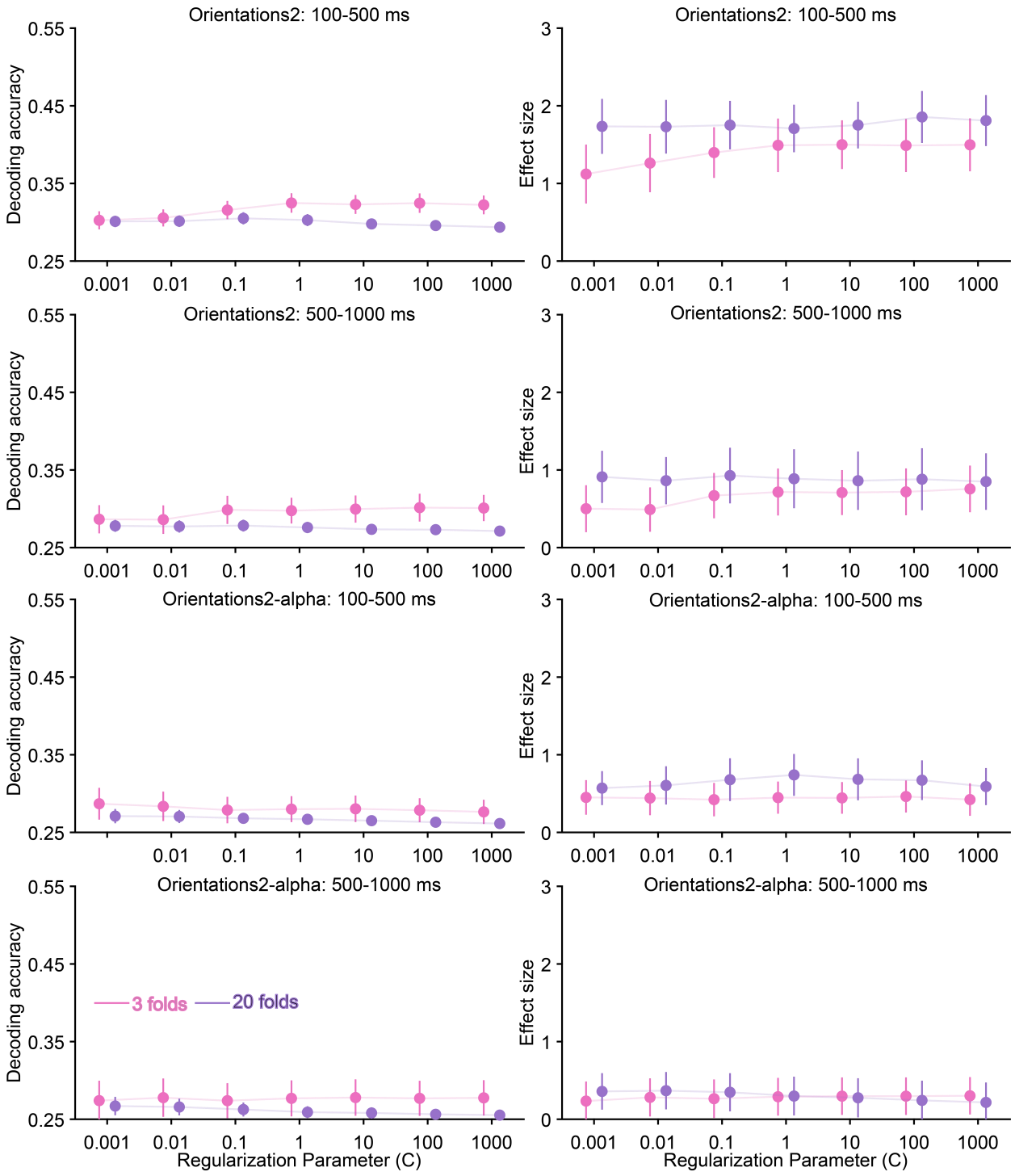

Figure S4. Decoding accuracy (left column) and effect size (right column) for different regularization strengths for Orientations2 dataset for different crossfold numbers (3 vs. 20). Error bars show the standard error for each value.