Behavior Research Methods
https://doi.org/10.3758/s13428-022-01946-w

Stimulus whitening improves the efficiency of reverse correlation

Alexis Compton[1] · Benjamin W. Roop[2] · Benjamin Parrell[3] · Adam C. Lammert[1,2]

Accepted: 26 July 2022

Abstract
Human perception depends upon internal representations of the environment that help to organize the raw information available from the senses by acting as reference patterns. Internal representations are widely characterized using reverse correlation, a method capable of producing unconstrained estimates of the representation itself, all on the basis of simple responses to random stimuli. Despite its advantages, reverse correlation is often infeasible to apply because of its inefficiency—a very large number of stimulus–response trials are required in order to obtain an accurate estimate. Here, we show that an important source of this inefficiency is small, yet nontrivial, correlations that occur by chance between randomly generated stimuli. We demonstrate in simulation that whitening stimuli to remove such correlations before eliciting responses provides greater than 85% improvement in efficiency for a given estimation quality, as well as a two- to fivefold increase in quality for a given sample size. Moreover, unlike conventional approaches, whitening improves the efficiency of reverse correlation without introducing bias into the estimate, or requiring prior knowledge of the target internal representation. Improving the efficiency of reverse correlation with whitening may enable a broader scope of investigations into the individual variability and potential universality of perceptual mechanisms.

Keywords Perceptual representations · Receptive fields · Classification images · Reverse correlation · Whitening

Reverse correlation is a powerful method for characterizing the underlying mechanisms of perception (Ahumada Jr & Lovell, 1971; De Boer & Kuyper, 1968). It has a long history of use in characterizing the latent representations encapsulated in neural tuning (e.g., receptive fields; Ringach & Shapley, 2004; Nishimoto et al., 2006), and has more recently become a primary method for inferring cognitive representations that drive the top-down processes of perception (e.g., face or phoneme recognition; Ahumada Jr & Lovell, 1971; Gosselin & Schyns, 2001; Jäkel et al., 2009; Neri & Levi, 2006; Smith et al., 2012; Varnet et al., 2013a, 2013b), and even to estimate representations associated with abstract psychosocial categories (e.g., "male" vs. "female" faces; Brinkman et al., 2017; Mangini & Biederman, 2004; Moon et al., 2020; Ponsot et al., 2018). Indeed, the method has broad applicability for characterizing many aspects of neurological, cognitive, or psychological function and is closely related to the widely used "white noise approach" to characterizing physiological (Marmarelis & Marmarelis, 1978) and engineering systems (Volterra, 1930; Wiener, 1958; Ljung, 1999).

In reverse correlation, stimulus–response data are elicited via the presentation of richly varying stimuli. For example, in psychophysical applications of reverse correlation, subjects may be presented with images composed of white noise and asked to make subjective "yes/no" responses about whether they perceived the presence of a specific signal, such as a face (e.g., Smith et al., 2012). Latent perceptual representations that optimally explain the pattern of responses can then be estimated by regressing subject responses against the stimuli over many trials, with the regression coefficients constituting an estimate of the representation itself.

However, current formulations of reverse correlation are widely known to be inefficient in the sense that many stimulus–response trials are required to achieve desirable estimation accuracies (Mineault et al., 2009). This inefficiency

✉ Adam C. Lammert
alammert@wpi.edu

[1] Biomedical Engineering Department, Worcester Polytechnic Institute, 100 Institute Rd, Worcester, MA 01609, USA

[2] Program of Neuroscience, Worcester Polytechnic Institute, Worcester, MA, USA

[3] Department of Communication Sciences and Disorders, University of Wisconsin-Madison, Madison, WI, USA

Published online: 29 August 2022

severely limits the feasibility of conducting reverse correlation studies to experimental protocols where subject participation can be maintained over extended timelines. Long protocols may mean that very few participants can be examined in any given study, and thus any analyses and inferences regarding possible universal aspects of human cognitive representation are severely limited. For example, in a notable study on representations of orthographic characters, Gosselin and Schyns (2003) collected 20,000 trials from three subjects over a period of two weeks. At the same time, inefficiency is an important consideration even for applications where collecting a large number of trials is feasible, because its existence implies that higher accuracies may be possible for a given number of trials if efficiency can be improved.

Attempts to improve the efficiency of reverse correlation can be broadly characterized as either retrospective or prospective. *Retrospective* approaches impose some constraints on the inferred representations at the time of estimation, after data collection is complete. One common example of this approach is smoothing (e.g., low-pass filtering) the raw estimates (Gosselin & Schyns, 2003), which stems from the assumption that high-frequency information in the estimate is irrelevant noise. It has also been shown that the assumption of sparsity—i.e., that the target representation can be sparsely represented in some basis—can lead to dramatic improvements in efficiency when methods that incorporate this assumption are employed in the estimation process. For example, Mineault et al. (2009) showed efficiency improvements using generalized linear models with sparsity priors, and Roop et al. (2021) employed a compressive sensing framework with L1 optimization.

On the other hand, *prospective* approaches to improving efficiency attempt to condition the stimuli in some way, prior to their presentation as part of the data collection. The most common example of this approach is to assume that the target has a certain form, even a very generalized one, and then construct stimuli that vary in relation to that form in specified ways. For example, approaches to psychosocial aspects of human faces (Mangini & Biederman, 2004; Dotsch & Todorov, 2012; Brinkman et al., 2017; Moon et al., 2020; Daube et al., 2021; Peterson et al., 2022; Zhan et al., 2021) have often proceeded from the assumption that representation of a trustworthy face is similar to a neutral face, and consequently generated stimuli by adding noise to an exemplar image of a neutral face. A similar approach has been taken in several auditory studies, in which stimuli were generated by adding noise to recordings of natural speech (Varnet et al., 2013a, 2013b; Varnet et al., 2015; Varnet et al., 2016). Incorporating prior knowledge about the target representation into the stimuli improves efficiency by limiting variation along dimensions that are assumed to be irrelevant to the representation.

Whether retrospective or prospective, existing approaches to improving the efficiency of reverse correlation all function on the basis of some assumed knowledge regarding the target representation—i.e., that it has some general form, or is smooth or sparse—which is then incorporated either into the stimuli, in prospective approaches, or into the estimation process, in retrospective approaches. The assumed knowledge incorporated into existing approaches, even if well justified, will exert a direct influence on estimates of the representation, limiting the essential power and promise of reverse correlation, which may be viewed as stemming from its ability to provide estimates that are unconstrained and unbiased. If, on the other hand, such assumed knowledge is not well justified, then it will compromise the quality or interpretation of the estimate by introducing bias a priori.

Rather than relying on assumed knowledge regarding the target representation, the present work attempts, for the first time, to develop a prospective approach that instead conditions reverse correlation stimuli such that their general statistical properties are more favorable for efficient estimation of any arbitrary representation, without any a priori assumptions about the nature of that target. The present approach begins only with the knowledge that a randomly generated set of stimuli will be expected to contain pairs of stimuli that are correlated by chance, especially under the conditions in which reverse correlation is typically applied—i.e., many stimuli that are low- to moderate-dimensional in size. Such correlation may be expected to decrease the effective sample size (Kish & Frankel, 1969; Liang & Zeger, 1993) of reverse correlation experiments using those stimuli by making observations overlapping and mutually predictable. Here, we prospectively employ *whitening* to improve the statistical properties of reverse correlation stimuli by eliminating covariance among the same. Whitening, sometimes called *sphering*, is a well-known statistical transformation—named in reference to white noise, which is composed of uncorrelated random variables—that eliminates covariance in multivariate data. We develop and present a mathematical justification for the effectiveness of stimulus whitening, and we demonstrate empirically that whitening can dramatically improve the efficiency of reverse correlation.

Background

Reverse correlation

Reverse correlation follows, in essence, a regression model (see Fig. 1). Subject responses, $y \in \{-1, 1\}$ are assumed to be generated by a process following

$$y = DX\beta + \varepsilon, \tag{1}$$

where the subject's internal representation is a p-by-1 vector β, the n stimuli corresponding to n trials are contained in an

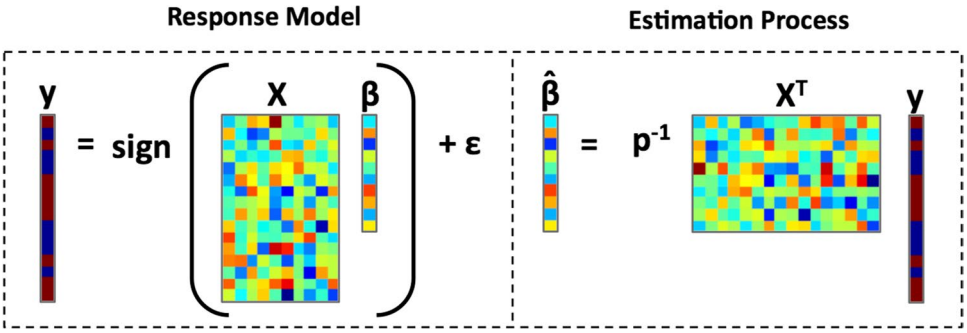

Fig. 1 In reverse correlation, the vector of subject responses (y) is modeled as resulting from the multiplication of a latent representation vector (β) and a stimulus matrix (X). This can be thought of as calculating the similarity between the latent representation and a vector representation of each presented stimulus. To estimate the latent representation, the responses are regressed against the stimuli

n-by-p matrix X, and ε is some noise. The matrix D, where $D_{ii} = 1/|X_i\beta|$ for row i of X, acts to binarize the responses to values -1 (a negative response) and 1 (a positive response), equivalent to applying the signum function. Estimates of β can be obtained using the normal equation,

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T y, \tag{2}$$

but are often made (e.g., Ahumada, 1996; Gosselin & Schyns, 2003; Murray, 2011) using the simplified formula

$$\hat{\beta} = \frac{1}{p} X^T y, \tag{3}$$

under the assumption that, across many stimuli, the columns of X will be nearly uncorrelated, meaning that $X^T X = I$, the identity matrix. Although this assumption may become inappropriate for small values of n, the value of n is typically large in reverse correlation experiments and, even so, the full normal equation can be used to compensate for any correlations that do exist among the columns of X.

Efficiency of reverse correlation

Crucial to improving the efficiency of reverse correlation is the ability to quantify estimation quality. Here, we quantify estimation quality by applying Pearson's product–moment correlation coefficient between the internal representation, β, and the estimate of that representation, $\hat{\beta}$, using the following equation:

$$r\left(\beta, \hat{\beta}\right) = \frac{(C\beta)^T \left(C\hat{\beta}\right)}{\sqrt{\frac{\beta^T C\beta}{p}} \sqrt{\frac{\hat{\beta}^T C\hat{\beta}}{p}}} \tag{4}$$

where $C = I - \frac{1}{p}1$ is the centering matrix and 1 is a matrix of all ones. Note that this metric of estimation quality assumes, as is typical in reverse correlation experiments, that internal representations encode only the relative values of the signal and not its overall magnitude. Given this metric of estimation quality, the goal of improving efficiency can be stated more precisely as maximizing the value of $r\left(\beta, \hat{\beta}\right)$ while limiting the number of trials n.

Expected correlation of random stimuli

In typical reverse correlation experiments, n stimuli of dimensionality p are randomly generated and presented to the subject in sequence. Here, we consider the case where all n stimuli are generated prior to initiation of the experiment as an n-by-p matrix X. If the stimuli are images, for instance, the rows of X represent images composed of p pixels, which can be reshaped into the desired two-dimensional format prior to presentation. Each element of X is drawn from a normal distribution with mean zero and variance 1:

$$X_{ij} \sim N(0, 1). \tag{5}$$

Equivalently, the stimuli are generated as the n rows of X, where each row is drawn from a p-variate normal distribution with mean zero and covariance matrix $V = I$:

$$X_i \sim N_p(0, V). \tag{6}$$

To examine the similarity of randomly generated stimuli, we begin by considering the row-wise scatter matrix S of X, which contains the inner product between all pairs of stimuli. The matrix S is known to follow a Wishart distribution:

$$S = XX^T \sim W_p(V, p), \tag{7}$$

where V is a scale matrix and p is the degrees of freedom of the distribution. Accordingly, the mean of S is $pV = pI$, which has off-diagonal elements of zero, indicating that the expected similarity between stimuli across experiments is nil. However, the variance of elements of S is

$$Var(S_{ij}) = p\left(v_{ij}^2 + v_{ii}v_{jj}\right), \tag{8}$$

for elements v_{ij} of V. Assuming once again that $V = I$, the above expression simplifies to

$$Var(S_{ij}) = p, \tag{9}$$

indicating that the similarities between stimuli will be expected to vary substantially between experiments, and therefore remain nontrivial for any given experiment.

Extending this analysis beyond the scatter matrix to the covariance matrix of X can help to clarify the expected magnitude of similarities between stimuli irrespective of their dimensionality. The covariance matrix can be written as:

$$\Sigma = \frac{S}{p-1} = \frac{XX^T}{p-1}, \tag{10}$$

assuming for the sake of simplicity that the rows of X are mean-zero. One can find the variance of an element of the covariance matrix, Σ_{ij}, in the case where $V = I$, as:

$$Var(\Sigma_{ij}) = \frac{p}{(p-1)^2}, \tag{11}$$

by observing that $Var\left(\frac{S_{ij}}{p-1}\right) = \frac{Var(S_{ij})}{(p-1)^2}$. The variance of elements Σ_{ij} indicates, as above, that the similarities between stimuli are expected to be nontrivial for any given experiment, when those stimuli are composed simply of elements drawn from the normal distribution. Examining the covariance matrix clarifies that this expectation is especially important when the dimensionality of the stimuli is low (i.e., p is small), which is often the case in reverse correlation experiments. The expected variance of elements C_{ij}, and associated similarity between stimuli, can be eliminated through the process of whitening. Below, we describe how stimuli can be whitened, and provide a mathematical justification for why whitening is expected to improve estimation quality and the efficiency of reverse correlation.

Method

Whitening and estimation quality: Mathematical justification

Our goal here is to show that whitening the rows of the n-by-p matrix of the stimuli, X, will maximize the correlation between β and $\hat{\beta}$. To clarify the role of X, we proceed to write Eq. 4 in terms of only X and β. Substituting values for $\hat{\beta}$ and y from the regression equations, stated above, the numerator can then be rewritten as $\frac{1}{p}(DX\beta)^T XC\beta$. Using these same observations, the denominator can be rewritten as $\sqrt{\frac{\beta^T C\beta}{p}}\sqrt{\frac{(DX\beta)^T XCX^T(DX\beta)}{p}}$. Together, these alterations yield the following equation:

$$r(\beta, \hat{\beta}) = \frac{(DX\beta)^T XC\beta}{p\sqrt{\frac{\beta^T C\beta}{p}}\sqrt{\frac{(DX\beta)^T XCX^T(DX\beta)}{p}}} \tag{12}$$

As mentioned above, we assume that internal representations encode only the relative values of the signal and not its overall magnitude. Therefore, it is reasonable to consider that $mean(\beta) = 0$ and that $\|\beta\|_2 = 1$, which allows us to further simply Eq. 12, becoming:

$$r(\beta, \hat{\beta}) = \frac{(DX\beta)^T X\beta}{\sqrt{(DX\beta)^T XCX^T(DX\beta)}} = \frac{\beta^T X^T D^T X\beta}{\sqrt{\left(\beta^T X^T D^T\right)XCX^T(DX\beta)}} \tag{13}$$

because $C\beta = \beta$ and $\beta^T C\beta = 1$ under the above assumptions, respectively. Again, the goal is to maximize the value of this equation for an arbitrary value of β, which can be done by maximizing the numerator and/or minimizing the denominator.

The numerator of Eq. 13 effectively compares the similarity, by way of taking the inner product, between β and each stimulus, and then sums the absolute values of those comparisons. This value is maximized when rows of X are equal to $\pm\beta$, and therefore cannot be optimized without prior knowledge of the value of β. Assumption of prior knowledge of the value of β is an approach often taken in the literature for improving the efficiency of reverse correlation experiments. However, such knowledge will always introduce estimation bias a priori.

The denominator of Eq. 13 clearly depends upon the similarity of rows of X, owing to the calculation of the centered row-wise scatter matrix XCX^T, but is more difficult to analyze than the numerator. To simplify the analysis, we assume that for each row i of X, $mean\left(X_i\right) = 0$, and that $\left\|X_i\right\|_2 = 1$, both of which are similar to the assumptions made above in that they are consistent with the idea that internal representations encode only the relative values of the signal and not its overall magnitude. Given these assumptions, it can be easily verified that when X has been whitened with respect to its rows, meaning that $XCX^T = I$, the denominator of Eq. 13 is equal to $\sqrt{n}$. It can also be easily verified that when X is anti-white with respect to its rows—i.e., all off-diagonal elements of XCX^T are equal to ± 1 (e.g., $XCX^T = 1$)—the value of the denominator in Eq. 13 is equal to $\sqrt{nn} = n$. Therefore, using whitened stimuli is much more favorable than using stimuli that have the opposite statistical properties, because $\sqrt{n} < n$.

The expected value of the denominator for typical stimuli—i.e., matrix X such that $X_{ij} \sim N(0, 1)$—was estimated in this work through a series of numerical simulations. In each

of these simulations, the value of the above denominator was evaluated for a randomly generated X and $\beta \sim N(0,1)$. The values of n (the number of stimuli) and p (the dimensionality of the stimuli) were assigned to 8, 16, 32, 64, 128, or 256, such that all combinations of n and p were considered. For each unique combination of n and p, 1000 total simulations were conducted. It was found that, for a unique value of n and p, the mean value of the denominator was well described by the formula $\sqrt{n\left(\frac{n}{p}+1\right)}$. Note that the value of this formula exists between $\sqrt{n}$ and n for most relevant values of n and p. It is higher than $\sqrt{n}$ for all values of n, $p \geq 1$, and lower than n for all values of n other than $n \gg p$, or more specifically all values of $\frac{n}{p} < n$, and then approximately equal to n. Furthermore, the denominator is equal to $\sqrt{n} \approx \sqrt{n\left(\frac{n}{p}+1\right)}$ when $p \gg n$, which is consistent with the expectation that rows with many elements (i.e., stimuli of high dimensionality) will be less covariant on average. Critically, the fact that the denominator value from this formula is higher than $\sqrt{n}$ for all reasonable values of n and p confirms that whitening the matrix X can be expected to maximize $r\left(\beta, \hat{\beta}\right)$, and therefore estimation quality.

Whitening and estimation quality: Empirical demonstration

To assess whether the theoretical efficiency improvements associated with whitening the stimuli could be observed empirically, a series of simulations were conducted in MATLAB to assess estimation quality as a function of the number of trials. The simulations were designed to follow Gosselin and Schyns (2003), in which one target of study was the internal representation of the printed letter "S." In this study, three subjects completed 20,000 trials, in which subjects were shown random images (i.e., with pixel values drawn from a Bernoulli distribution) and asked to indicate, with a simple yes/no response, whether the image contained the letter "S." Each subject's responses were used to generate an estimate using reverse correlation. One subject's estimated representation of "S" (shown in Fig. 3a) was used as the internal representation, β, in the simulations described here. The "S" was recreated by horizontally scaling a lowercase "s" in Verdana font, as described in Gosselin and Schyns (2003).

For each simulation, a stimulus matrix of normally distributed random values of size n-by-p was generated as described in Eq. 5. This stimulus matrix was then either whitened (see Eq. 15) or left unwhitened. Responses were generated using the assumed response-generating process described in Eq. 1. Representation estimates were obtained using the typical regression-based reverse correlation procedure described in Eq. 3. Simulations were conducted with values of n ranging from 100 to 10,000 (specifically, $n = 100$ to 500 in increments of 100, $n = 500$ to 5000 in increments of 500, and $n = 5000$ to 10,000 in increments of 1000), and values of p ranging from 100 to 10,000 (specifically, $p = 10^2, 20^2, 30^2, \ldots, 100^2$), for all combinations of values for n and p. At each combination of n and p, a total of 60 independent simulations were conducted (30 with unwhitened stimuli, and another 30 with whitened stimuli), and the mean estimation quality and 95% confidence intervals (CI) were calculated separately for whitened and unwhitened stimuli. As above, estimation quality was defined as $r\left(\beta, \hat{\beta}\right)$.

Whitening procedure

For an unwhitened stimulus matrix X_u of size n-by-p, the whitening matrix W may be defined as follows:

$$W = \left(\frac{X_u C X_u^T}{p-1}\right)^{-\frac{1}{2}} \tag{14}$$

where C is the centering matrix. Other whitening matrices are possible (see, e.g., Kessy et al., 2015). This specific whitening procedure is sometimes called Mahalanobis whitening, or ZCA (zero-phase component analysis) whitening, and can be seen as inverting the matrix square root of the row-wise covariance matrix of X_u. Using the whitening matrix, one can calculate the matrix $X_w = CX_u^T W$, which is the data matrix X_u with whitened rows.

Note that when the value of n becomes very large, numerical difficulties in the whitening process may arise from the need to invert the correspondingly large matrix $\left(\frac{X_u C X_u^T}{p-1}\right)$, which is of size n-by-n. This can be overcome, and the range of possible values of n expanded, by introducing a slight bias to the diagonal of the matrix to be inverted of the form:

$$W = \left(\frac{X_u C X_u^T}{p-1} + \varepsilon I\right)^{-\frac{1}{2}} \tag{15}$$

where the value of ε is small. This is conceptually similar to ridge regression.

The regularization approach presented in Eq. 15 can be justified as follows. As suggested in Eq. 6, the rows of X are consistent with a multivariate normal distribution, their values having been drawn from $N_p(0, I)$. Consider that the rows of X were drawn, instead, from $N_p(0, \Sigma)$, with unknown covariance Σ. In that case, we would like to determine the expected value of Σ given the stimuli X. The conjugate distribution for such data is commonly taken to be the inverse Wishart distribution, with the prior $p(\Sigma) = W^{-1}(\Psi, \upsilon)$, for scale matrix Ψ and degrees of freedom υ, and posterior $p(\Sigma|X) = W^{-1}(A + \Psi, n + \upsilon)$, where $A = XX^{\mathsf{T}}$. The mean of the inverse Wishart distribution is $\Psi/(\upsilon - p - 1)$, implying that the posterior expectation of Σ

is $(A+\Psi)/(\mathrm{n}+\upsilon-\mathrm{n}-1)$. If we know that $\upsilon = p$ and $\Psi = \mathrm{I}$, then it will be the case that the posterior expectation of Σ is $\left(XX^T + I\right)/(p-1) = XX^{\mathrm{T}}/(p-1) + \epsilon I$, where $\epsilon = 1/(p-1)$. Therefore, based on this argument, and for the sake of consistency, we implement whitening using Eq. 15 for all values of n, with $\epsilon = 1/(p-1)$.

Functions implemented in MATLAB code for stimulus whitening (Eq. 15), reverse correlation (Eq. 3), and simulated subject response generation (Eq. 1) are available at https://github.com/alammert/stimulus-whitening, along with a MATLAB script constituting a working example exemplifying their use.

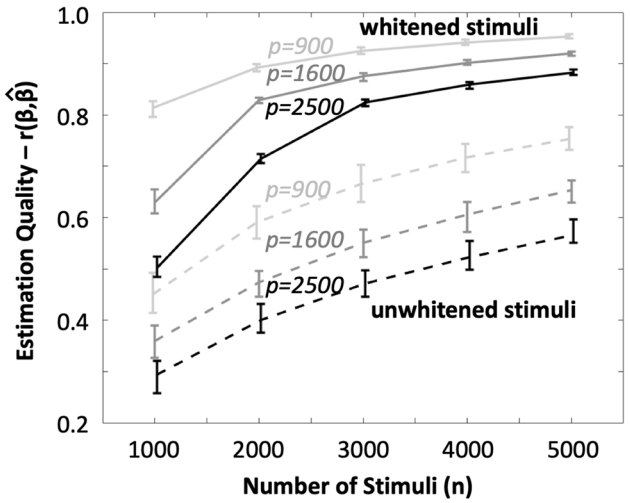

Fig. 2 Estimation quality (mean and 95% CI) for both random, unwhitened stimuli (dashed lines) and whitened stimuli (solid lines) as a function of number of stimuli presented (n). The dimensionality of the stimulus (p) has an effect that is shown at corresponding values of n, indicated by shaded lines and the value of p in text above each line

Results

Figure 2 shows estimation quality, $r\left(\beta, \hat{\beta}\right)$, as a function of number of trials (n = 1000, 2000, 3000, 4000, 5000) and dimensionality of the stimulus (p = 900, 1600, 2500) with mean and 95% CI shown. Accuracies using whitened and unwhitened stimuli are shown separately. Mean estimation quality can be seen to increase with the increasing value of n, while mean quality can be seen to decrease with the increasing value of p. Mean estimation quality was found to be higher and variability in quality was found to be lower using whitened stimuli versus unwhitened stimuli at all values of n.

Figure 3 shows estimation quality, $r\left(\beta, \hat{\beta}\right)$, as a function of the entire considered range for number of trials (n = 100 to 10,000) and dimensionality of the stimulus (p = 100 to 10,000), with colored panels and contour lines indicating the mean estimation quality for a given value of n and p. Accuracies using unwhitened and whitened stimuli are shown in figure panels a and b, respectively. Figure panel c shows the difference in estimation quality when using whitened versus unwhitened stimuli. As in Fig. 2, mean estimation quality can be seen to increase with the increasing value of n, and decrease with the increasing value of p. It was found that whitening increased mean estimation accuracy for all values of n and p. The increase in estimation quality due to whitening was found to be highest when the ratio of n and p was close to unity.

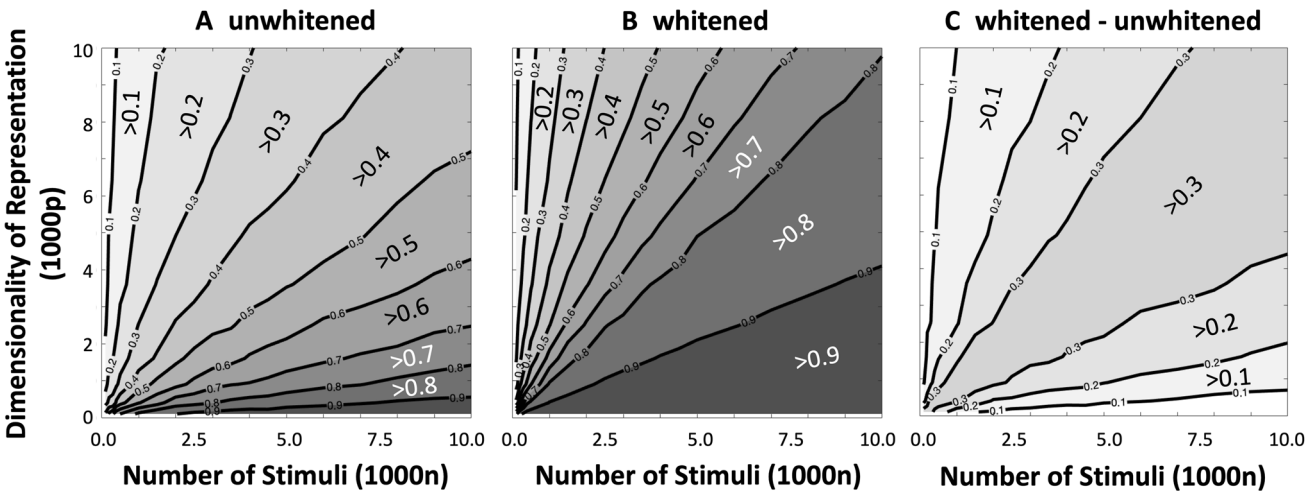

Fig. 3 Mean estimation quality for both random, unwhitened stimuli (panel **a**) and whitened stimuli (panel **b**) as a function of number of stimuli presented (n) and the dimensionality of the stimulus (p). The difference in estimation quality using whitened versus unwhitened stimuli is shown in panel **c**. Estimation quality at a given value of n and p is indicated by shading between equal-quality isolines, as well as text labels that indicate the minimum estimation quality within a shaded area.

Additional simulations were conducted to determine the number of trials required, using unwhitened stimuli, to reach mean quality equivalent to 5000 trials using whitened stimuli - i.e., when the value was at least $r = 0.89$. The simulation procedure was repeated at increasingly higher values of n, in increments of 1000, until the estimation quality using unwhitened stimuli reach this level. It was found that 40,000 trials were required to reach this level of quality, which represents an effective reduction in the number of trials of approximately 87.5%.

To facilitate a qualitative comparison of estimation quality relative to both the impact of stimulus whitening and the number of stimuli, additional simulations were conducted to contrast estimates obtained from three conditions: (1) using 5000 unwhitened stimuli, (2) using 1300 stimuli, which was found by interpolating the results in Fig. 2 to produce equivalent quality to 5000 unwhitened stimuli, and (3) using 5000 whitened stimuli. Estimates obtained from these conditions are shown in Fig. 4. The code available at https://github.com/alammert/stimulus-whitening conducts simulations under these conditions, and produces a version of Fig. 3 using randomly generated stimuli.

Discussion and conclusion

The mathematical justification provided above revealed that the two major sources of variability in estimation quality, as quantified by correlation between the estimate and the target representation, are (a) the degree of correlation among stimuli (the denominator in Eq. 13) and (b) the degree of correlation between the stimuli and the target (the numerator in Eq. 13). As such, lowering the degree of correlation among stimuli is expected in general to increase estimation quality. Whitening the stimuli is a process designed to accomplish exactly this goal, and should be expected, therefore, to improve estimation quality in reverse correlation.

Empirical results from the simulation study presented here indicate that the efficiency of reverse correlation is greatly improved by whitening stimuli before presenting them to subjects. For a given number of trials, estimation quality was observed to improve substantially when whitened stimuli were used, as compared with stimuli that were randomly generated and left unwhitened. This improvement was observed across the entire considered range for number of trials and dimensionality of the stimulus, and all

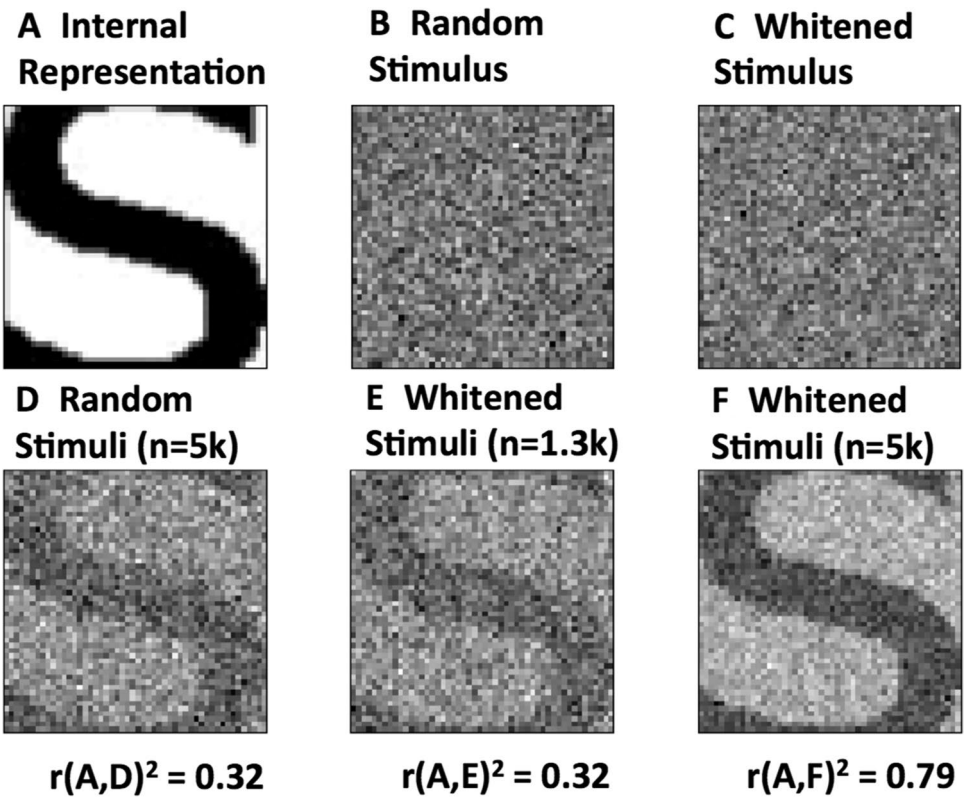

Fig. 4 Comparison of reconstruction quality using conventional random stimuli and whitened stimuli. Example random/unwhitened stimuli and whitened stimuli are shown in **b** and **c**, respectively. Estimates of the template image **a** are shown in **d**–**f**, with the stimulus type and number of stimuli used (n) indicated above those images, and the correlation coefficient between the template and the estimate (r^2, an indication of estimation quality) shown below

combinations thereof. Furthermore, when using whitened stimuli, the number of trials required to produce estimates of equivalent quality was substantially reduced, as well.

The effect of whitening for removing chance correlations between stimuli varies as a function of the dimensionality of the stimulus (p). Correlations of substantial magnitude between stimuli are more likely when p is small, as indicated by Eq. 11, in which the variance of elements of the row-wise covariance matrix is shown to increase as p decreases. Therefore, it is more likely that whitening will make a larger adjustment to randomly generated stimuli when the dimensionality of those stimuli is low. By contrast, the number of stimuli (n) is not expected to influence the size of adjustments to randomly generated stimuli, other than possibly by creating numerical problems—mentioned in the methods—associated with the need to invert an n-by-n matrix when n is very large. Based on this reasoning, one might expect the advantage of whitening regarding estimation accuracy to be maximized when p is small, regardless of the value of n. In practice, however, this is only true when n is also small. If p is quite small relative to n, then the estimation problem is not difficult, and estimation quality will be excellent using either whitened or unwhitened stimuli (i.e., "ceiling effects" are observed). Conversely, the advantage of whitening would be expected to shrink when p is large. In practice, this is only true when n is correspondingly large. If p is large relative to n, then the estimation problem is extremely difficult, and estimation quality will be poor whether whitening is employed or not. Thus, whitening performs best when p is small, unless n is considerably smaller than p. The advantage of whitening tends to disappear when n becomes much larger than p due to ceiling effects.

The empirical results also revealed that variance in estimation quality is sharply reduced by whitening stimuli. Again, the mathematical justification above revealed that the degree of random correlation among stimuli is a major source of variability in estimation quality. By eliminating any such correlation, whitening leaves only random correlation between the stimuli and the target as a source of variation in estimation quality.

Finally, the empirical results reinforce the notion, widely understood by reverse correlation practitioners, that estimation quality increases with the increasing number of trials and decreasing dimensionality of the stimulus. In broad terms, the best results from a reverse correlation experiment, regardless of whether whitening is employed, would therefore be expected when a large number of trials are for a target and stimuli that are low in dimension.

Reverse correlation has the potential to uncover latent representations underlying perception and transform our understanding of perceptual mechanisms at various levels of investigation: neural, cognitive, and psychological. However, in order for this potential to be fully realized, the fundamental inefficiency of reverse correlation paradigms must be overcome so that the breadth of its application may be increased. Whitening stimuli provides for more accurate estimates with fewer trials than simply using random stimuli, as in traditional approaches. Moreover, whitening does not impose any prior assumptions on the estimation process regarding the target representation. The dramatic improvements in efficiency demonstrated here can enable researchers to access the promise of reverse correlation by broadening its scope of application, allowing for studies to examine a wider array of representations within one individual, and also allowing deeper investigations into individual variability in, and potentially universal aspects of, perceptual representations.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13428-022-01946-w.

Acknowledgements The work described in this paper was supported by funding from the Pilot Projects Program from the University of Massachusetts Center for Clinical and Translational Science.

References

Ahumada Jr., A., & Lovell, J. (1971). Stimulus features in signal detection. *The Journal of the Acoustical Society of America, 49*(6B), 1751–1756.

Ahumada Jr, A. J. (1996). Perceptual classification images from Vernier acuity masked by noise. *Perception*, *25*(1_suppl), 2–2.

Brinkman, L., Todorov, A., & Dotsch, R. (2017). Visualising mental representations: A primer on noise-based reverse correlation in social psychology. *European Review of Social Psychology, 28*(1), 333–361.

Daube, C., Xu, T., Zhan, J., Webb, A., Ince, R. A., Garrod, O. G., & Schyns, P. G. (2021). Grounding deep neural network predictions of human categorization behavior in understandable functional features: The case of face identity. *Patterns, 2*(10), 100348.

De Boer, E., & Kuyper, P. (1968). Triggered correlation. *IEEE Transactions on Biomedical Engineering, 3*, 169–179.

Dotsch, R., & Todorov, A. (2012). Reverse correlating social face perception. *Social Psychological and Personality Science, 3*(5), 562–571.

Gosselin, F., & Schyns, P. G. (2001). Bubbles: A technique to reveal the use of information in recognition tasks. *Vision Research, 41*(17), 2261–2271.

Gosselin, F., & Schyns, P. G. (2003). Superstitious perceptions reveal properties of internal representations. *Psychological Science, 14*(5), 505–509.

Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2009). Does cognitive science need kernels? *Trends in Cognitive Sciences, 13*(9), 381–388.

Kessy, A., Lewin, A., & Strimmer, K. (2015). Optimal whitening and decorrelation: Eprint. *arXiv preprint arXiv:1512.00809*.

Kish, L., & Frankel, M. (1969). Balanced repeated replications (BRR) for analytical statistics. *Journal of the American Statistical Association, 64*(326), 698.

Liang, K. Y., & Zeger, S. L. (1993). Regression analysis for correlated data. *Annual Review of Public Health, 14*(1), 43–68.

Ljung, L. (1999). *System identification: theory for the user* (2nd ed.). Prentice Hall.

Mangini, M. C., & Biederman, I. (2004). Making the ineffable explicit: Estimating the information employed for face classifications. *Cognitive Science, 28*(2), 209–226.

Marmarelis, V. Z., & Marmarelis, P. D. (1978). *Analysis of physiological systems. The white noise approach*. Plenum Press.

Mineault, P. J., Barthelme, S., & Pack, C. C. (2009). Improved classification images with sparse priors in a smooth basis. *Journal of Vision, 9*(10), 17–17.

Moon, K., Kim, S., Kim, J., Kim, H., & Ko, Y. G. (2020). The Mirror of Mind: Visualizing Mental Representations of Self Through Reverse Correlation. *Frontiers in Psychology, 11*, 1149.

Murray, R. F. (2011). Classification images: A review. *Journal of vision, 11*(5), 2–2.

Neri, P., & Levi, D. M. (2006). Receptive versus perceptive fields from the reverse-correlation viewpoint. *Vision Research, 46*(16), 2465–2474.

Nishimoto, S., Ishida, T., & Ohzawa, I. (2006). Receptive field properties of neurons in the early visual cortex revealed by local spectral reverse correlation. *Journal of Neuroscience, 26*(12), 3269–3280.

Peterson, J. C., Uddenberg, S., Griffiths, T. L., Todorov, A., & Suchow, J. W. (2022). Deep models of superficial face judgments. *Proceedings of the National Academy of Sciences, 119*(17), e2115228119.

Ponsot, E., Burred, J. J., Belin, P., & Aucouturier, J. J. (2018). Cracking the social code of speech prosody using reverse correlation. *Proceedings of the National Academy of Sciences, 115*(15), 3972–3977.

Ringach, D., & Shapley, R. (2004). Reverse correlation in neurophysiology. *Cognitive Science, 28*(2), 147–166.

Roop, B. W., Parrell, B., & Lammert, A. C. (2021). A Compressive Sensing Approach to Inferring Cognitive Representations with Reverse Correlation. *bioRxiv*.

Smith, M. L., Gosselin, F., & Schyns, P. G. (2012). Measuring internal representations from behavioral and brain data. *Current Biology, 22*(3), 191–196.

Varnet, L., Knoblauch, K., Meunier, F., & Hoen, M. (2013a). Using auditory classification images for the identification of fine acoustic cues used in speech perception. *Frontiers in Human Neuroscience, 7*, 865.

Varnet, L., Knoblauch, K., Meunier, F., & Hoen, M. (2013b). Show me what you listen to! Auditory classification images can reveal the processing of fine acoustic cues during speech categorization. In: *INTERSPEECH* (pp. 3167-3171).

Varnet, L., Knoblauch, K., Serniclaes, W., Meunier, F., & Hoen, M. (2015). A psychophysical imaging method evidencing auditory cue extraction during speech perception: A group analysis of auditory classification images. *PLoS One, 10*(3)

Varnet, L., Meunier, F., Trollé, G., & Hoen, M. (2016). Direct viewing of dyslexics' compensatory strategies in speech in noise using auditory classification images. *PLoS One, 11*(4), e0153781.

Volterra, V. (1930). *Theory of functionals and of integral and integrodifferential equations*. Blakie.

Wiener, N. (1958). *Nonlinear problems in random theory*. John Wiley & Sons.

Zhan, J., Liu, M., Garrod, O. G., Daube, C., Ince, R. A., Jack, R. E., & Schyns, P. G. (2021). Modeling individual preferences reveals that face beauty is not universally perceived across cultures. *Current Biology, 31*(10), 2243–2252.