Comparing two basic approaches to decorrelation transforms

Kris Innanen, Marcelo Guarido and Daniel Trad

ABSTRACT

Statistical decorrelation transforms map clusters of multivariate data to domains in which they are uncorrelated. In 2020 an algorithm was introduced to decorrelate deterministic optimization problems. In the approach, a given model space is re-parameterized such that a quadratic objective function defined on that space maps to one whose Hessian matrix is the unit; this procedure is immediately applicable to statistical decorrelation problems. The approach is essentially geometrical, in that involves designing the re-parameterization as a coordinate transform involving oblique-rectilinear basis vectors. In this paper the approach, which is procedurally very different from other decorrelation approaches, is investigated to understand what relationship it bears to standard methods, which are generally based on factorization algorithms. The results are suggestive that the geometric approach and its various realizations are different from existing methods, they may represent a generalization of the ZCA approach. The algorithm meanwhile may have some advantages, in that once one instance of the transform is constructed, alternate versions can be computed with little additional calculation.

INTRODUCTION

Decorrelation (sometimes called sphering, or whitening) is a procedure whereby sets of multiple random variables are examined in altered coordinate systems, within which they are uncorrelated (e.g., Kessy et al., 2018). For problems of N dimensions (i.e., in which a single datum is specified with a column vector $\mathbf{x}$ of length N), the coordinate transforms take the form of $N \times N$ matrices $\mathbf{L}$ designed to act on the datum $\mathbf{x}$ such that the new datum is specified by $\mathbf{y} = \mathbf{L}\mathbf{x}$, whose covariance matrix is the unit.

Decorrelation has broad application in multivariate statistics, data processing, and machine learning (e.g., Li and Zhang, 1998; Ioffe and Szegedy, 2015; Santurkar et al., 2018), but its impact is in fact broader than this. Optimization methods which involve locally quadratic approximations of an objective function can benefit from decorrelation, after which the Hessian matrix is the unit; this is currently being examined as a means to suppress the phenomenon of cross-talk in seismic inversion (Lume et al., 2022). In Hamiltonian dynamics (both classical and quantum), in particular those involving harmonic oscillators, decorrelation procedures can render equations of motion in coordinate systems in which the energy of multiparticle systems is additive, simplifying analysis.

Innanen (2020) described an approach to decorrelation which emerged from an effort to develop a geometrical approach to combating cross-talk, i.e., by transforming a Hessian matrix to the unit. Because by substituting the covariance matrix for the Hessian this immediately applies to multivariate statistical decorrelation, the question arises of how the geometric approach relates to existing statistical decorrelation algorithms. In the concise and clear review of Kessy et al. (2018), five main transforms are reviewed, and their relations are discussed. Each of the PCA, ZCA, PCA-cor, ZCA-cor, and Cholesky methods

are set out and discussed in terms of the rotational ambiguity of the equations defining decorrelation.

The purpose of this report is to take several of these standard methods, and investigate how they relate to the geometrical approach. The approach taken is to examine the behaviour of the transforms in a maximally simple situation, which in this case means bivariate data, clustering of which is easy to plot and qualitatively examine. The answers arrived at are therefore also qualitative, of course, but that is what we were seeking — a way of understanding the transformations.

DECORRELATION BY FACTORIZATION

Let N random variables be stored in an $N \times 1$ column vector $\mathbf{x} = [x_1, x_2, ..., x_N]^T$, where the elements x_i are the components of $\mathbf{x}$ in a particular coordinate system. Let its covariance values (determined through analysis of many realizations of $\mathbf{x}$) be stored in the symmetric, invertible $N \times N$ covariance matrix $\mathbf{C}$, which is generally different from the unit matrix if there are correlations amongst the N variables. The $N \times N$ transformation matrix $\mathbf{L}$ is referred to as a "decorrelation transformation" if it satisfies

$$\mathbf{L}\mathbf{C}\mathbf{L}^T = \mathbf{I}, \tag{1}$$

that is, if it takes $\mathbf{x}$ over into a $\mathbf{y} = \mathbf{L}\mathbf{x}$ whose N variables are uncorrelated. By pre- and post-multiplying (1) by $\mathbf{L}^{-1}$ and $\left(\mathbf{L}^T\right)^{-1}$ respectively, this condition on $\mathbf{L}$ can be re-expressed as

$$\mathbf{L}^T\mathbf{L} = \mathbf{C}^{-1}, \tag{2}$$

which shows that designing a decorrelation transformation is ultimately a problem of factorization of $\mathbf{C}$. Most decorrelations are based on eigen-decompositions. We set up two eigen-systems,

$$\mathbf{C} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^T \text{ and } \mathbf{P} = \mathbf{G}\mathbf{\Theta}\mathbf{G}^T, \tag{3}$$

where $\mathbf{P}$ is an auxiliary matrix satisfying $\mathbf{C} = \mathbf{V}^{1/2}\mathbf{P}\mathbf{V}^{1/2}$, where $\mathbf{V}$ is the diagonal matrix containing the variances of $\mathbf{x}$. Amongst other things, these decompositions give computational meaning to the idea of raising $\mathbf{C}$ and $\mathbf{P}$ to rational powers, through

$$\mathbf{C}^\alpha = \mathbf{U}\mathbf{\Sigma}^\alpha\mathbf{U}^T, \text{ and } \mathbf{P}^\beta = \mathbf{G}\mathbf{\Theta}^\beta\mathbf{G}. \tag{4}$$

Standard decorrelation/whitening procedures may now be defined. Setting $\alpha = -1/2$ in (4) produces the symmetric matrix $\mathbf{C}^{-1/2} = \left(\mathbf{C}^{-1/2}\right)^T$. Because $\mathbf{C}$ and all its powers are symmetric, and because its powers commute*, we can construct

$$\mathbf{I} = \mathbf{C}^{-1}\mathbf{C} = \left(\mathbf{C}^{-1/2}\mathbf{C}^{-1/2}\right)\mathbf{C} = \mathbf{C}^{-1/2}\mathbf{C}\left(\mathbf{C}^{-1/2}\right)^T. \tag{5}$$

*This comes about by (4): any two operators with the same eigenvectors commute.

Comparing (1) and (5), evidently

$$\mathbf{L}_{\text{zca}} \equiv \mathbf{C}^{-1/2} \tag{6}$$

is a valid decorrelation transformation. Use of $\mathbf{L}_{\text{zca}}$ is called "zero-phase component analysis whitening", or ZCA whitening for short. Next, we observe that if

$$\mathbf{L}_{\text{pca}} \equiv \mathbf{\Sigma}^{-1/2}\mathbf{U}^T, \tag{7}$$

then

$$\mathbf{L}_{\text{pca}}\mathbf{C}\mathbf{L}_{\text{pca}}^T = \mathbf{\Sigma}^{-1/2}\mathbf{U}^T\big(\mathbf{U}\mathbf{\Sigma}\mathbf{U}^T\big)\mathbf{U}\mathbf{\Sigma}^{-1/2} = \mathbf{I}, \tag{8}$$

using the orthonormality of the eigenvectors which implies $\mathbf{U}^T\mathbf{U} = \mathbf{I}$, and also the fact that diagonal matrices commute with all other matrices. Comparing (1) and (8), we again see that (8) defines a valid decorrelation transform. Because $\mathbf{\Sigma}^{-1/2}\mathbf{U}^T$ is a rotation into the system of principal components, this is referred to as decorrelation based on principal component analysis, or "PCA whitening".

An alternative form of ZCA-whitening can be set up based on the auxiliary system $\mathbf{P}$ instead of $\mathbf{C}$, i.e., via

$$\mathbf{L}_{\text{zca-cor}} \equiv \mathbf{P}^{-1/2}\mathbf{V}^{-1/2}, \tag{9}$$

a process referred to as "correlation-adjusted" ZCA whitening, or ZCA-cor. Likewise, PCA-whitening can be enacted on the $\mathbf{P}$ system by replacing $\mathbf{P}^{-1/2}$:

$$\mathbf{L}_{\text{pca-cor}} \equiv \mathbf{\Theta}^{-1/2}\mathbf{G}^T\mathbf{V}^{-1/2}, \tag{10}$$

producing PCA-cor. The last whitening operation is based on Cholesky factorization as opposed to an eigendecomposition. We compute the Cholesky factorization of the inverse of $\mathbf{C}$, i.e., $\mathbf{H}\mathbf{H}^T \equiv \mathbf{C}^{-1}$. Setting

$$\mathbf{L}_{\text{chol}} \equiv \mathbf{H}^T, \tag{11}$$

such by using $\mathbf{H}^T = \mathbf{H}^{-1}\mathbf{C}^{-1}$ and $\mathbf{L}_{\text{chol}}^T = \mathbf{H}$ in

$$\mathbf{L}_{\text{chol}}\mathbf{C}\mathbf{L}_{\text{chol}}^T = \mathbf{H}^T\mathbf{C}\mathbf{H} = \mathbf{H}^{-1}\mathbf{C}^{-1}\mathbf{C}\mathbf{H} = \mathbf{I}, \tag{12}$$

we observe by comparing the left-most and right-most sides of (12) that (11) is another valid whitening operator. Summarizing, we have five standard decorrelation of whitening transforms, $\mathbf{L}_{\text{zca}}$, $\mathbf{L}_{\text{pca}}$, $\mathbf{L}_{\text{zca-cor}}$, $\mathbf{L}_{\text{pca-cor}}$, and $\mathbf{L}_{\text{chol}}$. The fact that at least a few different transformation matrices exist, each of which produces a different uncorrelated set of data, reflects the fact that for a symmetric covariance matrix $\mathbf{C}$, $\mathbf{L}\mathbf{C}\mathbf{L}^T = \mathbf{I}$ only provides $N(N+1)/2$ constraint equations. Kessy et al. (2018) discuss this in terms of rotational freedom.

GEOMETRIC DECORRELATION

The whitening matrix $\mathbf{L}$ can be understood geometrically as carrying out a coordinate transform — resolving the vector $\mathbf{x}$ in a new coordinate system, whose basis vectors are given by the columns of $\mathbf{L}^T$. We will refer to any methodology motivated by this view (such as the one to follow) as "geometric decorrelation", in contrast to the essentially algebraic "factorization decorrelation" discussed above.

The rotational ambiguity discussed above is based on the fact that decorrelation applies only $N(N+1)/2$ constraints onto N^2 degrees of freedom. In principle, the ambiguity can be resolved by supplying the remaining $N(N-1)/2$ equations externally. In practice, unfortunately, haphazardly supplying $N(N+1)/2$ elements of $\mathbf{L}$, and then calculating the rest, does not make for a computable scheme. However, it has been shown (Innanen, 2020) that supplying the lower-triangular elements of $\mathbf{T} = \mathbf{L}^T$, and then computing both its diagonal and upper triangular elements, column-by-column and left-to-right, embodies a stable and relatively efficient procedure.

To give a conceptual overview of the process, consider a 4×4 transformation matrix

$$\mathbf{T} = \begin{bmatrix} t_{11} & t_{12} & t_{13} & t_{14} \\ t^*_{21} & t_{22} & t_{23} & t_{24} \\ t^*_{31} & t^*_{32} & t_{33} & t_{34} \\ t^*_{41} & t^*_{42} & t^*_{43} & t_{44} \end{bmatrix}. \tag{13}$$

Given the t^*_{ij} entries a priori, the constraint equations can be applied sequentially to the columns. In this example, t_{11} is first solved for, followed by the 2 t_{i2} entries, then the 3 t_{i3} entries, and finally the 4 t_{i4} entries. The process propagates from left to right, i.e., the t_{ij}, $j > J$ calculations depend on the t_{ij}, $j \leq J$ entries, but not vice versa.

GEOMETRIC VERSUS FACTORIZATION APPROACHES

The distinction between factorization approaches based on rotational freedom discussed by Kessy et al. (2018) is quite general. It implies that any decorrelated output must be interpretable as a scaled and rotated version of every other decorrelated output. A "new" decorrelation scheme is one for which the scaling and/or rotation effect on a dataset is dissimilar to that of all other approaches.

To observe this qualitatively, we next compare different transforms, using a non-Gaussian dataset of low dimension. The low-dimensionality will make the effective rotation more obvious, and clusters of imperfectly Gaussian data will tend to contain structures remaining after decorrelation which make rotational changes apparent.

Well log data used in a previous CREWES machine learning project (Guarido and Trad, 2019) has many of the right features to test the approach. In Figure 1a, 10,816 well log data point pairs, corresponding to $s^1 =$ *porosity from density* and $s^2 =$ *neutron porosity*, are plotted. These particular data are useful for our purpose in that many of the points appear to cluster in a somewhat Gaussian, but highly correlated fashion. That is, they appear to be defining a roughly elliptical distribution, but one which is eccentric and misaligned

with the s^1 and s^2 axes. However, outlying data points are also present which are clearly non-Gaussian, giving the cluster a sort of star-shape.

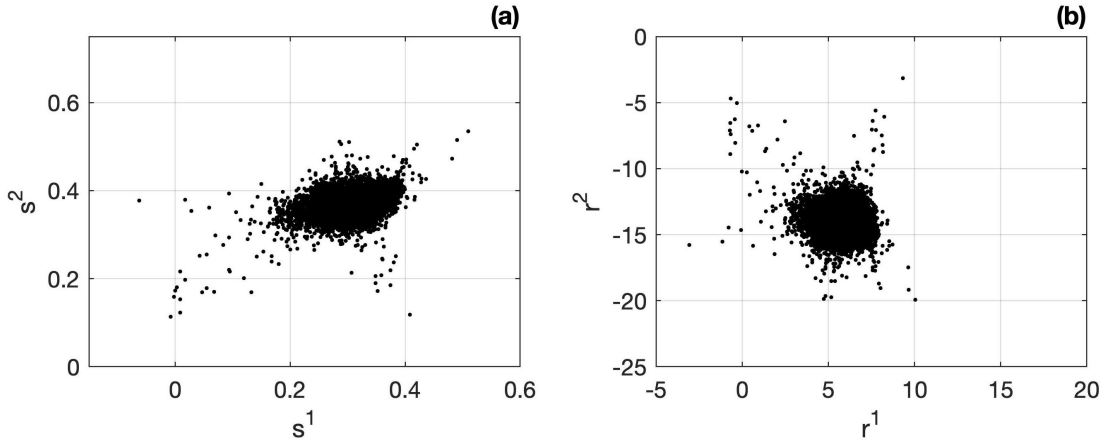

FIG. 1. Example bi-variate, on Gaussian data. (a) Well log data corresponding to $s^1 = \textit{porosity from density}$ and $s^2 = \textit{neutron porosity}$. (b) Data after decorrelation. Qualitatively, "sphering" has taken place, but the non-Gaussianity remains, characterizing the orientation of the cluster in the transformed space.

From these data the corresponding 2×2 covariance matrix is computed:

$$\mathbf{C} = 10^{-3} \begin{bmatrix} 1.811 & 0.285 \\ 0.285 & 0.628 \end{bmatrix}. \tag{14}$$

The decorrelation transformation matrix is set up with the single fixed entry $t^*_{21} = 0$:

$$\mathbf{T} = \begin{bmatrix} t_{11} & t_{12} \\ t^*_{21} & t_{22} \end{bmatrix} = \begin{bmatrix} t_{11} & t_{12} \\ 0 & t_{22} \end{bmatrix}, \tag{15}$$

and the algorithm described by Innanen (2020) is carried out, producing $\mathbf{L} = \mathbf{T}^T$

$$\mathbf{L} = \begin{bmatrix} 23.50 & 0.00 \\ 6.51 & -41.40 \end{bmatrix}. \tag{16}$$

We take each pair of points in Figure 1a, arrange them in a 2×1 column vector, and apply this $\mathbf{L}$. The resulting pairs $[r^1, r^2]^T$ are plotted in Figure 1b. Qualitatively, the impact of decorrelation is evident: the central cluster of the data points now looks largely "circular". The non-Gaussian points have been evidently rotated, but maintain the appearance of a warped version of the input – still somewhat star-shaped.

Similar procedures can be applied to the data, using the standard decorrelation transforms discussed earlier. Let us inspect ZCA, PCA, and Cholesky, as compared to the geometric decorrelation method. The clusters after decorrelation with these three standard methods are plotted in blue, khaki, and green, respectively in Figure 2. By inspection, we observe essentially common structures amongst the three clusters.

In contrast, we plot the transformed clusters three times to represent the geometric decorrelation approach. The cluster arising from setting t^*_{21} is plotted in red. Then, the

transformation is carried out over a quasi continuous range of input t_{21}^* values, ranging from -40 to 40. The first cluster, created by transforming with $t_{21}^* = -40$, and the last cluster, created by transforming with $t_{21}^* = +40$, are plotted in black. Six representative points from the cluster are tracked for each value of t_{21}^* in between, and the positions of these points are plotted as continuous dashed black lines. These trajectories characterize the way in which the geometric decorrelation resolves the rotational amibiguity: continuously varying the pre-selected entry of the transformation matrix selects one "angle" in the r coordinate system to map the cluster to. Individual cluster points sweep out circular trajectories in the transformed space as the pre-selected entry of $\mathbf{L}$ varies continously.

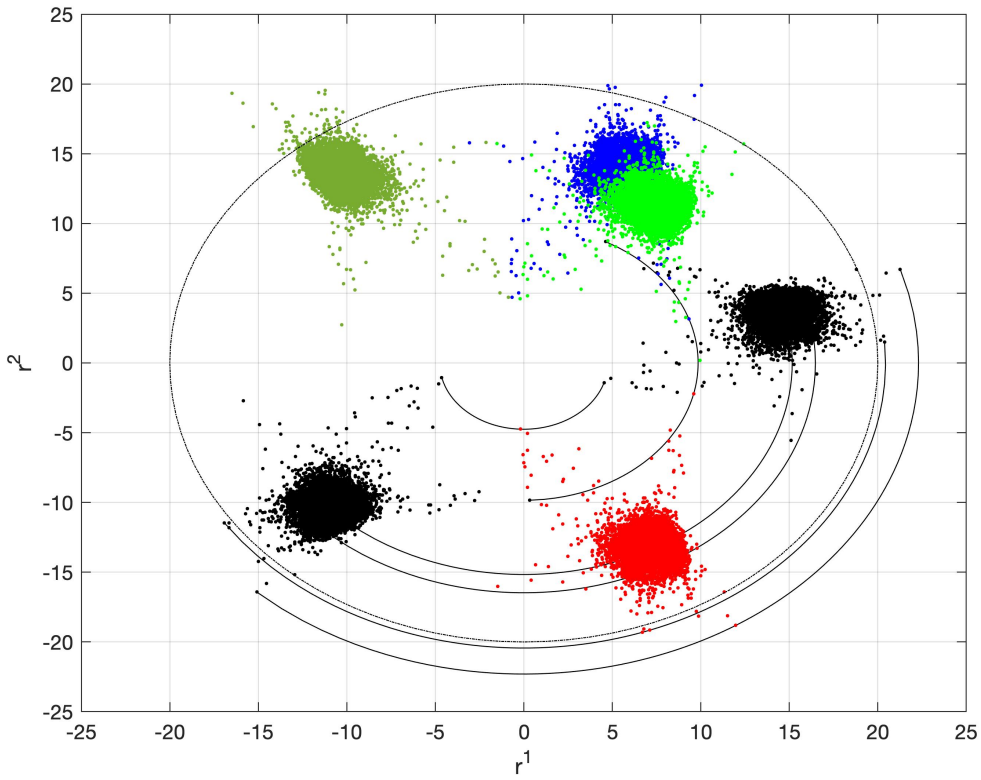

FIG. 2. Comparing various decorrelation transforms. Colour scheme: khaki=PCA; blue=ZCA; green=Cholesky; black = Geometric (for largest and smallest values of t_{21}^*); red = Geometric (for $t_{21}^* = 0$). Six points in the geometric clusters are tracked as t_{21}^* is continuously varied from -40 to 40, describing circular trajectories in the transformed space (a dashed circle is plotted for reference).

Of all of the standard transformations, it appears most reasonable to refer to the geometric approach as a version of ZCA. This is because while rigidly rotating the clusters under geometric transformation (by subjecting t_{21}^* to continuous variations), it is possible, by reflecting the cluster across the r^2 axis, to produce a cluster which almost exactly reproduce the ZCA output (see Figure 3). It is not currently clear what the significance is of the special value of t_{21}^* which brings this match about.

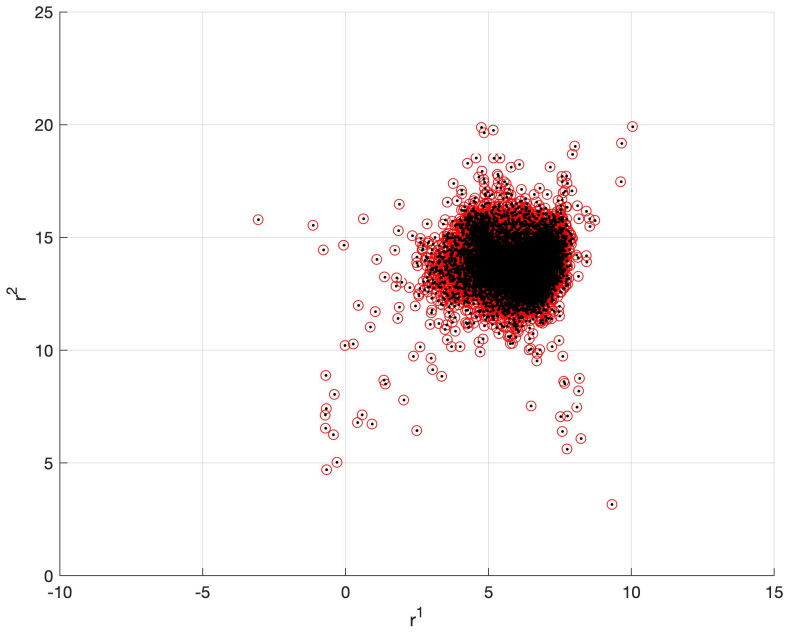

FIG. 3. Clusters under transformations associated with ZCA (red circles) and geometric (black dots) algorithms, in the latter case with t^*_{21} specially chosen to provide a close match.

DISCUSSION

The results in this report are qualitative (i.e., they are based on numerics rather than algebraic demonstrations), but they are suggestive that

1. The "geometric" approach to decorrelation is different from most standard approaches, in the sense of selecting generally different realizations within the rotational ambiguity inherent to the problem.

2. Although the approach defines a swath as opposed to a single transform, a continuous range of the pre-selected entries of the transformation matrix maps each data point to a circular arc in the transformed space; clusters generated by one choice of entries are rigidly rotated versions of clusters generated by all others.

3. This implies that exploring the degrees of freedom within the geometric decorrelation approach can be thought of as exploring different orientations of the cluster, which may reveal very different information, especially in higher dimensional data.

4. The geometric approach "resembles" the ZCA decorrelation most closely, in the sense that a value of the pre-selected transformation matrix entries can be found which reproduces the ZCA cluster. The significance of this, in regards to a geometric transform being capable of factorizing the covariance matrix in the style of ZCA, is not currently known.

5. The left-to-right propagation of information in the geometric algorithm means that re-computing certain alternative transformation matrices can occur with little additional computation. For instance, in equation (13), once the 16 entries of $\mathbf{T}$ are

calculated, perturbing t^*_{43} can only impact the calculation of the remaining elements of the 3rd column and the elements of the 4th column.

CONCLUSIONS

Statistical decorrelation transforms map clusters of multivariate data to domains in which they are uncorrelated. In 2020 an algorithm was introduced to decorrelate deterministic optimization problems. In the approach, a given model space is re-parameterized such that a quadratic objective function defined on that space maps to one whose Hessian matrix is the unit; this procedure is immediately applicable to statistical decorrelation problems. The approach is essentially geometrical, in that involves designing the re-parameterization as a coordinate transform involving oblique-rectilinear basis vectors. In this paper the approach, which is procedurally very different from other decorrelation approaches, is investigated to understand what relationship it bears to standard methods, which are generally based on factorization algorithms. The results are suggestive that the geometric approach and its various realizations are different from existing methods, they may represent a generalization of the ZCA approach. The algorithm meanwhile may have some advantages, in that once one instance of the transform is constructed, alternate versions can be computed with little additional calculation.

ACKNOWLEDGEMENTS

The sponsors of CREWES are gratefully thanked for continued support. This work was funded by CREWES industrial sponsors, NSERC (Natural Science and Engineering Research Council of Canada) through the grant CRDPJ 543578-19, and in part by an NSERC-DG.

REFERENCES

Guarido, M., and Trad, D., 2019, Machine learning as a tool to predict the mass of oil from well logs: CREWES Research Report, **31**.

Innanen, K. A., 2020, Numerical procedures for computing constrained coordinate transformation matrices: CREWES Research Report, **32**.

Ioffe, S., and Szegedy, C., 2015, Batch normalization: accelerating deep network training by reducing internal covariate shift: International conference on machine learning, 448–456.

Kessy, A., Lewin, A., and Strimmer, K., 2018, Optimal whitening and decorrelation: The American Statistician, **72**, No. 4, 309–314.

Li, G., and Zhang, J., 1998, Sphering and its properties: Sankhyā: the Indian Journal of Statistics, Series A, **60**, No. 1, 119–133.

Lume, M., Keating, S. D., and Innanen, K. A., 2022, Towards improving convergence and cross-talk suppression in multiparameter fwi by decorrelating parameter classes: CREWES Research Report, **34**.

Santurkar, S., Tsipras, D., Ilyas, A., and Madry, A., 2018, How does batch normalization help optimization?: Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS 2018), **32**, 2488–2498.