Optimal Whitening and Decorrelation

Agnan Kessy[1], Alex Lewin[2], and Korbinian Strimmer[3] *

2 December 2015

[1]Statistics Section, Dept. of Mathematics, Imperial College London, South Kensington Campus, London SW7 2AZ, UK.

[2]Dept. of Mathematics, Brunel University London, Kingstone Lane, Uxbridge UB8 3PH, UK.

[3]Epidemiology and Biostatistics, School of Public Health, Imperial College London, Norfolk Place, London W2 1PG, UK.

*To whom correspondence should be addressed. Email: k.strimmer@imperial.ac.uk

Abstract

Whitening, or sphering, is a common preprocessing step in statistical analysis to transform random variables to orthogonality. However, due to rotational freedom there are infinitely many possible whitening procedures. Consequently, there is a diverse range of sphering methods in use, for example based on principal component analysis, Cholesky matrix decomposition and Mahalanobis transformation, among others.

Here we provide an overview of the underlying theory and discuss five natural whitening procedures. Subsequently, we demonstrate that investigating the cross-covariance and the cross-correlation matrix between sphered and original variables allows to break the rotational invariance of whitening and to identify optimal transformations. As a result we recommended two particular whitening approaches: CAT-CAR whitening to produce sphered variables that are maximally similar to the original variables, and PCA-whitening based on the correlation matrix to obtain maximally compressed whitened variables.

Keywords: Whitening, decorrelation, PCA, ZCA, Cholesky decomposition, Mahalanobis transformation, CAT score, CAR score.

1 Introduction

In multivariate data analysis orthogonality among random variables is a highly desirable feature as it greatly simplifies analysis, both from a computational and a statistical standpoint. *Whitening*, or *sphering*, is a linear transformation that converts a d-dimensional random vector $\boldsymbol{x}$ with mean $\text{E}(\boldsymbol{x}) = \boldsymbol{\mu}$ and positive definite covariance matrix $\text{var}(\boldsymbol{x}) = \boldsymbol{\Sigma}$ into a new random vector

$$\boldsymbol{z} = \boldsymbol{W}\boldsymbol{x} \tag{1}$$

that has the same dimension d and unit diagonal covariance $\text{var}(\boldsymbol{z}) = \boldsymbol{I}$. To ensure that $\boldsymbol{z}$ is also centered, i.e. $E(\boldsymbol{z}) = 0$, the mean can be subtracted, either before or after the whitening transformation, however this is not necessary for whitening itself. The transformation to $\boldsymbol{z}$ is controlled by the whitening matrix $\boldsymbol{W}$ which is square but generally not symmetric. As $\text{var}(\boldsymbol{z}) = \boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^T = \boldsymbol{I}$ it needs to satisfy the condition

$$\boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{\Sigma}^{-1} \tag{2}$$

but can otherwise be freely chosen. Thus infinitely many possible whitening transformations exist, which begs the question how to further differentiate among these procedures and how to select an optimal whitening matrix in a particular situation.

In practice there will be an $n \times d$ data matrix $\boldsymbol{X}$ whose rows are implicitly assumed to be drawn from a distribution with expectation $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The data matrix is whitened by transformation to $\boldsymbol{Z} = \boldsymbol{X}\widehat{\boldsymbol{W}}^T$, where $\widehat{\boldsymbol{W}}$ is a function of an appropriate covariance matrix estimate $\widehat{\boldsymbol{\Sigma}}$.

For clarity, in the sections below the various whitening procedures will be discussed in terms of the random variables $\boldsymbol{x}$ and $\boldsymbol{z}$ and the population parameter $\boldsymbol{\Sigma}$.

2 Rotational freedom in whitening

First we define important quantities needed in the different procedures. In order to understand the different whitening procedures, we will make use of the decomposition of the covariance matrix into the correlation matrix $\boldsymbol{P}$ and the diagonal matrix $\boldsymbol{V}$ containing the variances: $\boldsymbol{\Sigma} = \boldsymbol{V}^{1/2}\boldsymbol{P}\boldsymbol{V}^{1/2}$, and also the eigen-decomposition of the covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$ and the eigen-decomposition of the correlation matrix $\boldsymbol{P} = \boldsymbol{G}\boldsymbol{\Theta}\boldsymbol{G}^T$, where $\boldsymbol{U}$, $\boldsymbol{G}$ contain the eigenvectors and $\boldsymbol{\Lambda}$, $\boldsymbol{\Theta}$ the eigenvalues of $\boldsymbol{\Sigma}$, $\boldsymbol{P}$ respectively.

The whitening constraint Eq. 2 allows rotational freedom in the choice of whitening matrix. This freedom becomes clear by representing it in its polar decomposition

$$\boldsymbol{W} = \boldsymbol{Q}_1\boldsymbol{\Sigma}^{-1/2}, \tag{3}$$

where $\boldsymbol{\Sigma}^{-1/2}$ is the unique inverse matrix square root of $\boldsymbol{\Sigma}$, and $\boldsymbol{Q}_1$ an arbitrary orthogonal matrix. This implies a geometrical interpretation of whitening as a combination of multivariate rescaling by $\boldsymbol{\Sigma}^{-1/2}$ and rotation by $\boldsymbol{Q}_1$. Regardless of $\boldsymbol{Q}_1$ all whitening matrices $\boldsymbol{W}$ have the same singular values $\boldsymbol{\Lambda}^{-1/2}$, which follows from the singular value decomposition $\boldsymbol{W} = (\boldsymbol{Q}_1\boldsymbol{U})\boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}^T$ with $\boldsymbol{Q}_1\boldsymbol{U}$ orthogonal. This highlights that the fundamental rescaling is via the square root of the eigenvalues $\boldsymbol{\Lambda}^{-1/2}$. Geometrically, the whitening transformation is a rotation ($\boldsymbol{U}^T$) followed by scaling, possibly followed by another rotation (depending on choice of $\boldsymbol{Q}_1$).

We can consider other decompositions of $\boldsymbol{W}$ satisfying the whitening constraint. Since in many situations it is desirable to work with standardized variables $\boldsymbol{V}^{-1/2}\boldsymbol{x}$, another useful decomposition is

$$\boldsymbol{W} = \boldsymbol{Q}_2\boldsymbol{P}^{-1/2}\boldsymbol{V}^{-1/2}, \tag{4}$$

where $\boldsymbol{P}^{-1/2}$ is the unique inverse matrix square root of the correlation matrix, and $\boldsymbol{Q}_2$ is a freely chosen orthogonal matrix. In this view the variables are first scaled by the square root of the diagonal variance matrix, then rotated by $\boldsymbol{G}^T$, then scaled again by the square root of the eigenvalues of the correlation matrix, and possibly rotated once more (depending on the choice of $\boldsymbol{Q}_2$).

Note that $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$ are different rotations that lead to the same whitening matrix $\boldsymbol{W}$. Since the eigen-decompositions of covariance and correlation are not readily related to each other, there is no simple connection between $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$ other than $\boldsymbol{Q}_1 = \boldsymbol{Q}_2\boldsymbol{A}$ where the matrix $\boldsymbol{A} = \boldsymbol{P}^{-1/2}\boldsymbol{V}^{-1/2}\boldsymbol{\Sigma}^{1/2}$ is itself orthogonal.

When we are considering the properties of the different whitening procedures, two particularly useful quantities are the *cross-covariance* and *cross-correlation* matrix between the whitened vector $\boldsymbol{z}$ and the original random vector $\boldsymbol{x}$. The cross-covariance matrix $\boldsymbol{\Phi}$ between $\boldsymbol{z}$ and $\boldsymbol{x}$ is given by

$$\begin{aligned}\boldsymbol{\Phi} &= \operatorname{cov}(\boldsymbol{z}, \boldsymbol{x}) = \operatorname{cov}(\boldsymbol{W}\boldsymbol{x}, \boldsymbol{x}) \\ &= \boldsymbol{W}\boldsymbol{\Sigma} = \boldsymbol{Q}_1\boldsymbol{\Sigma}^{1/2}.\end{aligned} \tag{5}$$

The corresponding cross-correlation matrix is

$$\begin{aligned}\boldsymbol{\Psi} &= \operatorname{cor}(\boldsymbol{z}, \boldsymbol{x}) = \boldsymbol{\Phi}\,\boldsymbol{V}^{-1/2} \\ &= \boldsymbol{Q}_2\boldsymbol{A}\,\boldsymbol{\Sigma}^{1/2}\boldsymbol{V}^{-1/2} = \boldsymbol{Q}_2\boldsymbol{P}^{1/2}.\end{aligned} \tag{6}$$

Both $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ are in general not symmetric, unless $\boldsymbol{Q}_1 = \boldsymbol{I}$ or $\boldsymbol{Q}_2 = \boldsymbol{I}$.

3 Five natural whitening procedures

In practical application of whitening there are three procedures most commonly used (e.g. Li and Zhang, 1998): the Mahalanobis transformation, a scaled principal components analysis (PCA) rotation and whitening using the Cholesky decomposition of the inverse covariance matrix.

The Mahalanobis whitening transformation employs the sphering matrix

$$\boldsymbol{W}^{\text{Maha}} = \boldsymbol{\Sigma}^{-1/2}. \tag{7}$$

In the machine learning literature this approach is also known as "zero-phase components analysis" (ZCA) whitening (Bell and Sejnowski, 1997). It is the unique whitening procedure with a symmetric sphering matrix $\boldsymbol{W}$.

Whitening based on scaled principal components analysis (PCA) uses

$$\boldsymbol{W}^{\text{PCA}} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}^T \tag{8}$$

(e.g. Friedman, 1987). This transformation first rotates the data using the eigenmatrix of the covariance $\boldsymbol{\Sigma}$ (as used in standard PCA). This results in orthogonal variables, but with in general different variances. To achieve whitened data the rotated variables are scaled by the square root of the eigenvalues $\boldsymbol{\Lambda}^{-1/2}$. This is probably the most widely applied whitening procedure due to its direct connection with PCA.

It can be seen that the scaled-PCA and Mahalanobis transformations are related by a rotation $\boldsymbol{U}^T$, so Mahalanobis whitening can be seen as rotation followed by scaling followed by rotation back to the original coordinate system. The Mahalanobis and the scaled PCA transformation both naturally follow the polar decomposition Eq. **3**, with $\boldsymbol{Q}_1$ equal to $\boldsymbol{I}$ and $\boldsymbol{U}^T$ respectively.

A third commonly used whitening procedure relies on Cholesky factorization of the precision matrix which yields the sphering matrix

$$\boldsymbol{W}^{\text{Chol}} = \boldsymbol{L}^T \tag{9}$$

where $\boldsymbol{L}$ is the unique lower triangular matrix with positive diagonal values from the Cholesky factorization $\boldsymbol{L}\boldsymbol{L}^T = \boldsymbol{\Sigma}^{-1}$. The same matrix $\boldsymbol{L}$ can also be obtained from a QR decomposition of $\boldsymbol{W}^{\text{Maha}} = (\boldsymbol{\Sigma}^{1/2}\boldsymbol{L})\,\boldsymbol{L}^T$.

Recently, a further natural whitening transformation was used implicitly in connection with the correlation-adjusted test statistics CAT and CAR for variable selection (Zuber and Strimmer, 2009; Ahdesmäki and Strimmer, 2010; Zuber and Strimmer, 2011; Zuber et al., 2012). This approach, which we call here CAT-CAR whitening, employs

$$\boldsymbol{W}^{\text{CAT-CAR}} = \boldsymbol{P}^{-1/2}\boldsymbol{V}^{-1/2} \tag{10}$$

as its sphering matrix. This method arises from first standardizing the random variable by multiplication with $\boldsymbol{V}^{-1/2}$ and subsequently employing Mahalanobis-ZCA whitening using the correlation rather than covariance matrix. Note that the resulting whitening matrix $\boldsymbol{W}^{\text{CAT-CAR}}$ is in general distinct from $\boldsymbol{W}^{\text{Maha}}$, and asymmetric.

In a similar fashion, applying scaled PCA-whitening to standardized variables leads to a fifth natural whitening procedure with

$$\boldsymbol{W}^{\text{PCA-cor}} = \boldsymbol{\Theta}^{-1/2}\boldsymbol{G}^T\boldsymbol{V}^{-1/2}. \tag{11}$$

Here the standardized variables are rotated by the eigenmatrix of the correlation matrix, followed by scaling using the correlation eigenvalues. Note that $\boldsymbol{W}^{\text{PCA-cor}}$ is different from $\boldsymbol{W}^{\text{PCA}}$. The CAT-CAR whitening has the same relation to this transformation as does Mahalanobis whitening to the scaled PCA, that is CAT-CAR whitening can be seen

Table 1: Summary of the properties of the five natural whitening transformations discussed in this paper.

	Sphering	Cross-	Cross-	Rotation	
	matrix	covariance	correlation	matrix	
	$\boldsymbol{W}$	$\boldsymbol{\Phi}$	$\boldsymbol{\Psi}$	$\boldsymbol{Q}_1$	$\boldsymbol{Q}_2$
Maha.-ZCA	$\boldsymbol{\Sigma}^{-1/2}$	$\boldsymbol{\Sigma}^{1/2}$	$\boldsymbol{\Sigma}^{1/2}\boldsymbol{V}^{-1/2}$	$\boldsymbol{I}$	$\boldsymbol{A}^T$
PCA	$\boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}^T$	$\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}^T$	$\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}^T\boldsymbol{V}^{-1/2}$	$\boldsymbol{U}^T$	$\boldsymbol{U}^T\boldsymbol{A}^T$
Cholesky	$\boldsymbol{L}^T$	$\boldsymbol{L}^T\boldsymbol{\Sigma}$	$\boldsymbol{L}^T\boldsymbol{\Sigma}\boldsymbol{V}^{-1/2}$	$\boldsymbol{L}^T\boldsymbol{\Sigma}^{1/2}$	$\boldsymbol{L}^T\boldsymbol{V}^{1/2}\boldsymbol{P}^{1/2}$
CAT-CAR	$\boldsymbol{P}^{-1/2}\boldsymbol{V}^{-1/2}$	$\boldsymbol{P}^{1/2}\boldsymbol{V}^{1/2}$	$\boldsymbol{P}^{1/2}$	$\boldsymbol{A}$	$\boldsymbol{I}$
PCA-cor	$\boldsymbol{\Theta}^{-1/2}\boldsymbol{G}^T\boldsymbol{V}^{-1/2}$	$\boldsymbol{\Theta}^{1/2}\boldsymbol{G}^T\boldsymbol{V}^{1/2}$	$\boldsymbol{\Theta}^{1/2}\boldsymbol{G}^T$	$\boldsymbol{G}^T\boldsymbol{A}$	$\boldsymbol{G}^T$

as a rotation of the standardized variables, followed by scaling and another rotation back to the frame of the standardized variables.

Finally, we also consider applying the Cholesky whitening to standardized variables, but in fact the resulting sphering matrix remains identical since the Cholesky factor of the inverse correlation matrix $\boldsymbol{P}^{-1}$ is $\boldsymbol{V}^{1/2}\boldsymbol{L}$, hence $\boldsymbol{W}^{\text{Chol-cor}} = (\boldsymbol{V}^{1/2}\boldsymbol{L})^T\boldsymbol{V}^{-1/2} = \boldsymbol{L}^T = \boldsymbol{W}^{\text{Chol}}$.

In Tab. **1** we compare the five natural whitening procedures and list their sphering matrices, the corresponding cross-covariance and cross-correlation matrices as well as the associated rotation matrices $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$.

4 Optimal whitening

We now demonstrate how an optimal sphering matrix $\boldsymbol{W}$ can be selected by evaluating suitable objective functions based on cross-covariance $\boldsymbol{\Phi}$ and cross-correlation $\boldsymbol{\Psi}$, thereby breaking the the rotational invariance in whitening.

4.1 Mahalanobis-ZCA whitening

In many applications a desirable outcome of whitening is to remove correlation among variables but otherwise to aim that the transformed data remains as similar as possible to the original data.

One possible implementation of this idea is to find the whitening transformation that minimizes the squared distance $||\boldsymbol{Z}_c - \boldsymbol{X}_c||^2$ between the mean-centered whitened data matrix $\boldsymbol{Z}_c = \boldsymbol{X}_c\boldsymbol{W}^T$ and the centered original data matrix $\boldsymbol{X}_c$ (e.g. Eldar and Oppenheim, 2003). Expressed in terms of mean-centered random vectors $\boldsymbol{z}_c$ and $\boldsymbol{x}_c$ this least squares objective becomes

$$\begin{aligned} \mathrm{E}\left((\boldsymbol{z}_c - \boldsymbol{x}_c)^T(\boldsymbol{z}_c - \boldsymbol{x}_c)\right) &= \mathrm{tr}(\boldsymbol{I}) - 2\,\mathrm{E}\left(\mathrm{tr}\left(\boldsymbol{z}_c\boldsymbol{x}_c^T\right)\right) + \mathrm{tr}(\boldsymbol{\Sigma}) \\ &= d - 2\mathrm{tr}(\boldsymbol{\Phi}) + \mathrm{tr}(\boldsymbol{V})\,. \end{aligned} \tag{12}$$

Therefore, instead of minimizing the expression in Eq. **12**, we may equivalently *maximize* the trace of the cross-covariance matrix

$$\operatorname{tr}(\boldsymbol{\Phi}) = \sum_{i=1}^{d} \operatorname{cov}(z_i, x_i) = \operatorname{tr}\left(\boldsymbol{Q}_1 \boldsymbol{\Sigma}^{1/2}\right) \equiv g_1(\boldsymbol{Q}_1)\,. \tag{13}$$

Proposition 1. *Maximization of $g_1(\boldsymbol{Q}_1)$ uniquely determines the optimal whitening matrix to be the symmetric sphering matrix $\boldsymbol{W}^{Maha}$.*

Proof. $g_1(\boldsymbol{Q}_1) = \operatorname{tr}(\boldsymbol{Q}_1 \boldsymbol{U} \boldsymbol{\Lambda}^{1/2} \boldsymbol{U}^T) = \operatorname{tr}(\boldsymbol{\Lambda}^{1/2} \boldsymbol{U}^T \boldsymbol{Q}_1 \boldsymbol{U}) \equiv \operatorname{tr}(\boldsymbol{\Lambda}^{1/2} \boldsymbol{B}) = \sum_i \boldsymbol{\Lambda}_{ii}^{1/2} B_{ii}$ (since $\boldsymbol{\Lambda}$ is diagonal). Since $\boldsymbol{Q}_1$ and $\boldsymbol{U}$ are both orthogonal $\boldsymbol{B} \equiv \boldsymbol{U}^T \boldsymbol{Q}_1 \boldsymbol{U}$ is also orthogonal. This implies diagonal entries $B_{ii} \leq 1$, with equality signs for all i occurring only if $\boldsymbol{B} = \boldsymbol{I}$, hence the maximum of $g_1(\boldsymbol{Q}_1)$ is assumed at $\boldsymbol{B} = \boldsymbol{I}$, or equivalently at $\boldsymbol{Q}_1 = \boldsymbol{I}$. From Eq. **3** it follows that the corresponding optimal sphering matrix is $\boldsymbol{W} = \boldsymbol{\Sigma}^{-1/2} = \boldsymbol{W}^{\text{Maha}}$. □

For related proofs see also Genizi (1993, p. 412) and Garthwaite et al. (2012, p. 789).

As a result, we recognize that Mahalanobis-ZCA whitening is the unique procedure that maximizes the average cross-*covariance* between each component of the whitened and original vectors. Furthermore, with $\boldsymbol{Q}_1 = \boldsymbol{I}$ it is also the unique whitening procedure with a symmetric cross-covariance matrix $\boldsymbol{\Phi}$.

4.2 CAT-CAR whitening

In particular for whitening with variable selection purposes in mind we may wish to employ a similarity criterion that is invariant with regard to scale.

Specifically, when computing the least squares distance we prefer to compare the whitened centered vector $\boldsymbol{z}_c$ with the corresponding *standardized* centered vector $\boldsymbol{V}^{-1/2}\boldsymbol{x}_c$, thus minimizing

$$\operatorname{E}\left((\boldsymbol{z}_c - \boldsymbol{V}^{-1/2}\boldsymbol{x}_c)^T (\boldsymbol{z}_c - \boldsymbol{V}^{-1/2}\boldsymbol{x}_c)\right) = 2d - 2\operatorname{tr}(\boldsymbol{\Psi})\,. \tag{14}$$

Correspondingly, the objective to be maximized is the trace of the cross-*correlation* matrix

$$\operatorname{tr}(\boldsymbol{\Psi}) = \sum_{i}^{d} \operatorname{cor}(z_i, x_i) = \operatorname{tr}\left(\boldsymbol{Q}_2 \boldsymbol{P}^{1/2}\right) \equiv g_2(\boldsymbol{Q}_2)\,. \tag{15}$$

Proposition 2. *Maximization of $g_2(\boldsymbol{Q}_2)$ uniquely determines the whitening matrix to be the asymmetric sphering matrix $\boldsymbol{W}^{CAT\text{-}CAR}$.*

Proof. Completely analogous to Proposition 1, we can write $g_2(\boldsymbol{Q}_2) = \operatorname{tr}(\boldsymbol{Q}_2 \boldsymbol{G} \boldsymbol{\Theta}^{1/2} \boldsymbol{G}^T) = \sum_i \boldsymbol{\Theta}_{ii}^{1/2} C_{ii}$ where $\boldsymbol{C} \equiv \boldsymbol{G}^T \boldsymbol{Q}_2 \boldsymbol{G}$ is orthogonal. By the same argument as before it follows that $\boldsymbol{Q}_2 = \boldsymbol{I}$ maximizes $g_2(\boldsymbol{Q}_2)$. From Eq. **4** it follows that $\boldsymbol{W} = \boldsymbol{P}^{-1/2} \boldsymbol{V}^{-1/2} = \boldsymbol{W}^{\text{CAT-CAR}}$. □

As a result, we identify CAT-CAR whitening as the unique procedure guaranteeing that the components of the whitened vector $\boldsymbol{z}$ remain maximally correlated with the corresponding components of the original variables $\boldsymbol{x}$. In addition, with $\boldsymbol{Q}_2 = \boldsymbol{I}$ it is also the unique whitening transformation exhibiting a symmetric cross-correlation matrix $\boldsymbol{\Psi}$.

4.3 PCA whitening

Another frequent objective in whitening is the generation of new orthogonal variables that are useful for dimension reduction and compression.

The most widespread way to orthogonalize variables *without whitening* is via the PCA transformation, which rotates (or rotates and reflects) variables using the eigenmatrix of an estimate of the covariance matrix $\boldsymbol{\Sigma}$, where the eigenvectors are ordered so that the corresponding eigenvalues are in decreasing order. This is well-known to be the optimal orthogonal transformation for successively maximizing the variances of the transformed variables: $\text{var}(z_1), \text{var}(z_2|z_1), \cdots, \text{var}(z_d|z_1, \cdots, z_{d-1})$.

Since the whitening transformation leads to $\text{var}(z_i) = 1$ for all i, the variances cannot be optimized in this case. Instead, we search for a whitened vector $\boldsymbol{z}$ whose components z_i are maximally linked not just to the corresponding x_i as in Mahalanobis-ZCA or CAT-CAR whitening, but simultaneously to all components $x_1, \ldots, x_d$. We therefore successively optimize $\text{cov}(z_1, \boldsymbol{x}), (\text{cov}(z_2, \boldsymbol{x})|z_1), \cdots, (\text{cov}(z_d, \boldsymbol{x})|z_1, \cdots, z_{d-1})$. Since these are vectors (they are the successive rows of the cross-covariance matrix $\boldsymbol{\Phi}$), we optimize their moduli. As objective function we use the vector

$$\text{diag}\left(\boldsymbol{\Phi}\boldsymbol{\Phi}^T\right) = \text{diag}\left(\boldsymbol{Q}_1\boldsymbol{\Sigma}\boldsymbol{Q}_1^T\right) \equiv \boldsymbol{h}_1(\boldsymbol{Q}_1)\,. \tag{16}$$

containing the sum of squared covariance $\sum_{j=1}^{d} \text{cov}(z_i, x_j)^2$ between each whitened component z_i and the original variables x_j.

Proposition 3. *Optimal compression under $\boldsymbol{h}_1(\boldsymbol{Q}_1)$ is achieved by the whitening matrix $\boldsymbol{W}^{PCA}$.*

Proof. First we recall the fundamental property of PCA that principal components are optimally ordered with respect to dimension reduction, with the top k eigenvectors of $\boldsymbol{\Sigma}$ corresponding to the largest k eigenvalues providing an optimal reduced rank approximation (Jolliffe, 2002). The vector $\boldsymbol{h}_1(\boldsymbol{Q}_1)$ can be written as $\text{diag}(\boldsymbol{Q}_1\boldsymbol{\Sigma}\boldsymbol{Q}_1^T) = \text{diag}(\boldsymbol{Q}_1\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T\boldsymbol{Q}_1^T)$. The i-th element of $\boldsymbol{h}_1(\boldsymbol{Q}_1)$ is $\sum_j \Lambda_{jj} D_{ij}^2$ where $\boldsymbol{D} \equiv \boldsymbol{Q}_1\boldsymbol{U}$ is orthogonal. This is maximized when $\boldsymbol{D} = \boldsymbol{I}$, or equivalently $\boldsymbol{Q}_1 = \boldsymbol{U}^T$. In order to maintain the order of the squared rows of $\boldsymbol{\Phi}$, $\boldsymbol{\Lambda}$ must contain the eigenvalues in decreasing order, and $\boldsymbol{U}$ is ordered correspondingly. With Eq. 3 the corresponding optimal sphering matrix is $\boldsymbol{W} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}^T = \boldsymbol{W}^{\text{PCA}}$. □

Hence, we see that PCA whitening is the optimal approach if the cross-covariance is used as a measure of similarity and the whitened variables are to be maximally linked with all original variables.

It is interesting to note that the usual PCA transformation also optimizes the successive rows of the cross-covariance Φ, but subject to the transformation being orthogonal rather than whitening.

4.4 PCA-cor whitening

For reasons similar as with the CAT-CAR procedure we may also prefer to optimize cross-correlations rather than cross-covariances for whitening with compression in mind. This leads to the objective

$$\text{diag}\left(\boldsymbol{\Psi}\boldsymbol{\Psi}^T\right) = \text{diag}\left(\boldsymbol{Q}_2\boldsymbol{P}\boldsymbol{Q}_2^T\right) = \boldsymbol{h}_2(\boldsymbol{Q}_2) \tag{17}$$

containing the sum of squared correlations $\sum_{j=1}^{d} \text{cor}(z_i, x_j)^2$ between each z_i and the original x_j.

Proposition 4. *Optimal compression under $\boldsymbol{h}_2(\boldsymbol{Q}_2)$ leads to $\boldsymbol{W}^{PCA\text{-}cor}$ as the sphering matrix.*

Proof. Analogous to Proposition 3 we find $\boldsymbol{Q}_2 = \boldsymbol{G}^T$ to be optimal and with Eq. **4** we arrive at $\boldsymbol{W} = \boldsymbol{\Theta}^{-1/2}\boldsymbol{G}^T\boldsymbol{V}^{-1/2} = \boldsymbol{W}^{\text{PCA-cor}}$. □

Hence, the PCA-cor whitening transformation is optimal if we wish to maximize the sum of squared correlations between each component of the whitened vector and all components of the original untransformed random vector.

Closely related to the components of the vector $\boldsymbol{h}_2(\boldsymbol{Q}_2)$ are the multiple correlation coefficients $\text{diag}\left(\boldsymbol{\Psi}\boldsymbol{P}^{-1}\boldsymbol{\Psi}^T\right)$ between each z_i and $\boldsymbol{x}$ that are easily seen to reduce to $\boldsymbol{I}$ regardless of the choice of $\boldsymbol{Q}_2$, thereby confirming that all z_i are perfectly explained by a linear transformation of $\boldsymbol{x}$ (Eq. **1**).

4.5 Cholesky whitening

As shown above, most natural whitening procedures listed in Tab. **1** can be derived by optimizing functions of the cross-covariances and cross-correlations.

In contrast, for Cholesky whitening the optimality criterion is based on a symmetry constraint. Specifically, we note that for Cholesky whitening the cross-covariance matrix $\boldsymbol{\Phi}$ is lower-triangular with positive diagonal elements. Similarly, the same is true for the cross-correlation matrix $\boldsymbol{\Psi}$. This is a consequence of the Cholesky factorization with $\boldsymbol{L}$ being subject to the same constraint. Conversely, as $\boldsymbol{L}$ is unique, Cholesky whitening is thus the unique whitening procedure resulting from lower-triangular positive diagonal cross-covariance and cross-correlation matrices.

5 Data example

For illustrative comparison of the five discussed natural whitening transforms we applied them on the well-known iris data of Fisher and Anderson, which comprises $d = 4$

Table 2: Whitening transforms applied to Fisher's and Anderson's iris data set.

	Maha.-ZCA	PCA	Cholesky	CAT-CAR	PCA-cor
$\widehat{\text{cor}}(z_1, x_1)$	0.7137	0.8974	0.3760	0.8082	0.8902
$\widehat{\text{cor}}(z_2, x_2)$	0.9018	0.8252	0.8871	0.9640	-0.8827
$\widehat{\text{cor}}(z_3, x_3)$	0.8843	-0.0121	0.2700	0.6763	0.0544
$\widehat{\text{cor}}(z_4, x_4)$	0.5743	0.1526	1.0000	0.7429	0.0754
$\text{tr}(\widehat{\boldsymbol{\Phi}})$	**2.9829**	1.1978	1.9368	*2.8495*	0.5059
$\text{tr}(\widehat{\boldsymbol{\Psi}})$	*3.0742*	1.8632	2.5331	**3.1914**	0.1373
$\max \text{diag}(\widehat{\boldsymbol{\Phi}}\widehat{\boldsymbol{\Phi}}^T)$	3.1163	**4.2282**	3.9544	1.7437	*4.1885*
$\max \text{diag}(\widehat{\boldsymbol{\Psi}}\widehat{\boldsymbol{\Psi}}^T)$	1.9817	*2.8943*	2.7302	1.0000	**2.9185**

Bold font indicates best method, and italic font the second best method.

correlated variables (x_1: sepal length, x_2: sepal width, x_3: petal length, x_4: petal width) and $n = 150$ observations.

The results are shown in Tab. **2** with estimates based on the standard unbiased empirical estimator $\hat{\boldsymbol{\Sigma}}$ of the covariance. The upper half shows the estimated cross-correlations $\text{diag}(\boldsymbol{\Psi})$ between each component of the whitened and original vector for the five methods, and the lower half the values of the various objective functions discussed above.

As expected the Mahalanobis-ZCA and the CAT-CAR whitening produce sphered variables that are most correlated to the original data on a component-wise level, with the former achieving the best fit for the covariance-based and the latter for the correlation-based objective.

In contrast, the PCA and PCA-cor method are best at producing whitened variables that are maximally simultaneously linked with all components of the original variables. Consequently, as can be seen from the top half of Tab. **2** only the first two components of $\boldsymbol{z}$ are highly correlated with the corresponding two components in $\boldsymbol{x}$, the subsequent two components are essentially orthogonal. The last line of Tab. **2** shows that PCA-cor achieves higher maximum correlation of the first component z_1 with all components of $\boldsymbol{x}$ than PCA whitening, indicating better compression.

Finally, in all cases Cholesky whitening is the third best compromise approach, and is the only approach where one component (z_4 and x_4) is — by construction — perfectly correlated.

6 Conclusion

In this note we have investigated procedures for whitening of random variables. In general there is a continuum of valid whitening transforms, all satisfying the required constraint Eq. **2**. However, as we have demonstrated here, the rotational freedom inherent in whitening can be broken by considering cross-covariance $\boldsymbol{\Phi}$ and cross-correlations $\boldsymbol{\Psi}$

between whitened and original variables. Specifically, we have investigated five natural whitening transforms, all of which can be interpreted as either optimizing a suitable function of $\boldsymbol{\Phi}$ or $\boldsymbol{\Psi}$, or satisfying a symmetry constraint on $\boldsymbol{\Phi}$ or $\boldsymbol{\Psi}$.

As a result, we recommend two particular whitening approaches, depending on the context of application. If the aim is to obtain sphered variables that are maximally similar to the original ones, for example for reasons of interpretability, we suggest to employ cross-correlation as basis for optimization, which results in the CAT-CAR whitening procedure (Eq. **10**). Similarly, if maximal compression is desirable we suggest to use the PCA-based whitening approach, preferably also on the level of correlation (Eq. **11**).

Appendix A: Pure decorrelation transforms

Closely related to whitening procedures are pure decorrelation transforms that remove correlation but leave variances and means intact. As with whitening there are infinitely many possible pure decorrelation procedures due to rotational freedom. For any given whitening matrix $\boldsymbol{W}$ such a transformation is obtained by

$$\boldsymbol{x}^* = \boldsymbol{V}^{1/2}\boldsymbol{W}(\boldsymbol{x} - \boldsymbol{\mu}) + \boldsymbol{\mu} \tag{18}$$

which first centers, then performs whitening, and finally restores variances and means by applying a corresponding location-scale transform.

As correlation is unaffected by scale and location transformations we see that $\text{cor}(x_i^*, x_i) = \text{cor}(z_i, x_i)$, hence $\boldsymbol{\Psi}$ is also the cross-correlation between $\boldsymbol{x}^*$ and $\boldsymbol{x}$. Thus, Eq. **17** and Eq. **15** may also be used to determine optimal pure decorrelation transforms, which are obtained by substituting $\boldsymbol{W}^{\text{CAT-CAR}}$ and $\boldsymbol{W}^{\text{PCA-cor}}$ in Eq. **18**.

Appendix B: Interpretation of CAT-CAR scores

The optimality property of CAT-CAR whitening (Eq. **10** and Eq. **15**) also allows to better understand the mechanism behind variable selection and ranking using CAT and CAR scores (Zuber and Strimmer, 2009, 2011). CAT scores $\boldsymbol{P}^{-1/2}\boldsymbol{\tau}$ are 'Correlation-Adjusted T' scores where $\boldsymbol{\tau}$ is the vector of t-statistics, and CAR scores $\boldsymbol{P}^{-1/2}\boldsymbol{P}_{xy}$ are 'Correlation-Adjusted (marginal) coRrelations'. Both quantities are proposed for ranking variables in the presence of correlation and their squares as measure of variable importance.

It is straightforward to see that ranking variables by CAT and CAR scores is equivalent to first applying CAT-CAR whitening (or pure decorrelation) to the data, followed by ranking the transformed predictors with conventional t-scores $\boldsymbol{\tau}$ and marginal correlations $\boldsymbol{P}_{xy}$, respectively. The legitimacy of this procedure, and at the same time its approximating character, can thus be understood from the optimality of the underlying whitening transform.

Interestingly, estimation of correlation-adjusted test statistics can be done very effectively in high dimensions without ever computing or storing the full correlation matrix or explicitly performing whitening (Zuber et al., 2012).

References

Ahdesmäki, M. and Strimmer, K. (2010). Feature selection in omics prediction problems using cat scores and false non-discovery rate control. *Ann. Appl. Statist.*, 4:503–519.

Bell, A. J. and Sejnowski, T. J. (1997). The "independent components" of natural scenes are edge filters. *Vision Res.*, 37:3327–3338.

Eldar, Y. C. and Oppenheim, A. V. (2003). MMSE whitening and subspace whitening. *IEEE Trans. Inf. Theory*, 49:1846–1851.

Friedman, J. H. (1987). Exploratory projection pursuit. *J. Am. Stat. Assoc.*, 82:249–266.

Garthwaite, P. H., Critchley, F., Anaya-Izquierdo, K., and Mubwandarikwa, E. (2012). Orthogonalization of vectors with minimal adjustment. *Biometrika*, 99:787–798.

Genizi, A. (1993). Decomposition of R^2 in multiple regression with correlated regressors. *Statistica Sinica*, 3:407–420.

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer, New York, 2nd edition.

Li, G. and Zhang, J. (1998). Sphering and its properties. *Sankhya A*, 60:119–133.

Zuber, V., Duarte Silva, A. P., and Strimmer., K. (2012). A novel algorithm for simultaneous SNP selection in high-dimensional genome-wide association studies. *BMC Bioinformatics*, 13:284.

Zuber, V. and Strimmer, K. (2009). Gene ranking and biomarker discovery under correlation. *Bioinformatics*, 25:2700–2707.

Zuber, V. and Strimmer, K. (2011). High-dimensional regression and variable selection using CAR scores. *Stat. Appl. Genet. Molec. Biol.*, 10:34.