Spatio-Chromatic Adaptation via Higher-Order Canonical Correlation Analysis of Natural Images

Michael U. Gutmann[1,2]*, **Valero Laparra**[3], **Aapo Hyvärinen**[4,1,2,5], **Jesús Malo**[3]

1 Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland, **2** Helsinki Institute for Information Technology, University of Helsinki, Helsinki, Finland, **3** Image Processing Laboratory, Universitat de València, València, Spain, **4** Department of Computer Science, University of Helsinki, Helsinki, Finland, **5** Cognitive Mechanisms Laboratories, ATR, Kyoto, Japan

Abstract

Independent component and canonical correlation analysis are two general-purpose statistical methods with wide applicability. In neuroscience, independent component analysis of chromatic natural images explains the spatio-chromatic structure of primary cortical receptive fields in terms of properties of the visual environment. Canonical correlation analysis explains similarly chromatic adaptation to different illuminations. But, as we show in this paper, neither of the two methods generalizes well to explain both spatio-chromatic processing and adaptation at the same time. We propose a statistical method which combines the desirable properties of independent component and canonical correlation analysis: It finds independent components in each data set which, across the two data sets, are related to each other via linear or higher-order correlations. The new method is as widely applicable as canonical correlation analysis, and also to more than two data sets. We call it higher-order canonical correlation analysis. When applied to chromatic natural images, we found that it provides a single (unified) statistical framework which accounts for both spatio-chromatic processing and adaptation. Filters with spatio-chromatic tuning properties as in the primary visual cortex emerged and corresponding-colors psychophysics was reproduced reasonably well. We used the new method to make a theory-driven testable prediction on how the neural response to colored patterns should change when the illumination changes. We predict shifts in the responses which are comparable to the shifts reported for chromatic contrast habituation.

Citation: Gutmann MU, Laparra V, Hyvärinen A, Malo J (2014) Spatio-Chromatic Adaptation via Higher-Order Canonical Correlation Analysis of Natural Images. PLoS ONE 9(2): e86481. doi:10.1371/journal.pone.0086481

Editor: Daniel Osorio, University of Sussex, United Kingdom

Received November 15, 2013; **Accepted** December 6, 2013; **Published** February 12, 2014

Funding: MUG and AH acknowledge financial support by the Academy of Finland, Computational Science Program, LASTU, the Finnish Centre-of-Excellence in Algorithmic Data Analysis, the Finnish Centre-of-Excellence in Inverse Problems Research, and the Finnish Centre-of-Excellence in Computational Inference Research COIN (251170). VL and JM acknowledge financial support from the Spanish CICYT research projects TEC2009-13696 and TIN2012-38102-C03-01. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: michael.gutmann@helsinki.fi

Introduction

In this paper, we propose a new method to analyze several data sets jointly and use it to relate properties of chromatic natural images to properties of the primary visual cortex: We show that the new method provides a parsimonious statistical explanation of both spatio-chromatic processing and its adaptation to changes in illumination.

Statistical modeling of natural images under fixed, or uncontrolled, illumination reveals that "Gabor-like" features (oriented, local, bandpass features) are basic building blocks of natural images. These features are robustly obtained if statistical methods are used that take higher than second-order statistical information into account, for instance sparse coding [1], independent component analysis (ICA) and its extensions [2], k-means or restricted Boltzmann machines [3], or maximal causes analysis [4,5]. If the database of natural images contains chromatic images, features are obtained which are in addition color-opponent, that is blue-yellow, red-green, and white-black [6–10]. Color opponency is consistently obtained from tristimulus or hyperspectral images, using both second-order or higher-order approaches [11,12]. When using ICA, the spatio-chromatic tuning of the learned features was found to be similar to cells in the primary visual cortex (V1) [13]. Depending on the exact assumptions made, some methods yield features which fit experimental data better than others [5,14,15].

However, the statistical methods in [1–14] are not concerned with changing lighting conditions. The same object in daylight radiates a physically different stimulus than indoors under yellowish light. We conducted a simple motivating experiment on how ICA representations are affected by a change in illumination. Figure 1 shows that ICA filters which are optimal for daylight produced less sparse outputs for the same images under yellowish light. This shows that an efficient representation for one illuminant is not necessarily efficient for another one: To maintain efficiency, adaptation of the filters is needed [16].

Statistical modeling of tristimulus pixel values of images under different illuminations provides an explanation of chromatic adaptation for spatially flat stimuli [17]. The cited work explains adaptation in terms of mean and covariance shifts of the tristimulus pixel values. It combines an extension of measurements performed earlier [18] with a decorrelation-oriented explanation of adaptation [19].

However, the statistical methods in [17,19] are not concerned with the spatial domain, and model second-order chromatic structure (mean and covariance) only. Even after inclusion of

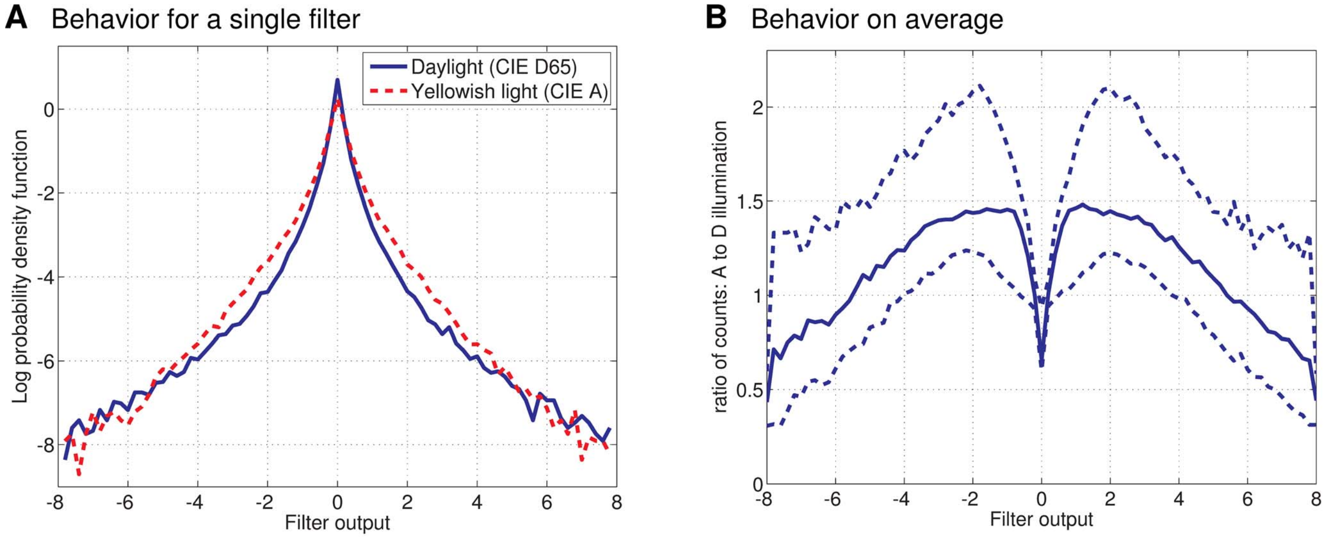


Figure 1. Efficient representations are illumination-dependent. We took ICA filters optimized to illumination CIE D65, daylight, and computed their outputs when the input images are taken under the same illuminant and under illumination CIE A, yellowish light. Each set of images was whitened with optimally adapted whitening filters. We computed histograms for all filter outputs and for both conditions. (a) For a single, randomly chosen filter, we show the log probability density functions (scaled histogram in the log domain) for daylight (blue solid) and yellowish light (red dashed). For yellowish light, the filter output takes more often intermediate values and less often very small ones; the output is less sparse. (b) For each filter, we took the ratio between the histogram obtained for yellowish and daylight illumination. This ratio allows us to read out a loss of efficiency as illumination changes: Since the ratio is smaller than one at zero and for large outputs, the response is less sparse under yellowish light than under daylight. The plot shows the median (solid curve) and the 0.1 and 0.9 quantiles (dashed curves) of the ratios of all filters.
doi:10.1371/journal.pone.0086481.g001

spatial information, modeling second-order structure does not yield biologically plausible representations, see Chapter 15 of [2].

Thus, the aforementioned statistical methods account for the different aspects of neural processing in V1 with which they are primary concerned, but neither of these approaches generalizes well to explain both aspects at the same time. We aim here at explaining both spatio-chromatic processing and adaptation using a single statistical method.

In this paper, we present a novel statistical method to jointly analyze multiple data sets (a preliminary version was presented before at a conference [20] and applied on video and magnetoencephalography data). The method is a generalization of canonical correlation analysis (CCA) that is sensitive to higher-order statistical structure: It finds independent components in each data set which, across the two data sets, are related to each other via linear or higher-order correlations. The new method is as widely applicable as CCA. We call it higher-order canonical correlation analysis (HOCCA). HOCCA is applied to a recently established database of natural images which were captured under two different lighting conditions, namely illumination CIE A, yellowish light, and illumination CIE D65, daylight [21]. Figure 2 depicts example images from the database. We show that the new statistical method allows to link both spatio-chromatic processing and adaptation in V1 to properties of natural images.

Results

Matlab code and data to reproduce the results are available at the homepage of the first author.

Higher-order canonical correlation analysis

First, we introduce HOCCA, our new statistical method to analyze multiple data sets jointly. We present HOCCA in line with the other parts of the paper: We consider the analysis of two data sets of natural images under different illumination. HOCCA is applicable to other kinds of data as well, and also to more than two data sets. More details on HOCCA can be found in Materials and Methods and Text S1.

Purpose of HOCCA. Given two data sets, the purpose of HOCCA is to efficiently represent the data as a superposition of meaningful features which are related to each other.

We denote the random vector corresponding to the first data set by $\mathbf{x}^A \in \mathbb{R}^n$, in this paper natural images under illumination CIE A; the random vector corresponding to the second data set is denoted by $\mathbf{x}^D \in \mathbb{R}^n$, here natural images under illumination CIE D65. We assume that the means have been removed. We also assume that preprocessing consists of individual whitening and, possibly, dimension reduction, both by principal component analysis (see Text S2). We denote the preprocessed data by $\mathbf{z}^A \in \mathbb{R}^m$ and $\mathbf{z}^D \in \mathbb{R}^m$, with $m \leq n$.

With these basic assumptions, the purpose of HOCCA is to represent $\mathbf{z}^A$ and $\mathbf{z}^D$ as superpositions of features $\mathbf{q}_k^A$ and $\mathbf{q}_k^D$,

$$\mathrm{z}^A = \sum_{k=1}^{m} \mathrm{q}_k^A s_k^A = \mathbf{Q}^A \mathbf{s}^A, \qquad \mathrm{z}^D = \sum_{k=1}^{m} \mathrm{q}_k^D s_k^D = \mathbf{Q}^D \mathbf{s}^D, \qquad (1)$$

such that, firstly, the canonical coordinates $\mathbf{s}^A \in \mathbb{R}^m$ and $\mathbf{s}^D \in \mathbb{R}^m$ represent the data efficiently and that, secondly, their k-th elements s_k^A and s_k^D are related to each other. We use the terms "efficient" and "related" here rather loosely. The $m \times m$ matrices $\mathbf{Q}^A$ and $\mathbf{Q}^D$ are orthonormal and contain the features as column vectors. Figure 3 summarizes the representation of the data $\mathbf{x}^A$ and $\mathbf{x}^D$ in terms of the canonical coordinates $\mathbf{s}^A$ and $\mathbf{s}^D$, respectively.

Related features exist naturally for the data considered in this paper since the images taken under the different illuminants depict the same physical objects. The statistical dependencies between $\mathbf{z}^A$

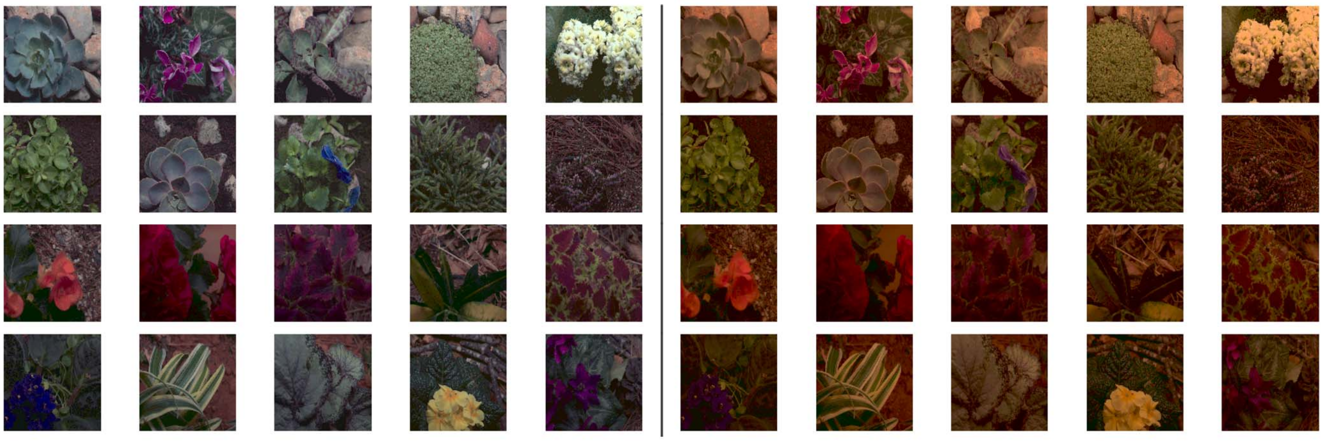


Figure 2. Examples of chromatic images from which we extracted the two data sets used in this paper. The data are image patches $\mathbf{x}^{\mathrm{A}}$ and $\mathbf{x}^{\mathrm{D}}$ of size 15×15 pixels. Left: scenes under CIE D65 illumination from where $\mathbf{x}^{\mathrm{D}}$ was obtained. Right: the same scenes under CIE A illumination from where $\mathbf{x}^{\mathrm{A}}$ was obtained. Each pair of patches was extracted at the same randomly chosen position.
doi:10.1371/journal.pone.0086481.g002

and $\mathbf{z}^{\mathrm{D}}$ are due the similar reflectance properties of the objects contained in the data sets.

In Text S1 we deal with the more general case where $\mathbf{z}^{\mathrm{A}}$ and $\mathbf{z}^{\mathrm{D}}$ can have different dimensionalities. That is, $\mathbf{z}^{\mathrm{A}}$ is assumed to have dimension m^{A} and $\mathbf{z}^{\mathrm{D}}$ dimension m^{D}. The purpose of HOCCA stays the same. Since we assume that only one s_k^{A} is related to one s_k^{D}, there can only be $m=\min(m^{\mathrm{A}},m^{\mathrm{D}})$ coupled canonical coordinates. The remaining canonical coordinates are "free" and can be used to maximize representation efficiency.

Key properties of HOCCA. In order to find a both efficient and related representation of the data, we constructed HOCCA so that higher-order statistical dependencies both within and across the data sets are taken into account. The construction of HOCCA is based on a probabilistic generative model of the data which is explained in Materials and Methods. In brief, the model couples two ICA models, one for $\mathbf{z}^{\mathrm{A}}$ and one for $\mathbf{z}^{\mathrm{D}}$, together by assuming that the independent components have statistical dependencies across the two data sets.

HOCCA has the following two key properties:

1. (Efficiency of representation) Sparsity of the estimated canonical coordinates $\langle \mathbf{q}_k^{\mathrm{A}},\mathbf{z}^{\mathrm{A}}\rangle$ and $\langle \mathbf{q}_k^{\mathrm{D}},\mathbf{z}^{\mathrm{D}}\rangle$ is taken into account when the features $\mathbf{q}_k^{\mathrm{A}}$ and $\mathbf{q}_k^{\mathrm{D}}$ are learned.
2. (Relation between data sets) The canonical coordinates can have linear or higher-order (variance) correlations across the data sets.

In addition to the coupled features $\mathbf{q}_k^{\mathrm{A}}$ and $\mathbf{q}_k^{\mathrm{D}}$, HOCCA yields estimates for the correlation coefficients ρ_k between the canonical coordinates s_k^{A} and s_k^{D}. HOCCA also estimates the degree of sparsity v_k (non-Gaussianity) of the canonical coordinates. Values close to two indicate strong non-Gaussianity while large values indicate an almost Gaussian distribution.

The above properties are in stark contrast to canonical correlation analysis (CCA). CCA represents the data using related features as in (1), but sparsity of the canonical coordinates is not a criterion, and CCA is sensitive to linear correlations between the two data sets only, see Text S2 or Chapter 3 of [22]. CCA has been extended in many ways. While extensions exist which are sensitive to higher-order correlations across the two data sets (for example kernel CCA, see the Discussion section), we are not aware

$$\begin{tikzcd}
\mathbf{x}^{\mathrm{A}} \arrow[d, shift right, "\mathbf{V}^{\mathrm{A}}"'] & & \mathbf{x}^{\mathrm{D}} \arrow[d, shift right, "\mathbf{V}^{\mathrm{D}}"'] & \begin{array}{l}\text{Original coordinates}\\ n \text{ dimensional}\end{array} \\
\mathbf{z}^{\mathrm{A}} \arrow[u, shift right, "\mathbf{V}^{\mathrm{A}\dagger}"'] \arrow[d, shift right, "(\mathbf{Q}^{\mathrm{A}})^{\top}"'] & & \mathbf{z}^{\mathrm{D}} \arrow[u, shift right, "\mathbf{V}^{\mathrm{D}\dagger}"'] \arrow[d, shift right, "(\mathbf{Q}^{\mathrm{D}})^{\top}"'] & \begin{array}{l}\text{White coordinates}\\ m \leq n \text{ dimensional}\end{array} \\
\mathbf{s}^{\mathrm{A}} \arrow[u, shift right, "\mathbf{Q}^{\mathrm{A}}"'] \arrow[rr, leftrightarrow, dashed, "\text{related}"] & & \mathbf{s}^{\mathrm{D}} \arrow[u, shift right, "\mathbf{Q}^{\mathrm{D}}"'] & \begin{array}{l}\text{Canonical coordinates}\\ m \leq n \text{ dimensional}\end{array}
\end{tikzcd}$$

Figure 3. Representing data in terms of coupled canonical coordinates. In this paper, random vectors $\mathbf{x}^{\mathrm{A}}$ and $\mathbf{x}^{\mathrm{D}}$ denote natural images under illumination CIE A (yellowish light) and under illumination CIE D65 (daylight), respectively. The whitening matrices $\mathbf{V}^{\mathrm{A}}$ and $\mathbf{V}^{\mathrm{D}}$ are determined from their covariance matrices. The symbol † denotes a (pseudo)inverse. See Text S2, Section S2.1, for formulae of these matrices. The purpose of HOCCA is to find the orthogonal matrices $\mathbf{Q}^{\mathrm{A}}$ and $\mathbf{Q}^{\mathrm{D}}$ such that, firstly, $\mathbf{x}^{\mathrm{A}}$ and $\mathbf{x}^{\mathrm{D}}$ are efficiently represented via the canonical coordinates $\mathbf{s}^{\mathrm{A}}$ and $\mathbf{s}^{\mathrm{D}}$, respectively, and that, secondly, the elements s_k^{A} and s_k^{D} of the vectors $\mathbf{s}^{\mathrm{A}}$ and $\mathbf{s}^{\mathrm{D}}$ are in a pairwise manner related to each other. We call each row of the compound matrix $(\mathbf{Q}^{\mathrm{A}})^{\mathrm{T}}\mathbf{V}^{\mathrm{A}}$ a filter or a sensor, and each column of $(\mathbf{V}^{\mathrm{A}})^{\dagger}\mathbf{Q}^{\mathrm{A}}$, and of $\mathbf{Q}^{\mathrm{A}}$ alone, a feature or optimal stimulus. The same naming convention is used for the quantities related to D65 illumination.
doi:10.1371/journal.pone.0086481.g003

of an extension which combines sensitivity to nonlinear correlations with efficiency of representation.

Performing HOCCA. HOCCA is performed by solving an optimization problem. The features $\mathbf{q}_k^{\mathrm{A}}$ and $\mathbf{q}_k^{\mathrm{D}}$, the correlation coefficients ρ_k between the canonical coordinates, and the non-Gaussianity indices ν_k are obtained by maximizing the objective f,

$$f(\mathbf{q}_1^{\mathrm{A}},\ldots,\mathbf{q}_m^{\mathrm{D}},\rho_1,\ldots,\rho_m,\nu_1,\ldots,\nu_m)=\sum_{k=1}^{m}\hat{\mathrm{E}}\left\{\log G(\mathbf{y}_k^{\mathrm{T}}\boldsymbol{\Lambda}_k\mathbf{y}_k;\nu_k,\rho_k)\right\}, \quad (2)$$

under the constraint that the features of each data set are orthogonal and of unit norm, $\langle\mathbf{q}_i^{\mathrm{A}},\mathbf{q}_j^{\mathrm{A}}\rangle=\langle\mathbf{q}_i^{\mathrm{D}},\mathbf{q}_j^{\mathrm{D}}\rangle=0$ if $i\neq j$ and one if $i=j$. The objective function is based on the log-likelihood of the probabilistic model underlying HOCCA, see Materials and Methods and Text S1. The symbol $\hat{\mathrm{E}}$ denotes the sample average over the whitened data. The vector $\mathbf{y}_k=(\langle\mathbf{q}_k^{\mathrm{A}},\mathbf{z}^{\mathrm{A}}\rangle,\langle\mathbf{q}_k^{\mathrm{D}},\mathbf{z}^{\mathrm{D}}\rangle)^{\mathrm{T}}$ contains the two inner products between the feature vectors and the whitened data $\mathbf{z}^{\mathrm{A}}$ and $\mathbf{z}^{\mathrm{D}}$. The matrix $\boldsymbol{\Lambda}_k$ is the precision matrix of the two random variables s_k^{A} and s_k^{D} which have unit variance and correlation coefficient $\rho_k\in(-1\ \ 1)$,

$$\boldsymbol{\Lambda}_k=\begin{pmatrix}1 & \rho_k\\ \rho_k & 1\end{pmatrix}^{-1}=\frac{1}{1-\rho_k^2}\begin{pmatrix}1 & -\rho_k\\ -\rho_k & 1\end{pmatrix}. \quad (3)$$

The parametrized function $G(u;\nu_k,\rho_k)$ is

$$G(u;\nu_k,\rho_k)=\frac{\Gamma\left(\frac{\nu_k+2}{2}\right)}{\pi(\nu_k-2)\Gamma\left(\frac{\nu_k}{2}\right)}\frac{1}{\sqrt{1-\rho_k^2}}\left(1+\frac{u}{\nu_k-2}\right)^{-\frac{(\nu_k+2)}{2}}, \quad (4)$$

which is valid for $u\geq 0$ and $\nu_k>2$.

The objective function f is a sum of m terms where each term only depends on a specific pair of features $\mathbf{q}_k^{\mathrm{A}}$ and $\mathbf{q}_k^{\mathrm{D}}$. This allows for an optimization scheme where the m terms are subsequently optimized, under the constraint that the new features $\mathbf{q}_k^{\mathrm{A}}$ and $\mathbf{q}_k^{\mathrm{D}}$ have unit norm and are orthogonal to the previous ones: $\langle\mathbf{q}_k^{\mathrm{A}},\mathbf{q}_i^{\mathrm{A}}\rangle=\langle\mathbf{q}_k^{\mathrm{D}},\mathbf{q}_i^{\mathrm{D}}\rangle=0,\ i<k$. In the simulations in this paper, we used such a sequential optimization.

We show in Text S1 that the objective function f stays valid in the more general setting where the dimensionality of $\mathbf{z}^{\mathrm{A}}$ and $\mathbf{z}^{\mathrm{D}}$ may differ. Maximizing f yields the $m=\min(m^{\mathrm{A}},m^{\mathrm{D}})$ coupled features $\mathbf{q}_k^{\mathrm{A}}\in\mathbb{R}^{m^{\mathrm{A}}}$ and $\mathbf{q}_k^{\mathrm{D}}\in\mathbb{R}^{m^{\mathrm{D}}}$, as well as the corresponding ν_k and ρ_k.

HOCCA as a nonlinear generalization of CCA. We show here that HOCCA is a nonlinear generalization of CCA: For large values of ν_k, the features which maximize the objective f in (2) are those which are obtained with CCA.

The objective in (1) considered as a function of the features is

$$f(\mathrm{q}_1^{\mathrm{A}},\ldots,\mathrm{q}_m^{\mathrm{D}})=const-\sum_{k=1}^{m}\hat{\mathrm{E}}\left\{\frac{\nu_k+2}{2}\log\left(1+\frac{1}{\nu_k-2}\mathrm{y}_k^{\mathrm{T}}\boldsymbol{\Lambda}_k\mathrm{y}_k\right)\right\}. \quad (5)$$

For large ν_k the term $1/(\nu_k-2)\mathbf{y}_k^{\mathrm{T}}\boldsymbol{\Lambda}_k\mathbf{y}_k$ is small so that we can use the first-order Taylor expansion $\log(1+x)=x+O(x^2)$. Taking further into account that the data is white and that the features have unit norm, we show in Text S1, Section S1.3, that

$$f(\mathrm{q}_1^{\mathrm{A}},\ldots,\mathrm{q}_m^{\mathrm{D}})\approx\mathrm{const}+\sum_{k=1}^{m}\frac{1}{1-\rho_k^2}\left(\rho_k\mathrm{q}_k^{\mathrm{DT}}\hat{\mathbf{K}}_{\mathrm{DA}}\mathrm{q}_k^{\mathrm{A}}\right), \quad (6)$$

where $\hat{\mathbf{K}}_{\mathrm{DA}}$ is the sample cross-correlation matrix between $\mathbf{z}^{\mathrm{D}}$ and $\mathbf{z}^{\mathrm{A}}$. Since $1-\rho_k^2$ is positive, the objective in (6) is maximized when $|\mathbf{q}_k^{\mathrm{DT}}\hat{\mathbf{K}}_{\mathrm{DA}}\mathbf{q}_k^{\mathrm{A}}|$ is maximized for all k under the orthonormality constraint for the features of each data set. We need the absolute value since ρ_k can be positive or negative. This set of optimization problems is the one solved by CCA, up to a possible difference in the signs, see Text S2, Section S2.3. CCA maximizes $\mathbf{q}_k^{\mathrm{DT}}\hat{\mathbf{K}}_{\mathrm{DA}}\mathbf{q}_k^{\mathrm{A}}$ so that for negative ρ_k, one of the features obtained with the maximization of f has switched signs compared to the one obtained with CCA.

Illustration of HOCCA. We illustrate here properties of HOCCA and provide some intuition by means of a simple example. We assume that $\mathbf{z}^{\mathrm{A}}$ is two dimensional and $\mathbf{z}^{\mathrm{D}}$ only one dimensional. The example thus demonstrates the applicability of HOCCA to data sets of different dimensionalities. Since the features are orthogonal, $\mathbf{q}_1^{\mathrm{A}}$ is of the form $(\cos(\alpha)\ \sin(\alpha))^{\mathrm{T}}$, for a certain angle α, and $\mathbf{q}_2^{\mathrm{A}}$ is the vector orthogonal to $\mathbf{q}_1^{\mathrm{A}}$. Feature $\mathbf{q}_1^{\mathrm{D}}$ is the scalar one (the sign is arbitrary). In this simple example, $m=1$ and the sum in (2) collapses to a single term.

We generated data according to the probabilistic model underlying HOCCA (see Materials and Methods and Text 1) with $\alpha=2,\rho=0.5$, and $\nu=2.5$. For illustration purposes, the sample size was chosen to be rather large, we used 50000 samples. A scatter plot of $\mathbf{z}^{\mathrm{A}}$ is shown in Figure 4(a). Two features $\mathbf{q}_1^{\mathrm{A}}$ are overlaid on the plot. Feature i was learned by HOCCA. Feature ii is an arbitrary alternative feature. Figure 4(b) shows scatter plots of the canonical coordinates, $\langle\mathbf{q}_1^{\mathrm{A}},\mathbf{z}^{\mathrm{A}}\rangle$ against $\langle\mathbf{q}_1^{\mathrm{D}},\mathbf{z}^{\mathrm{D}}\rangle=\mathbf{z}^{\mathrm{D}}$, and Figure 4(c) shows the distributions of $\langle\mathbf{q}_1^{\mathrm{A}},\mathbf{z}^{\mathrm{A}}\rangle$ for the features in Figure 3(a). Feature i corresponds better to the goals of HOCCA than feature ii since it yields a canonical coordinate which is sparser and more strongly statistically dependent on $\langle\mathbf{q}_1^{\mathrm{D}},\mathbf{z}^{\mathrm{D}}\rangle$. The learned correlation coefficient was $\rho=0.497$. Feature ii gave a correlation coefficient of 0.37.

Computing derivatives shows that $\log G(u;\nu,\rho)$ is monotonically decreasing and strictly convex in u. Figure 4(d) shows $\log G(u;\nu,\rho)$ for different values of ν and for ρ fixed to 0.5. According to the definition of G in Equation (4), ρ affects $\log G(u;\nu,\rho)$ only through the additive offset $-1/2\log(1-\rho^2)$ which is increasing as ρ tends to ±1. The offset is the mutual information between two Gaussian random variables with correlation coefficient ρ (see Materials and Methods). It provides a mechanism which allows HOCCA to find correlated features.

The argument of $\log G$ is the quadratic form $\mathbf{y}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{y}$ where $\mathbf{y}$ depends on α and $\boldsymbol{\Lambda}$ on ρ. The elements of $\mathbf{y}$ are the estimated canonical coordinates, and $\boldsymbol{\Lambda}$ is an estimate of their inverse covariance matrix. The quadratic form $\mathbf{y}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{y}$ corresponds thus to the squared norm of the estimated canonical coordinates after decorrelation (it is the squared Mahalanobis distance of $\mathbf{y}$ from the origin). Since $\log G(u;\nu,\rho)$ is convex, maximizing f for a fixed value of ν consists in finding features for which the norm of the decorrelated $\mathbf{y}$ is sparse, see Chapter 6 of [2]. The sum of two squared values is large or close to zero if each of the two decorrelated canonical coordinates are large or close to zero at the same time. This provides a mechanisms which allows HOCCA to find sparse canonical coordinates with possible variance correlations.

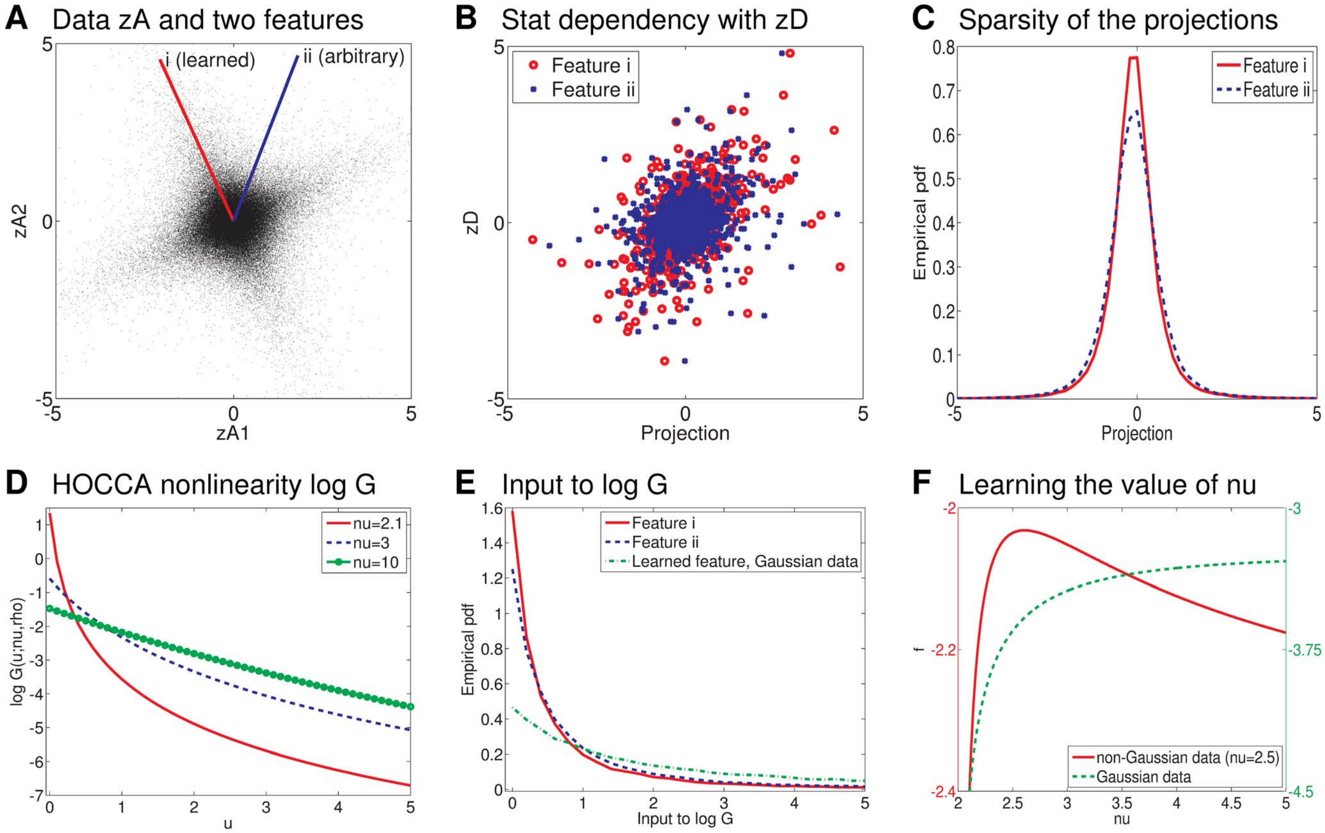


Figure 4. Illustrating HOCCA with a simple example where $\mathbf{z}^{\mathrm{A}} \in \mathbb{R}^2$ and $\mathbf{z}^{\mathrm{D}} \in \mathbb{R}$. In (a), we show two features $\mathbf{q}_1^{\mathrm{A}}$ overlaid on the scatter plot. Feature i was learned by HOCCA. Feature ii is an arbitrary alternative feature. Feature i corresponds better to the goals of HOCCA than feature ii since it yields a canonical coordinate (projection) $\langle \mathbf{q}_1^{\mathrm{A}}, \mathbf{z}^{\mathrm{A}} \rangle$ which is more strongly statistically dependent on $\langle \mathbf{q}_1^{\mathrm{D}}, \mathbf{z}^{\mathrm{D}} \rangle = \mathbf{z}^{\mathrm{D}}$ (subfigure b) and also sparser (subfigure c). (d) The nonlinearity $\log G(u;\nu,\rho)$ for different values of ν with ρ fixed to 0.5. Changing ρ does only lead to an additive offset, it does not change the shape of the nonlinearity. (e) The distribution of the input to $\log G$, $\mathbf{y}^{\mathrm{T}} \wedge \mathbf{y}$, is shown for the two features in (a). We also show the distribution for the feature learned for Gaussian data. In this case, the inputs to the nonlinearity $\log G(u;\nu,\rho)$ are less often close to zero. (f) The HOCCA objective f as a function of ν for both non-Gaussian and Gaussian data. For the non-Gaussian data, maximizing f identifies the correct value of ν. For the Gaussian data, f is increasing as ν increases (this holds also beyond the range of ν shown here). For large ν the nonlinearity in (d) is less peaked at zero, which corresponds well to the less peaked distribution for Gaussian data in (e).
doi:10.1371/journal.pone.0086481.g004

Figure 4(e) shows the distribution of $\mathbf{y}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{y}$ for the two features depicted in Figure 4(a). From a comparison with Figure 4(c), it can be seen that the feature which produces sparser canonical coordinates is also the feature which produces inputs to $\log G$ which are more often close to zero, in line with our reasoning above. The figure also shows the distribution of $\mathbf{y}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{y}$ for the learned feature when the data is Gaussian (with $\alpha = 2$ and $\rho = 0.5$ as for the non-Gaussian data above). It can be seen that the inputs to $\log G(u;\nu,\rho)$ are less often close to zero for that data. The different curves of $\log G(u;\nu,\rho)$ in Figure 4(d) suggest that, for the Gaussian data, the objective f will be larger for $\nu = 10$ than for $\nu = 2.1$.

In HOCCA, the parameter ν is learned from the data by maximizing f. Figure 4(f) shows the HOCCA objective f as a function of ν, with ρ and α fixed to their true values. We see that for the generated non-Gaussian data, $\nu \approx 2.5$ is maximizing f (red solid curve, left axis). The same figure also shows f for the Gaussian data (green dashed curve, right axis), where f increases as ν increases. In our numerical optimization, we obtained a value of $\nu = 59$, which was the value where our stopping criterion was satisfied. In this regime of ν, the approximation from the previous section becomes valid, and the features which maximize f are given by the CCA-features.

Validating HOCCA on artificial data. We used artificially generated data to validate HOCCA and to compare it with CCA. We generated data according to (1), with variable levels of correlation and sparsity of the canonical coordinates $\mathbf{s}^{\mathrm{A}}$ and $\mathbf{s}^{\mathrm{D}}$, and for randomly generated mixing matrices $\mathbf{Q}_{\mathrm{true}}^{\mathrm{A}}$ and $\mathbf{Q}_{\mathrm{true}}^{\mathrm{D}}$ of dimension $m = 10$. We constructed fifty random estimation problems and used 10000 samples to solve them (see Materials and Methods for details). In order to recover the mixing matrices, and thus the features which form their columns, we optimized the objective f in (2) for HOCCA. For CCA, we solved the singular value problem (S2–7) in Text S2.

We analyzed the results using three measures of performance (see Materials and Methods for details). First, we analyzed how well the mixing matrices (features) are recovered. Figure 5(a) shows that HOCCA led to a better recovery of the mixing matrices. The pointwise comparison in the third panel in the figure shows that HOCCA performed better for each of the fifty random estimation problems.

Second, we analyzed the efficiency of the representation, both from a sparsity and from a related information theoretical point of view. Figure 5(b) shows that the canonical coordinates recovered by HOCCA were mostly sparser than those recovered by CCA – thanks to the active sparsification inherent in HOCCA (the

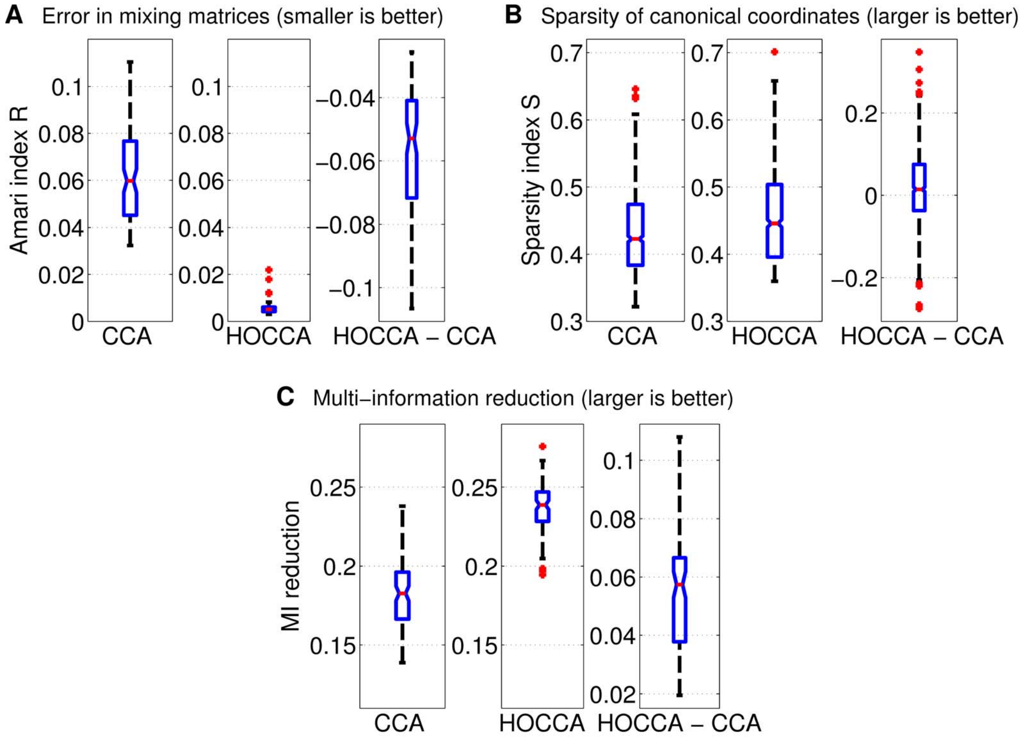


Figure 5. Validating HOCCA on artificial data: Feature identification and representation efficiency. The error of an estimated mixing matrix was measured by the Amari index $\mathcal{R}$ defined in (14). Sparsity of an estimated canonical coordinate was measured using the index $\mathcal{S}$ defined in (15). Multi-information reduction was measured by comparing the marginal entropies of the (whitened) data and the estimated canonical coordinates. We show the results for the estimation of 50 random $\mathbf{Q}^{\mathrm{A}}_{\mathrm{true}}$ and $\mathbf{Q}^{\mathrm{D}}_{\mathrm{true}}$ of dimension $m=10$: The boxplots in (a) and (c) contain 100 data points each, while the boxplots in (b) show the distribution of all 1000 estimated canonical coordinates. The first and second panel in each subfigure show the distribution of the performance indices for CCA and HOCCA, respectively. The third panel shows the distribution of the difference of the indices. HOCCA recovered the features more accurately, and led to representations with sparser and more independent canonical coordinates than CCA.

doi:10.1371/journal.pone.0086481.g005

average in the point-wise comparison is larger than zero (one-sided t-test, p-value $=10^{-11}$)). In line with this result, Figure 5(c) shows that HOCCA led to a stronger multi-information reduction than CCA. The third panel shows that HOCCA led to a more efficient representation for each of the fifty random estimation problems.

Third, we analyzed how well the coupling (correspondence) between the two data sets was identified. For that purpose, we measured the mutual information between the coupled pairs of sources. Figure 6(a) shows that HOCCA recovered in most cases almost all of the mutual information. In some rare cases, however, it failed. In preliminary work, we found that the objective has local optima [20]. The observed failures are presumably due to the fact that the optimization scheme did not find the global maximum. For CCA, such failures were more rare. The mode of the CCA distribution, on the other hand, is smaller than for HOCCA indicating that the general level of recovery was also smaller. In some cases, CCA recovered more mutual information per source-pair than what was actually available. Because the total amount of mutual information between all source-pairs is preserved, this means that CCA over-allocated mutual information for some sources while, consequently, having to allocate less to other sources.

In Figure 6(b), we investigate how much mutual information per estimation problem was recovered. While Figure 6(a) dealt with a comparison per source-pair, this figure is a comparison which takes all the source-pairs per estimation problem into account. The boxplot in the figure shows the difference between the fraction of total mutual information which HOCCA recovered per estimation problem and the fraction which CCA recovered. The distribution is skewed towards positive values which indicates that HOCCA recovered more often more mutual information between the corresponding sources than CCA.

The results reported above validate the theoretical properties of HOCCA: We found that HOCCA led to a more efficient representation of the two data sets than CCA, as measured by sparsity or gain in independence, and that the recovery of the correspondence between the two data sets was also better, as measured by mutual information.

From natural images to spatio-chromatic adaptation

Next, we apply HOCCA to chromatic natural images that were acquired under two different illumination conditions, daylight and yellowish light. We analyze the learned coupled representations, show that they account for known experimental results and make a theory-driven prediction. Two properties of the learned representations are of particular interest: First, the representation of the two data sets individually, that is, the spatio-chromatic processing for a given illuminant. Second, the coupling (correspondence) between the representations across the two data sets, that is, the adaptation to changes in the illumination. We also compare the representations learned by HOCCA with those from other statistical methods, namely ICA, CCA, and whitening by principal component analysis, see Materials and Methods for details and Tables 1 and 2 for an overview.

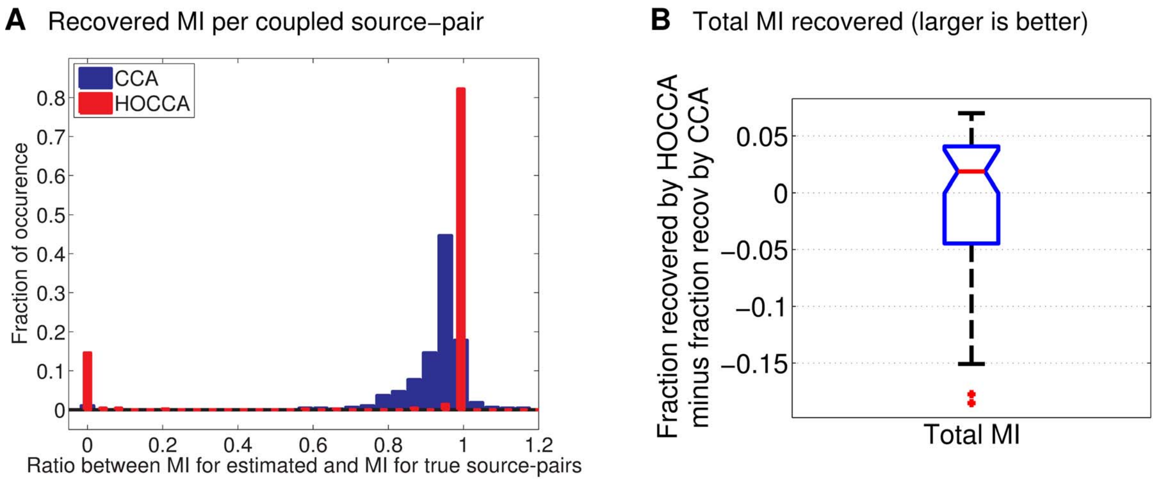


Figure 6. Validating HOCCA on artificial data: Identification of the coupling. (a) We computed the mutual information (MI) between the source-pairs for both the true and the estimated sources, and took their ratio. The distribution of the ratio is bimodal for HOCCA: While the recovery was very accurate in most cases, in some rare cases, the recovered sources were not dependent (local optima). For CCA, the distribution is unimodal: A large amount of the MI was recovered, but the recovered amount was usually smaller than for HOCCA. (b) The boxplot shows the difference between the fraction of total MI that HOCCA can recover per estimation problem and the fraction which CCA can recover. On average, HOCCA recovered more MI between the corresponding sources than CCA. Results for 50 random estimation problems are shown.
doi:10.1371/journal.pone.0086481.g006

Statistical approach to spatio-chromatic adaptation. Applying HOCCA, or one of the alternative methods considered, to the two sets of images produces two sets of coupled filters (sensors), each one adapted to one of the two lighting conditions. The filter outputs yield an internal representation of the images in terms of canonical coordinates, see Figure 3. There is a one-to-one correspondence between the canonical coordinates of each condition, and the corresponding coordinates are statistically dependent. The same one-to-one correspondence applies to the filters and features. The learned correspondence provides a model for spatio-chromatic adaptation: As the illumination changes, the filters should optimally change into their counterparts. The two corresponding filters may be considered to be instances of the same (hypothetical) physical sensor when adapted to the two different illuminants. The internal representation of an image can be adapted to changing lighting conditions by moving from one set of canonical coordinates to the other one.

Statistical properties of the learned representations. We analyzed the learned representations of natural images statistically using the same measure as for the artificial data. We used multi-information reduction and sparsity to assess the individual representation of each data set; to assess the coupling we used mutual information between the coupled canonical coordinates.

Figure 7(a) shows the amount by which multi-information was reduced by ICA, CCA, and HOCCA after whitening and dimensionality reduction. This means that we compared the reduction achieved by the different methods relative to the reduction achieved by whitening. The figure shows that ICA and HOCCA yielded similar results in multi-information reduction, with ICA being slightly better than HOCCA. Both methods led to a larger reduction than CCA. For CCA, we obtained negative values of multi-information reduction which means that it actually increased the statistical dependencies (redundancy) among the canonical coordinates.

Figure 7(b) shows the sparsity of the canonical coordinates. HOCCA led to a sparser representation than CCA or whitening, and to a slightly less sparse representation than ICA. With another measure of sparsity, robust kurtosis KR_2 due to J.J.A. Moors [23],

Table 1. Overview of the methods used to determine the matrices $\mathbf{Q}^{\mathrm{A}}$ and $\mathbf{Q}^{\mathrm{D}}$ in Figure 3.

Method	**Statistics used to determine $\mathbf{Q}^{\mathrm{A}}$ and $\mathbf{Q}^{\mathrm{D}}$**
HOCCA	(1) Sparsity of s_k^{A} and s_k^{D}
	(2) Correlation and variance dependencies between s_k^{A} and s_k^{D}
CCA	Correlation between s_k^{A} and s_k^{D}
ICA	Sparsity of s_k^{A} and s_k^{D} (correspondence determined by postprocessing)
Whitening by PCA	$\mathbf{Q}^{\mathrm{A}}$ and $\mathbf{Q}^{\mathrm{D}}$ are both the identity matrix (correspondence determined by postprocessing)

The variables s_k^{A} and s_k^{D} denote the canonical coordinates (feature outputs) of the representations. Higher-order canonical correlation analysis (HOCCA) generalizes canonical correlation analysis (CCA) in terms of the detected dependencies between the canonical coordinates. Moreover, it makes the canonical coordinates sparse which results in an efficient representation of the data. Independent component analysis (ICA) is maximizing the representation efficiency of the individual data sets without taking possible correspondences into account. Whitening by principal component analysis (PCA) is the first processing step in all methods. CCA and HOCCA yield coupled representations. For ICA and whitening, the correspondence between the filter outputs must be determined as part of a postprocessing step. We used mutual information maximization for the matching, see Materials and Methods for details.
doi:10.1371/journal.pone.0086481.t001

Table 2. Overview of our comparison of the learned coupled representations of natural images.

Criterion of comparison	Target property	Results
Independence and sparsity of the canonical coordinates	①	Figure 7
Mutual information between corresponding coordinates	②	Figure 8
Biological plausibility of the features	①	Figures 10, 11
Similarity of the coupled features	②	Figures 10, 11, Table 3
Psychophysics of corresponding colors	②	Figures 12, 13, Table 3
Noise-distortion curves	①+②	Figure 14

The representations learned by HOCCA, CCA, ICA, and whitening were compared from both statistical and biological points of view using multiple criteria. Two properties of the learned representations are of particular interest ① The individual representations of the two data sets, which is related to spatio-chromatic processing for a given illumination condition. ② The coupling (correspondence) between the two representations, which is related to adaptation to changes in the illumination. The different criteria measure different aspects of these two properties.

doi:10.1371/journal.pone.0086481.t002

we obtained similar results but HOCCA had higher values than ICA (results not shown). The two measures of efficiency shown in Figure 7 are consistent with each other: HOCCA resulted in a similarly efficient representation as ICA, and in a more efficient one than CCA, or whitening.

Figure 8(a) shows the mutual information between the coupled canonical coordinates. The correspondence learned by CCA and HOCCA resulted in coupled canonical coordinates which are more related to each other than the coupled coordinates obtained via ICA or whitening, as measured by mutual information. This suggests that learning the coupling and the features jointly leads to a stronger coupling than first learning the features and then selecting corresponding pairs by greedily maximizing mutual information.

Figure 8(b) shows a scatter plot of the learned correlation coefficient ρ_k and shape parameter v_k of HOCCA. According to the probabilistic model underlying HOCCA, the mutual information between the coupled canonical coordinates $(\mathbf{s}_k^{\mathrm{A}},\mathbf{s}_k^{\mathrm{D}})$ is a function of these parameters, see (13) in Materials and Methods and Figure 9. As the referenced equation and figure show, the two parameters contribute to the mutual information separately. The color of the markers in the figure indicates the value which the mutual information takes for each (ρ_k,v_k). This measurement of mutual information is based on the statistical model underlying HOCCA while in Figure 8(a), mutual information is measured in a nonparametric way. We found that the parametric and nonparametric measurements are consistent with each other (detailed analysis not shown). More importantly, most shape parameters v_k are between 2 and 2.5. With Figure 9, the shape parameters contribute around 0.4 bits to the mutual information, which corresponds to a correlation coefficient of about 0.65 for Gaussian variables. The values of v_k imply, first, that canonical coordinates for which ρ_k is close to zero are not statistically independent, and second, that their marginal distribution has heavier tails than a Gaussian. This is in line with the sparsity results shown in Figure 7(b).

Taken together, Figures 7 and 8 illustrate that HOCCA combines the desirable efficiency property of ICA with the desirable correspondence property of CCA.

The features of the learned representations. Figures 10 and 11 show the first 152 pairs of features which were learned with the different methods. For each pair, the upper feature is for CIE D65 illumination while the lower feature is for illumination CIE A. The feature-pairs are sorted by mutual information between the corresponding canonical coordinates: More related feature-pairs come first. The values of mutual information displayed in the

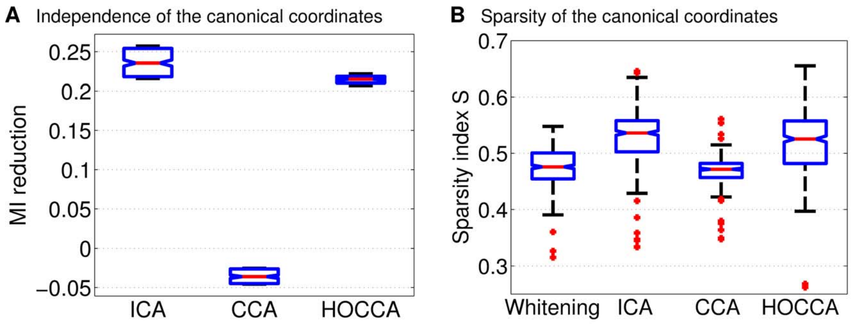


Figure 7. Analyzing the efficiency of the learned representations of natural images using independence and sparsity. (a) Independence was measured using reduction in multi-information (in bits per dimension, relative to whitened and dimensionality reduced data). The boxplot shows the distribution of the reduction for 100 pairs of bootstrapped data sets of size 150000. For reference, the multi-information reduction per dimension obtained by whitening without dimensionality reduction was 6.05 bits/dimension. Since dimensionality reduction introduces some information loss, the total reduction with regard to the pixel domain is not the sum of 6.05 bits/dimension plus the reductions reported in the figure. (b) Sparsity was measured using $\mathcal{S}$ in (15). A Gaussian has a value of $\mathcal{S}=0.2$. The boxplot shows the distribution of the sparsity of the 236 filters learned from natural images under illumination CIE A and CIE D65. The reported sparsity is the average value obtained for the above 100 bootstrapped data sets. The results for the CIE A and CIE D65 data are shown in the same boxplot. The figure suggest that HOCCA resulted in a similarly efficient representation as ICA, and in a more efficient one than CCA, or whitening.

doi:10.1371/journal.pone.0086481.g007

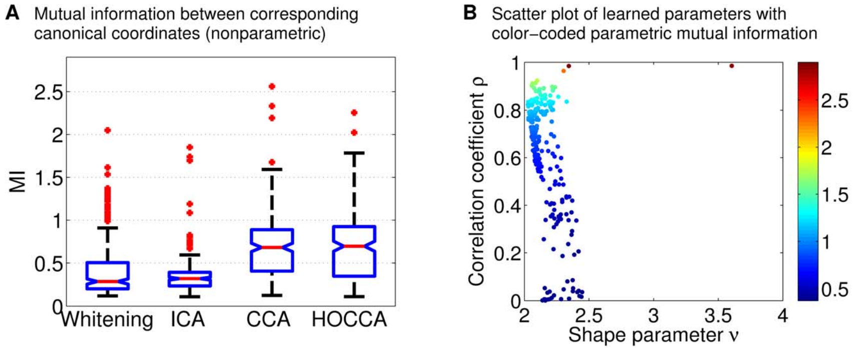


Figure 8. Analyzing the coupling of the learned representations of natural images using mutual information. (a) The nonparametric mutual information (MI) measurement was performed as for the artificial data, using CIE A and CIE D65 data sets of size 150000. (b) The parametric measurement was performed using (13), see Materials and Methods. The correspondence learned by CCA and HOCCA resulted in coupled canonical coordinates which are more related to each other than the coupled coordinates obtained via ICA or whitening.
doi:10.1371/journal.pone.0086481.g008

figures indicate the range for the filters in each row. We first analyze the features per data set. Then, we analyze the coupling between the features.

Regarding the features per data set, we use the finding that neurons in V1 are dominantly tuned to spatially localized oriented Gabor-like features with achromatic, red-green and yellow-blue chromatic content as plausibility baseline [24–26]. For all methods, the features learned from images under illumination CIE A have oscillations around a yellowish mean, which is reasonable given the yellowish illumination. For the images under illumination CIE D65, the learned features have achromatic averages. Whitening yielded spatially extended gratings of different orientation, frequency and opponent chromatic content, similar to a discrete cosine transform. CCA yielded some low-frequency features, the remaining features show non-localized high frequency oscillations, and have a quite undefined spatial structure. ICA and HOCCA yielded localized Gabor-like features of different orientation, frequency and opponent chromatic content.

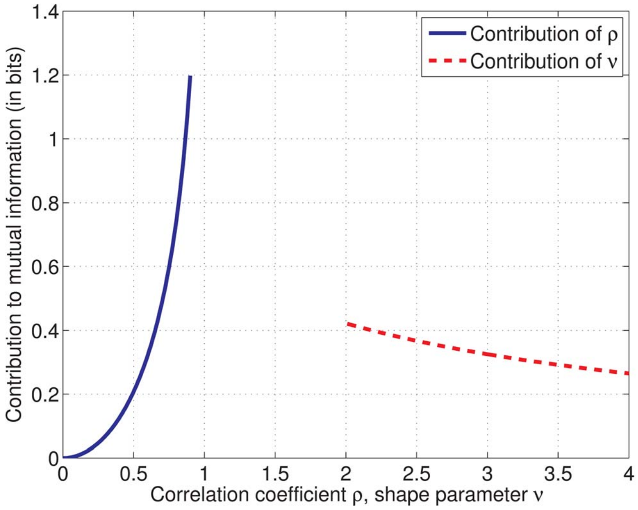


Figure 9. Mutual information for a bivariate student's t-distribution. The correlation coefficient $\rho \in (-1\ 1)$ and the shape parameter $\nu > 2$ contribute separately to the mutual information, see (13). The contribution of ρ is symmetric around zero and shown in blue for $\rho \geq 0$ (solid curve), the contribution of ν is shown in red (dashed curve). The mutual information of the bivariate student's t-distribution is given by the sum of the two contributions. The contribution of ν reflects the higher-order statistical dependencies between the two random variables.
doi:10.1371/journal.pone.0086481.g009

Visual inspection of the features shows that for whitening and CCA, the achromatic and the chromatic oscillations in the high frequency features do often not match spatially. They have different fundamental frequencies. For ICA and HOCCA, however, there is no such mismatch between achromatic and chromatic parts. Further, the ICA and HOCCA filters seem chromatically less saturated than the whitening and CCA filters. This means that in order to elicit a comparable response, ICA and HOCCA filters would require a stronger amplitude for chromatic than for achromatic gratings.

Regarding the coupling, the sorting according to mutual information shows that for ICA, HOCCA and CCA, low-frequency features are more related than high-frequency ones. Further, for non-zero frequencies, the achromatic features are more related than the chromatic ones.

We analyzed the similarity of the corresponding features, using the mean squared error as distance measure. Direct application of this distance would, however, be strongly biased by the chromatic shift due to the different illuminations. Therefore, Von-Kries color compensation [27] was applied to the features of illumination CIE A before computing the mean squared error. The resulting spatio-chromatic distances for the different learning methods are shown in Table 3 (first row). HOCCA yielded feature-pairs which are more similar to each other than the other methods. The same result was also obtained using other color compensations than Von-Kries before computation of the distance, such as CIELab [27].

Reproducing psychophysics of corresponding colors. We further investigated the learned coupling by assessing the ability of the different representations to reproduce psychophysical data on color corresponding pairs (color constancy). In the color psychophysics literature, physically different stimuli are referred to as corresponding if they give rise to the same perceived color when viewed under different conditions [28–30]. Corresponding colors illustrate the (purely) chromatic adaptation ability of the human visual system and form a standard benchmark for chromatic adaptation models, see, for example, [21].

We show in Figure 12 the experimentally corresponding colors [30]. Figure 13, left column, shows the same corresponding colors in the CIE xy chromaticity diagram. Each point in the lower and upper diagram denotes one color in Figure 12(a) for illumination A

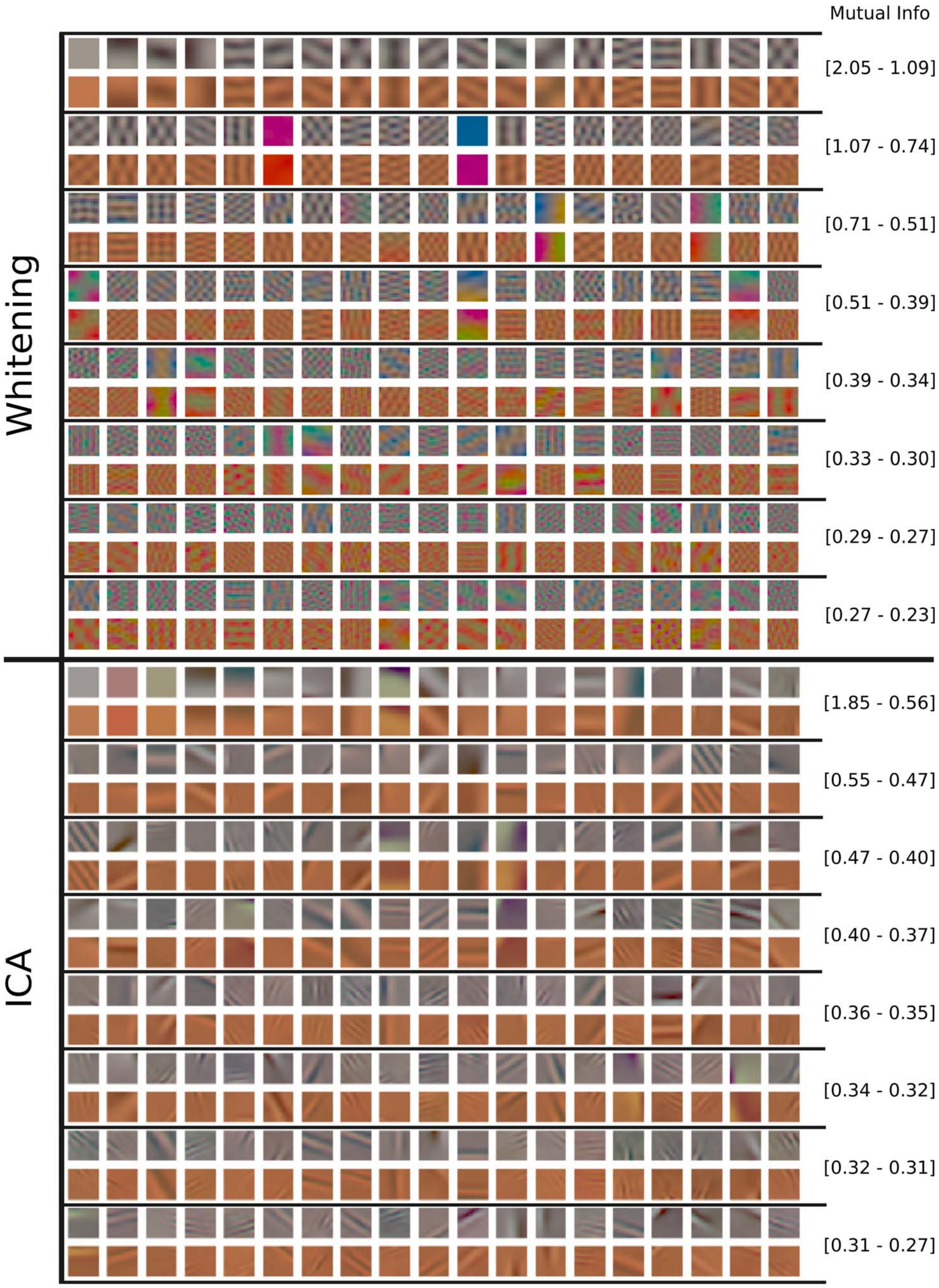


Figure 10. Features learned by whitening and ICA from natural images. After learning, the features from the two data sets were matched so that the mutual information between the corresponding canonical coordinates is maximized, see Materials and Methods for details. In each row, the upper feature is for CIE D65 illumination while the lower feature is the corresponding one for illumination CIE A. Only the first 152 feature-pairs are shown. The feature-pairs are sorted by mutual information between the canonical coordinates. The numbers on the right indicate the range of the mutual information (in bits) for the feature-pairs in each row. ICA features are biologically plausible while whitening features are not.
doi:10.1371/journal.pone.0086481.g010

and in Figure 12(b) for illumination D65, respectively. In this (standard) visualization, the correspondence between the colors is not made explicit. Qualitative comparisons of different chromatic adaptation models are based on the arrangement of the points in the diagram [21].

Figure 13, right column, shows (linear) predictions of the color-corresponding pairs from the learned representations of the data, performed as described in Materials and Methods. The top row shows the predictions for illumination D65 obtained from the sample colors under illumination A, the bottom row shows the predictions for illumination A obtained from the samples under illumination D65.

A qualitative comparison of these predictions with the experimental data in the left column shows that HOCCA and CCA led to a better performance than whitening or ICA-based correspondence methods: For whitening, the arrangement of the points is rather different from the experimental data. For ICA, the predictions are often over-saturated such that many of the predicted colors fall outside the chromaticity diagram, which

Table 3. Quantification of the results in Figures 10 to 13.

	Whitening	CCA	ICA	HOCCA
Similarity of coupled features (RMSE)	18.7±4.2	17.5±3.6	17.7±4.6	15.7±4.4
Corresponding colors (relative RMSE)	0.29±0.91	0.046±0.13	0.18±0.60	–

The numbers indicate the (relative) root mean squared error (RMSE, average ± std). First row: Spatio-chromatic similarity between the coupled features in Figures 10 and 11 after Von-Kries color compensation. On average, HOCCA yielded a smaller RMSE than the other methods (two-sample t-test, largest p-value $<9{\cdot}10^{-7}$). Second row: Prediction error for the color-corresponding pairs in Figures 12 and 13. The RMSE of the different methods is computed relative to the error of HOCCA. A positive relative difference indicates that the alternative method has a larger error than HOCCA. On average, the relative difference is positive for all alternative methods, and significantly larger than zero (one-sided t-test, largest p-value was 0.0030).

doi:10.1371/journal.pone.0086481.t003

means that they are physically unrealistic. For HOCCA and CCA, however, this was much more rarely the case.

A quantitative comparison was performed by computing the root mean squared error between the predicted and the experimental colors in the XYZ color-space. The results are given in Table 3 (second row). Averaging over all color-points, we found that HOCCA gave the best results, followed by CCA. In more detail, we assessed for each color how well the different methods are doing relative to HOCCA. A positive relative difference indicates that the alternative method has a larger root mean squared error. On average, the relative difference was found to be significantly larger than zero for all alternative methods.

Adaptation with neural noise constraints. In the previous sections, we used different criteria to analyze the learned representations per data set and the coupling across the data sets. The criteria used assessed the two aspects of the learned representations separately. Here, we consider both aspects at the same time.

The learned filters map the images $\mathbf{x}^A$ and $\mathbf{x}^D$ into a canonical domain where they are represented by the coordinates $\mathbf{s}^A$ and $\mathbf{s}^D$. This transformation was considered to be free of noise. Real systems, however, are intrinsically noisy and the noise-level may depend on the signal. A measure of noisiness is the Fano factor F which is the variance of the noise divided by its mean, see Chapter 1 of [31]. In alert Macaque monkeys, the Fano factor in V1 was found to be less than one for optimal stimuli (with an average value of 0.33), and around one for stimuli close to the visual threshold [32].

We investigated how the different representations perform in an adaptation task in the presence of neural noise: We compared the different methods in their ability to predict the representation of an image under illumination CIE D65 from its representation under CIE A. The mean squared prediction error is derived in Materials and Methods and Text S3. It equals

$$\mathrm{E}\left(\|\mathbf{s}^D-\hat{\mathbf{s}}^D\|^2\right)+\mathrm{F}\sum_k|\varrho_k|\mathrm{E}\left(|s_k^A|\right), \tag{7}$$

where E denotes expectation, and where $\hat{\mathbf{s}}^D$ is the prediction when there is no neural noise. The first term is the squared noise-free prediction error. The second term is a weighted sum of sparsity penalties $\mathrm{E}(|s_k^A|)$. The weighting depends on the Fano factor (noise-level) F and the correlation coefficients ϱ_k between the canonical coordinates s_k^A and s_k^D. For HOCCA, $\varrho_k=\rho_k$. If s_k^A is sparse, the sparsity penalty is small. For F close to zero, the noise-free prediction error dominates but as F increases, sparsity becomes relevant. For a small overall prediction error, the representations should be sparse and have a good correspondence (coupling).

Figure 14 shows the root mean squared error (RMSE) of the prediction for the different methods as F varies (noise-distortion curves). For $F<0.1$, CCA gives the smallest error, which is reasonable because it minimizes the noise-free prediction error. For larger F, the importance of sparsity becomes visible. In a sparse representation, the introduced neural noise is lower on average. HOCCA has the smallest error from $F\approx0.1$ to $F\approx1.5$ because it combines sparsity with good coupling. ICA is better than CCA for $F>0.7$, where the sparsity penalty starts to offset the noise-free prediction error. For $F>1.5$, ICA yields the smallest error among all methods since its representation is the sparsest one. However, this regime of F does not seem realistic for neurons in V1 [32]. Since the difference between CCA and HOCCA is rather small for $F<0.1$, we conclude that HOCCA compares favorably to the other methods in the relevant regime of F: It combines good prediction accuracy with robustness to noise.

HOCCA in comparison to the alternative methods. We analyzed the learned representations of HOCCA, ICA, CCA, and whitening from both statistical and biological points of view using multiple criteria, see Table 1 for an overview. The different points of view yielded the same picture: While ICA performed well with regard to the individual representations (assessed by independence, sparsity, and plausibility of features), and CCA well with regard to correspondence (assessed by mutual information, similarity of features, and color psychophysics), HOCCA performed well in both aspects. The noise-distortion curves exemplified this favorable performance of HOCCA.

Predicting response-adaptation for spatio-chromatic inputs. The previous sections showed that HOCCA provides a single (unified) statistical framework to study both efficient representations and adaptation. In this section, we use HOCCA to make a testable prediction about the response of human spatio-chromatic sensors (neurons) to colored patterns under change of illumination.

HOCCA produced pairs of filters optimized for illumination CIE A and D65. Considering the two corresponding filters to be instances of the same (hypothetical) physical sensor when adapted to two different illuminants, we can investigate how adaptation changes the response to the same stimulus.

For the prediction, we used six representative HOCCA filter-pairs, three pairs with chromatic content in the red-green (RG) direction (feature-pairs 109, 134 and 152 with red frames in Figure 11) and three in the yellow-blue (YB) direction (feature-pairs 18, 67 and 78 with blue frames in Figure 11). With this choice, we consider filters of different spatial frequency and orientation for each chromatic content.

For each sensor considered, we determined its optimal stimulus under illumination D65 and changed its chromatic contrast and its color content through rotations in the RG-YB plane, as done in [24] and [33], see Materials and Methods. Figures 15 and 16 show the obtained stimuli for the RG and YB filters, respectively. We used these stimuli both for the sensors adapted to illumination D65 and for the sensors adapted to illumination A. This allowed us to make a prediction of what should happen to the response to the same colored pattern when a (biological) sensor is adapted to illumination A instead of D65. To the best of our knowledge, there are no such measurements in the experimental literature.

Figure 17 shows the average response of the considered RG and YB filters to the spatio-chromatic stimuli in Figures 15 and 16.

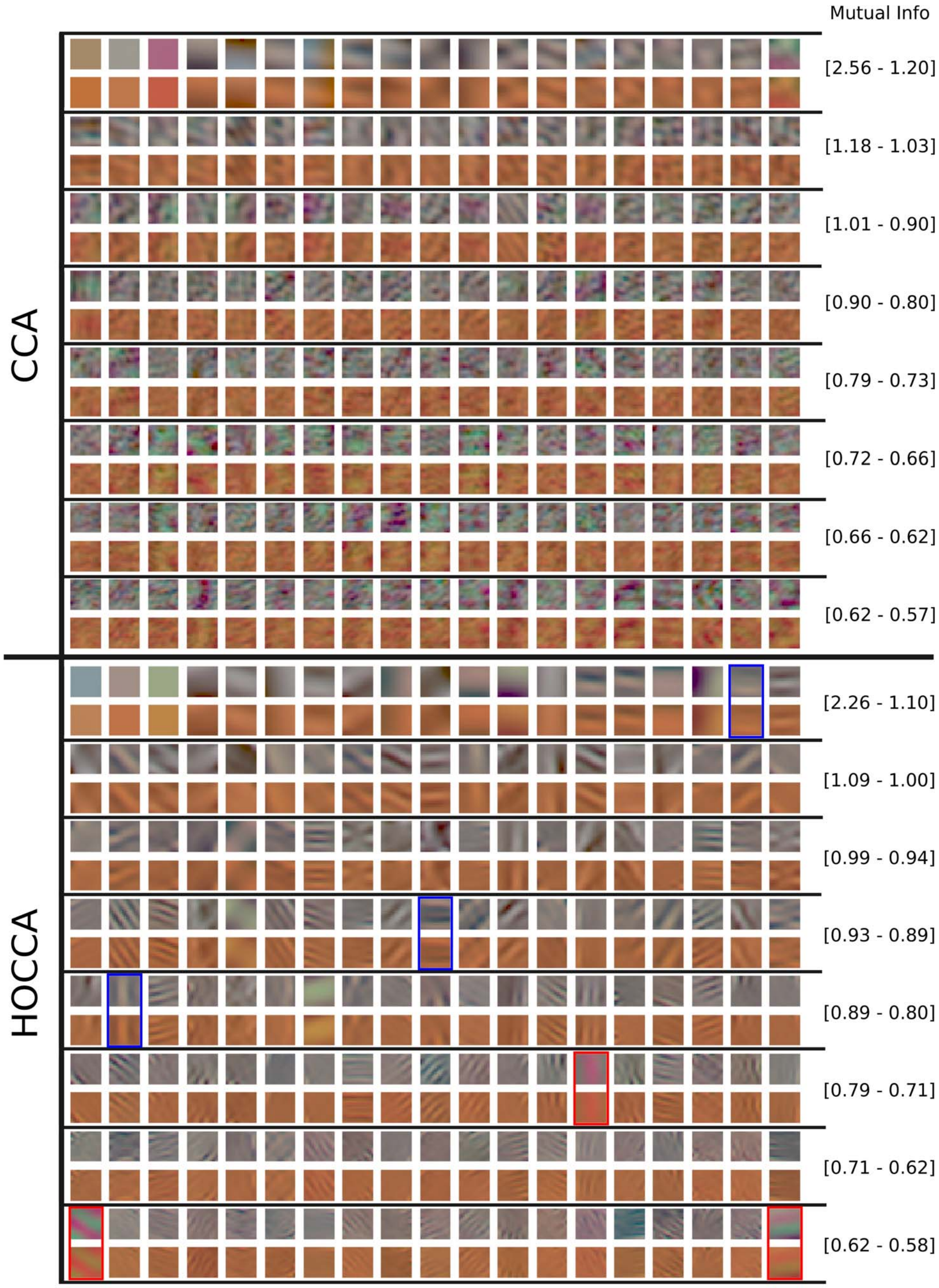


Figure 11. Features learned by CCA and HOCCA from natural images. Only the first 152 feature-pairs are shown. The numbers on the right indicate the range of the mutual information (in bits) for the features in each row. The feature-pairs are arranged as in Figure 10. HOCCA features are biologically plausible while CCA features are not. The pairs of HOCCA features marked with red and blue frames are used to make a prediction about response-adaptation for spatio-chromatic inputs.
doi:10.1371/journal.pone.0086481.g011

Solid curves show the response of the sensor adapted to illumination D65, dashed curves the response when adapted to illumination A. HOCCA predicts sinusoidal oscillations as a function of the rotation angle of the chromatic modulation in the RG-YB plane. By definition of the stimuli used, the maximal responses are obtained for D65 illumination. The linear behavior implies linear reduction of the oscillation as the chromatic contrast decreases to zero. The response curves have an offset. This is due to the presence of an achromatic modulation in the stimuli, the filters learned by HOCCA are not purely chromatic. Interestingly, the solid and dashed curves do not have their optimum at the same angle. The optimal stimulus of a sensor adapted to illumination D65 is no longer optimal when the sensor is adapted to illumination A: We predict a shift in the responses as adaptation to the new illumination occurs.

Discussion

We reported two sets of results in this paper. First, we proposed a new statistical method, called higher-order canonical correlation analysis (HOCCA), to jointly analyze multiple data sets. HOCCA combines desirable properties of canonical correlation analysis

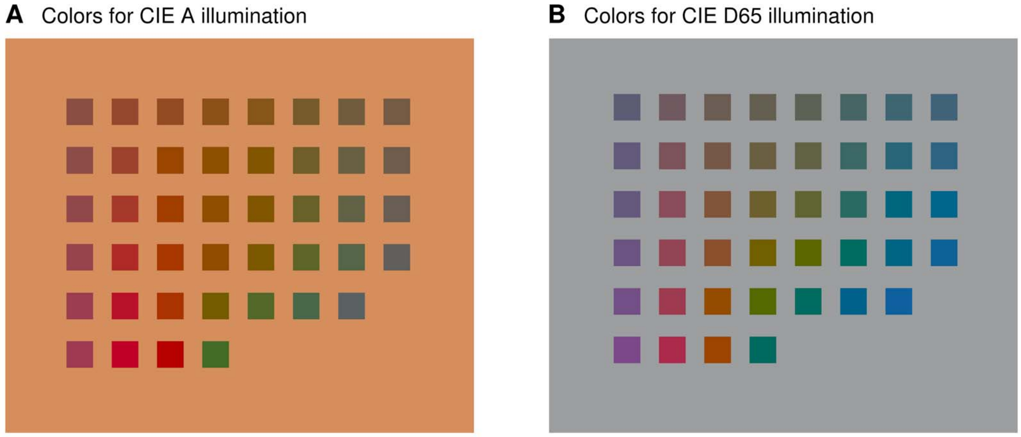

A Colors for CIE A illumination
B Colors for CIE D65 illumination

Figure 12. Corresponding-colors psychophysics. For humans, the color of a patch in (a), when seen under CIE A illumination, appears to be the same as the color of the patch in (b) at the same location on the grid, when seen under CIE D65 illumination. Two colors which give rise to the same perception under two different viewing conditions are said to be corresponding. The experimental findings visualized in the figure are due to [30].
doi:10.1371/journal.pone.0086481.g012

(CCA) and independent component analysis (ICA). HOCCA seeks independent and sparse sources inside each data set which have linear or variance correlations across the data sets. HOCCA is as widely applicable as CCA. Moreover, it generalizes CCA because it is not only sensitive to linear correlations but also to higher-order dependencies. We validated HOCCA on artificial data and proved that CCA emerges as a special case.

Second, we showed that HOCCA provides a single (unified) statistical framework to study visual processing under fixed lighting conditions and adaptation to new ones. Results on chromatic natural images demonstrated the benefits of jointly maximizing efficiency of representation and coupling across the data sets, as opposed to first maximizing efficiency and then finding a suitable coupling, or focusing on coupling only. We found that HOCCA features are consistent with the spatio-chromatic tuning properties of neurons in the primary visual cortex and that HOCCA reproduces corresponding colors psychophysics reasonably well. HOCCA provided us with a specific, experimentally testable prediction on how the response to colored patterns should change when the illumination changes.

Relation to other statistical methods

We showed that HOCCA provides a generalization of CCA. CCA has been extended in many ways. Kernel CCA is a nonlinear extension of CCA that is sensitive to nonlinear dependencies across the data sets, see Section 3.2 of [34] and [35,36]. One difference to our work is that kernel CCA does not yield an efficient representation of the data in terms of sparse canonical coordinates. Sparsity was incorporated in CCA [37,38] but this was done on the level of the features and not on the level of the canonical coordinates as we do here.

CCA was also combined with ICA [39]. In that work, however, ICA serves more as a preprocessing step, and after the ICA rotation, the independent components are subject to a further

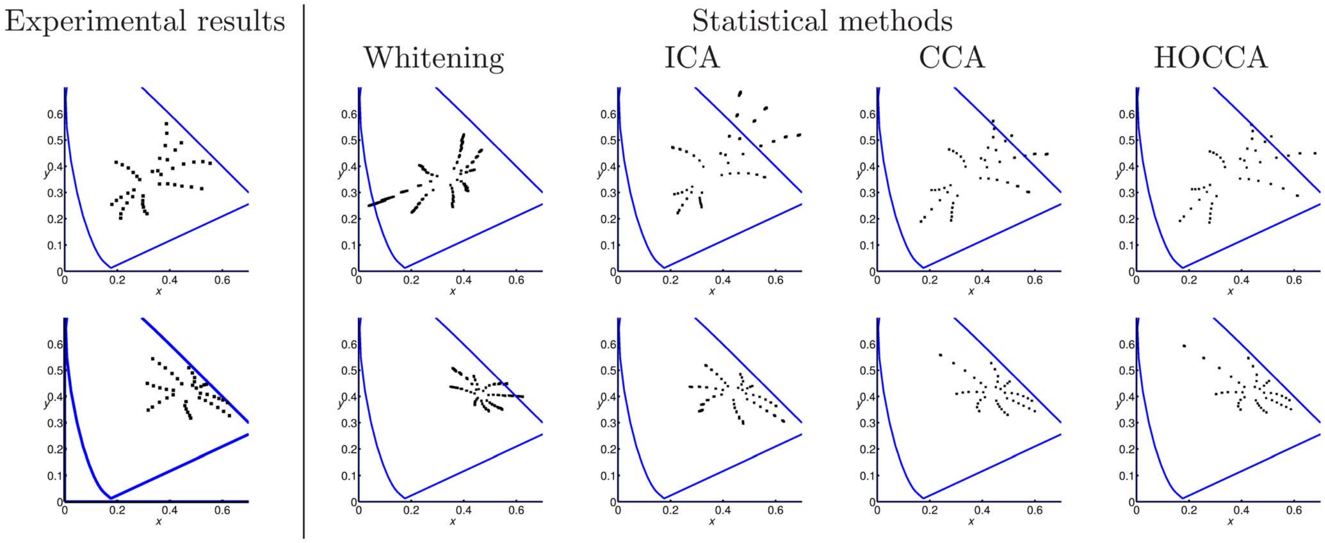


Figure 13. Using the learned representations to reproduce corresponding-colors psychophysics. Left: Experimentally corresponding colors of Figure 12 in the CIE xy diagram. Right, top row: Predictions of the corresponding colors under illumination D65 from samples under illumination A. Right, bottom row: Predictions of the corresponding colors under illumination A from samples under illumination D65.
doi:10.1371/journal.pone.0086481.g013

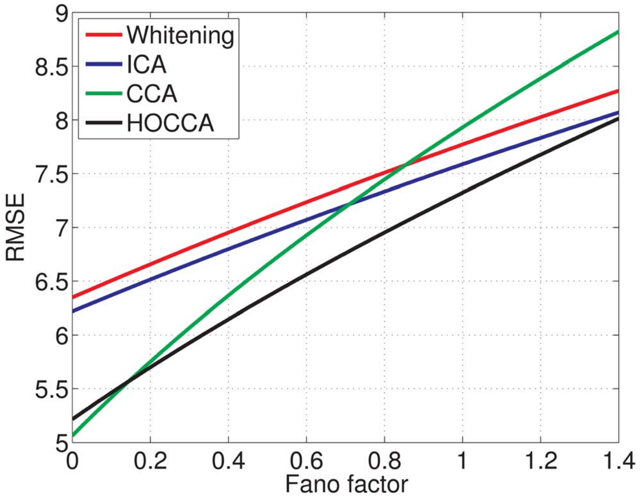


Figure 14. Noise-distortion curves. We compared the different methods in their ability to predict the representation of an image in daylight (CIE D65) from an image under yellowish light (CIE A) in the presence of neural noise. The curves show the root mean squared error (RMSE) of the predicted representation as the Fano factor (neural noise level) F varies. The Fano factor in V1 is typically less than one, on average around $F=0.33$ [32]. Representations which are sparse are less affected by the noise, representations which have a good correspondence give a low error for zero noise. HOCCA compares favorably to the other methods in the relevant regime of F because it combines sparsity with good correspondence.
doi:10.1371/journal.pone.0086481.g014

rotation to maximize the nonlinear correlation across the data sets. In our context, such a rotation would, however, be suboptimal since rotating sparse independent components decreases their sparsity. In very recent work [40], the authors reversed the order of analysis (first analysis across the data sets, then finding independent sources within each data set). In our context, such an approach would, however, also be suboptimal since it does not seem to yield coupled canonical coordinates but only coupled subspaces.

In ICA, there is an ambiguity in the ordering of the column vectors of the mixing matrix. The joint estimation of $\mathbf{Q}^{\mathrm{A}}$ and $\mathbf{Q}^{\mathrm{D}}$ in HOCCA reduces this ambiguity because the ordering in both matrices must be the same: Due to the correspondence between canonical coordinates across the data sets, the ordering for one matrix cannot be changed without changing the ordering of the other in the same way.

In our simulations on natural images, we used a postprocessing stage after ICA where the mixing matrices $\mathbf{Q}^{\mathrm{A}}$ and $\mathbf{Q}^{\mathrm{D}}$ are reordered to obtain a correspondence. For natural image data, HOCCA was found to yield better results than this simple strategy. For other data, however, in particular if the individual data sets follow ICA models exactly, simple postprocessing of individual ICA results may work very well.

Coupled representations learned from natural images

We applied HOCCA to two sets of images acquired with different illuminants, namely daylight-like CIE D65 and yellowish CIE A. HOCCA produced two sets of coupled filters, each one adapted to one of the illuminants. We compared HOCCA with three other statistical methods: Whitening by principal component analysis, ICA, and CCA. For HOCCA and CCA, the filters are learned together with their correspondence. For whitening and ICA, however, the filters are learned separately for each data set. We sought a correspondence after learning of the filters by finding pairs which had maximal mutual information.

Regarding the representations per data set, the mutual information reduction achieved by ICA and HOCCA is consistent with previously reported reductions for ICA in achromatic images [41,42]. The filters learned by whitening and ICA are in line with previously reported results [13,43,44]. Further, our finding that ICA and HOCCA filters are less sensitive to chromatic than to achromatic gratings is consistent with sensitivity results in human vision [45].

Regarding correspondence, we found that, as HOCCA, CCA yielded a large amount of mutual information between the corresponding canonical coordinates even though CCA is only sensitive to linear correlations. The reason for this is two-fold: First, linear correlations contribute strongly to mutual information, see Figure 9, and CCA finds canonical coordinates which are maximally correlated. Second, even though CCA is only sensitive to linear correlations, this does not mean that the canonical coordinated obtained by CCA are Gaussian. In fact, Figure 7 shows that the marginals of the canonical coordinates of CCA are sparser than Gaussian random variables. This non-Gaussianity also contributes to mutual information.

Corresponding colors have been inferred from properties of natural images before [21]. The approach in the cited paper differs from the one in this paper in two main aspects: First, spatial information was not taken into account. Only the properties of the tristimulus pixel values were modeled. Second, the prediction of the corresponding colors was nonlinear. Compared to the linear methods used in this paper, nonlinear prediction is better suited to keep colors inside the chromatic diagram. Perceptually, this means that the nonlinear method avoids over-saturation of the predicted colors. Inspection of the predicted points in the chromatic diagram shows that the hues of the prediction, however, correspond better to the experimental data for HOCCA than for the nonlinear method.

In the joint learning of the features and the correspondence between them, HOCCA (and CCA) had access to the same images under two different illuminations. That is, the input data came labeled in terms of the illuminant (we used the superscripts $^{\mathrm{A}}$ and $^{\mathrm{D}}$ for the labeling). Furthermore, the objective function optimized in HOCCA consists of a sample average over several such observations. The visual system, however, is exposed to only one scene under one illumination at a time. While a sample average can be computed in an online fashion, assuming that the visual system has access to labeled input data is more problematic. However, information about the labels is often implicitly available and the labels can be inferred from it. For the inference of the labels, it is enough that the brain "knows" that an object under either of the two illuminations is the same (kind of) object. Such information about the identity of an object could be provided by top-down processes. For instance, when leaving a house with a red apple in the hand, the brain "knows" that the apple was not switched out but that the same object is in the hand both inside and outside, even though the radiance sensed by the eyes is different. This means that HOCCA should not be considered a mechanistic model of visual processing and adaptation; its neural implementation is left unspecified. Instead, HOCCA should be considered a normative theory based on statistical principles. It tells us what we can expect if efficiency and correspondence are optimized jointly, in case the same images under two different illuminations were available.

Adaptation to changes in illumination is related to illuminant compensation or color constancy. To compensate for the illuminant, additional measurements can be used, like measure-

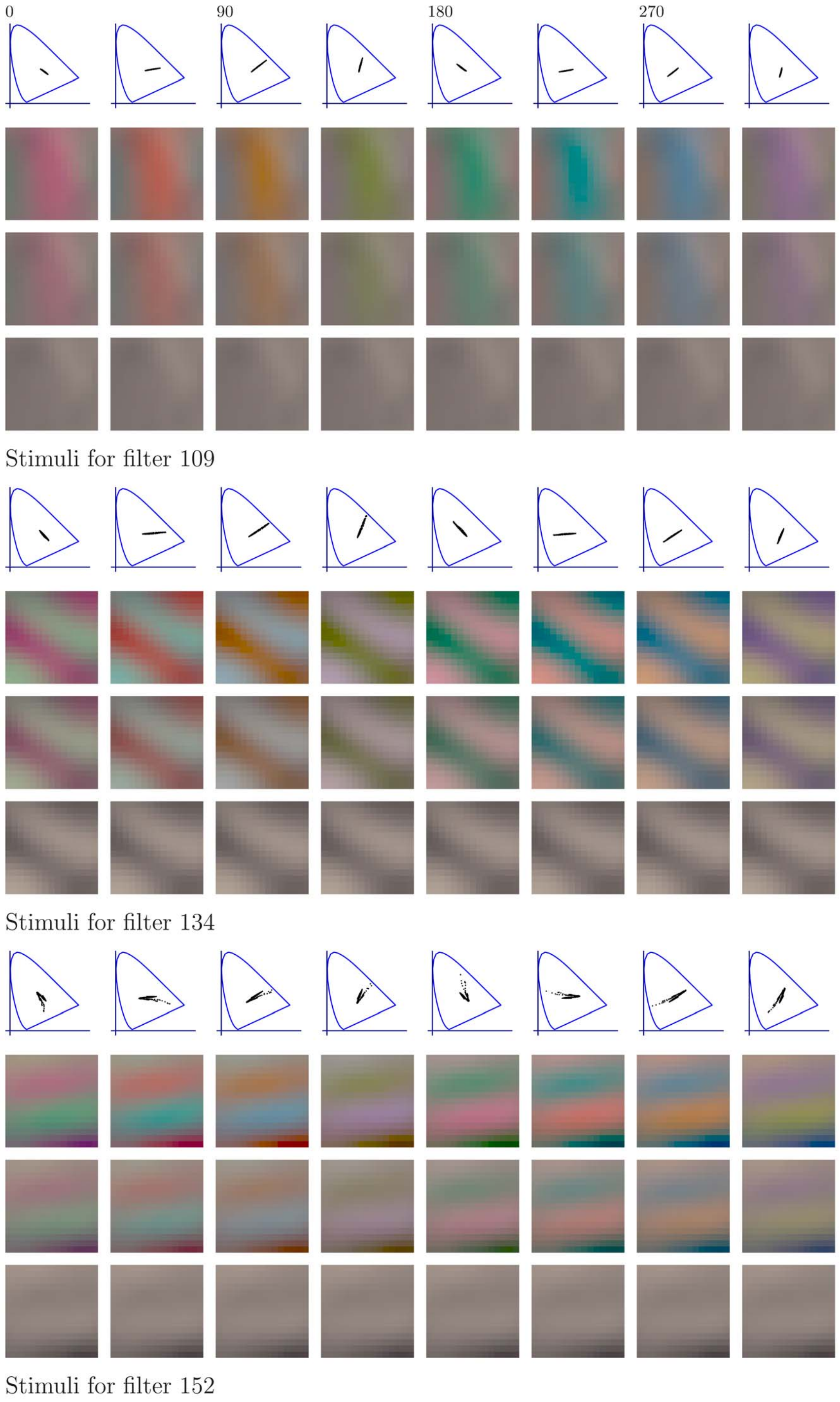
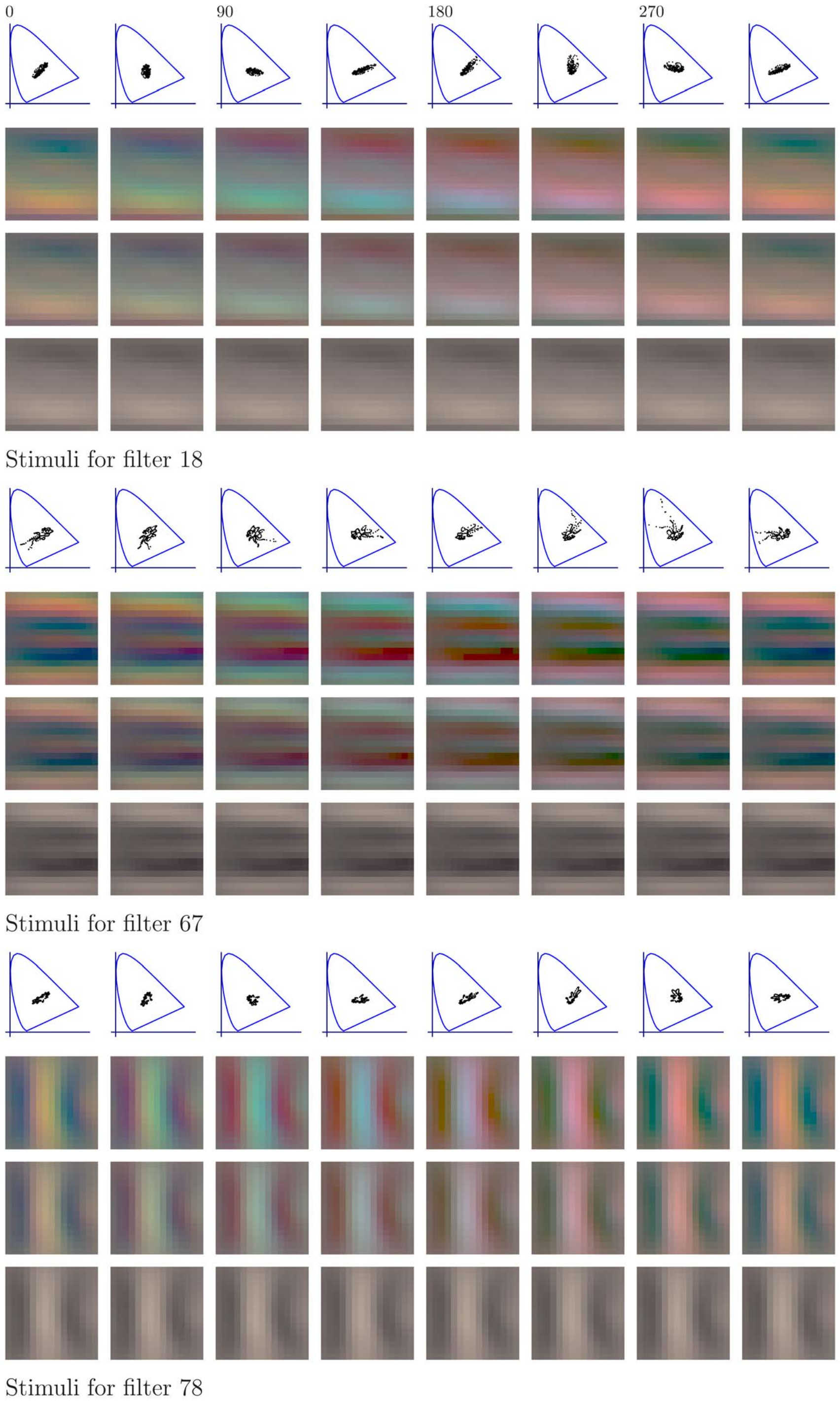


Figure 15. Stimuli used to predict the response-adaptation of RG sensors. Each 4-row panel corresponds to the stimuli used for one sensor. For each sensor, the stimuli were obtained by rotating and scaling the chromatic part of the optimal stimulus (the top left image in each panel). The color content changes in constant steps from left to right, and the scaling factor varies linearly from top to bottom, see Materials and Methods for details. The top row in each panel shows the chromatic diagrams for the first row of images.
doi:10.1371/journal.pone.0086481.g015

ments from a white object in the surround [27,46,47], or measurements from a wide ensemble of neighboring surfaces [48–51]. Another approach consists in mapping illuminant-dependent images to a domain which is illuminant independent [17–19,21,52]. The mappings can be seen as transforms which, like HOCCA, take into account the different statistical properties of the images in the different acquisition conditions.

HOCCA allowed us to make a testable prediction about the response to spatio-chromatic stimuli when adapted to CIE D65 or CIE A illumination. The prediction can be thought to correspond to the best-case scenario where labeled data has shaped the



Stimuli for filter 18

Stimuli for filter 67

Stimuli for filter 78

Figure 16. Stimuli used to predict the response-adaptation of YB sensors. The stimuli were generated as those in Figure 15.
doi:10.1371/journal.pone.0086481.g016

properties of the neurons. Next, we discuss the relation of our prediction to the experiments performed in [24] and [33]. In [24], responses to patterns with chromatic modulation in rotated directions of the red-green (RG), yellow-blue (YB) plane using a fixed white adaptation point similar to D65 were measured. The responses were found to oscillate sinusoidally as the stimulus rotated in the RG-YB plane. In [33], similar measurements were used to investigate the effect of habituation to high chromatic contrast stimuli modulated in certain directions of the color space. Again, a D65-like white average was used. In the control situation of zero contrast habituation stimuli, sinusoidal responses as in the aforementioned results were obtained. For non-zero habituation, these oscillations were found to shift and scale depending on the presence of linear or non-linear interactions between the basic RG-YB sensors.

Adaptation to illumination CIE D65 or CIE A is not exactly habituation to high chromatic contrast stimuli. Moreover, the linear nature of our filtering (computation of the canonical

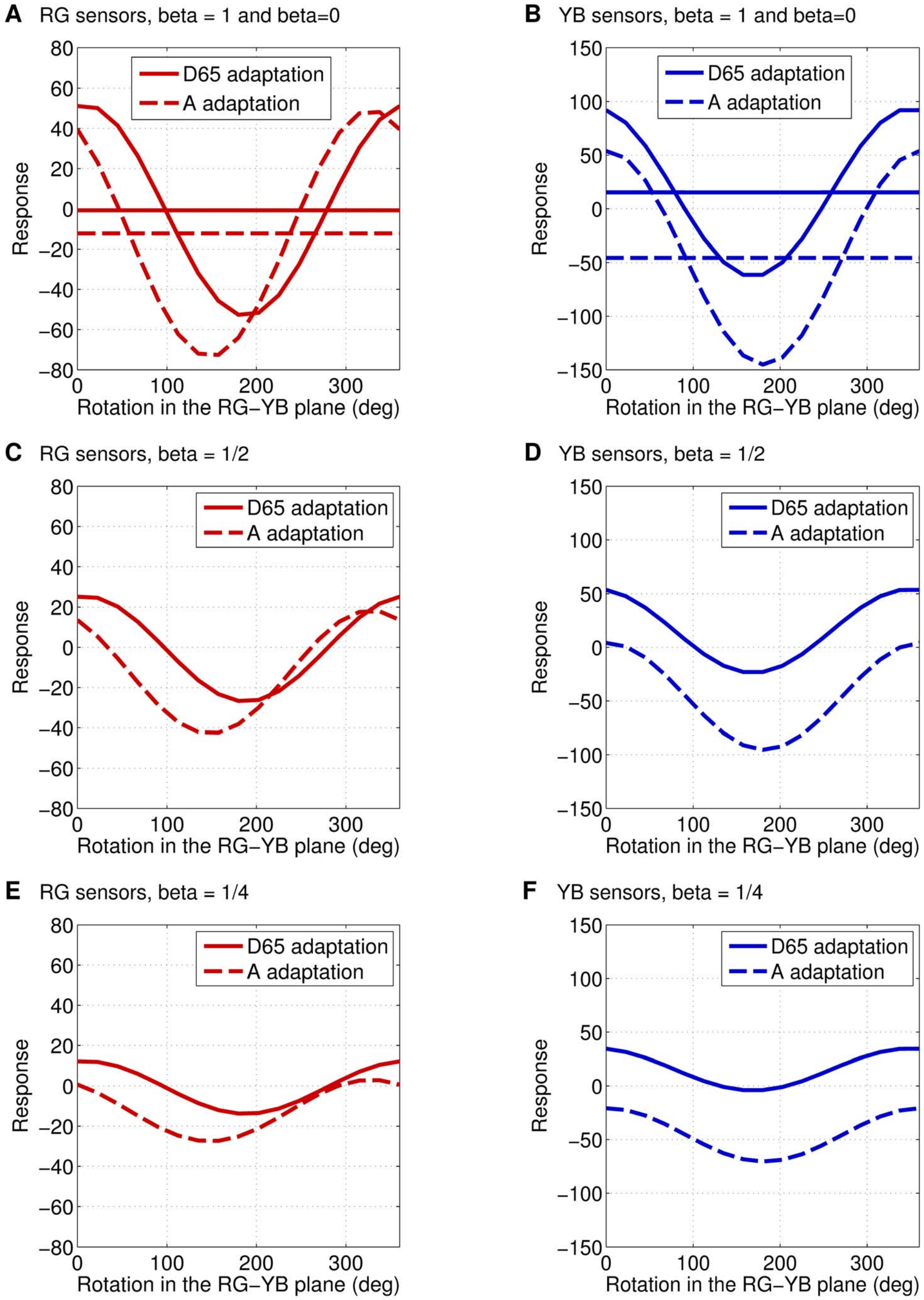


Figure 17. A testable prediction about response-adaptation for spatio-chromatic inputs. The figures show the average response of RG sensors and YB sensors when stimulated with the stimuli in Figures 15 and 16, respectively. Solid lines display responses of sensors adapted to CIE D65 illumination while dashed lines indicate adaptation to illumination CIE A. The constant curves in (a) and (b) are obtained for $\beta = 0$. The optimal stimulus of a sensor adapted to illumination D65 is no longer optimal when the sensor is adapted to illumination A. We predict a shift in the response as adaptation to the new illumination occurs.
doi:10.1371/journal.pone.0086481.g017

coordinates) cannot reproduce effects from eventual non-linear interactions. Therefore, our adaptation predictions cannot straightforwardly be compared to the habituation results reported in [33]. Nevertheless, we can notice interesting connections: First, both in our setup and in the aforementioned experimental work, smooth oscillations of the responses are obtained when the chromatic content of the optimal stimuli is changed, see Figure 16. Second, the offset of the curves is also similar to the reported experimental behavior. Third, the shifts in the responses which we predict as adaptation to the changed illumination occurs, are qualitatively similar to the shifts reported for contrast habituation.

Materials and Methods

Probabilistic generative model of HOCCA

We construct here HOCCA such that it takes higher-order statistical dependencies both within and across the data sets into account, in contrast to CCA. This allows us to find a both related and efficient representation of the data. The new method is based on a probabilistic generative model which we outline next. In Text S1, the model is generalized to the case where there are more than two data sets, each possibly of different dimensionality.

In order to find an efficient representation for each of the two data sets, we assume that each of the two vectors of canonical coordinates $\mathbf{s}^{\mathrm{A}}$ and $\mathbf{s}^{\mathrm{D}}$ in (1) consists of statistically independent sparse random variables. The independence assumption concerns the elements within each vector only. In order to find features that are related across the data sets, we assume that the k-th random variable of $\mathbf{s}^{\mathrm{A}}$ and the k-th random variable of $\mathbf{s}^{\mathrm{D}}$ are statistically dependent.

The independence assumption for the canonical coordinates within a data set makes the whitened data $\mathbf{z}^{\mathrm{A}}$ and $\mathbf{z}^{\mathrm{D}}$ in (1) follow an ICA model with mixing matrices $\mathbf{Q}^{\mathrm{A}}$ and $\mathbf{Q}^{\mathrm{D}}$. In this context, we call the canonical coordinates also sources.

Let $\mathbf{s}_k = (s_k^{\mathrm{A}} \;\; s_k^{\mathrm{D}})^{\mathrm{T}}$ denote the column vector which contains the k-th canonical coordinate (source) from both data sets. With the above independence assumptions, the joint probability density function (pdf) of all the sources $\mathbf{s} = (\mathbf{s}^{\mathrm{A}}; \mathbf{s}^{\mathrm{D}})$ factorizes into m factors,

$$p_{\mathbf{s}}(\mathbf{s}) = \prod_{k=1}^{m} p_{\mathbf{s}_k}(\mathbf{s}_k), \tag{8}$$

where $p_{\mathbf{s}_k}$ denotes the pdf of $\mathbf{s}_k$. With the ICA models in (0) and the orthogonality of the mixing matrices, the joint pdf $p_{\mathbf{z}}$ of the random variables $\mathbf{z}^{\mathrm{A}}$ and $\mathbf{z}^{\mathrm{D}}$ is

$$p_{\mathbf{z}}(\mathbf{z}^{\mathrm{A}}, \mathbf{z}^{\mathrm{D}}) = \prod_{k=1}^{m} p_{\mathbf{s}_k}(\langle \mathbf{q}_k^{\mathrm{A}}, \mathbf{z}^{\mathrm{A}} \rangle, \langle \mathbf{q}_k^{\mathrm{D}}, \mathbf{z}^{\mathrm{D}} \rangle), \tag{9}$$

where $\langle \mathbf{a}, \mathbf{b} \rangle$ denotes the inner product between the two vectors $\mathbf{a}$ and $\mathbf{b}$. If $p_{\mathbf{s}_k}$ was known, $p_{\mathbf{z}}$ would be properly defined. We then could maximize the (rescaled) log-likelihood ℓ,

$$\ell = \hat{\mathrm{E}} \left\{ \sum_{k=1}^{m} \log p_{\mathbf{s}_k}(\langle \mathbf{q}_k^{\mathrm{A}}, \mathbf{z}^{\mathrm{A}} \rangle, \langle \mathbf{q}_k^{\mathrm{D}}, \mathbf{z}^{\mathrm{D}} \rangle) \right\}, \tag{10}$$

to estimate the features $\mathbf{q}_k^{\mathrm{A}}$ and $\mathbf{q}_k^{\mathrm{D}}$, $k = 1 \ldots m$. In the above equation, $\hat{\mathrm{E}}$ denotes the sample average over the joint observations of the whitened data sets $\mathbf{z}^{\mathrm{A}}$ and $\mathbf{z}^{\mathrm{D}}$.

While $p_{\mathbf{s}_k}$ is generally unknown, we define it now such that we are capturing two possible types of dependencies between the data sets: linear correlation and variance dependencies. Linear correlation is presumably the simplest form of statistical dependency, and coupling in variance is the next simplest one. Variables which are linearly uncorrelated but correlated in variance tend to have high or low energies (squared values) at the same time. Modeling such dependencies proved useful when modeling the statistical dependencies within a given data set of natural images, see Chapter 10 of [2].

Sources $\mathbf{s}_k$ with linear and variance dependencies can be generated via

$$\mathbf{s}_k = \sigma_k \left(\tilde{s}_k^{\mathrm{A}} \;\; \tilde{s}_k^{\mathrm{D}} \right)^{\mathrm{T}}, \tag{11}$$

where $\tilde{s}_k^{\mathrm{A}}$ and $\tilde{s}_k^{\mathrm{D}}$ are two zero mean Gaussian random variables with correlation coefficient ρ_k, and $\sigma_k > 0$ is the variance variable responsible for the scaling. We prove in Text S1, Section S1.1, that the distribution $p_{\mathbf{s}_k}$ has the form

$$p_{\mathbf{s}_k}(\mathbf{s}_k) = G_k(\mathbf{s}_k^{\mathrm{T}} \mathbf{\Lambda}_k \mathbf{s}_k), \tag{12}$$

where $G_k(u), u \geq 0$ is a monotonically decreasing, strictly convex function which depends on the prior for σ_k and the correlation coefficient ρ_k. It is further shown that the same also holds for $\log G_k(u)$. Direct calculations, or the derivation in the supporting text, show that the correlation coefficient between s_k^{A} and s_k^{D} is given by ρ_k. The matrix $\mathbf{\Lambda}_k$ is the precision matrix (inverse covariance matrix) of $\mathbf{s}_k$. Since the sources in ICA are commonly assumed to have variance one, $\mathbf{\Lambda}_k$ is given by (3).

While different choices are possible for G_k, an interesting family of functions is obtained by assuming that σ_k^2 follows an inverse Gamma distribution. As shown in Text S1, Section S1.2, the functions G_k are then given by $G(u; \nu_k, \rho_k)$ in (4). The resulting pdf $p_{\mathbf{s}_k}$ is bivariate student's t. The HOCCA objective function in (2) follows from (10) with this choice for G_k.

The family $\{G(u; \nu_k, \rho_k)\}_{\nu_k, \rho_k}$ is interesting since the shape parameter ν_k controls the extent of higher-order statistical dependencies between s_k^{A} and s_k^{D} while the correlation coefficient ρ_k captures their linear correlation. This can best be seen by considering the mutual information between s_k^{A} and s_k^{D}. Mutual information measures the amount of information about s_k^{A} that one can obtain from s_k^{D}, and vice versa [53], see (S1–34) in Text S1, Section S1.2, for the formal definition. For the bivariate student's t distribution $p_{\mathbf{s}_k}$, the mutual information MI consists of two parts [54],

$$\mathrm{MI}(\nu_k, \rho_k) = \Omega(\nu_k) - \frac{1}{2} \log\left(1 - \rho_k^2\right). \tag{13}$$

The analytical expression for the first part, $\Omega(\nu_k)$, is given in (S1–36) in Text S1. The function $\Omega(\nu_k)$ decreases to zero as ν_k increases. The second part depends only on ρ_k and corresponds to the mutual information between two Gaussian random variables with correlation coefficient ρ_k. Hence, for large ν_k when Ω becomes small, the correlation coefficient ρ_k captures already most of the dependency between s_k^{A} and s_k^{D}. If ρ_k goes to zero and ν_k is large, s_k^{A} and s_k^{D} become statistically independent. If ν_k is small, on the other hand, there are higher-order statistical dependencies. Figure 8 shows the non-Gaussian part Ω and the Gaussian part as a function of ν_k and ρ_k, respectively. The figure shows that a value of ν_k close to two contributes to the mutual information like a correlation coefficient ρ_k of about 0.65, $\nu_k \approx 3$ corresponds to $\rho_k \approx 0.55$. Furthermore, the shape parameter ν_k affects the non-Gaussianity (sparsity) of the marginal distributions of s_k^{A} and s_k^{D}: The marginal distributions are univariate student's t distributions with the same shape parameters ν_k as $p_{\mathbf{s}_k}$ [55]. As ν_k decreases, the distributions become more heavy-tailed and peaked around zero, that is, the random variables are more sparse.

Simulations on artificial data

We used artificial data to validate HOCCA. We give here details for the data generation and the performance measures used in the assessment and in the comparison with CCA.

The data was generated according to (1). The dimensionality was $m = 10$ and the mixing matrices $\mathbf{Q}_{\mathrm{true}}^{\mathrm{A}}$ and $\mathbf{Q}_{\mathrm{true}}^{\mathrm{D}}$ were randomly generated by independently drawing the elements of the matrices from a standard normal distribution, followed by orthonormalization of each matrix. The correlation coefficients ρ_k^{true} between s_k^{A} and s_k^{D} were drawn from an uniform distribution on $(-1\ \ -0.1]\bigcup[0.1\ \ 1)$, and the parameters v_k^{true} from an uniform distribution on $[2.1\ \ 3]$. The true canonical coordinates were thus sparse and linearly correlated. We avoided sampling correlation coefficients close to zero since CCA is sensitive to linear correlation only.

We analyzed the estimation results of HOCCA and CCA using three measures of performance. The first measure assesses how well the mixing matrices (features) were recovered, the second the efficiency of the learned representation, and the third how well the coupling (correspondence) between the two data sets was identified. We assessed the efficiency of the representation from a sparsity and a related information theoretical point of view. Note that the first two measures are insensitive to the coupling between the data sets.

In order to quantify the accuracy of the estimated matrices $\mathbf{Q}^{\mathrm{A}}$ and $\mathbf{Q}^{\mathrm{D}}$, we used the Amari index $\mathcal{R}$ [56],

$$\mathcal{R}(\mathbf{P})=\sum_{i=1}^{m}\left(\sum_{j=1}^{m}\frac{|p_{ij}|}{\max_k|p_{ik}|}-1\right)+\sum_{j=1}^{m}\left(\sum_{i=1}^{m}\frac{|p_{ij}|}{\max_k|p_{kj}|}-1\right),\quad(14)$$

applied to $\mathbf{P}=(\mathbf{Q}^{\mathrm{A}})^{\mathrm{T}}\mathbf{Q}^{\mathrm{A}}_{\mathrm{true}}$ and $\mathbf{P}=(\mathbf{Q}^{\mathrm{D}})^{\mathrm{T}}\mathbf{Q}^{\mathrm{D}}_{\mathrm{true}}$. The entry in row i and column j of the matrix $\mathbf{P}$ is denoted by p_{ij}. The index is zero if $\mathbf{P}$ is a permutation matrix. For 10 dimensional random matrices formed by independent standard normal random variables, the index takes typically values around 0.37 ± 0.06 (average ± two standard deviations).

In order to quantify the sparsity of the recovered canonical coordinates $\langle\mathbf{q}_k^{\mathrm{A}},\mathbf{z}^{\mathrm{A}}\rangle$ and $\langle\mathbf{q}_k^{\mathrm{D}},\mathbf{z}^{\mathrm{D}}\rangle$, we used the index [57]

$$\mathcal{S}(s_k)=\left(\frac{\sqrt{T}}{\sqrt{T}-1}\right)\frac{\hat{\mathrm{E}}(|s_k|)}{\sqrt{\hat{\mathrm{E}}(s_k^2)}},\quad(15)$$

applied to $s_k=\langle\mathbf{q}_k^{\mathrm{A}},\mathbf{z}^{\mathrm{A}}\rangle$ and $s_k=\langle\mathbf{q}_k^{\mathrm{D}},\mathbf{z}^{\mathrm{D}}\rangle$, after removal of their mean. As before, $\hat{\mathrm{E}}$ denotes the sample average and we took $T=100000$ data points to compute it. The index $\mathcal{S}$ is non-decreasing with increasing sparsity; it takes zero as minimal and one as maximal value. A Gaussian has a value of $\mathcal{S}=0.2$.

In order to measure the efficiency of the learned representation from an information theoretical point of view, we computed by which extent the mutual information between the recovered coordinates was smaller than the mutual information between the original (white) data. This difference in mutual information is called multi-information reduction. We can here compute it by comparing the entropies of the marginal pdfs of the (white) data and the recovered canonical coordinates [41]. In our context, multi-information reduction is related to sparsity maximization since sparse variables have a smaller entropy than Gaussian variables of the same variance. We computed the multi-information reduction using 100000 data points.

In order to assess how well the coupling between the two data sets was identified, we computed the mutual information between the inferred corresponding sources $\left(\langle\mathbf{q}_k^{\mathrm{A}},\mathbf{z}^{\mathrm{A}}\rangle,\langle\mathbf{q}_k^{\mathrm{D}},\mathbf{z}^{\mathrm{D}}\rangle\right)$, and compared it to the mutual information of the "true" corresponding sources $(s_k^{\mathrm{A}},s_k^{\mathrm{D}})$. We measured mutual information using the maximum likelihood estimator with Miller-Maddow correction [58]. For computation of the mutual information, we used five million data points, and 1000 bins for the joint histogram after uniformization of the marginals.

Natural image data and preprocessing

The data used for the learning consists of pairs of images (image patches) that we extracted from a set of 50 larger natural images acquired under CIE D65, daylight, and under CIE A, yellowish light. Figure 2 shows pairs of example images from the database. The database is publicly available at http://isp.uv.es/data_color.htm and a detailed description was given before [21]. The images of the database are given in standard CIE XYZ tristimulus values. This makes it an appropriate data set to reproduce classical psychophysical results since they were obtained with these standard illuminants.

We used $1.5\cdot10^5$ corresponding image patches of size 15×15 pixels which we extracted from the pairs of larger images at the same randomly chosen position. After removal of the mean, the patches from images taken under illumination D65 give $\mathbf{x}^{\mathrm{D}}$, and the patches from images under illumination A are $\mathbf{x}^{\mathrm{A}}$. Each pair of images $(\mathbf{x}^{\mathrm{A}},\mathbf{x}^{\mathrm{D}})$ shows the same extract of the larger visual scene under two different illuminants. The dimension of $\mathbf{x}^{\mathrm{A}}$ and $\mathbf{x}^{\mathrm{D}}$ is $n=3\cdot15\cdot15=675$.

We then performed whitening and reduced the dimensionality of each individual data set by principal component analysis (PCA). Dimension reduction is worthwhile if there are strong correlations in the data, that is, if the data is essentially located in a subspace of lower dimensionality than n. Reducing the dimension of $\mathbf{x}^{\mathrm{A}}$ and $\mathbf{x}^{\mathrm{D}}$ can then reduce the average prediction error when trying to linearly predict $\mathbf{x}^{\mathrm{D}}$ from $\mathbf{x}^{\mathrm{A}}$, or vice versa (see Text S2, Section S2.2). In order to objectively decide about the amount of dimension reduction, we used the fraction of variance accounted for by the prediction (coefficient of determination R^2) when image patches under D65 illumination are linearly predicted from PCA truncated patches under illumination A. Figure 18 shows the coefficient of determination as a function of the retained dimension m of the data. According to the behavior in Figure 18 we decided to reduce the dimension of $\mathbf{x}^{\mathrm{A}}$ and $\mathbf{x}^{\mathrm{D}}$ from $n=675$ to $m=236$. Retaining 236 dimensions removes only $1.5\cdot10^{-3}\%$ and $2.3\cdot10^{-3}\%$ of the variance of $\mathbf{x}^{\mathrm{A}}$ and $\mathbf{x}^{\mathrm{D}}$, respectively.

Learning representations of natural images

We used HOCCA to learn the coupled representation by maximizing the objective f in (2). Other statistical methods can also be used to learn coupled representations, that is, the matrices $\mathbf{Q}^{\mathrm{A}}$ and $\mathbf{Q}^{\mathrm{D}}$ in Figure 3. We compared HOCCA to three alternative methods: canonical correlation analysis (CCA), a method based on whitening by principal component analysis, and a method based on independent component analysis (ICA). Table 1 provides an overview of the methods used.

CCA is briefly reviewed in Text S2. HOCCA and CCA naturally lead to coupled canonical representations. Whitening and ICA, however, are specific to each data set itself. After initial whitening or ICA, separately performed on each data set, we thus matched the learned filters across the data sets by greedily choosing pairs of components which had maximal mutual information. In this way, we obtained a coupled representation that can be used in the comparison with HOCCA.

Comparing HOCCA with the whitening-based approach is interesting since whitening is the first step in all methods. Comparing HOCCA with CCA and the ICA-based approach is interesting since these methods can be considered to represent limiting cases of HOCCA: ICA features are obtained by maximizing the efficiency (sparsity) of the representation of each data set individually, without concern for a possible correspondence between them. CCA features are obtained by maximization

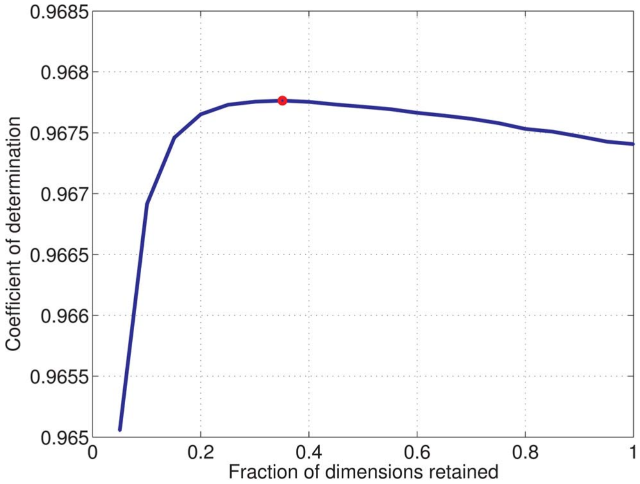


Figure 18. Choosing the amount of dimension reduction based on the performance when $\mathbf{x}^{\mathrm{D}}$ is linearly predicted from $\mathbf{x}^{\mathrm{A}}$. The plot shows the coefficient of determination (fraction of explained variance) as a function of the retained dimension m of $\mathbf{x}^{\mathrm{D}}$ and $\mathbf{x}^{\mathrm{A}}$. Retaining about 35% of the dimensions (236 out of 675) gives the best performance. Retaining 236 dimensions removes $1.5{\cdot}10^{-3}\%$ and $2.3{\cdot}10^{-3}\%$ of the total variance of $\mathbf{x}^{\mathrm{A}}$ and $\mathbf{x}^{\mathrm{D}}$, respectively.
doi:10.1371/journal.pone.0086481.g018

of correspondence (measured by linear correlation), without concern for the efficiency (sparsity) of the individual representations. HOCCA, on the other hand, is jointly maximizing the efficiency of the individual representations and the correspondence between them.

Analyzing the learned representations of natural images

The representations were statistically analyzed by assessing their efficiency and the coupling between the corresponding canonical coordinates (feature outputs) s_k^{A} and s_k^{D}. Efficiency was measured using multi-information reduction and sparsity, coupling was measured using mutual information. These measurements were performed as in the analysis of the results on artificial data.

For the visualization of the learned features, the features were first scaled to have unit norm and then contrast-normalized by applying a global scaling factor to the deviation from the average. The scaling was chosen so that the feature colors are reproducible in conventional displays: Too small scaling factors lead to chromatically uniform features while too large factors give rise to non-reproducible imaginary colors, that is, to negative luminance or to colors outside the reproducible gamut.

We reproduced corresponding-colors based on the learned representations as follows: Given a spatially uniform patch of a certain color under illumination CIE A, we identified it with $\mathbf{x}^{\mathrm{A}}$ in Figure 2 and represented it using the canonical coordinates $\mathbf{s}^{\mathrm{A}}$. Then, we predicted the k-th canonical coordinates s_k^{D} from s_k^{A}, and transformed back to the original pixel-representation, that is, to $\mathbf{x}^{\mathrm{D}}$, which yielded the corresponding color under illumination CIE D65. Given colors under illumination D65, the procedure was reversed. The prediction of s_k^{D} from s_k^{A} (and vice versa) was constrained to be linear, even though nonlinear prediction would be possible too. In more detail, since the canonical coordinates have zero mean and unit variance, the prediction of s_k^{D} is $\hat{s}_k^{\mathrm{D}}=\varrho_k s_k^{\mathrm{A}}$, where ϱ_k is the correlation coefficient between s_k^{D} and s_k^{A}.

For the noise-distortion curves, the setup of the corresponding-colors was modified in two aspects: First, we used image data with spatial structure. Second, the internal representation by means of the canonical coordinates was subject to noise. The noisy version of $\hat{s}_k^{\mathrm{D}}$ is denoted by $\tilde{s}_k^{\mathrm{D}}$,

$$\tilde{s}_k^{\mathrm{D}}=\hat{s}_k^{\mathrm{D}}+\sigma(\hat{s}_k^{\mathrm{D}})n_k. \tag{16}$$

The random variables n_k are independent from each other and from the canonical coordinates, and have zero mean and unit variance. For a fixed image $\mathbf{x}^{\mathrm{A}}$, the noisy response $\tilde{s}_k^{\mathrm{D}}$ fluctuates around the mean $\hat{s}_k^{\mathrm{D}}$ with a variance σ^2. The variance was assumed to be signal dependent,

$$\sigma^2(\hat{s}_k^{\mathrm{D}})=F|\hat{s}_k^{\mathrm{D}}|, \tag{17}$$

where F is the Fano factor (index of dispersion) of the noisy response. We measured to which extent the noisy inferred representation deviates from the noise-free representation of the same image under illumination D65. That is, we measured how much $\tilde{s}_k^{\mathrm{D}}$ deviates from s_k^{D}. We used the root mean squared error (RMSE), $\mathrm{RMSE}=\sqrt{E||\mathbf{s}^{\mathrm{D}}-\tilde{\mathbf{s}}^{\mathrm{D}}||^2}$ as error metric. The analytical expression for the squared error reported in (7) is derived in Text S3.

For the stimuli used in the prediction, we changed the chromatic contrast and color content as follows: In a achromatic red-green yellow-blue representation, the color $\mathbf{c}\in\mathbb{R}^3$ of each pixel can be seen as an achromatic and a chromatic departure from the average $\bar{\mathbf{c}}$: $\mathbf{c}=\bar{\mathbf{c}}+\boldsymbol{\delta}_a+\boldsymbol{\delta}_c$. The average can further be divided into an achromatic ($\bar{\mathbf{c}}_a$) and chromatic part ($\bar{\mathbf{c}}_c$). The stimuli were obtained through rotations of $\boldsymbol{\delta}_c$, via a 3×3 rotation matrix $\mathbf{R}_\alpha$, and by scaling the resulting chromatic part: $\mathbf{c}'(\alpha,\beta)=(\bar{\mathbf{c}}_a+\boldsymbol{\delta}_a)+\beta(\mathbf{R}_\alpha\boldsymbol{\delta}_c+\bar{\mathbf{c}}_c)$. The color content was rotated in constant steps, and the scaling factor β was varied linearly from one to zero.

Supporting Information

Text S1 Details of Higher-Order Canonical Correlation Analysis.
(PDF)

Text S2 Background Material from Multivariate Analysis.
(PDF)

Text S3 Calculations for the Noise-Distortion Analysis.
(PDF)

Author Contributions

Conceived and designed the experiments: MUG VL JM. Performed the experiments: MUG. Analyzed the data: MUG VL JM. Wrote the paper: MUG. Assisted in writing: VL AH JM. Developed HOCCA: MUG. Assisted in developing HOCCA: AH. Relation to perception: VL JM.

References

1. Olshausen B, Field D (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381: 607–609.
2. Hyvärinen A, Hurri J, Hoyer P (2009) Natural Image Statistics. Springer.
3. Coates A, Ng A, Lee H (2011) An analysis of single-layer networks in unsupervised feature learning. In: Gordon G, Dunson D, Dudik M, editors, JMLR Workshop and Conference Proceedings. volume 15, pp. 215–223.

4. Puertas J, Bornschein J, Lücke J (2010) The maximal causes of natural scenes are edge filters. In: Lafferty J, Williams CKI, Zemel R, Shawe-Taylor J, Culotta A, editors, Advances in Neural Information Processing Systems 23. pp. 1939–1947.
5. Bornschein J, Henniges M, Lücke J (2013) Are V1 simple cells optimized for visual occlusions? A comparative study. PLoS Comput Biol 9: e1003062–.
6. Hoyer P, Hyvärinen A (2000) Independent component analysis applied to feature extraction from colour and stereo images. Network: Computation in Neural Systems, 11(3): 191–210.
7. Tailor D, Finkel L, Buchsbaum G (2000) Color-opponent receptive fields derived from independent component analysis of natural images. Vision Research 40: 2671–2676.
8. Wachtler T, Lee T, Sejnowski T (2001) Chromatic structure of natural scenes. Journal of the Optical Society of America A 18: 65–77.
9. Lee T, Wachtler T, Sejnowski T (2002) Color opponency is an efficient representation of spectral properties in natural scenes. Vision Research 42: 2095–2103.
10. Doi E, Inui T, Lee TW, Wachtler T, Sejnowski T (2003) Spatiochromatic receptive field properties derived from information-theoretic analyses of cone mosaic responses to natural scenes. Neural Computation 15: 397–417.
11. Ruderman D, Cronin T, Chiao C (1998) Statistics of cone responses to natural images: implications for visual coding. Journal of the Optical Society of America A 15: 2036–2045.
12. Chakrabarti A, Zickler T (2011) Statistics of real-world hyperspectral images. In: Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). pp. 193–200.
13. Caywood M, Willmore B, Tolhurst D (2004) Independent components of color natural scenes resemble V1 neurons in their spatial and color tuning. J Neurophysiol 91: 2859–2873.
14. Rehn M, Sommer F (2007) A network that uses few active neurones to code visual input predicts the diverse shapes of cortical receptive fields. Journal of Computational Neuroscience 22: 135–146.
15. Ringach D (2002) Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. J Neurophysiol 88: 455–63.
16. Clifford C, Webster M, Stanley G, Stocker A, Kohn A, et al. (2007) Visual adaptation: Neural, psychological and computational aspects. Vision Research 47: 3125–3131.
17. Webster M, Mollon J (1997) Adaptation and the color statistics of natural images. Vision Research 37: 3283–3298.
18. Webster M, Mollon J (1991) Changes in colour appearance following post-receptoral adaptation. Nature 349: 235–238.
19. Atick J, Li Z, Redlich A (1993) What does post-adaptation color appearance reveal about cortical color representation? Vision Res 33: 123–129.
20. Gutmann M, Hyvärinen A (2011) Extracting coactivated features from multiple data sets. In: Honkela T, editor, Proc. Int. Conf. on Artificial Neural Networks (ICANN). Berlin, Heidelberg: Springer, volume 6791 of Lecture Notes in Computer Science, pp. 323–330.
21. Laparra V, Jiménez S, Camps-Valls G, Malo J (2012) Nonlinearities and adaptation of color vision from sequential principal curves analysis. Neural Computation 24: 2751–2788.
22. Hastie T, Tibshirani R, Friedman J (2009) The Elements of Statistical Learning. Springer.
23. Kim TH, White H (2004) On more robust estimation of skewness and kurtosis. Finance Research Letters 1: 56–73.
24. Lennie P, Krauskopf J, Sclar G (1990) Chromatic mechanisms in striate cortex of macaque. J Neurosci 10: 649–669.
25. Conway B (2001) Spatial structure of cone inputs to color cells in alert macaque primary visual cortex V1. J Neurosci 21: 2768–2783.
26. Johnson E, Hawken M, Shapley R (2001) The spatial transformation of color in the primary visual cortex of the macaque monkey. Nat Neurosci 4: 409–416.
27. Fairchild M (2005) Color Appearance Models. Chichester, UK: Wiley-IS&T, 2nd edition.
28. Breneman E (1987) Corresponding chromaticities for different states of adaptation to complex visual fields. Journal of the Optical Society of America A 4: 1115–1129.
29. Luo M, Clarke A, Rhodes P, Scrivener S, Schappo A, et al. (1991) Quantifying colour appearance. part I. LUTCHI colour appearance data. Color Res Appl 16: 166–180.
30. Luo M, Rhodes P (1999) Corresponding-colour datasets. Color Res Appl 24: 295–296.
31. Dayan P, Abbott L (2001) Theoretical Neuroscience. The MIT Press.
32. Gur M, Snodderly D (2006) High response reliability of neurons in primary visual cortex (V1) of alert, trained monkeys. Cerebral Cortex 16: 888–895.
33. Tailby C, Solomon S, Dhruv N, Lennie P (2008) Habituation reveals fundamental chromatic mechanisms in striate cortex of macaque. Journal of Neuroscience 28: 1131–1139.
34. Bach F, Jordan M (2002) Kernel independent component analysis. Journal of Machine Learning Research 3: 1–48.
35. Akaho S (2001) A kernel method for canonical correlation analysis. In: Proceedings of the International Meeting of the Psychometric Society (IMPS). Springer-Verlag.
36. Melzer T, Reiter M, Bischof H (2003) Appearance models based on kernel canonical correlation analysis. Pattern Recognition 36: 1961–1971.
37. Archambeau C, Bach F (2009) Sparse probabilistic projections. In: Koller D, Schuurmans D, Bengio Y, Bottou L, editors, Advances in Neural Information Processing Systems 21. pp. 73–80.
38. Witten D, Tibshirani R, Hastie T (2009) A penalized matrix decomposition, with applications to sparse canonical correlation analysis and principal components. Biostatistics 10.
39. Karhunen J, Ukkonen T (2007) Extending ICA for finding jointly dependent components from two related data sets. Neurocomputing 70: 2969–2979.
40. Karhunen J, Hao T, Ylipaavalniemi J (2013) Finding dependent and independent components from related data sets: A generalized canonical correlation analysis based method. Neurocomputing 113: 153–167.
41. Bethge M (2006) Factorial coding of natural images: How effective are linear models in removing higher-order dependencies? Journal of the Optical Society of America A 23: 1253–1268.
42. Malo J, Laparra V (2010) Psychophysically tuned divisive normalization approximately factorizes the PDF of natural images. Neural Computation 22: 3179–3206.
43. Clarke R (1981) Relation between the Karhunen-Loeve transform and cosine transforms. Communications, Radar and Signal Processing, IEE Proceedings F 128: 359–360.
44. Hancock P, Baddeley R, Smith L (1992) The principal components of natural images. Network 3: 61–72.
45. Mullen K (1985) The CSF of human colour vision to red-green and yellow-blue chromatic gratings. J Physiol 359: 381–400.
46. Moroney N, Fairchild M, Hunt R, Li C, Luo M, et al. (2002) The CIECAM02 color appearance model. In: IS&T/SID 10th Color Imaging Conference. pp. 23–27.
47. Verdu F, Luque M, Malo J, Felipe A, Artigas J (1997) Implementations of a novel algorithm for colour constancy. Vision Research 37: 1829–1844.
48. Marimont D, Wandell B (1992) Linear models of surface and illuminant spectra. Journal of the Optical Society of America A 9: 1905–1913.
49. D'Zmura M, Iverson G (1993) Color constancy. I. Basic theory of two-stage linear recovery of spectral descriptions for lights and surfaces. Journal of the Optical Society of America A 10: 2148–2165.
50. D'Zmura M, Iverson G (1993) Color constancy. II. Results for two-stage linear recovery of spectral descriptions for lights and surfaces. Journal of the Optical Society of America A 10: 2166–2180.
51. Abrams A, Hillis J, Brainard D (2007) The relation between color discrimination and color constancy: when is optimal adaptation task dependent? Neural Computation 19: 2610–2637.
52. Tuia D, Muñoz-Marí J, Gómez-Chova L, Malo J (2013) Graph matching for adaptation in remote sensing. IEEE T Geoscience and Remote Sensing 51: 329–341.
53. Cover T, Thomas J (2006) Elements of Information Theory. Wiley-Interscience, 2nd edition.
54. Guerrero-Cusumano JL (1996) An asymptotic test of independence for multivariate t and Cauchy random variables with applications. Information Sciences 92: 33–45.
55. Nadarajah S, Kotz S (2005) Mathematical properties of the multivariate t-student distribution. Acta Applicandae Mathematicae 89: 53–84.
56. Amari S, Cichocki A, Yang H (1996) A new learning algorithm for blind signal separation. In: Advances in Neural Information Processing Systems. MIT Press, pp. 757–763.
57. Hoyer P (2004) Non-negative matrix factorization with sparseness constraints. Journal of Machine Learning Research 5: 1457–1469.
58. Miller G (1955) Note on the bias of information estimates. Information Theory in Psychology 2b: 95–100.

S1 Details of Higher-Order Canonical Correlation Analysis

We present here the mathematical details of higher-order canonical correlation analysis (HOCCA). Section S1.1 deals with the (general) probabilistic data model that underlies HOCCA. In Section S1.2, we insert more assumptions and derive the parametric formulation presented in the main text. Section S1.3 contains the details on the relation to canonical correlation analysis.

S1.1 General nonparametric case

We consider here the generalized case of n_c coupled data sets of possibly different dimensionality. Equation (1) generalizes to

$$\mathbf{z}^i = \mathbf{Q}^i \mathbf{s}^i \quad (i = 1, \ldots, n_c), \tag{S1-1}$$

where $\mathbf{z}^i, \mathbf{s}^i \in \mathbb{R}^{m_i}$. We assume that the first $m = \min(m_1, \ldots, m_{n_c})$ canonical coordinates of the data sets are possibly coupled with each other while the remaining coordinates are independent. Thus, the joint pdf of the sources decomposes as

$$p_{\mathbf{s}}(s_1^1, \ldots, s_{m_{n_c}}^{n_c}) = \prod_{k=1}^{m} p_{\mathbf{s}_k}(\mathbf{s}_k) \prod_{i=1}^{n_c} \prod_{k=m+1}^{m_i} p_{ik}(s_k^i) \tag{S1-2}$$

where the m vectors $\mathbf{s}_k$,

$$\mathbf{s}_k = (s_k^1, \ldots, s_k^{n_c})^\top, \tag{S1-3}$$

contain the possibly coupled coordinates, and the p_{ik} are the pdfs of the non-coupled sources.

We next specify $p_{\mathbf{s}_k}$. We assume that the elements s_k^i of the vector $\mathbf{s}_k$ are coupled via

$$s_k^1 = \sigma_k \tilde{s}_k^1, \qquad s_k^2 = \sigma_k \tilde{s}_k^2, \qquad s_k^3 = \sigma_k \tilde{s}_k^3, \qquad \ldots \qquad s_k^{n_c} = \sigma_k \tilde{s}_k^{n_c}, \tag{S1-4}$$

where the random variable $\sigma_k > 0$ sets the variance, and the $\tilde{s}_k^i$ are zero mean Gaussian random variables. The distribution of σ_k affects the strength of the coupling. Inverting the linear transform in (S1-4) gives

$$\tilde{\mathbf{s}}_k = \frac{1}{\sigma_k} \mathbf{s}_k, \tag{S1-5}$$

where $\tilde{\mathbf{s}}_k = (\tilde{s}_k^1, \ldots, \tilde{s}_k^{n_c})$. The determinant of this linear transformation is $1/\sigma_k^{n_c}$. Integrating out the variable σ_k with density p_{σ_k} leads to an expression for the density of $\mathbf{s}_k$,

$$p_{\mathbf{s}_k}(\mathbf{s}_k) = \int \frac{p_{\sigma_k}(\sigma_k)}{\sigma_k^{n_c}} p_{\tilde{\mathbf{s}}_k}\left(\frac{\mathbf{s}_k}{\sigma_k}\right) \mathrm{d}\sigma_k. \tag{S1-6}$$

Equivalently, we can specify a prior p_{ω_k} for $\omega_k = \sigma_k^2$. The density $p_{\mathbf{s}_k}$ is then

$$p_{\mathbf{s}_k}(\mathbf{s}_k) = \int \frac{p_{\omega_k}(\omega_k)}{\omega_k^{n_c/2}} p_{\tilde{\mathbf{s}}_k}\left(\frac{\mathbf{s}_k}{\sqrt{\omega_k}}\right) \mathrm{d}\omega_k. \tag{S1-7}$$

The variables $\tilde{\mathbf{s}}_k$ are assumed jointly Gaussian with density $p_{\tilde{\mathbf{s}}_k}$,

$$p_{\tilde{\mathbf{s}}_k}(\tilde{\mathbf{s}}_k) = \frac{1}{(2\pi)^{\frac{n_c}{2}} |\tilde{\mathbf{\Sigma}}_k|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} \tilde{\mathbf{s}}_k^\top \tilde{\mathbf{\Sigma}}_k^{-1} \tilde{\mathbf{s}}_k\right), \tag{S1-8}$$

where $\tilde{\mathbf{\Sigma}}_k$ is the covariance matrix. The covariance matrix $\mathrm{V}(\mathbf{s}_k)$ of $\mathbf{s}_k$ is

$$\begin{aligned}
\mathrm{V}(\mathbf{s}_k) &= \int_0^\infty p_{\omega_k}(\omega_k) \, \mathrm{V}(\mathbf{s}_k|\omega_k) \mathrm{d}\omega_k & \text{(S1-9)} \\
&= \mathrm{V}(\tilde{\mathbf{s}}_k) \int_0^\infty \omega_k p_{\omega_k}(\omega_k) \mathrm{d}\omega_k & \text{(S1-10)} \\
&= \tilde{\mathbf{\Sigma}}_k \mu_k, & \text{(S1-11)}
\end{aligned}$$

where μ_k denotes the mean of ω_k. For the second equality, we have used that $\mathrm{V}(\mathbf{s}_k|\omega_k) = \omega_k \mathrm{V}(\tilde{\mathbf{s}}_k)$. The covariance matrix $\mathrm{V}(\mathbf{s}_k)$ is proportional to $\tilde{\boldsymbol{\Sigma}}_k$, which means that the correlation coefficient between s_k^i and s_k^j is the same as the correlation coefficient between $\tilde{s}_k^i$ and $\tilde{s}_k^j$, $i \neq j$. Further,

$$\tilde{\boldsymbol{\Sigma}}_k^{-1} = \mu_k \boldsymbol{\Lambda}_k, \tag{S1-12}$$

where $\boldsymbol{\Lambda}_k = \mathrm{V}(\mathbf{s}_k)^{-1}$ is the precision matrix of $\mathbf{s}_k$. Hence, the prior $p_{\mathbf{s}_k}$ is

$$p_{\mathbf{s}_k}(\mathbf{s}_k) = G_k(\mathbf{s}_k^\top \boldsymbol{\Lambda}_k \mathbf{s}_k), \tag{S1-13}$$

where the function G_k is defined via the one-dimensional integral

$$G_k(u) = \frac{1}{(2\pi)^{\frac{n_c}{2}} |\tilde{\boldsymbol{\Sigma}}_k|^{\frac{1}{2}}} \int_0^\infty \exp\left(-\frac{\mu_k}{2\omega_k} u\right) \frac{p_{\omega_k}(\omega_k)}{\omega_k^{n_c/2}} \mathrm{d}\omega_k \quad (u \geq 0), \tag{S1-14}$$

which depends on the prior p_{ω_k} and the covariance matrix of $\tilde{\mathbf{s}}_k$. Taking the derivative under the integral sign, and using that $\mu_k > 0$, we find that that $G_k'(u) < 0$. Taking the second derivative shows further that $G_k''(u) > 0$. Hence, G_k is monotonically decreasing and strictly convex for $u > 0$. The same also holds for $\log G_k$: $(\log G_k(u))' = G_k'(u)/G_k(u) < 0$ since G_k is positive, and $(\log G_k(u))'' > 0$ follows from a development as in Section 10.8 of [2] using the Cauchy-Schwarz inequality.

By orthogonality of $\mathbf{Q}^i$, the joint distribution of $\mathbf{z} = (\mathbf{z}^1, \ldots, \mathbf{z}^{n_c})^\top$ is

$$\begin{aligned} p_{\mathbf{z}}(\mathbf{z}^1, \ldots, \mathbf{z}^{n_c}) &= p_{\mathbf{s}}(\langle \mathbf{q}_1^1, \mathbf{z}^1 \rangle, \ldots, \langle \mathbf{q}_{m_{n_c}}^{n_c}, \mathbf{z}^{n_c} \rangle) && \text{(S1-15)} \\ &= \prod_{k=1}^m p_{\mathbf{s}_k}(\langle \mathbf{q}_k^1, \mathbf{z}^1 \rangle, \ldots, \langle \mathbf{q}_k^{n_c}, \mathbf{z}^{n_c} \rangle) \prod_{i=1}^{n_c} \prod_{k=m+1}^{m_i} p_{ik}(\langle \mathbf{q}_k^i, \mathbf{z}^i \rangle). && \text{(S1-16)} \end{aligned}$$

Denoting the n_c-dimensional vector $(\langle \mathbf{q}_k^1, \mathbf{z}^1 \rangle, \ldots, \langle \mathbf{q}_k^{n_c}, \mathbf{z}^{n_c} \rangle)^\top$ by $\mathbf{y}_k$, the t-th observation of $\mathbf{z}^i$ by $\mathbf{z}^i(t)$, and the t-th observation of $\mathbf{y}_k$ by $\mathbf{y}_k(t)$, we obtain the log-likelihood ℓ,

$$\ell = \sum_{t=1}^T \sum_{k=1}^m \log G_k(\mathbf{y}_k(t)^\top \boldsymbol{\Lambda}_k \mathbf{y}_k(t)) + \sum_{t=1}^T \sum_{i=1}^{n_c} \sum_{k=m+1}^{m_i} \log p_{ik}(\langle \mathbf{q}_k^i, \mathbf{z}^i(t) \rangle). \tag{S1-17}$$

Here, T denotes the total number of observations and we tacitly assume that we can easily evaluate G_k and p_{ik}.

The log-likelihood separates into two parts: The first part with the G_k contains the possibly coupled features while the second part with the p_{ik} contains the remaining ones. The two parts are independent from each other up to the orthogonality constraint that $\langle \mathbf{q}_k^i, \mathbf{q}_j^i \rangle = 0$ for $k \neq j$. Moreover, the second part separates into n_c independent sub-parts. Hence, to maximize the log-likelihood, it is possible to maximize f,

$$\begin{aligned} f &= \frac{1}{T} \sum_{k=1}^m \log G_k(\mathbf{y}_k(t)^\top \boldsymbol{\Lambda}_k \mathbf{y}_k(t)) && \text{(S1-18)} \\ &= \sum_{k=1}^m \hat{\mathrm{E}} \log G_k(\mathbf{y}_k^\top \boldsymbol{\Lambda}_k \mathbf{y}_k), && \text{(S1-19)} \end{aligned}$$

in a first step, and afterwards the remaining terms J_i,

$$\begin{aligned} J_i &= \frac{1}{T} \sum_{t=1}^T \sum_{k=m+1}^{m_i} \log p_{ik}(\langle \mathbf{q}_k^i, \mathbf{z}^i(t) \rangle) && \text{(S1-20)} \\ &= \sum_{k=m+1}^{m_i} \hat{\mathrm{E}} \log p_{ik}(\langle \mathbf{q}_k^i, \mathbf{z}^i \rangle) && \text{(S1-21)} \end{aligned}$$

for $i = 1 \ldots n_c$. In the equations, the symbol $\hat{\mathrm{E}}$ denotes the sample average. Optimizing the terms J_i corresponds to doing ordinary ICA on the individual $\mathbf{z}^i$ under the constraint that the $\mathbf{q}_k^i$, $k = m+1 \ldots m_i$, are orthogonal to the $\mathbf{q}_k^i$, $k = 1 \ldots m$, which are obtained in the optimization of f. If we are interested in the possibly coupled features only, it suffices to maximize f.

S1.2 Convenient parametrization

We derive here a convenient family of functions for the G_k. We consider the case where the variance variable $\omega_k = \sigma_k^2$ follows the inverse Gamma distribution with parameters $\alpha_k > 1$, $\beta_k > 0$,

$$p_{\omega_k}(\omega_k; \alpha_k, \beta_k) = \frac{\beta_k^{\alpha_k}}{\Gamma(\alpha_k)} \omega_k^{-\alpha_k - 1} \exp\left(-\frac{\beta_k}{\omega_k}\right). \tag{S1-22}$$

Here, $\Gamma(\alpha_k)$ is the gamma function,

$$\Gamma(\alpha_k) = \int_0^\infty u^{\alpha_k - 1} \exp(-u) \mathrm{d}u. \tag{S1-23}$$

The mean μ_k of ω_k is $\beta_k/(\alpha_k - 1)$. The function $G_k(u)$ in (S1-14) becomes thus

$$G_k(u) = \frac{\beta_k^{\alpha_k}}{\Gamma(\alpha_k)} \frac{1}{(2\pi)^{\frac{n_c}{2}} |\tilde{\mathbf{\Sigma}}_k|^{\frac{1}{2}}} \int_0^\infty \omega_k^{-\alpha_k - 1 - \frac{n_c}{2}} \exp\left(-\left(\beta_k + \frac{\beta_k}{2(\alpha_k - 1)} u\right) \frac{1}{\omega_k}\right) \mathrm{d}\omega_k. \tag{S1-24}$$

Making the change of variables

$$\omega_k = \left(\beta_k + \frac{\beta_k}{2(\alpha_k - 1)} u\right) \frac{1}{v} \tag{S1-25}$$

we obtain

$$\begin{aligned} G_k(u) &= \frac{\beta_k^{\alpha_k}}{\Gamma(\alpha_k)} \frac{1}{(2\pi)^{\frac{n_c}{2}} |\tilde{\mathbf{\Sigma}}_k|^{\frac{1}{2}}} \int_0^\infty \left(\beta_k + \frac{\beta_k}{2(\alpha_k - 1)} u\right)^{-\alpha_k - \frac{n_c}{2}} v^{\alpha_k + \frac{n_c}{2} - 1} \exp(-v) \mathrm{d}v \\ &= \frac{\beta_k^{\alpha_k}}{\Gamma(\alpha_k)} \frac{1}{(2\pi)^{\frac{n_c}{2}} |\tilde{\mathbf{\Sigma}}_k|^{\frac{1}{2}}} \left(\beta_k + \frac{\beta_k}{2(\alpha_k - 1)} u\right)^{-\alpha_k - \frac{n_c}{2}} \Gamma\left(\alpha_k + \frac{n_c}{2}\right) && \text{(S1-26)} \\ &= \frac{\Gamma\left(\alpha_k + \frac{n_c}{2}\right)}{\Gamma(\alpha_k)} \frac{1}{(2\pi\beta_k)^{\frac{n_c}{2}} |\tilde{\mathbf{\Sigma}}_k|^{\frac{1}{2}}} \left(1 + \frac{1}{2(\alpha_k - 1)} u\right)^{-\alpha_k - \frac{n_c}{2}}. && \text{(S1-27)} \end{aligned}$$

From (S1-12), we have

$$|\tilde{\mathbf{\Sigma}}_k|^{-\frac{1}{2}} = \mu_k^{\frac{n_c}{2}} |\mathbf{\Lambda}_k|^{\frac{1}{2}}, \tag{S1-28}$$

and as $\mu_k = \beta_k/(\alpha_k - 1)$, we obtain

$$|\tilde{\mathbf{\Sigma}}_k|^{-\frac{1}{2}} = \left(\frac{\beta_k}{\alpha_k - 1}\right)^{\frac{n_c}{2}} |\mathbf{\Lambda}_k|^{\frac{1}{2}}, \tag{S1-29}$$

so that

$$G_k(u) = \frac{\Gamma\left(\alpha_k + \frac{n_c}{2}\right)}{\Gamma(\alpha_k)} \frac{1}{(2\pi(\alpha_k - 1))^{\frac{n_c}{2}}} |\mathbf{\Lambda}_k|^{\frac{1}{2}} \left(1 + \frac{1}{2(\alpha_k - 1)} u\right)^{-\alpha_k - \frac{n_c}{2}}, \tag{S1-30}$$

which does not depend on the parameter β_k. Introducing the parameter $\nu_k = 2\alpha_k > 2$, the function $G_k(u)$ is

$$G_k(u) = \frac{\Gamma\left(\frac{\nu_k + n_c}{2}\right)}{\Gamma\left(\frac{\nu_k}{2}\right)} \frac{|\mathbf{\Lambda}_k|^{\frac{1}{2}}}{(\pi(\nu_k - 2))^{\frac{n_c}{2}}} \left(1 + \frac{u}{\nu_k - 2}\right)^{-\frac{\nu_k + n_c}{2}}, \tag{S1-31}$$

which is (4) in the main text of the paper for $n_c = 2$.

The random vector $\mathbf{s}_k$ has the density $p_{\mathbf{s}_k}(\mathbf{s}_k) = G_k(\mathbf{s}_k^\top \mathbf{\Lambda}_k \mathbf{s}_k)$, see (S1-13). The re-scaled random vector $\mathbf{t}$,

$$\mathbf{t} = \mathbf{s}_k \sqrt{\frac{\nu_k}{\nu_k - 2}}, \tag{S1-32}$$

has the density $p_\mathbf{t}$,

$$p_\mathbf{t}(\mathbf{t}) = \frac{\Gamma\left(\frac{\nu_k + n_c}{2}\right)}{\Gamma\left(\frac{\nu_k}{2}\right)} \frac{|\mathbf{\Lambda}_k|^{\frac{1}{2}}}{(\pi \nu_k)^{\frac{n_c}{2}}} \left(1 + \frac{\mathbf{t}^\top \mathbf{\Lambda}_k \mathbf{t}}{\nu_k}\right)^{-\frac{\nu_k + n_c}{2}}, \tag{S1-33}$$

which is the parametrization of a n_c-variate student's t distribution.

Mutual information MI between n random variables $y_1, \ldots, y_n$ is defined as the Kullback Leibler divergence between their joint pdf $p(y_1, \ldots, y_n)$ and the product of their marginal pdfs $\prod_i p(y_i)$,

$$\mathrm{MI} = \int p(y_1, \ldots, y_n) \log \frac{p(y_1, \ldots, y_n)}{\prod_i p(y_i)} \mathrm{d}y_1 \ldots \mathrm{d}y_n. \tag{S1-34}$$

Mutual information between several random variables is also known as multi-information [41]. For the n_c-variate student's t distribution, the mutual information MI is [54]

$$\mathrm{MI} = \Omega(\nu_k) + \frac{1}{2} \log |\mathbf{\Lambda}_k|, \tag{S1-35}$$

where

$$\begin{aligned} \Omega(\nu) \quad = \quad & \log \left[\frac{\Gamma\left(\frac{n_c}{2}\right)}{\pi^{\frac{n_c}{2}}} \frac{\left(\beta\left(\frac{1+\nu}{2}, \frac{1}{2}\right)\right)^{n_c}}{\beta\left(\frac{n_c+\nu}{2}, \frac{n_c}{2}\right)} \right] + \frac{n_c(1+\nu)}{2} \left[\psi\left(\frac{1+\nu}{2}\right) - \psi\left(\frac{\nu}{2}\right) \right] \\ & - \frac{n_c + \nu}{2} \left[\psi\left(\frac{n_c + \nu}{2}\right) - \psi\left(\frac{\nu}{2}\right) \right], \end{aligned} \tag{S1-36}$$

with β being the beta-function, and ψ the digamma-function. Note that the matrix $\mathbf{\Lambda}$ is the inverse of the matrix A used by [54]. Since mutual information is scale invariant, $\mathbf{s}_k$ has the mutual information in (S1-35). For the case of $n_c = 2$, $\log |\mathbf{\Lambda}_k| = -\log(1 - \rho_k^2)$ which, together with (S1-35), yields (13) in the main text.

S1.3 Relation to canonical correlation analysis

We consider here the case of two data sets and derive (6). We start with computing $1/(\nu_k - 2)\mathbf{y}_k^\top \mathbf{\Lambda}_k \mathbf{y}_k$. Using the definition of $\mathbf{y}_k$,

$$\mathbf{y}_k = (\langle \mathbf{q}_k^\mathrm{A}, \mathbf{z}^\mathrm{A} \rangle,\ \langle \mathbf{q}_k^\mathrm{D}, \mathbf{z}^\mathrm{D} \rangle)^\top, \tag{S1-37}$$

and the definition of $\mathbf{\Lambda}_k$ in (3), we obtain

$$\frac{\mathbf{y}_k^\top \mathbf{\Lambda}_k \mathbf{y}_k}{\nu_k - 2} = \frac{1}{\nu_k - 2} \frac{1}{1 - \rho_k^2} \left[\langle \mathbf{q}_k^\mathrm{A}, \mathbf{z}^\mathrm{A} \rangle^2 + \langle \mathbf{q}_k^\mathrm{D}, \mathbf{z}^\mathrm{D} \rangle^2 - 2\rho_k \langle \mathbf{q}_k^\mathrm{A}, \mathbf{z}^\mathrm{A} \rangle \langle \mathbf{q}_k^\mathrm{D}, \mathbf{z}^\mathrm{D} \rangle \right]. \tag{S1-38}$$

For large ν_k the term $1/(\nu_k - 2)\mathbf{y}_k^\top \mathbf{\Lambda}_k \mathbf{y}_k$ is small. Hence,

$$\log\left(1 + \frac{1}{\nu_k - 2} \mathbf{y}_k^\top \mathbf{\Lambda}_k \mathbf{y}_k\right) = \frac{1}{\nu_k - 2} \mathbf{y}_k^\top \mathbf{\Lambda}_k \mathbf{y}_k + O\left(\frac{1}{\nu_k^2}\right), \tag{S1-39}$$

where we have used the first-order Taylor expansion of $\log(1+x)$ around $x=0$. Dropping terms of order $1/\nu_k^2$ and smaller, we have for $f(\mathbf{Q}^{\mathrm{A}}, \mathbf{Q}^{\mathrm{D}})$ in (5)

$$\begin{aligned} f(\mathbf{q}_1^{\mathrm{A}}, \dots, \mathbf{q}_m^{\mathrm{D}}) &\approx \mathrm{const} - \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{m} \frac{\nu_k+2}{2\nu_k-4}\frac{1}{1-\rho_k^2}\left[\langle \mathbf{q}_k^{\mathrm{A}}, \mathbf{z}^{\mathrm{A}}(t)\rangle^2 + \right. \\ &\qquad \left. \langle \mathbf{q}_k^{\mathrm{D}}, \mathbf{z}^{\mathrm{D}}(t)\rangle^2 - 2\rho_k \langle \mathbf{q}_k^{\mathrm{A}}, \mathbf{z}^{\mathrm{A}}(t)\rangle \langle \mathbf{q}_k^{\mathrm{D}}, \mathbf{z}^{\mathrm{D}}(t)\rangle\right]. \end{aligned} \tag{S1-40}$$

Since ν_k is assumed large,

$$\frac{\nu_k+2}{2\nu_k-4} \approx \frac{1}{2} \tag{S1-41}$$

and thus

$$\begin{aligned} f(\mathbf{q}_1^{\mathrm{A}}, \dots, \mathbf{q}_m^{\mathrm{D}}) &\approx \mathrm{const} - \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{m} \frac{1}{2}\frac{1}{1-\rho_k^2}\left[\langle \mathbf{q}_k^{\mathrm{A}}, \mathbf{z}^{\mathrm{A}}(t)\rangle^2 + \right. \\ &\qquad \left. \langle \mathbf{q}_k^{\mathrm{D}}, \mathbf{z}^{\mathrm{D}}(t)\rangle^2 - 2\rho_k \langle \mathbf{q}_k^{\mathrm{A}}, \mathbf{z}^{\mathrm{A}}(t)\rangle \langle \mathbf{q}_k^{\mathrm{D}}, \mathbf{z}^{\mathrm{D}}(t)\rangle\right]. \end{aligned} \tag{S1-42}$$

The sum over the samples is

$$\sum_{t=1}^{T} \langle \mathbf{q}_k^{\mathrm{A}}, \mathbf{z}^{\mathrm{A}}(t)\rangle^2 + \langle \mathbf{q}_k^{\mathrm{D}}, \mathbf{z}^{\mathrm{D}}(t)\rangle^2 - 2\rho_k \langle \mathbf{q}_k^{\mathrm{A}}, \mathbf{z}^{\mathrm{A}}(t)\rangle \langle \mathbf{q}_k^{\mathrm{D}}, \mathbf{z}^{\mathrm{D}}(t)\rangle,$$

which equals

$$T\left[\mathbf{q}_k^{\mathrm{A}\top}\hat{\mathbf{K}}_{\mathrm{A}}\mathbf{q}_k^{\mathrm{A}} + \mathbf{q}_k^{\mathrm{D}\top}\hat{\mathbf{K}}_{\mathrm{D}}\mathbf{q}_k^{\mathrm{D}} - 2\rho_k \mathbf{q}_k^{\mathrm{D}\top}\hat{\mathbf{K}}_{\mathrm{DA}}\mathbf{q}_k^{\mathrm{A}}\right],$$

where the matrices

$$\hat{\mathbf{K}}_{\mathrm{A}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{z}^{\mathrm{A}}(t)\mathbf{z}^{\mathrm{A}}(t)^\top, \qquad \hat{\mathbf{K}}_{\mathrm{D}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{z}^{\mathrm{D}}(t)\mathbf{z}^{\mathrm{D}}(t)^\top, \qquad \hat{\mathbf{K}}_{\mathrm{DA}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{z}^{\mathrm{D}}(t)\mathbf{z}^{\mathrm{A}}(t)^\top \tag{S1-43}$$

are the sample covariance matrices and the cross-correlation matrix of $\mathbf{z}^{\mathrm{D}}$ and $\mathbf{z}^{\mathrm{A}}$. The matrices $\hat{\mathbf{K}}_{\mathrm{A}}$ and $\hat{\mathbf{K}}_{\mathrm{D}}$ are the identity by the assumed preprocessing. Since $\mathbf{q}_k^{\mathrm{A}}$ and $\mathbf{q}_k^{\mathrm{D}}$ are the columns of an orthonormal matrix, we obtain for all k

$$\mathbf{q}_k^{\mathrm{A}\top}\hat{\mathbf{K}}_{\mathrm{A}}\mathbf{q}_k^{\mathrm{A}} = 1, \qquad \mathbf{q}_k^{\mathrm{D}\top}\hat{\mathbf{K}}_{\mathrm{D}}\mathbf{q}_k^{\mathrm{D}} = 1. \tag{S1-44}$$

We obtain (6) by plugging these relations into (S1-42),

$$f(\mathbf{q}_1^{\mathrm{A}}, \dots, \mathbf{q}_m^{\mathrm{D}}) \approx \mathrm{const} - \sum_{k=1}^{m}\frac{1}{1-\rho_k^2}\frac{1}{2}\left[2 - 2\rho_k \mathbf{q}_k^{\mathrm{D}\top}\hat{\mathbf{K}}_{\mathrm{DA}}\mathbf{q}_k^{\mathrm{A}}\right] \tag{S1-45}$$

$$\approx \mathrm{const} + \sum_{k=1}^{m}\frac{1}{1-\rho_k^2}\left[\rho_k \mathbf{q}_k^{\mathrm{D}\top}\hat{\mathbf{K}}_{\mathrm{DA}}\mathbf{q}_k^{\mathrm{A}}\right]. \tag{S1-46}$$

S2 Background Material from Multivariate Analysis

We present here background material from multivariate analysis that is relevant for this paper. In Section S2.1, we briefly review whitening and dimension reduction by principal component analysis (PCA). Section S2.2 is on regression and how to use it in order to determine the amount of dimension reduction. In Section S2.3, we review canonical correlation analysis. For more background than provided here, we refer the reader to Chapters 3 and 14 of [22].

S2.1 Whitening and PCA dimension reduction

Both whitening of zero mean $\mathbf{x}^{\mathrm{A}} \in \mathrm{R}^{n^{\mathrm{A}}}, \mathbf{x}^{\mathrm{D}} \in \mathrm{R}^{n^{\mathrm{D}}}$ and reducing their dimension by PCA to $m^{\mathrm{A}}, m^{\mathrm{D}}$ can be performed by

$$\mathbf{z}^{\mathrm{A}} = \mathbf{V}^{\mathrm{A}}\mathbf{x}^{\mathrm{A}}, \qquad \mathbf{z}^{\mathrm{D}} = \mathbf{V}^{\mathrm{D}}\mathbf{x}^{\mathrm{D}}, \tag{S2-1}$$

where $\mathbf{V}^{\mathrm{A}}$ and $\mathbf{V}^{\mathrm{D}}$ are $m^{\mathrm{A}} \times n^{\mathrm{A}}$ and $m^{\mathrm{D}} \times n^{\mathrm{D}}$ whitening matrices,

$$\mathbf{V}^{\mathrm{A}} = (\mathbf{D}^{\mathrm{A}})^{-1/2}(\mathbf{E}^{\mathrm{A}})^{\top}, \qquad \mathbf{V}^{\mathrm{D}} = (\mathbf{D}^{\mathrm{D}})^{-1/2}(\mathbf{E}^{\mathrm{D}})^{\top}. \tag{S2-2}$$

The diagonal matrices $\mathbf{D}^{\mathrm{A}}$ and $\mathbf{D}^{\mathrm{D}}$ contain the m^{A} and m^{D} largest eigenvalues of the covariance matrix of $\mathbf{x}^{\mathrm{A}}$ and $\mathbf{x}^{\mathrm{D}}$, respectively. The matrices $\mathbf{E}^{\mathrm{A}}$ and $\mathbf{E}^{\mathrm{D}}$ have as columns the corresponding eigenvectors. The (pseudo) inverses of $\mathbf{V}^{\mathrm{A}}$ and $\mathbf{V}^{\mathrm{D}}$ are the matrices $(\mathbf{V}^{\mathrm{A}})^{\dagger} = \mathbf{E}^{\mathrm{A}}(\mathbf{D}^{\mathrm{A}})^{1/2}$ and $(\mathbf{V}^{\mathrm{D}})^{\dagger} = \mathbf{E}^{\mathrm{D}}(\mathbf{D}^{\mathrm{D}})^{1/2}$, respectively.

S2.2 Regression to determine the degree of dimension reduction

For zero mean random variables, the linear prediction of $\mathbf{x}^{\mathrm{D}}$ from $\mathbf{x}^{\mathrm{A}}$ which minimizes the expected squared error is given by $\hat{\mathbf{x}}^{\mathrm{D}} = \mathbf{B}\mathbf{x}^{\mathrm{A}}$, with regression matrix $\mathbf{B}$,

$$\mathbf{B} = \mathrm{E}(\mathbf{x}^{\mathrm{D}}\mathbf{x}^{\mathrm{A}\top})\left[\mathrm{E}(\mathbf{x}^{\mathrm{A}}\mathbf{x}^{\mathrm{A}\top})\right]^{-1}, \tag{S2-3}$$

where we assume that the covariance matrix $\mathrm{E}(\mathbf{x}^{\mathrm{A}}\mathbf{x}^{\mathrm{A}\top})$ is invertible. If $\mathrm{E}(\mathbf{x}^{\mathrm{A}}\mathbf{x}^{\mathrm{A}\top})$ is invertible but badly conditioned, taking the inverse is a nonrobust operation. The matrix is badly conditioned if the components of $\mathbf{x}^{\mathrm{A}}$ are strongly correlated, so that only few data points lie outside a subspace of lower dimensionality than n^{A}. This means that only few data points determine the behavior of $\mathbf{B}$ outside that subspace. As a consequence, the variance of the prediction (the prediction error) can get large. Reducing the dimension of the data prior to the regression may reduce the prediction error. However, if too many dimensions are omitted the prediction error increases. There is thus an optimal amount of dimension reduction. It can be found empirically by comparing the prediction error for different numbers of retained dimensions.

In the main part of the paper, $\mathbf{x}^{\mathrm{A}}$ and $\mathbf{x}^{\mathrm{D}}$ are of the same dimensionality n. Since they show the same physical objects, we reduced the dimension of both data sets by the same amount using (S2-1), with $m^{\mathrm{A}} = m^{\mathrm{D}} = m$. The regression matrix for the whitened and dimension reduced data is the $m \times m$ matrix $\mathbf{K}_{\mathrm{DA}}$,

$$\mathbf{K}_{\mathrm{DA}} = \mathrm{E}(\mathbf{z}^{\mathrm{D}}\mathbf{z}^{\mathrm{A}\top})\left[\mathrm{E}(\mathbf{z}^{\mathrm{A}}\mathbf{z}^{\mathrm{A}\top})\right]^{-1} = \mathrm{E}(\mathbf{z}^{\mathrm{D}}\mathbf{z}^{\mathrm{A}\top}), \tag{S2-4}$$

which is the cross-correlation matrix between $\mathbf{z}^{\mathrm{D}}$ and $\mathbf{z}^{\mathrm{A}}$. Including the whitening matrices into the formula yields the rank m regression matrix $\mathbf{B}_m$ for the prediction of $\mathbf{x}^{\mathrm{D}}$ from $\mathbf{x}^{\mathrm{A}}$,

$$\mathbf{B}_m = (\mathbf{V}^{\mathrm{D}})^{\dagger}\mathbf{K}_{\mathrm{DA}}\mathbf{V}^{\mathrm{A}}. \tag{S2-5}$$

We measured the prediction error using the coefficient of determination R^2 on test data,

$$R^2(m) = 1 - \frac{\text{average squared prediction error of } \mathbf{x}^{\mathrm{D}} \text{ using } \mathbf{B}_m}{\text{total variance of } \mathbf{x}^{\mathrm{D}}}, \tag{S2-6}$$

and set the number of dimensions retained in (S2-1) to the value of m which minimized R^2.

S2.3 Canonical correlation analysis

Canonical correlation analysis (CCA) is a classical method to find related features in two data sets, that is, the matrices $\mathbf{Q}^{\mathrm{A}}$ and $\mathbf{Q}^{\mathrm{D}}$ in Figure 3. In CCA, related means correlated. After whitening and, possibly, dimension reduction, CCA rotates the individual coordinate systems of data $\mathbf{z}^{\mathrm{A}}$ and $\mathbf{z}^{\mathrm{D}}$ such that the corresponding coordinates s_k^{A} and s_k^{D} are maximally correlated. If $\mathbf{z}^{\mathrm{A}}$ and $\mathbf{z}^{\mathrm{D}}$ have dimensions m^{A} and m^{D}, respectively, CCA allows to find $m = \min(m^{\mathrm{A}}, m^{\mathrm{D}})$ related features. The features are found by the singular value decomposition of the cross-correlation matrix $\mathbf{K}_{\mathrm{DA}}$ between $\mathbf{z}^{\mathrm{D}}$ and $\mathbf{z}^{\mathrm{A}}$,

$$\mathbf{K}_{\mathrm{DA}} = \mathbf{Q}^{\mathrm{D}}\mathbf{S}(\mathbf{Q}^{\mathrm{A}})^{\top}. \tag{S2-7}$$

The matrix $\mathbf{S}$ is diagonal and contains the correlation coefficients between the canonical coordinates s_k^{A} and s_k^{D}. The $m^{\mathrm{D}} \times m$ and $m^{\mathrm{A}} \times m$ matrices $\mathbf{Q}^{\mathrm{D}}$ and $\mathbf{Q}^{\mathrm{A}}$ contain the features which have maximally correlated canonical coordinates. CCA is insensitive to statistical dependencies beyond correlation, both across and within the data sets. From Section S2.2, it follows that CCA is closely connected to linear regression.

S3 Calculations for the Noise-Distortion Analysis

This section contains the detailed calculations to derive the analytical expression in (7) for the squared prediction error $\mathrm{E}\,||\mathbf{s}^{\mathrm{D}} - \tilde{\mathbf{s}}^{\mathrm{D}}||^2$.

The squared error of the prediction is

$$\begin{aligned} \mathrm{E}\left(||\mathbf{s}^{\mathrm{D}} - \tilde{\mathbf{s}}^{\mathrm{D}}||^2\right) &= \mathrm{E}\,\Big(\sum_k (s_k^{\mathrm{D}} - \hat{s}_k^{\mathrm{D}} - \sigma(\hat{s}_k^{\mathrm{D}}) n_k)^2\Big) && \text{(S3-1)} \\ &= \sum_k \mathrm{E}\left[(s_k^{\mathrm{D}} - \hat{s}_k^{\mathrm{D}})^2 + \sigma^2(\hat{s}_k^{\mathrm{D}}) n_k^2 - 2(s_k^{\mathrm{D}} - \hat{s}_k^{\mathrm{D}})\sigma(\hat{s}_k^{\mathrm{D}}) n_k\right] && \text{(S3-2)} \\ &= \sum_k \mathrm{E}\left[(s_k^{\mathrm{D}} - \hat{s}_k^{\mathrm{D}})^2 + \sigma^2(\hat{s}_k^{\mathrm{D}}) n_k^2\right], && \text{(S3-3)} \end{aligned}$$

where the last equation follows from the zero mean and independence assumption for n_k. Using the definition of $\sigma^2(\hat{s}_k^{\mathrm{D}})$, and the independence assumption for n_k, we have

$$\mathrm{E}\left(||\mathbf{s}^{\mathrm{D}} - \tilde{\mathbf{s}}^{\mathrm{D}}||^2\right) = \sum_k \mathrm{E}\left[(s_k^{\mathrm{D}} - \hat{s}_k^{\mathrm{D}})^2\right] + F\,\mathrm{E}\left(|\hat{s}_k^{\mathrm{D}}|\right)\mathrm{E}(n_k^2). \tag{S3-4}$$

The variance of n_k is one so that

$$\mathrm{E}\left(||\mathbf{s}^{\mathrm{D}} - \tilde{\mathbf{s}}^{\mathrm{D}}||^2\right) = \mathrm{E}\left(||\mathbf{s}^{\mathrm{D}} - \hat{\mathbf{s}}^{\mathrm{D}}||\right) + F \sum_k \mathrm{E}\left(|\hat{s}_k^{\mathrm{D}}|\right). \tag{S3-5}$$

Using that $\hat{s}_k^{\mathrm{D}} = \varrho_k s_k^{\mathrm{A}}$, we obtain (7),

$$\mathrm{E}\left(||\mathbf{s}^{\mathrm{D}} - \tilde{\mathbf{s}}^{\mathrm{D}}||^2\right) = \mathrm{E}\left(||\mathbf{s}^{\mathrm{D}} - \hat{\mathbf{s}}^{\mathrm{D}}||^2\right) + F \sum_k |\varrho_k|\,\mathrm{E}\left(|s_k^{\mathrm{A}}|\right). \tag{S3-6}$$