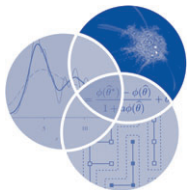
Focus Article



Tensor sufficient dimension reduction

Wenxuan Zhong,[1]* Xin Xing[1] and Kenneth Suslick[2]

Tensor is a multiway array. With the rapid development of science and technology in the past decades, large amount of tensor observations are routinely collected, processed, and stored in many scientific researches and commercial activities nowadays. The colorimetric sensor array (CSA) data is such an example. Driven by the need to address data analysis challenges that arise in CSA data, we propose a tensor dimension reduction model, a model assuming the nonlinear dependence between a response and a projection of all the tensor predictors. The tensor dimension reduction models are estimated in a sequential iterative fashion. The proposed method is applied to a CSA data collected for 150 pathogenic bacteria coming from 10 bacterial species and 14 bacteria from one control species. Empirical performance demonstrates that our proposed method can greatly improve the sensitivity and specificity of the CSA technique.

How to cite this article:
WIREs Comput Stat 2015, 7:178–184. doi: 10.1002/wics.1350

Keywords: tensor analysis; dimension reduction; iterative estimation; sliced inverse regression

INTRODUCTION

Tensor data analysis has a long history in social and behavioral sciences, and is now becoming a burgeoning research area in numerous other scientific disciplines, including life science, econometrics, chemistry as well as image and signal processing.[1,2] With the rapid development of image technology, the usage of tensor variate in quantitative models is becoming more and more popular. For example, tensor has been extensively used in understanding human brain's functional mechanism using fMRI technique.[3]

In general, a color image can be naturally digitalized as a random matrix (second-order tensor) with each row corresponding to a pixel's three color coordinates on Red Green and Blue axises where their values range from 0 to +255. When multiple replicates of a random matrix are observed, such as the EEG example used in Refs 4, 5, we can use the matrix observations for image classification and regression.

Though tensor analysis is more and more important, research on this area is still lacking due to the intrinsic mathematical complications. For example, tensor is not invariant under rotations, and spectral decomposition for a tensor has not been well established mathematically. The existing analytical models largely ignore the two-way tensor structure by simply vectorizing each 2-way observation into a vector and offering solutions using the vector-based statistical methods, such as linear regression analysis and linear discriminant analysis (LDA).[6] These solutions, however, are far from satisfactory. First, the simple vectorization destroys the original design information and leads to interpretation difficulties. Second, the simple vectorization significantly aggravates the curse of dimensionality, which refers to various difficulties a large number of variables (or dimensions) can cause to function approximation, model fitting, information extraction as well as to computation. For example, simple vectorization of a 36×3 matrix can generate 108 parameters for a conventional linear regression model, which renders the classical least square

*Correspondence to: wenxuan@uga.edu

[1]Department of Statistics, University of Georgia, Athens, GA, USA

[2]Department of Chemistry, University of Illinois, Urbana, IL, USA

Conflict of interest: The authors have declared no conflicts of interest for this article.

approach inappropriate for data with small sample size. Moreover, even if we have fairly large sample size, both the computational efficiency and the estimation accuracy of the classical vector-based analysis will be compromised by simple vectorization. There is an obvious pressing need for new statistical methods that can be used to incorporate intrinsic tensor structure.

Regression analysis is probably the most popular statistical tool for modeling the relationship between a response Y and a series of predictor variables X. In general, regression can be considered as an inference about the conditional distribution of Y given X, often with the mean response $E(Y|X)$ of particular interest. When X is a vector, various models and methods have been developed for regression analysis in the literature, ranging from classic linear regression to nonparametric regression. Among the large amount of literatures, there is a popular class of regression models, which assume that the response only depends on a lower dimensional projection of the predictors. To identify the lower dimensional projection is clearly critical in regression. A wide range of methods have been proposed to facilitate the estimation of the projection in the literature. For linear regression, two examples are principal component regression[7] and partial least squares regression,[8] both of which regress Y on some selected linear combinations of X (or components). For nonparametric regression, some examples include projection pursuit regression,[9] generalized additive models,[10] multivariate adaptive regression splines[11] and sufficient dimension reduction regression.[12] These methods use lower dimensional functions to approximate the relationship between Y and X. In this article, we follow the approach of sufficient dimension reduction and generalize our recent work on tensor classification[13] to tensor sufficient dimension reduction. Our proposed method is different from the dimension folding method as defined in Ref 4 or a modified version of it called tensor sliced inverse regression as defined in Ref 5. The previous mentioned methods target on a space that is larger than the dimension reduction subspace that is defined in our model.

MODEL DESCRIPTION

Sufficient Dimension Reduction in Vector Space

Let us first consider the vector-based dimension reduction model in $\mathbb{R}^p$. The generalization of the model to the tensor variates will be discussed later. Let $X \in \mathbb{R}^p$ be a random vector and $Y \in \mathbb{R}$ be a random scalar. Suppose S is a subspace of $\mathbb{R}^p$ and P_{S} is the projection operator from $\mathbb{R}^p$ to S in the standard inner product. If

$$Y \perp X | P_{\mathsf{S}} X, \tag{1}$$

where $\perp$ means 'independent of', then it is said that $P_{\mathsf{S}}X$ is sufficient for the dependence of Y on X. In other words, the projection $P_{\mathsf{S}}X$ captures all the information contained in X regarding Y. Model (1) was formally proposed by Ref 14 and was further discussed by Refs 12, 15. Model (1) is equivalent to several other formulations, such as the general index model

$$Y = h\left(\beta_1^\tau X, \cdots, \beta_d^\tau X, \varepsilon\right), \tag{2}$$

that was proposed in Ref 16. Here h is an unknown function, β_i's are p -dimensional vectors of unit length, d is an integer less than p, τ denotes transpose, ε is independent of X and $E(\varepsilon) = 0$. Given $(\beta_1^\tau X, \cdots, \beta_d^\tau X)$, Y and X are independent, therefore the subspace spanned by β_i's can serve as the subspace in model (1). Conversely, if (1) holds, then there exist h and ε such that (2) holds. A brief proof of the equivalence between the two models can be found in Ref 17.

Models (1) and (2) are referred to as the dimension reduction regression (DR) model and S is referred to as a dimension reduction subspace. Dimension reduction subspace may not be unique. Cook[15] introduced an important concept called *central space*, which is defined as the intersection of all dimension reduction subspaces when it is a dimension reduction subspace itself. The central subspace is denoted by $\mathsf{S}_{Y|X}$, and the dimension of $\mathsf{S}_{Y|X}$ is called the structural dimension of regressing Y on X. Under mild conditions, it can be shown that $\mathsf{S}_{Y|X}$ exists; see Ref 14 for details.

The DR model is a very general formulation and covers a wide range of parametric and semi-parametric models. For example, if Y is a discrete variable taking values in $\{1, 2, \cdots, K\}$, the DR model covers classification models. Let $P_{\mathsf{S}_{Y|X}}X$ be the projection operator from $\mathbb{R}^p$ to $\mathsf{S}_{Y|X}$. When X satisfies some mild conditions, such as the so-called linearity condition, i.e., $\mathrm{E}\left(\beta X | P_{\mathsf{S}_{Y|X}} X\right)$ is linear in $P_{\mathsf{S}_{Y|X}}X$, the $(\beta_1, \cdots, \beta_d)$ in model (2) can be obtained by recursively maximizing

$$L(\eta) = \max_T \mathrm{corr}^2\left(T(Y), \eta^\tau X\right), \tag{3}$$

subject to the constraint that $\beta_i var(X)\beta_j = I_{\{i=j\}}$, where $I_{\{\cdot\}}$ is the indicator function, $\eta \in \mathbb{R}^p$ and T are any possible transformations of Y including non-monotone ones.[18,19] In general, the $L(\eta)$ reflects the largest possible squared correlation between a transformed response $T(Y)$ and the projection $\eta^\tau X$. At the

population level,[18] showed that $L(\eta)$ has an explicit form

$$L(\eta) = \frac{\eta^\tau var[E(X|Y)]\eta}{\eta^\tau var(X)\eta} = \frac{\eta^\tau M\eta}{\eta^\tau \Sigma_X \eta}, \quad (4)$$

where $M \overset{\Delta}{=} var[E(X|Y)]$ and $\Sigma_X \overset{\Delta}{=} var(X)$. Therefore, $\beta_1, \cdots, \beta_d$ are the eigenvectors of $\Sigma_X^{-1}M$ corresponding to the largest d eigenvalues.

Based on the definition of M, a computationally stable and fast procedure called sliced inverse regression (SIR) was proposed in Ref 16 to generate an estimate of M, which is further used to generate an estimate of $P_{\mathsf{S}_{Y|X}}$. Observing $\{(\mathbf{x}_i, y_i)\}_{i=1,\ldots,n}$, where $\mathbf{x}_i \in \mathbb{R}^p$, the SIR algorithm is described as follows: (1) Divide the range of $\{y_i\}_{i=1,\ldots,n}$ into several disjoint intervals $I_1, \cdots, I_H$; (2) Estimate $E(X|Y \in I_h)$, $var(X)$ and M by their sample version $\overline{\mathbf{x}}_h = \frac{1}{n_h}\sum_{i=1}^n \mathbf{x}_i I_{\{y_i \in I_h\}}$, $\widehat{\Sigma}_X = \frac{1}{n}\sum_{i=1}^n (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})'$ and $\widehat{M} = \frac{1}{n}\sum_{h=1}^H n_h (\overline{\mathbf{x}}_i - \overline{\mathbf{x}})(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})'$ respectively, where $I_{\{y_i \in I_h\}}$ equals to 1 if $y_i \in I_h$ and 0 otherwise and $n_h = \sum_{i=1}^n I_{\{y_i \in I_h\}}$; (3) Apply the spectral decomposition to $\widehat{\Sigma}_X^{-1}\widehat{M}$ to obtain its eigenvalue-eigenvector pairs $\left(\widehat{\lambda}_i, \widehat{\beta}_i\right)$ where $\widehat{\lambda}_1 \leq \cdots \leq \widehat{\lambda}_d$. The $\widehat{\beta}_i$ is the SIR estimate of β_i and is referred to as the ith SIR direction. Empirical studies demonstrate that SIR is fairly successful in achieving dimension reduction for the high dimensional regression. The properties and asymptotic behaviors of SIR have been well studied in the literature, see Refs 20 and 21 among others.

Tensor DR and Central Tensor Dimension Reduction Set

When observation is a second or higher order tensor, we consider a tensor dimension reduction regression (TDR) model. To avoid the notation confusion, in the rest of the artical, we assume $\mathbf{X} \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_m}$ to be a mth-order tensor and vec($\mathbf{X}$) to be the vectorized $\mathbf{X}$. Let $\gamma_j = \beta_j^{(1)} \otimes \beta_j^{(2)} \otimes \cdots \otimes \beta_j^{(m)}$ be the kronecker product of vector $\beta_j^{(1)}, \cdots, \beta_j^{(m)}$ where $\beta_j^{(i)} \in \mathbb{R}^{p_i}$. Then, given the tensor predictor $\mathbf{X} \in \mathbb{R}^{p_1 \times \ldots, p_m}$ and the response $Y \in \mathbb{R}$, the TDR model is of the form

$$Y = g\left(\gamma_1^\tau \mathrm{vec}(\mathbf{X}), \cdots, \gamma_D^\tau \mathrm{vec}(\mathbf{X}), \varepsilon\right), \quad (5)$$

where g is an unknown function, γ_j for $j = 1, \cdots, D$ are regression indexes, and ε is a random error that is independent of vec($\mathbf{X}$). Clearly, TDR model assumes that Y is independent of vec($\mathbf{X}$) given $\left(\gamma_1^\tau \mathrm{vec}(\mathbf{X}), \ldots, \gamma_D^\tau \mathrm{vec}(\mathbf{X})\right)$. Let S be the linear space spanned by $\{\gamma_1, \ldots, \gamma_D\}$. With a little abuse of notation, we use S to denote the tensor dimension reduction space.

It is worth noting that not all the elements in S can be written in the form of a Kronecker product of vectors. For example, $\beta_1^{(1)} \otimes \beta_1^{(2)} + \beta_2^{(1)} \otimes \beta_2^{(2)} \in \mathsf{S}$ but cannot be expressed in a Kronecker product of any vectors in S_1 and S_2, unless $\beta_1^{(1)} = \beta_2^{(1)}$ or $\beta_1^{(2)} = \beta_2^{(2)}$. To keep the tensor structure and the interpretability of the indexes in model (5), we borrow the concept of decomposable tensor. A vector γ is said to be mth-order decomposable if it can be written in the form of $\eta_1 \otimes \eta_2 \otimes \cdots \otimes \eta_m$, where $\eta_i \in \mathbb{R}^{p_i}$ for $i = 1, \ldots, m$. Let $\mathbb{S}$ be the set that consists of all the decomposable tensors in S. It is easy to show that $\mathbb{S}$ exists and is unique if S exists and is unique. From now on, we define $\mathbb{S}$ as the tensor dimension reduction set (TDRS) and the intersection of all $\mathbb{S}$ that satisfy model (5) as the central tensor dimension reduction set (CTDRS). The rank of the CTDRS is referred to as the central dimension of regressing Y on X.

We also want to emphasize that model (5) is substantially different from model (2), although they look similar in expression. First, the TDR model can naturally alleviate the curse of dimensionality without increasing the estimation bias. For example, assuming $d = 1$, the index β_1 in model (2) with vec($\mathbf{X}$) has $p_1 p_2$ parameters, while γ_1 in model (5) has only $p_1 + p_2$ parameters. Second, the indexes in the TDR model are not necessarily the basis of the central subspace that is defined in (1) because of the decomposable constraint that we require on the indexes.

It is also important to know that model (5) is significantly different from the dimension folding model,[4] which assumes

$$Y \perp \mathrm{vec}(\mathbf{X}) | P_{\mathsf{S}_1} \otimes P_{\mathsf{S}_2} \otimes \cdots \otimes P_{\mathsf{S}_m} \mathrm{vec}(\mathbf{X}), \quad (6)$$

where S_i is the d_i dimensional sufficient dimension reduction space of $\mathbb{R}^{p_i}$ and P_{S_i} for $i = 1, \cdots, m$ are projection operator from $\mathbb{R}^{p_i}$ to S_i in a standard inner product. Let $\mathsf{S}_1 \otimes \cdots \otimes \mathsf{S}_m$ be the space that spanned by $\{v_1 \otimes \ldots \otimes v_m | \forall\, v_j \in S_j\}$. Then $\mathsf{S}_1 \otimes \cdots \otimes \mathsf{S}_m$ is referred to as the dimension folding subspace in Ref 4 and central tensor subspace in Ref 5. It's easy to see that S in model (5) is in general much smaller than $\mathsf{S}_1 \otimes \cdots \otimes \mathsf{S}_m$, i.e., $\mathsf{S} \subseteq \mathsf{S}_1 \otimes \cdots \otimes \mathsf{S}_m$. For example, for model $Y = g\left(\gamma_1^\tau \mathrm{vec}(\mathbf{X}), \gamma_2^\tau \mathrm{vec}(\mathbf{X}), \varepsilon\right)$, where $\gamma_1 = \beta_1^{(1)} \otimes \beta_1^{(2)}$ and $\gamma_2 = \beta_2^{(1)} \otimes \beta_2^{(2)}$, S is the space spanned by γ_1 and γ_2, while the central dimension folding subspace is spanned by $\left(\beta_1^{(1)} \otimes \beta_1^{(2)}, \beta_1^{(1)} \otimes \beta_2^{(2)}, \beta_2^{(1)} \otimes \beta_1^{(2)}, \beta_2^{(1)} \otimes \beta_2^{(2)}\right)$. Thus, model (5) can significantly improve the dimension folding model in terms of the downstream statistical

analysis, such as estimating the unknown function form g. Moreover, in order to obtain an initial estimate of $\mathsf{S}_1 \otimes \cdots \otimes \mathsf{S}_m$, dimension folding needs to estimate the inverse of $var\,(\mathrm{vec}(\mathbf{X}))$, which may not be estimable if the sample size is fairly small. This complication can be bypassed by using our sequential iterative dimension reduction algorithm provided below.

Sequential Iterative Dimension Reduction Algorithm

From now on, we focus on the estimate of the γ_j, $1 \le j \le D$ while leaving g unspecified. Let $\mathscr{R}$ be the set of all the mth-order decomposable tensors in $\mathbb{R}^{p_1} \otimes \cdots \otimes \mathbb{R}^{p_m}$. With a little abuse of notation, we let $L(\eta) = \max_T corr^2(T(Y), \eta^\tau \mathrm{vec}(\mathbf{X}))$. Then, similar to model (2),

$$\gamma_1 = \underset{\eta \in \mathscr{R}}{\mathrm{argmax}}\, L(\eta), \tag{7}$$

$$= \underset{\eta \in \mathscr{R}}{\mathrm{argmax}} \frac{\eta^\tau var\left[E\left(\mathrm{vec}\left(\mathbf{X}\right)|Y\right)\right]\eta}{\eta^\tau var\left(\mathrm{vec}\left(\mathbf{X}\right)\right)\eta}. \tag{8}$$

Comparing (8) to (4), we may naturally attempt the optimization of (8) using the tensor spectral decomposition as what we did for the vector predictors. However, tensor spectral decomposition, such as the PARAFAC[22] or Tucker model,[23,24] unlike its vector sibling, cannot provide us the maximizer of $L(\eta)$. Beyond this, the definition and algorithm of the tensor spectral decomposition are far from mature and have many intrinsic problems. For example, the orthogonality on each mode is not assumed and decomposition on the same mode is not unique using different algorithms.[2,25] Thus, we are keen on an alternative proposal as follows. To ease the presentation, we assume $m = 2$ in this session.

Recall that $\mathrm{vec}(\mathbf{X})$ is the vectorization of $\mathbf{X}$, where $\mathbf{X} \in \mathbb{R}^{p_2 \times p_1}$, and for any $\eta \in \mathscr{R}$, we have $\eta^\tau \mathrm{vec}(\mathbf{X}) = \left(\beta_1^{(2)^\tau} \mathbf{X} \beta_1^{(1)}\right)$, where $\beta_1^{(j)} \in \mathbb{R}^{p_j}$ for $j = 1$ and 2. Maximizing (8) is equivalent to maximizing

$$L\left(\beta_1^{(1)}, \beta_1^{(2)}\right) = \frac{var\left[E\left(\beta_1^{(2)^\tau} \mathbf{X} \beta_1^{(1)} | Y\right)\right]}{var\left(\beta_1^{(2)^\tau} \mathbf{X} \beta_1^{(1)}\right)}, \tag{9}$$

with respect to $\beta_1^{(1)}$ and $\beta_1^{(2)}$. Practically, the maximizer of (9) can be obtained by iteratively maximizing the following two functions

$$L_1(\eta) = \frac{\eta^\tau var\left[E\left(\mathbf{X}^\tau \beta_1^{(2)} | Y\right)\right]\eta}{\eta^\tau var\left(\mathbf{X}^\tau \beta_1^{(2)}\right)\eta}, \tag{10}$$

for given $\beta_1^{(2)}$ and,

$$L_2(\eta) = \frac{\eta^\tau var\left[E\left(\mathbf{X}\beta_1^{(1)} | Y\right)\right]\eta}{\eta^\tau var\left(\mathbf{X}\beta_1^{(1)}\right)\eta}, \tag{11}$$

for given $\beta_1^{(1)}$ respectively.

Observing $\{(\mathbf{X}_i, y_i)\}_{i=1,\ldots,n}$, where $\mathbf{X}_i \in \mathbb{R}^{p_1 \times p_2}$, we can estimate $var\left(\mathbf{X}^\tau \beta_1^{(2)}\right)$, $var\left(\mathbf{X}^{\beta_1^{(1)}}\right)$, $var\left[E\left(\mathbf{X}^\tau \beta_1^{(2)} | Y\right)\right]$ and $var\left[E\left(\mathbf{X}\beta_1^{(1)} | Y\right)\right]$ by their sample version in the following way. Let $\mathbf{z}_i^{(1)} = \mathbf{X}_i^\tau \beta_1^{(1)}$, $\mathbf{z}_i^{(2)} = \mathbf{X}_i \beta_1^{(2)}$, and divide the range of $\{y_i\}_{i=1,\ldots,n}$ into several disjoint intervals $I_1, \cdots, I_H$. Calculate $\overline{\mathbf{z}}_h^{(j)} = \frac{1}{n_h} \sum_{i=1}^n \mathbf{z}_i^{(j)} I_{\{y_i \in I_h\}}$ for $j = 1, 2$. Then, the sample version of $var\,(\mathbf{z}^{(j)})$ and $var\,(E(\mathbf{z}^{(j)}|Y))$ can be estimated by $\widehat{\Sigma}_j = \frac{1}{n} \sum_{i=1}^n \left(\mathbf{z}_i^{(j)} - \overline{\mathbf{z}}^{(j)}\right)\left(\mathbf{z}_i^{(j)} - \overline{\mathbf{z}}^{(j)}\right)'$ and $\widehat{M}_j = \frac{1}{n} \sum_{i=1}^H n_h \left(\overline{\mathbf{z}}_i^{(j)} - \overline{\mathbf{z}}^{(j)}\right)\left(\overline{\mathbf{z}}_i^{(j)} - \overline{\mathbf{z}}^{(j)}\right)'$ for $j = 1, 2$.

Notice that equation (11) is a quadratic form on the vector space $\mathbb{R}^{p_2}$. Thus, if $\beta_1^{(1)}$ is given, $\beta_1^{(2)}$ can be obtained by SIR, and vise versa. Thus γ_1 can be obtained by iteratively maximize (10) and (11) respectively. Notice that $L_j(\cdot)$ for $j = 1, 2$ are bounded above by one and the iterative maximization approach ensures the nondecreasing of $L(\cdot)$ in each iteration. Thus the convergence of iteration is guaranteed.

More generally, given $B_k = (\gamma_1, \cdots, \gamma_k)$, we can estimate γ_{k+1} in a sequential way. Let $P_k = \Sigma_{\mathrm{vec}(\mathbf{X})}^{\frac{1}{2}} B_k \left(B_k^\tau \Sigma_{\mathrm{vec}(\mathbf{X})} B_k\right)^{-1} B_k^\tau \Sigma_{\mathrm{vec}(\mathbf{X})}^{\frac{1}{2}}$ be the projection matrix from $\mathbb{R}^{p_2} \otimes \mathbb{R}^{p_1}$ onto the space that is spanned by B_k, say S_{B_k}, with respect to $\Sigma_{\mathrm{vec}(\mathbf{X})}$, where $\Sigma_{\mathrm{vec}(\mathbf{X})}$ is defined as $var\,(\mathrm{vec}(\mathbf{X}))$. Let $\mathrm{vec}(\mathbf{X}^{(k)})$ be the projection of $\mathrm{vec}(\mathbf{X})$ in the complementary space of S_{B_k}, i.e., $\mathrm{vec}(\mathbf{X}^{(k)}) = (I - P_k)\mathrm{vec}(\mathbf{X})$. Let $\mathscr{R}_k$ denote the set of all decomposable tensors of $\mathbb{R}^{p_2} \otimes \mathbb{R}^{p_1}$ that are orthogonal to S_{B_k}. Then we have

$$-\gamma_{k+1} = \underset{\eta \in \mathscr{R}_k}{\mathrm{argmax}} \frac{\eta^\tau \mathrm{var}\left[\mathrm{E}\left(\mathrm{vec}\left(\mathbf{X}^{(k)}\right)|Y\right)\right]\eta}{\eta^\tau \mathrm{var}(\mathrm{vec}\left(\mathbf{X}^{(k)}\right)\eta}. \tag{12}$$

The observation of $\mathrm{vec}(\mathbf{X}^{(k)})$ can be obtained by $(I - P_k)\mathrm{vec}(\mathbf{X})_i$. Clearly, γ_{k+1} can be estimated in the same fashion as γ_1. For given D, the SIDRA algorithm is summarized in the flowchart in Figure 1. In general, D can be determined by a χ^2 test. The details of this test are beyond the scope of this paper.

CASE STUDY

Rapid and accurate detection of pathogenic bacteria is important not only for containing its potential spread,

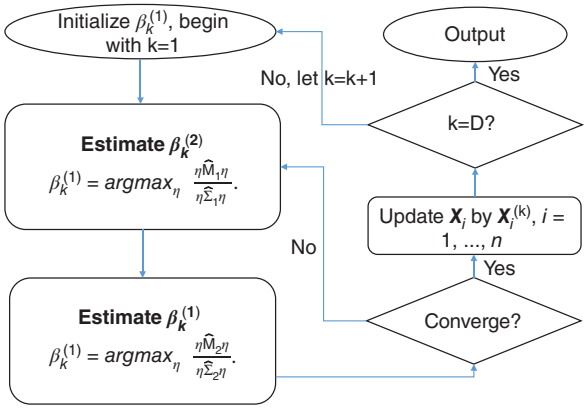


FIGURE 1 | The flowchart of SIDRA algorithm.

but also for determining potential medical remedies. Existing bacterial identification methods are severely limited by the necessity of long culturing times, the need for highly trained laboratory personnel, and the requirement of expensive and high-maintenance equipment. There is a pressing need for a simple but powerful technology for pathogenic bacterial detection and identification. Motivated by the fact that human pathogenic bacteria can be discriminated from the volatile metabolites (VMs) they produce, we have invented a chemical sensor: colorimetric sensor arrays (CSA).

CSA is essentially a chemically engineered chip with 36 chemo-responsive dyes printed in a 6×6 array, where each dye is essentially a 'litmus paper' that can change color when interacting with certain chemical volatile. The composite color difference of the 36 dyes measured before and after exposure to a certain volatile can be recorded as a multiway array and will serve as the volatile's unique fingerprint. Plotted in Figure 2 are the chemo images for ten pathogenic bacteria collected using the CSA technology.

A bacterial color difference map can be digitalized as a 36×3 matrix (2nd-order tensor), where each row represents the color change of one of the 36 chemo-responsive dyes and each column represents its color coordinates on the RGB axis of a color cube respectively. In our preliminary studies, the color difference maps of 150 pathogenic bacteria from 10 bacterial species and 14 bacteria from one control species have been collected every 30 min using CSA technology with a time range from 120 to 600 min. For each bacterium, a $36 \times 3 \times 16$ array is generated with the first dimension being the dye effect, second dimension being the RGB-color effect and the third dimension being the time effect. Because, the change of color for the same VMs is not continuous as the time increases, we treat the time effect as a multivariate vector rather than a continuous variable.

Plotted in Figure 3(a) is the projection of the 164 color difference maps on the first and second directions obtained using SIDRA method. Though, some of the bacteria such as *E. nterococcusaecalis* and *Enterococcus faecium* cannot be clearly separated using two directions, most bacteria can be well separated. To further evaluate the accuracy of our algorithm in image recognition, we separate the data by randomly sampling one difference map from each class to form the testing sample and use the rest difference maps to form the training sample. SIDRA is applied to the training data and the misclassification error is calculated using the testing data. The procedure is repeated 50 times. In Figure 3(b), we plot the average prediction error of the testing set along with different number of directions. The prediction error rate is 2.55% with four directions and below 1% with 7 or more directions. We also find that including more than seven directions

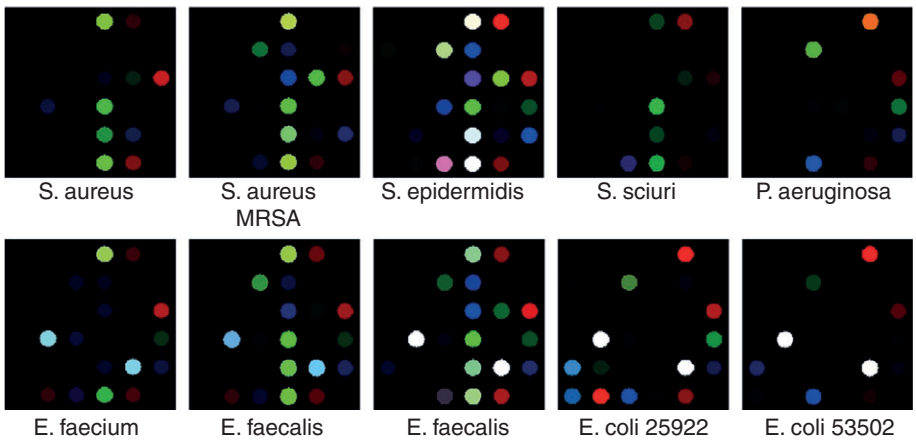


FIGURE 2 | Digital images of 10 pathogenic bacteria at full vapor pressure at 300 K.

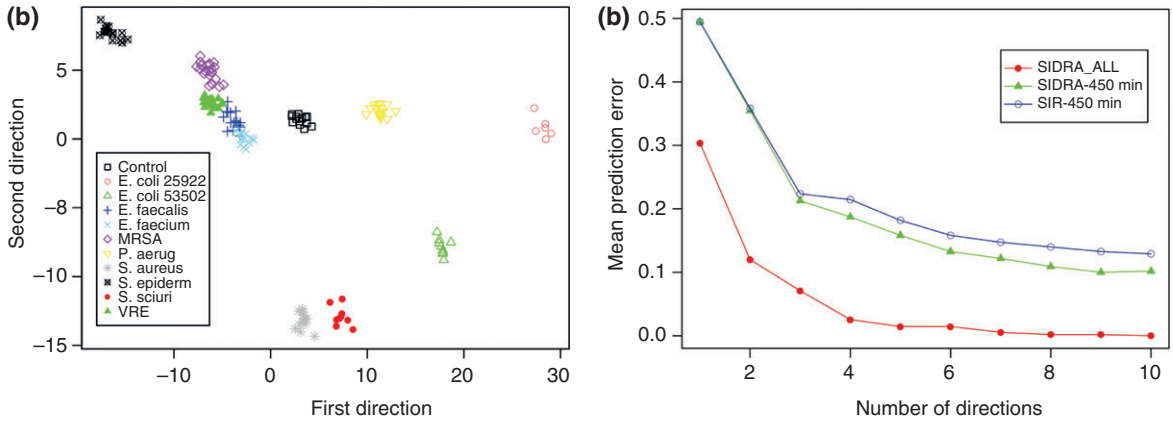


FIGURE 3 | Plotted in the left panel is the projection of 150 pathogenic bacteria on the first two dimension reduction directions. In the right panel, we compare the prediction error of SIR-450 min (blue line), SIDRA-450 min (green line) using data collected 450 min after exposing colorimetric sensor array (CSA) to the bacteria and SIDRA_ALL (red line) using all the time points from 120 min to 600 min, where measurement was taken every 30 min.

does not increase the prediction accuracy significantly, which implies that the 164 bacteria can be separated in a seven dimensional subspace. Our preliminary studies on 10 strains of bacteria, including *E. faecalis* and *Staphylococcus aureus* and their antibiotic resistant forms, demonstrate that CSA can clearly discriminate the 10 strains with 98.8% accuracy within 10 h, a clinically important timeframe.

This study supports our hypothesis that the bacteria produce VMs that can be well detected using CSA. It is worth noting that existing vector-based discriminant analysis is not applicable for this data if we consider all the time points, because the data has a severe $p \gg n$ problem (36 dyes×3RGB-color×17 time points = 1836 parameters $\gg$ sample size 164). To compare SIR and SIDRA method in bacterial detection, we compute the misclassification error of SIR and SIDRA only at the 450-min. It is easy to see that MDA consistently outperform SIR for all low dimensional projections.

The initial studies reported herein are being used to guide development of arrays with greater sensitivity and classification capabilities for bacteria. These arrays are currently being applied to blood culture systems using liquid growth media with very low initial inoculum concentrations. In addition, we are applying the same technology for rapid diagnosis of fungal infections, which too has become increasingly important biomedically, as the recent debacle over spinal injectable steroids has demonstrated.

DISCUSSION

Like all the iterative estimation approaches, the SIDRA procedure proposed in this article may encounter issues typical to all the other iterative estimation approaches. One major limitation of the iterative estimation approach is that the estimate may practically be attracted to a local optimal estimate and fail to reach the global optimal estimate. To solve this issue, we advocate trying multiple starting points for iteration and choosing the best estimate. The advantage of this approach is currently under investigation and the simulation results will be reported in the future publication.

ACKNOWLEDGMENTS

This research was supported by NIH R01 GM113242, NSF DMS 1440038 and DMS 1120368 to WZ.

REFERENCES

1. Kroonenberg PM. *Applied Multiway Data Analysis*, vol. 702. Hoboken, NJ: John Wiley & Sons; 2008.
2. Smilde A, Bro R, Geladi P. *Multi-Way Analysis: Applications in the Chemical Sciences*. West Sussex, England: John Wiley & Sons; 2005.
3. Zhou H, Li L. Regularized matrix regression. *J R Stat Soc Ser B Stat Methodol* 2014, 76:463–483.
4. Li B, Kim MK, Altman N. On dimension folding of matrix-or array-valued statistical objects. *Ann Stat* 2010, 38:1094–1121.

5. Ding S, Cook RD. Tensor sliced inverse regression. *J Multivar Anal* 2015, 133:216–231.
6. Friedman JH. Regularized discriminant analysis. *J Am Stat Assoc* 1989, 84:165–175.
7. Jolliffe IT. A note on the use of principal components in regression. *Appl Stat* 1982, 31:300–303.
8. Wold S, Sjöström M, Eriksson L. Pls-regression: a basic tool of chemometrics. *Chemom Intell Lab Syst* 2001, 58:109–130.
9. Friedman JH, Stuetzle W. Projection pursuit regression. *J Am Stat Assoc* 1981, 76:817–823.
10. Hastie TJ, Tibshirani RJ. *Generalized Additive Models*, vol. 43. Boca Raton, FL: CRC Press; 1990.
11. Friedman JH. Multivariate adaptive regression splines. *Ann Stat* 1991, 19:1–67.
12. Cook RD. *Regression Graphics: Ideas for Studying Regressions Through Graphics*, vol. 482. New York, NY: John Wiley & Sons; 2009.
13. Zhong W, Suslick K. Penalized classification for matrix predictors with application to colorimetric sensor arrays. *Technometrics* 2014. doi: 10.1080/00401706.2014.965347.
14. Cook RD, Weisberg S. *An Introduction to Regression Graphics*, vol. 405. New York, NY: John Wiley & Sons; 2009.
15. Cook RD. Graphics for regressions with a binary response. *J Am Stat Assoc* 1996, 91:983–992.
16. Li K-C. Sliced inverse regression for dimension reduction. *J Am Stat Assoc* 1991, 86:316–327.
17. Zhu Y, Zeng P. Fourier methods for estimating the central subspace and the central mean subspace in regression. *J Am Stat Assoc* 2006, 101:1638–1651.
18. Chen C-H, Li K-C. Can SIR be as popular as multiple linear regression? *Stat Sinica* 1998, 8:289–316.
19. Zhong W, Zhang T, Zhu Y, Liu JS. Correlation pursuit: forward stepwise variable selection for index models. *J R Stat Soc Ser B Stat Methodol* 2012, 74:849–870.
20. Duan N, Li K-C. Slicing regression: a link-free regression method. *Ann Stat* 1991, 19:505–530.
21. Hsing T, Carroll RJ. An asymptotic theory for sliced inverse regression. *Ann Stat* 1992, 20:1040–1061.
22. Harshman RA, Lundy ME. The PARAFAC model for three-way factor analysis and multidimensional scaling. In: Law HG, Snyder CW Jr., Hattie J, McDonald RP, eds. *Research Methods for Multimode Data Analysis*. New York: Praeger; 1984, 122–215.
23. Tucker LR. Some mathematical notes on three-mode factor analysis. *Psychometrika* 1966, 31:279–311.
24. Tucker LR. A method for synthesis of factor analysis studies. Technical report, DTIC Document; 1951.
25. Kolda TG, Bader BW. Tensor decompositions and applications. *SIAM Rev* 2009, 51:455–500.

