Covariance-free Partial Least Squares:
An Incremental Dimensionality Reduction Method

Artur Jordao, Maiko Lie, Victor Hugo Cunha de Melo and William Robson Schwartz
Smart Sense Laboratory, Computer Science Department
Universidade Federal de Minas Gerais, Brazil
Email: {arturjordao, maikolie, victorhcmelo, william}@dcc.ufmg.br

Abstract

Dimensionality reduction plays an important role in computer vision problems since it reduces computational cost and is often capable of yielding more discriminative data representation. In this context, Partial Least Squares (PLS) has presented notable results in tasks such as image classification and neural network optimization. However, PLS is infeasible on large datasets (e.g., ImageNet) because it requires all the data to be in memory in advance, which is often impractical due to hardware limitations. Additionally, this requirement prevents us from employing PLS on streaming applications where the data are being continuously generated. Motivated by this, we propose a novel incremental PLS, named Covariance-free Incremental Partial Least Squares (CIPLS), which learns a low-dimensional representation of the data using a single sample at a time. In contrast to other state-of-the-art approaches, instead of adopting a partially-discriminative or SGD-based model, we extend Nonlinear Iterative Partial Least Squares (NIPALS) — the standard algorithm used to compute PLS — for incremental processing. Among the advantages of this approach are the preservation of discriminative information across all components, the possibility of employing its score matrices for feature selection, and its computational efficiency. We validate CIPLS on face verification and image classification tasks, where it outperforms several other incremental dimensionality reduction methods. In the context of feature selection, CIPLS achieves comparable results when compared to state-of-the-art techniques.

1. Introduction

Dimensionality reduction is widely used in computer vision applications from image classification \cite{11,2} to neural network optimization \cite{3}. The idea behind this technique is to estimate a transformation matrix that projects the high-dimensional feature space onto a low-dimensional latent space \cite{20,7}. Previous works have demonstrated that dimensionality reduction can improve not only computational cost but also the effectiveness of the data representation \cite{18} \cite{31} \cite{29}. In this context, Partial Least Squares (PLS) has presented remarkable results when compared to other dimensionality reduction methods \cite{29}. This is mainly due to the criterion through which PLS finds the low dimensional space, which is by capturing the relationship between independent and dependent variables. Another interesting aspect of PLS is that it can operate as a feature selection method, for instance, by employing Variable Importance in Projection (VIP) \cite{21}. The VIP technique employs score matrices yielded by NIPALS (the standard algorithm used for traditional PLS) to compute the importance of each feature based on its contribution in the generation of the latent space.

Despite achieving notable results, PLS is not suitable for large datasets (e.g., ImageNet \cite{5}) since it requires all the data to be in memory in advance, which is often impractical due to hardware limitations. Additionally, this requirement prevents us from employing PLS on streaming applications, where the data are being generated continuously. Such limitation is not particular to PLS, many dimensionality reduction methods, such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA), also suffer from this problem \cite{32} \cite{2}.

To handle the aforementioned problem, many works have proposed incremental versions of traditional dimensionality reduction methods. The idea behind these methods is to estimate the projection matrix using a single data sample (or a subset) at a time while keeping some properties of the traditional dimensionality reduction methods. A well-known class of incremental methods is the one based on Stochastic Gradient Descent (SGD) \cite{3} \cite{2}. These methods interpret dimensionality reduction as a stochastic optimization problem of an unknown distribution. As shown by Weng et al. \cite{32}, incremental methods based on SGD are computationally expensive, present convergence problems and require many parameters that depend on the nature of
the data. To address this problem, Zeng et al. [34] proposed an efficient and low-cost incremental PLS (IPLS). In their work, the first dimension (component) of the latent space is found incrementally, while the other dimensions are estimated by projecting the first component onto the reconstructed covariance matrix, which is employed to address the issue of impractical memory requirements of a full covariance matrix.

Even though IPLS achieves better performance than SGD-based and other state-of-the-art incremental methods, the discriminability of its higher-order components (i.e., all except the first) is not preserved, as shown in Figure 1(a). This behavior appears because the higher-order components are estimated using only the independent variables, that is, it is based on an approximation of the covariance matrix \( X^\top X \) (similar to PCA) instead of \( X^\top Y \) employed in PLS. This can degrade the discriminability of the latent model since preserving the relationship between independent and dependent variables is an important property of the original PLS [7]. It is important to emphasize that, for high-dimensional data, employing several components often provides better results [29] [9] [10] [17], hence, IPLS might not be suitable on these cases.

Motivated by limitations and drawbacks in incremental PLS-based approaches, we propose a novel incremental method \(^1\). Our method relies on the hypothesis that the estimation of higher-order components using the covariance matrix, as proposed by Zeng et al. [34], is inadequate since the relationship between independent and dependent variables is lost. Therefore, to preserve this characteristic, we extend NIPALS \(^1\) to avoid the computation of \( X^\top Y \) and, consequently, enable it for incremental operation. Since our proposed extension is based on a simple algebraic decomposition, we preserve the simplicity and efficiency that makes NIPALS popular, and we ensure that the relationship between independent and dependent variables is propagated to all components, differently from other methods.

As shown in Figure 1, our method is capable of separating data classes better than IPLS, mainly on the second component (i.e., y-axis). Since the proposed method does not use the covariance matrix \( X^\top X \) to estimate higher-order components, we refer to it as Covariance-free Incremental Partial Least Squares (CIPLS). Besides providing superior performance, our method can easily be extended as a feature selection technique since it provides all the requirements to execute VIP. Existing incremental PLS methods, on the other hand, require more complex techniques to operate as feature selection [21].

We compare the proposed method on the tasks of face verification and image classification, where it outperforms several other incremental methods in terms of accuracy and efficiency. In addition, in the context of feature selection, we evaluate and compare the proposed method to state-of-the-art methods, where it achieves competitive results.

2. Related Work

To enable PCA to operate in an incremental scheme, Weng et al. [32] proposed to compute the principal components without estimating the covariance matrix, which is unknown and impossible to be calculated in incremental methods. For this purpose, their method (CCIPCA) updates the projection matrix for each sample \( x \), replacing the unknown covariance matrix by the sample covariance matrix \( (xx^\top) \). While CCIPCA provides a minimum reconstruction error of the data, it might not improve or even preserve the discriminability of the resulting subspace since label information is ignored (similarly to traditional PCA) [20].

To achieve discriminability, incremental methods based on Linear Discriminant Analysis (LDA) have been proposed [12] [19]. In particular, this class of methods is less
explored since they present issues such as the sample size problem [13], which makes them infeasible for some tasks. Different from incremental LDA methods, incremental PLS methods are more flexible and present better results [32]. Motivated by this, Arora et al. [3] proposed an incremental PLS based on stochastic optimization (SGDPLS), where the idea is to optimize an objective function using a single sample at a time. Similarly to Arora et al. [3], Stott et al. [30] proposed applying stochastic gradient maximization on NIP ALS, extending it for incremental processing. Even though they present promising results on synthetic data, their approach presented convergence problems when evaluated on real-world datasets. Thus, in this work, we consider only the approach by Arora et al. [3], which was the one that converged for several of the datasets evaluated and presented better results.

While SGDPLS is effective, as demonstrated by Weng et al. [32] and Zeng et al. [34], SGD-based methods applied to dimensionality reduction are computationally expensive and present convergence problems. In addition, this class of approaches requires careful parameter tuning and their results are often sensitive to the type of dataset [32].

To address convergence problems in SGD-based PLS, Zeng et al. [34] proposed to decompose the relationship between independent and dependent matrices (variables) into a sample relationship (i.e., a single sample with its label). This process is performed only to compute the first component, the higher-order components are estimated by projecting the first component onto an approximated covariance matrix using a few PCA components. As we mentioned earlier, since traditional PCA cannot be employed in incremental methods, Zeng et al. [34] used CCIPCA to reconstruct the principal components of the covariance matrix.

In contrast to existing incremental PLS methods, our method presents superior performance in both accuracy and execution time for estimation of the projection matrix, method presents superior performance in both accuracy and execution time for estimation of the projection matrix. As we mentioned earlier, since traditional PCA cannot be employed in incremental methods, Zeng et al. [34] used CCIPCA to reconstruct the principal components of the covariance matrix.

In this section, we start by describing the traditional Partial Least Squares (PLS). Then, we present the proposed Covariance-free Incremental Partial Least Squares (CIPLS) and the Variable Importance in Projection (VIP) technique, which enables PLS and CIPLS to be employed for feature selection. Unless stated otherwise, let $X \in \mathbb{R}^{n \times m}$ be the matrix of independent variables denoting $n$ training samples in a $m$-dimensional space. Furthermore, let $Y \in \mathbb{R}^{n \times 1}$ be the matrix of dependent variables representing the binary class label. Finally, let $x_n \in \mathbb{R}^{1 \times m}$ and $y_n \in \mathbb{R}^{1 \times 1}$ be a single sample of $X$ and $Y$, respectively. We highlight that, in the context of streaming data, $x_n$ is a data sample acquired at time $n$.

### 3. Proposed Approach

In this section, we start by describing the traditional Partial Least Squares (PLS). Then, we present the proposed Covariance-free Incremental Partial Least Squares (CIPLS) and the Variable Importance in Projection (VIP) technique, which enables PLS and CIPLS to be employed for feature selection. Unless stated otherwise, let $X \in \mathbb{R}^{n \times m}$ be the matrix of independent variables denoting $n$ training samples in a $m$-dimensional space. Furthermore, let $Y \in \mathbb{R}^{n \times 1}$ be the matrix of dependent variables representing the binary class label. Finally, let $x_n \in \mathbb{R}^{1 \times m}$ and $y_n \in \mathbb{R}^{1 \times 1}$ be a single sample of $X$ and $Y$, respectively. We highlight that, in the context of streaming data, $x_n$ is a data sample acquired at time $n$.

#### 3.1. Partial Least Squares

Given a high $m$-dimensional space, PLS finds a projection matrix $W(w_1, w_2, ..., w_c)$, which projects this space onto a low $c$-dimensional space, where $c \ll m$. For this purpose, PLS aims at maximizing the covariance between the independent and dependent variables. Formally, PLS constructs $W$ such that

$$w_i = \maximize(\text{cov}(Xw, Y)), \text{s.t.} \|w\| = 1, \quad (1)$$

where $w_i$ denotes the $i$th component of the $c$-dimensional space. The exact solution to Equation 1 is given by

$$w_i = \frac{X^\top Y}{\|X^\top Y\|}. \quad (2)$$

From Equation 2 we can compute all the $c$ components using either Nonlinear Iterative Partial Least Squares (NIPALS) 1 or Singular Value Decomposition (SVD). Most works employ NIPALS since it is capable of finding only the $c$ first components, while SVD always finds all the $m$ components, being computationally inefficient compared to NIPALS 1.
3.2. Covariance-free Incremental PLS

The core idea in our method is to ensure that, as in traditional PLS, the relationship between independent and dependent variables (Equation 2) is kept on all the c components. To achieve this goal, our method works as follows. First, we need to center the data to the mean of the training samples. To achieve this goal, our method works as follows. To face this problem, we center the current data sample using an approximate centralization process [32], which consists of estimating an incremental mean using the nth sample. According to Weng et al. [32], we can compute the incremental mean µn w.r.t. the nth data sample as

\[
\mu_n = \frac{n-1}{n}\mu_{(n-1)} + \frac{1}{n}x_n. \tag{3}
\]

Once we have centralized the sample, the next step in our method is to compute the component w_i following Equation 2. As we mentioned, X and its respective Y are unknown or are not in memory in advance, which prohibits us to apply Equation 2 directly. However, as suggested by Zeng et al. [34], we employ the following decomposition:

\[
X^TY = \sum_{k=1}^{n-1} x_ky_k + x_ny_n. \tag{4}
\]

By replacing \(X^TY\) in Equation 2 by Equation 4, it is possible to calculate the ith component of PLS considering a single sample at a time. In other words, Equation 4 enables us to apply Equation 2 directly. However, as suggested by Zeng et al. [34], we employ the following decomposition:

\[
X^TY = \sum_{k=1}^{n-1} x_ky_k + x_ny_n. \tag{4}
\]

To compute the higher-order components (w_i, i > 1), we employ a deflation process, which consists of subtracting the contribution of the current component on the sample before estimating the next component. Following the NIPALS algorithm, the deflation process works as follows

\[
t = Xw_i, \tag{5}
\]

\[
p = X^Tt, q = Y^Tt, \tag{6}
\]

\[
X = X - tp^T, Y = Y - tq^T, \tag{7}
\]

where t denotes the projected samples onto the current component w_i, and p and q represent the loadings of this projection. It should be noted that while t works in an incremental scheme (since we can project one sample at a time), p and q cannot be computed since X and Y are neither known nor are in memory in advance. However, in light of Equation 4, we can decompose p and q as

\[
p = \sum_{k=1}^{n-1} x_kt_k + x_nt_n, q = \sum_{k=1}^{n-1} y_kt_k + y_nt_n. \tag{8}
\]

Algorithm 1: CIPLS Algorithm.

| Input | nth data sample \(x_n\) and its label \(y_n\) |
|-------|------------------------------------------|
|       | Number of components c                     |
|       | Projection matrix \(W_{(n-1)} \subset \mathbb{R}^{c \times m}\) |
|       | Loading matrix \(P_{(n-1)} \subset \mathbb{R}^{c \times m}\) |
|       | Loading matrix \(Q_{(n-1)} \subset \mathbb{R}^{c \times 1}\) |
| Output| Updated \(W\), \(P\) and \(Q\)            |

1. Update \(\mu_n\) using Equation 3
2. \(\bar{x}_n = x_n - \mu_n\)
3. for \(i = 1\) to \(c\) do
   4. \(w_i = \bar{x}_ny_n + w_{i(n-1)}\), where \(w_i \in W\)
   5. \(t_n = \frac{\bar{x}_n}{\|\bar{x}_n\|}\)
   6. \(p_i = \bar{x}_nt_n + p_{i(n-1)}\), where \(p_i \in P\)
   7. \(q_i = y_n t_n + q_{i(n-1)}\), where \(q_i \in Q\)
   8. \(\bar{x}_n = \bar{x}_n - t_n p_i^T\)
   9. \(y_n = y_n - t_n q_i^T\)
10. end

By embedding Equation 8 on the deflation process, we can remove the contribution of the current component and repeat the process to compute a single component \(w_i\). Observe that Equation 7 deflates each sample by its reconstructed value, therefore, Equation 7 can be computed sample-by-sample, working in an incremental scheme. With this formulation, we are now capable of computing the \(c\) components incrementally. Algorithm 1 summarizes the steps of the proposed method. It should be mentioned that the matrices \(W\), \(P\) and \(Q\) are initialized with zeros.

According to Algorithm 1, the proposed method maintains the property of capturing the relationship between \(X\) and \(Y\) for all components (step 4 in Algorithm 1). In addition, since we compute all components at once, our method has a time complexity of \(O(ncm)\), where \(n\), \(c\) and \(m\) denote the number of samples, number of components, and dimensionality of the data, respectively.

3.3. CIPLS for Feature Selection

An advantage of PLS is that, after estimating the projection matrix \(W\), it is possible to estimate the importance of each feature, enabling PLS to operate as a feature selection method. For this purpose, it is possible to employ Variable Importance in Projection (VIP), which estimates the importance of each feature \(f_j\) w.r.t. its contribution to yield the low dimensional space. According to [21], VIP is defined as

\[
f_j = \sqrt{\sum_{i=1}^{c} q_i^2 t_i^T t_i (w_{ij} / \|w_i\|^2)} / \sum_{i=1}^{c} q_i^2 t_i^T t_i. \tag{9}
\]
Once we have estimated the score of each feature, we can remove a percentage of features based on their scores. As can be verified in Algorithm 1, CIPLS preserves the ability of traditional PLS to be employed as a feature selection method via VIP (Equation 9). On the other hand, it is important to emphasize that IPLS and SGDPLS cannot be used to compute VIP since they do not provide the loading matrix $Q (q_1, q_2, ..., q_c)$.

4. Experimental Results

In this section, we first introduce the experimental setup and the tasks employed to validate the proposed method. Then, we present the procedure conducted to calibrate the parameters of the methods. Next, we compare the proposed method with other incremental partial least squares methods as well as with the traditional PLS. Afterwards, we present the influence of higher-order components on the classification performance. Finally, we discuss the time complexity of the methods, their performance on a streaming scenario and compare our method on the feature selection context.

Experimental Setup. Throughout the experiments, we use a linear SVM for binary classification (face verification) because, according to Zeng et al. [34], a linear SVM coupled with dimensionality reduction is able to achieve remarkable results while being computationally efficient. In addition, it has been shown that simple classifiers when fed by features from convolutional networks are able to achieve results comparable to more sophisticated classifiers [6] [25]. For multi-class problems (image classification), on the other hand, we prefer to use a multilayer perceptron because it handles the multi-class problem naturally, avoiding the need for employing a binary classifier on a one-versus-rest fashion, which would be computationally expensive. All experiments and methods were executed on an Intel Core i5-8400, 2.4 GHz processor with 16 GB of RAM.

To assess the differences in efficacy and efficiency among the compared methods, throughout the experiments we follow the approach by Jain et al. [15] and perform statistical tests based on a paired t-test using 95% of confidence. We highlight that the statistical tests were conducted only for face verification due to the computational cost of retraining (i.e., fine-tuning) the convolutional neural network for image classification, which is considerably high since we employ large-scale datasets in our assessment.

Face Verification. Given a pair of face images, face verification determines whether this pair belongs to the same person. For this purpose, we use a three-stage pipeline [24] [4], which works as follows. First, we extract a feature vector of each face using a deep learning model. In this work, we use the feature maps from the last convolutional layer of the VGG16 model, learned on the VGGFaces dataset [23], as feature vector. Then, we compute the distance between the two feature vectors employing the $\ell_1$-distance metric and present the result of the distance metric to a classifier.

We conduct our evaluation on two face verification datasets, namely Labeled Faces in the Wild (LFW) [14] and Youtube Faces (YTF) [33].

Image Classification. Image classification consists of deciding to which one of a given set of categories an image belongs. Traditionally, this is done by extracting features from the samples and feeding these features to a classifier, which determines the category to which each image belongs. For this purpose, we use the feature maps from the last convolutional layer of the VGG16 model as features.

For the image classification task, we consider two versions of the ImageNet dataset, with images of size $224 \times 224$ and $32 \times 32$ pixels. The former is used since it is the original version of the dataset, while the latter is used because it has been demonstrated to be more challenging than the original version [26] [22].

Number of Components. One of the most important aspects in dimensionality reduction methods is the number of components $c$ of the resulting latent space. Therefore, to choose the best number of components for each method, we vary $c$ from 1 to 10 and select the value for which the method achieved the highest accuracy on the validation set (10% of the training set). Once the best $c$ is chosen, we use the training and validation set to learn the projection method and the classifier. We repeat this process for each dataset.

Comparison with Incremental Methods. This experiment compares our CIPLS with other incremental dimensionality reduction methods. Table 1 summarizes the results and shows that, on LFW, our method outperformed SGDPLS and IPLS by 1.08 and 1.38 percentage points (p.p.), respectively. Similarly, on the YTF dataset, CIPLS outperformed SGDPLS and IPLS by 0.88 and 1.88 p.p., in this order.

Finally, on the ImageNet dataset, the difference in accuracy compared to IPLS was of 0.07 and 1.35 p.p., for the $32 \times 32$ and $224 \times 224$ versions, respectively. It is important to mention that we do not consider SGDPLS on these datasets due to convergence problems and high computa-

| Method  | LFW | YTF | ImageNet 32x32 | ImageNet 224x224 |
|---------|-----|-----|----------------|------------------|
| CCIPCA  | 89.87 | 81.48 | 40.30 | 52.58 |
| SGDPLS  | 90.60 | 83.22 | – | – |
| IPLS    | 90.30 | 82.22 | 43.24 | 65.74 |
| CIPLS   | 91.68 | 84.10 | 43.31 | 67.09 |
| PLS     | 92.47 | 85.96 | – | – |
Comparison with Partial Least Squares. As suggested by Weng et al. [32], we compare the incremental methods with the traditional approach (in our case, traditional PLS), in which the closer to the accuracy of the baseline, the better.

According to Table 1 besides providing better results than IPLS and SGDPLS, CIPLS achieved the closest results to traditional PLS. For instance, on LFW, the difference in accuracy between PLS and CIPLS was 0.79 p.p. while on the YTF dataset it was 1.86 p.p. In contrast, the difference in accuracy between PLS and IPLS is higher – 1.87 p.p. on LFW and 2.74 p.p. on the YTF dataset. In addition, the difference in accuracy between PLS and IPLS is among the highest, 2.17 and 3.74 p.p. for the LFW and YTF, respectively. In particular, the results for PLS and CIPLS are statistically equivalent, while IPLS and SGDPLS present results statistically inferior compared to PLS.

It should be noted that the results of IPLS are closer to CCIPCA than PLS since only the first component of IPLS maintains the relationship between independent and dependent variables. On the other hand, the proposed method preserves this relation along higher-order components, which provides better discriminability, as seen in our results.

Higher-order Components. This experiment assesses the discriminability of the higher-order components of CIPLS compared to each of the other incremental methods. For this purpose, we follow a process suggested by Martinez [20], which consists of removing the first component of the latent space before presenting the projected data to the classifier. This evaluates the performance of the remaining components, not only the first one which tends to be better.

Table 2 shows the results. According to Table 2 CIPLS outperformed IPLS by 32.48 p.p. Observe that when all the components are used, CIPLS outperformed IPLS by 1.17 p.p. This larger difference when removing the first component is an effect of the better discriminability achieved by the components extracted by CIPLS. As we have argued, CIPLS preserves the relationship between dependent and independent variables across higher-order components, yielding more accurate results. Compared to SGDPLS, CIPLS outperforms it by 24.83 p.p.

Table 3. Comparison of incremental dimensionality reduction methods in terms of time complexity and execution time (in seconds) for estimating the projection matrix. $m$, $n$ denote dimensionality of the original data and number of samples, while $c$, $L$ and $T$ denote number of PLS components, number of PCA components and convergence steps, respectively.

| Method       | Time Complexity | Time (in seconds) |
|--------------|-----------------|-------------------|
| CCIPCA [32]  | $O(nLm)$        | 0.015             |
| SGDPLS [3]   | $O(Tcn)$        | 0.01              |
| IPLS [34]    | $O(nLm + c^2m)$ | 0.005             |
| CIPLS (Ours) | $O(ncm)$        | 0.005             |

Time Issues. To demonstrate the efficiency of CIPLS, in this experiment, we compare its time complexity to compute the projection matrix with the incremental methods evaluated. Following Weng et al. [32] and Zeng et al. [34], we report this complexity w.r.t. dimensionality of the original data ($m$), number of samples ($n$), number of components ($c$) and number of PCA components ($L$ — required only by IPLS and CCIPCA). Table 3 shows the time complexity of the methods.

According to Table 3 CIPLS presents a low time complexity for estimating the projection matrix. The complexity of CIPLS is not only on the same class as CCIPCA, which is the fastest among the compared methods, but it also has a very small constant factor. This constant factor is the number of components, $c$ for CIPLS and $L$ for CCIPCA. Experimentally, we found that the optimal constant factor for PLS is negligible, $c = 2$ resulted in the highest accuracies. While, for fairness, the same number of components was adopted for all methods in Table 3 typically $c < L$ on practical applications. This is a known advantage of PLS, it has been shown to require substantially less components to achieve its optimal accuracy than PCA [29].

Figure 2. Average prediction time (in seconds) for estimating the projection matrix, lower values are better. Black bars denote the confidence interval.
Finally, we report the average computation time (considering 30 executions) of the methods for estimating the projection matrix for one new sample. To make a fair comparison, we set $c = 4$ for all methods and for the other parameters we use the values where the methods achieved the best results in validation. As shown in Figure 2, SGDPLS is the slowest incremental PLS method, which is a consequence of its strategy for estimating the projection matrix, where for each sample the convergence step is run $T$ times. Our experiments showed that $T \geq 100$ is required for good results. The computation time for estimating the projection matrix of our method was statistically equivalent (according to a paired t-test) to that of CCIPCA, which is the fastest among the incremental dimensionality reduction methods assessed. Moreover, CIPLS was statistically faster than IPLS and SGDPILS, demonstrating that it is the fastest among the compared incremental PLS methods.

**Incremental Methods on the Streaming Scenario.** As we argued before, incremental methods can be employed on streaming applications, where the training data are continuously generated. To demonstrate the robustness of our method on these scenarios, in this experiment, we evaluate the methods on a synthetic streaming context, as proposed by Zeng et al. [34]. The procedure works as follows. First, the training data is divided into $k$ blocks, where $k = 20$. The idea behind this process is to interpret each block as a new instance of arriving data. Then, we create a new training set and insert each block at a time. Each time we insert a new block, we learn the projection method and evaluate its accuracy on the testing set. For instance, when adding the tenth block, all the $1, 2, ..., 10$ blocks are being used as training. It is important to mention that a block contains more than one sample, however, this does not modify the strategy of the incremental methods, which is to estimate the projection matrix by using a single sample at a time.

Figure 3 (a) and (b) show the results on the LFW and YTF datasets, respectively. On LFW, until the fifth block, it is not possible to determine the best method since the accuracy presents high variance, however, from the sixth block onwards, CIPLS outperforms all other methods. On YTF, our method achieves the highest accuracy for all blocks. These results show that the proposed method is more adequate for streaming applications than existing incremental PLS methods.

**Comparison with Feature Selection Methods.** Our last experiment evaluates the performance of CIPLS as a feature selection method. Table 4 shows the results for different percentages of kept features on LFW and YTF. According to Table 4, CIPLS achieves comparable results when compared to state-of-the-art feature selection techniques. For example, on LFW the difference in accuracy, on average, from CIPLS to infFS and ilFS is of 0.15 and 0.25 p.p., respectively. In contrast, on YTF for some percentages of kept features (e.g., 15% and 20%), CIPLS outperforms infFS and ilFS. We highlight that these methods were designed specifically for feature selection. Additionally, the difference, on average, between CIPLS and PLS is of 0.27, 0.14 and 0.50 p.p. on the LFW and YTF datasets, respectively. Moreover, the largest accuracy difference between PLS and CIPLS is of 0.4 p.p., on LFW dataset with 10% of features kept. This result reinforces that the proposed decompositions to extend the NIPALS and enable the employment of VIP are a good approximation of the original method.

Based on the results shown, it is possible to conclude that, besides dimensionality reduction, CIPLS achieves state-of-the-art results in the context of feature selection.

![Figure 3. Labeled Faces in the Wild (LFW).](image)

![Figure 3. Youtube Faces (YTF).](image)
Table 4. Comparison of feature selection methods using different percentages of kept features.

|               | LFW |                  | YTF |                  |
|---------------|-----|------------------|-----|------------------|
|               | 10  | 15  | 20  | 50  | 10  | 15  | 20  | 50  |
| infFS [28]    | 91.92 | 91.58 | 92.03 | 92.23 | 86.64 | 86.68 | 87.14 | 87.30 |
| ilFS [27]     | 92.03 | 91.67 | 92.25 | 92.23 | 86.60 | 86.94 | 86.84 | 87.54 |
| PLS+VIP       | 92.05 | 91.67 | 92.13 | 92.38 | 86.70 | 86.82 | 87.18 | 87.68 |
| CIPLS (Ours)+VIP | 91.63 | 91.55 | 91.80 | 92.18 | 86.48 | 86.92 | 87.02 | 87.40 |

5. Conclusions

This work presented a novel incremental partial least squares method, named Covariance-free Incremental Partial Least Squares (CIPLS). The method extends the NI-PALS algorithm for incremental operation and enables computation of the projection matrix using one sample at a time while still presenting the main property of traditional PLS, namely preserving the relation between dependent and independent variables. Compared to existing incremental partial least squares methods, CIPLS attains superior performance besides being computationally efficient. In addition, different from previous incremental partial least squares, CIPLS can easily operate as a feature selection method. In this context, the proposed method is able to achieve comparable results to the state-of-the-art.

Acknowledgments

The authors would like to thank the Brazilian National Research Council – CNPq (Grants #311053/2016-5 and #438629/2018-3), the Minas Gerais Research Foundation – FAPEMIG (Grants APQ-00567-14 and PPM-00540-17) and the Coordination for the Improvement of Higher Education Personnel – CAPES (DeepEyes Project). The authors would like to thank Xue-Qiang, Zeng Guo-Zheng Li, Raman Arora, Poorya Mianji, Alexander Stott and Teodor Marinov for sharing their source code.

References

[1] H. Abdi. Partial least squares regression and projection on latent structure regression (pls regression). Wiley Interdisciplinary Reviews: Computational Statistics, 2010.
[2] S. Alakkar and J. Dingliana. An acceleration scheme for mini-batch, streaming pca. British Machine Vision Conference (BMVC), 2019.
[3] R. Arora, P. Mianji, and T. V. Marinov. Stochastic optimization for multiview representation learning using partial least squares. In Proceedings of the 33rd International Conference on International Conference on Machine Learning, volume 48 of International Conference on Machine Learning (ICML), 2016.
[4] J. Chen, R. Ranjan, S. Sankaranarayanan, A. Kumar, C. Chen, V. M. Patel, C. D. Castillo, and R. Chellappa. Unconstrained still/video-based face verification with deep convolutional neural networks. International Journal of Computer Vision, 126, 2018.
[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In IEEE Computer Vision and Pattern Recognition (CVPR), pages 248–255, 2009.
[6] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In International Conference on Machine Learning (ICML), 2014.
[7] P. Geladi and B. R. Kowalski. Partial least-squares regression: a tutorial. Analytica Chimica Acta, 185:1 – 17, 1986.
[8] T. Yamada, N. Zhang. PLSP+. In IEEE International Conference on Computer Vision (ICCV), pages 1440–1448, 2015.
[9] G. Guo and G. Mu. Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011.
[10] G. Guo and G. Mu. Joint estimation of age, gender and ethnicity: CCA vs. PLS. In IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2013.
[11] R. Hasegawa and K. Hotta. Plsnet: A simple network using partial least squares regression for image classification. In International Conference on Pattern Recognition (ICPR), pages 1601–1606, 2016.
[12] K. Hiraoka, S. Yoshizawa, K. Hidai, M. Hamahira, H. Mizoguchi, and T. Mishima. Convergence analysis of online linear discriminant analysis. In IEEE International Joint Conference on Neural Network (IJCNN), pages 387–391, 2000.
[13] P. Howland, J. Wang, and H. Park. Solving the small sample size problem in face recognition using generalized discriminant analysis. Pattern Recognition, 39(2):277–287, 2006.
[14] G. B. Huang, M. A. Mattar, H. Lee, and E. G. Learned-Miller. Learning to align from scratch. In Neural Information Processing Systems (NIPS), pages 773–781, 2012.
[15] R. Jain. The art of computer systems performance analysis: techniques for experimental design, measurement, simulation, and modeling. Wiley professional computing. John Wiley & Sons, 1990.
[16] A. Jordan, R. Kloss, F. Yamada, and W. R. Schwartz. Pruning deep neural networks using partial least squares. In British Machine Vision Conference (BMVC) Workshops, pages 1–9, 2019.
[17] R. B. Kloss, A. Jordao, and W. R. Schwartz. Face verification strategies for employing deep models. In IEEE International Conference on Automatic Face & Gesture Recognition (FG), pages 258–262, 2018.

[18] M. T. Law, J. Snell, A. Farahmand, R. Urtasun, and R. S. Zemel. Dimensionality reduction for representing the knowledge of probabilistic models. In International Conference on Learning Representations (ICLR), 2019.

[19] G. Lu, J. Zou, and Y. Wang. Incremental learning of complete linear discriminant analysis for face recognition. Knowledge-Based Systems, 31:19–27, 2012.

[20] A. M. Martínez and A. C. Kak. PCA versus LDA. IEEE Pattern Analysis and Machine Intelligence (PAMI), 2001.

[21] T. Mehmood, K. H. Liland, L. Snipen, and S. Sb. A review of variable selection methods in partial least squares regression. Chemometrics and Intelligent Laboratory Systems, 2012.

[22] Z. A. Milacski, B. Poczos, and A. Lorincz. Differentiable unrolled alternating direction method of multipliers for onenet. British Machine Vision Conference (BMVC), 2019.

[23] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In British Machine Vision Conference (BMVC), pages 41.1–41.12, 2015.

[24] R. Ranjan, S. Sankaranarayanan, A. Bansal, N. Bodla, J. Chen, V. M. Patel, C. D. Castillo, and R. Chellappa. Deep learning for understanding faces: Machines may be just as good, or better, than humans. Signal Processing Magazine, 35, 2018.

[25] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: An astounding baseline for recognition. In Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2014.

[26] S. Rebuffi, H. Bilen, and A. Vedaldi. Learning multiple visual domains with residual adapters. In Neural Information Processing Systems (NIPS), pages 506–516, 2017.

[27] G. Roffo, S. Melzi, U. Castellani, and A. Vinciarelli. Infinite latent feature selection: A probabilistic latent graph-based ranking approach. In IEEE International Conference on Computer Vision (ICCV), pages 1407–1415, 2017.

[28] G. Roffo, S. Melzi, and M. Cristani. Infinite feature selection. In IEEE International Conference on Computer Vision (ICCV), pages 4202–4210, 2015.

[29] W. R. Schwartz, A. Kembhavi, D. Harwood, and L. S. Davis. Human detection using partial least squares analysis. In IEEE International Conference on Computer Vision (ICCV), pages 24–31, 2009.

[30] A. E. Stott, S. Kanna, D. P. Mandic, and W. T. Pike. An online NIPALS algorithm for partial least squares. In International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 4177–4181, 2017.

[31] B. Su and Y. Wu. Learning low-dimensional temporal representations. In International Conference on Machine Learning (ICML), 2018.

[32] J. Weng, Y. Zhang, and W. Hwang. Candid covariance-free incremental principal component analysis. IEEE Pattern Analysis and Machine Intelligence (PAMI), 25(8):1034–1040, 2003.

[33] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In Computer Vision and Pattern Recognition (CVPR), pages 529–534, 2011.

[34] X. Zeng and G. Li. Incremental partial least squares analysis of big streaming data. Pattern Recognition, 47:3726–3735, 2014.