Robust Data Whitening as an Iteratively Re-weighted Least Squares Problem

Arun Mukundan(✉), Giorgos Tolias, and Ondřej Chum

Visual Recognition Group, Czech Technical University in Prague,
Prague, Czech Republic
{arun.mukundan,giorgos.tolias,chum}@cmp.felk.cvut.cz

Abstract. The entries of high-dimensional measurements, such as image or feature descriptors, are often correlated, which leads to a bias in similarity estimation. To remove the correlation, a linear transformation, called whitening, is commonly used. In this work, we analyze robust estimation of the whitening transformation in the presence of outliers. Inspired by the Iteratively Re-weighted Least Squares approach, we iterate between centering and applying a transformation matrix, a process which is shown to converge to a solution that minimizes the sum of ℓ_2 norms. The approach is developed for unsupervised scenarios, but further extend to supervised cases. We demonstrate the robustness of our method to outliers on synthetic 2D data and also show improvements compared to conventional whitening on real data for image retrieval with CNN-based representation. Finally, our robust estimation is not limited to data whitening, but can be used for robust patch rectification, *e.g.* with MSER features.

1 Introduction

In many computer vision tasks, visual elements are represented by vectors in high-dimensional spaces. This is the case for image retrieval [3,14], object recognition [17,23], object detection [9], action recognition [20], semantic segmentation [16] and many more. Visual entities can be whole images or videos, or regions of images corresponding to potential object parts. The high-dimensional vectors are used to train a classifier [19] or to directly perform a similarity search in high-dimensional spaces [14].

Vector representations are often post-processed by mapping to a different representation space, which can be higher or lower dimensional. Such mappings or embeddings can be either non-linear [2,5] or linear [4,6]. In the non-linear case, methods that directly evaluate [2] or efficiently approximate [5] non-linear kernels are known to be perform better. Typical applications range from image classification [5] and retrieval [4] to semantic segmentation [8]. Examples of the linear kind are used for dimensionality reduction in which dimensions carrying the most meaningful information are kept. Dimensionality reduction with Principal

P. Sharma and F.M. Bianchi (Eds.): SCIA 2017, Part I, LNCS 10269, pp. 234–247, 2017.
DOI: 10.1007/978-3-319-59126-1_20

Component Analysis (PCA) is very popular in numerous tasks [4,6,15]. In the same vein as PCA is data whitening, which is the focus of this work[1].

A whitening transformation is a linear transformation that performs correlation removal or suppression by mapping the data to a different space such that the covariance matrix of the data in the transformed space is identity. It is commonly learned in an unsupervised way from a small sample of training vectors. It is shown to be quite effective in retrieval tasks with global image representations, for example, when an image is represented by a vector constructed through the aggregation of local descriptors [13] or by a vector of Convolutional Neural Network (CNN) activations [11,22]. In particular, PCA whitening significantly boosts the performance of CNN compact image vectors, *i.e.* 256 to 512 dimensions, due to handling of inherent co-occurrence phenomena [4]. Principal components found are ordered by decreasing variance, allowing for dimensionality reduction at the same time [12]. Dimensionality reduction may also be performed in a discriminative, supervised fashion. This is the case in the work by Cai *et al.* [6], where the covariance matrices are constructed by using information of pairs of similar and non-similar elements. In this fashion, the injected supervision performs better separation between matching and non-matching vectors and has better chances to avoid outliers in the estimation. It has been shown [10] that an unsupervised approach based on least squares minimization is likely to be affected by outliers: even a single outlier of high magnitute can significantly deviate the solution.

In this work, we propose an unsupervised way to learn the whitening transformation such that the estimation is robust to outliers. Inspired by the Iteratively Re-weighted Least Squares of Aftab and Hartley [1], we employ robust M-estimators. We perform minimization of robust cost functions such as ℓ_1 or Cauchy. Our approach iteratively alternates between two minimizations, one to perform the centering of the data and one to perform the whitening. In each step a weighted least squares problem is solved and is shown to minimize the sum of the ℓ_2 norms of the training vectors. We demonstrate the effectiveness of this approach on synthetic 2D data and on real data of CNN-based representation for image search. The method is additionally extended to handle supervised cases, as in the work of Cai *et al.* [6], where we show further improvements. Finally, our methodology is not limited to data whitening. We provide a discussion on applying it for robust patch rectification of MSER features [18].

The rest of the paper is organized as follows: In Sect. 2 we briefly review conventional data whitening and give our motivation, while in Sect. 3 we describe the proposed iterative whitening approach. Finally, in Sects. 4 and 5 we compare our method to the conventional approach on synthetic and real data, respectively.

[1] The authors were supported by the MSMT LL1303 ERC-CZ grant, Arun Mukundan was supported by the SGS17/185/OHK3/3T/13 grant.

2 Data Whitening

In this section, we first briefly review the background of data whitening and then give a geometric interpretation, which forms our motivation for the proposed approach.

2.1 Background on Whitening

A whitening transformation is a linear transformation that transforms a vector of random variables with a known covariance matrix into a set of new variables whose covariance is the identity matrix. The transformation is called "whitening" because it changes the input vector into a white noise vector.

We consider the case where this transformation is applied on a set of zero centered vectors $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_N\}$, with $\mathbf{x}_i \in \mathbb{R}^d$, where $\Sigma = \sum_i \mathbf{x}_i \mathbf{x}_i^\top$. The whitening transformation P is given by

$$P^\top P = \Sigma^{-1}. \tag{1}$$

In Fig. 1 we show a toy example of 2D points and their whitened counterpart.

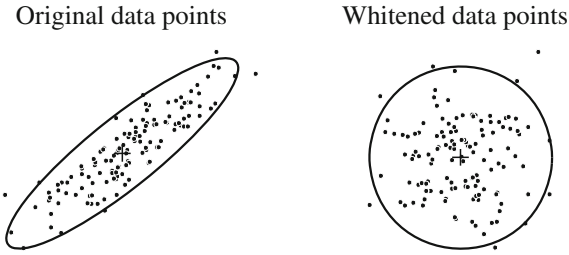

Fig. 1. Left: Points in 2D and their covariance shown with an ellipse. Right: The corresponding whitened 2D point set.

Assumption. In the following text, we assume that the points of $\mathcal{X}$ do *not* lie in a linear subspace of dimensionality $d' < d$. If this is the case, a solution is to first identify the d'-dimensional subspace and perform the proposed algorithms on this subspace. The direct consequence of the assumption is that the sample covariance matrix Σ is full rank, in particular $\det(\Sigma) > 0$.

It is clear from (1) that the whitening transformation is given up to an arbitrary rotation $R \in \mathbb{R}^{d \times d}$, with $R^\top R = I$. The transformation matrix P of the whitening is thus given by

$$P = R\Sigma^{-1/2}. \tag{2}$$

2.2 Geometric Interpretation

We provide a geometric interpretation of data whitening, which also serves as our motivation for the proposed method in this work.

Observation. Assuming zero-mean points, the whitening transform P in (2) minimizes the sum of squared ℓ_2 norms among all linear transforms T with $\det(T) = \det(\Sigma)^{-1/2}$.
Proof.

$$\begin{aligned} C_{\ell_2}(P) &= \sum_i ||P\mathbf{x}_i||^2 \\ &= \sum_i tr\left(\mathbf{x}_i^\top P^\top P \mathbf{x}_i\right) \\ &= \sum_i tr\left(\left(\mathbf{x}_i\mathbf{x}_i^\top\right) P^\top P\right) \\ &= tr\left(\left(\sum_i \mathbf{x}_i\mathbf{x}_i^\top\right) P^\top P\right) \\ &= tr\left(\Sigma P^\top P\right) \\ &= \sum_{j=1}^{d} \lambda_j, \end{aligned} \tag{3}$$

where λ_i are the eigenvalues of $\Sigma P^\top P$ and $||\cdot||$ is denoting ℓ_2 norm. Upon imposing the condition $\det(T) = \det(\Sigma)^{-1/2}$, we get that $\det(\Sigma P^\top P) = \prod_{j=1}^{d} \lambda_j$ is constant with respect to P. It follows from the arithmetic and geometric mean inequality, that the sum in (3) is minimized when $\lambda_i = \lambda_j, \forall i = j$. Equality of all eigenvalues allows us to show that

$$\begin{aligned} \Sigma P^\top P &= I \\ P^\top P &= \Sigma^{-1} \\ P &= R\Sigma^{-1/2} \end{aligned} \tag{4}$$

which is exactly the solution in (2) that also minimizes (3). The need for the existence of Σ^{-1} justifies the stated full rank assumption.

We have just shown that learning a whitening transformation reduces to a least squares problem.

3 Robust Whitening

In this section we initially review the necessary background on the the iteratively re-weighted least squares (IRLS) method recently proposed by Aftab and Hartley [1], which is the starting point for our method. Then, we present the robust whitening and centering procedures, which are posed as weighted least squares problems and performed iteratively. Finally, the extension to the supervised case is described.

3.1 Background on IRLS

In the context of distance minimization the IRLS method minimizes the cost function

$$C_h(\theta) = \sum_{i=1}^{N} h \circ f(\theta, \mathbf{x}_i), \tag{5}$$

where f is a distance function that is defined on some domain, h is a function that makes the cost less sensitive to outliers, and $\mathbf{x}_i \in \mathcal{X}$. Some examples of robust h functions are ℓ_1, Huber, pseudo-Huber, *etc.* as described in [1]. For instance, assume the case of the geometric median of the points in $\mathcal{X}$. Setting $f(\boldsymbol{\mu}, \mathbf{x}_i) = ||\boldsymbol{\mu} - \mathbf{x}_i||$ and $h(z) = z$, we get the cost (5) as the sum of ℓ_2 norms. The minimum of this cost is attained when $\boldsymbol{\mu}$ is equal to the geometric median.

It is shown [1] that a solution for $\operatorname{argmin}_\theta C_h(\theta)$ may be found by solving a sequence of weighted least squares problems. Given some initial estimate θ^0, the parameters θ are iteratively estimated

$$\theta^{t+1} = \operatorname*{argmin}_{\theta} \sum_{i=1}^{N} w(\theta^t, \mathbf{x}_i) f(\theta, \mathbf{x}_i)^2, \tag{6}$$

where for brevity $w(\theta^t, \mathbf{x}_i)$ is denoted w_i^t in the following. Provided $h(\sqrt{z})$ is differentiable at all points and concave, for certain values of w_i^t and conditions on f this solution minimizes $C_h(\theta)$. In some cases, it may even be possible to find a simple and anlytic solution.

Given that the iterative procedure indeed converges to a minimum cost of (5), we get the following condition on the weights:

$$\begin{aligned} \nabla_\theta (h \circ f(\theta, \mathbf{x}_i)) &= 0, \\ \nabla_\theta (w_i^t f(\theta, \mathbf{x}_i)^2) &= 0. \end{aligned} \tag{7}$$

This results in the following weights

$$w_i^t = \frac{h'(f(\theta^t, \mathbf{x}_i))}{2 f(\theta^t, \mathbf{x}_i)}. \tag{8}$$

Geometric median. The geometric median $\boldsymbol{\mu}$ of a set of points $\{\mathbf{x}_i\}$ is the point that minimizes the sum of ℓ_2 distances to the points. As shown in one of the cases in the work by Aftab and Hartley [1], the problem of finding the geometric median can be cast in an IRLS setting for certain value of weights. Setting $f(\boldsymbol{\mu}, \mathbf{x}_i) = ||\boldsymbol{\mu} - \mathbf{x}_i||$ and $h(z) = z$, the IRLS algorithm minimizes the sum of distances at each iteration, thus converging to the geometric median.

3.2 Method

From the observation in Sect. 2.2, we know that there is a closed-form solution to the problem of finding a linear transformation P so that $\sum_i ||P\mathbf{x}_i||^2$ is minimized subject to a fixed determinant $\det(P)$. The idea of the robust whitening is to use

this least squares minimizer in a framework similar to the iterative re-weighted least squares to minimize a robust cost.

Robust transformation estimation. In contrast to the conventional whitening and the minimization of (3), we now propose the estimation of a whitening transform (transformation matrix P) in a way that is robust to outliers. We assume zero mean points and seek the whitening transformation that minimizes the robust cost function of (5). We set $f(P, \mathbf{x}_i) = ||P\mathbf{x}_i||$ and use the ℓ_1 cost function $h(z) = z$. Other robust cost functions can be used, too[2].

We seek to minimize the sum of ℓ_2 norms in the whitened space

$$C_{\ell_1}(P) = \sum_{i=1}^{N} f(P, \mathbf{x}_i) = \sum_{i=1}^{N} ||P\mathbf{x}_i||. \tag{9}$$

The corresponding iteratively re-weighted least squares solution is given by

$$P^{t+1} = \underset{P}{\operatorname{argmin}} \sum_{i=1}^{N} w_i^t ||P\mathbf{y}_i^t||^2, \tag{10}$$

where $\mathbf{y}_i^t = P^t \mathbf{y}_i^{t-1}$ and $\mathbf{y}_i^0 = \mathbf{x}_i$. This means that each time transformation P^t is estimated and applied to whiten the data points. In the following iteration, the estimation is performed on data points in the whitened space. The effective transformation at iteration t with respect to the initial points $\mathbf{x}_i$ is given by

$$\hat{P}^t = \prod_{i=1}^{t} P^i. \tag{11}$$

Along the lines of proof (3) we find a closed form solution that minimizes (9) as

$$\begin{aligned} &\sum_i w_i^t ||P\mathbf{y}_i^t||^2 \\ &= tr\left(\left(\sum_i w_i^t \mathbf{y}_i^t \mathbf{y}_i^{t\top}\right) P^\top P\right) \\ &= tr\left(\tilde{\Sigma} P^\top P\right) \end{aligned} \tag{12}$$

where $\tilde{\Sigma} = \sum_i w_i^t \mathbf{y}_i^t \mathbf{y}_i^{t\top}$ is a *weighted covariance*. Therefore, P is given, up to a rotation, as

$$P = R\tilde{\Sigma}^{-1/2}. \tag{13}$$

[2] We also use Cauchy cost in our experiments. It is defined as $h(z) = b^2 log(1 + z^2/b^2)$.

Joint centering and transformation matrix estimation. In this section we describe the proposed approach for data whitening. We propose to jointly estimate a robust mean $\boldsymbol{\mu}$ and a robust transformation matrix P by alternating between the two previously described procedures: estimating the geometric median and estimating the robust transformation. In other words, in each iteration, we first find $\boldsymbol{\mu}$ keeping P fixed and then find P keeping $\boldsymbol{\mu}$ fixed. In this way the assumption for centered points when finding P is satisfied. Given that each iteration of the method outlined above reduces the cost, and that the cost must be non-negative, we are assured convergence to a local minimum.

We propose to minimize cost

$$C_{\ell_1}(P, \boldsymbol{\mu}) = \sum_{i=1}^{N} ||P(\mathbf{x}_i - \boldsymbol{\mu})||. \tag{14}$$

In order to reformulate this as an IRLS problem, we use $h(z) = z$, and $f(P, \boldsymbol{\mu}, \mathbf{x}_i) = ||P(\mathbf{x}_i - \boldsymbol{\mu})||$. Now, at iteration t the minimization is performed on points $\mathbf{y}_i^t = \hat{P}^t(\mathbf{x}_i - \hat{\boldsymbol{\mu}}^t)$ and the conditions for convergence with respect to $\boldsymbol{\mu}$ (skipping t and notation for effective parameters for brevity) are

$$\begin{aligned} \nabla_{\boldsymbol{\mu}}(h \circ f) &= \nabla_{\boldsymbol{\mu}} ||P(\mathbf{x}_i - \boldsymbol{\mu})|| \\ &= \nabla_{\boldsymbol{\mu}} \sqrt{(\mathbf{y}_i - \boldsymbol{\mu})^\top P^\top P(\mathbf{y}_i - \boldsymbol{\mu})} \\ &= \frac{1}{2||P(\mathbf{y}_i - \boldsymbol{\mu})||} \cdot \nabla_{\boldsymbol{\mu}} M \end{aligned} \tag{15}$$

$$\nabla_{\boldsymbol{\mu}}(w_i \cdot f^2) = w_i \cdot \nabla_{\boldsymbol{\mu}} M$$

where we have $M = (\mathbf{y}_i - \boldsymbol{\mu})^\top P^\top P(\mathbf{y}_i - \boldsymbol{\mu})$. This gives the expression for the weight

$$w_i^t = \frac{1}{2||\hat{P}^t(\mathbf{x}_i - \hat{\boldsymbol{\mu}}^t)||}. \tag{16}$$

A similar derivation gives us the weights for the iteration step of P. Therefore in each iteration, we find the solutions to the following weighted least squares problems,

$$\boldsymbol{\mu}^{t+1} = \underset{\boldsymbol{\mu}}{\operatorname{argmin}} \sum_{i=1}^{N} w_i(P^t, \boldsymbol{\mu}^t)||P^t(\mathbf{y}_i - \boldsymbol{\mu})||^2, \tag{17}$$

$$P^{t+1} = \underset{P}{\operatorname{argmin}} \sum_{i=1}^{N} w_i(P^t, \boldsymbol{\mu}^{t+1})||P(\mathbf{y}_i^t - \boldsymbol{\mu}^{t+1})||^2. \tag{18}$$

The effective centering and transformation matrix at iteration t are given by

$$\hat{\boldsymbol{\mu}}^t = \sum_{i=1}^{t} \left(\prod_{j=1}^{i-1} P_j^{-1} \right) \boldsymbol{\mu}^i \quad , \quad \hat{P}^t = \prod_{i=1}^{t} P^i. \tag{19}$$

Algorithm 1. Robust Whitening

procedure ROBUST WHITENING($\mathcal{X}$)
  $\mathbf{z}_0 \leftarrow \mathcal{X}$
  $\boldsymbol{\mu}_0 \leftarrow$ Initialize centre to mean of $\mathbf{z}_0$
  $P_0 \leftarrow$ Initialize transform to identity matrix
  for t≤niter **do**
    $\boldsymbol{\mu}^t \leftarrow \frac{1}{N}\sum_{i=1}^{N} w_i(P^{t-1}, \boldsymbol{\mu}^{t-1})\mathbf{z}_i^{t-1}$
    $\tilde{\Sigma}^t \leftarrow \sum_{i=1}^{N} w_i(P^{t-1}, \boldsymbol{\mu}^t)(\mathbf{z}_i^{t-1} - \boldsymbol{\mu}^t)(\mathbf{z}_i^{t-1} - \boldsymbol{\mu}^t)^\top$
    $P^t \leftarrow \frac{\mathrm{chol}(\tilde{\Sigma}^t)}{\det\left(\mathrm{chol}(\tilde{\Sigma}^t)\right)^{1/d}}$
    $\mathbf{z}^t \leftarrow P^t\left(\mathbf{z}^{t-1} - \boldsymbol{\mu}^t\right)$
    $\hat{\boldsymbol{\mu}}^t \leftarrow \sum_{i=1}^{t}\left(\prod_{j=1}^{i-1} P^{j\,-1}\right)\boldsymbol{\mu}^i$
    $\hat{P}^t \leftarrow \prod_{i=0}^{t} P^i$
  end for
  return $\hat{\boldsymbol{\mu}}^t$, $\hat{P}^t$
end procedure

The whole procedure is summarized in Algorithm 1, where chol is used to denote the Cholesky decomposition.

3.3 Extension with Supervision

We firstly review the work of Cai *et al.* [6] who perform supervised descriptor whitening and then present our extension for robust supervised whitening.

Background on linear discriminant projections [6]. The linear discriminant projections (LDP) are learned via supervision of pairs of similar and dissimilar descriptors. A pair (i, j) is similar if $(i, j) \in \mathcal{S}$ while dissimilar if $(i, j) \in \mathcal{D}$. The projections are learned in two parts. Firstly, the whitening part is obtained as the square-root of the intra-class covariance matrix $C_{\mathcal{S}}^{-1/2}$, where

$$C_{\mathcal{S}} = \sum_{(i,j\in\mathcal{S})} (x_i - x_j)(x_i - x_j)^\top. \tag{20}$$

Then, the rotation part is given by the PCA of the inter-class covariance matrix which is computed in the space of the whitened descriptors. It is computed as $\mathrm{eig}\left(C_{\mathcal{S}}^{-1/2} C_{\mathcal{D}} C_{\mathcal{S}}^{-1/2}\right)$, where

$$C_{\mathcal{D}} = \sum_{(i,j\in\mathcal{D})} (x_i - x_j)(x_i - x_j)^\top. \tag{21}$$

Algorithm 2. Supervised Robust Whitening

procedure SUPERVISED ROBUST WHITENING($\mathcal{X}, \mathcal{S}$)
  $\mathcal{X}_\mathcal{S} = \{d : d = x_i - x_j, x_i \in \mathcal{X}, x_j \in \mathcal{X}, (i,j) \in \mathcal{S}\}$
  $\mathcal{X}_\mathcal{S} = \{\mathcal{X}_\mathcal{S} \cup -\mathcal{X}_\mathcal{S}\}$
  $\boldsymbol{\mu}_1, P_1 \leftarrow$ Robust Whitening($\mathcal{X}_\mathcal{S}$)
  $\boldsymbol{\mu} \leftarrow$ Geometric Median($\mathcal{X}$)
  $\bar{\mathcal{X}} \leftarrow \mathcal{X} - \boldsymbol{\mu}$
  $\boldsymbol{\mu}_2, P_2 \leftarrow$ Robust Whitening($P_1 \bar{\mathcal{X}}$)
  $R_2 \leftarrow \text{eig}((P_2^\top P_2)^{-1})$
  $\hat{\boldsymbol{\mu}} \leftarrow \boldsymbol{\mu} + \boldsymbol{\mu}_2$
  $\hat{P} \leftarrow P_1 R_2$
  return $\hat{P}, \hat{\boldsymbol{\mu}}$
end procedure

The final whitening is performed by $P_{\mathcal{SD}}^\top (x - m)$, where m is the mean descriptor and $P_{\mathcal{SD}} = C_\mathcal{S}^{-1/2} \cdot \text{eig}\left(C_\mathcal{S}^{-1/2} C_\mathcal{D} C_\mathcal{S}^{-1/2}\right)$. It is noted [6] that, if the number of descriptors is large compared to the number of classes (two in this case), then $C_\mathcal{D} \approx C_{\mathcal{S}\cup\mathcal{D}}$ since $|\mathcal{S}| \ll |\mathcal{D}|$. This is the approach we follow.

Robust linear discriminant projections. The proposed method uses the provided supervision in a robust manner by employing the method introduced in Sect. 3.2. The whitening is estimated in a robust manner by Algorithm 1 on the intra-class covariance. In this manner, small weights are assigned to pairs of descriptors that are found to be outliers. Then, the mean and covariance are estimated in a robust manner in the whitened space. The whole procedure is summarized in Algorithm 2. Mean μ_1 is zero due to the including the pairs in a symmetric manner.

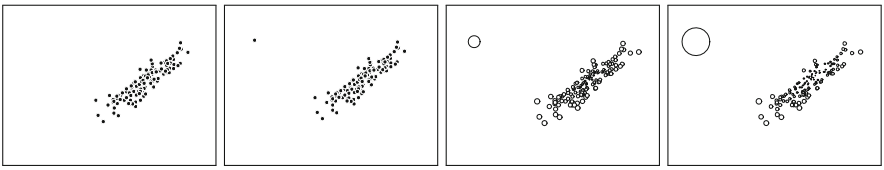

Fig. 2. (a) Set of 2D points drawn from a Gaussian distribution with zero mean. (b) Same set as (a) with an additional point (outlier) placed at a distance equal to 2 times the maximum distance from the center of the initial set. (c) Visualization of the weights assigned in the set of (b) with the robust whitening which uses the ℓ_1 cost function. Note that the size of the circles is inversely proportional to the weight. (d) Same as (c), but using the Cauchy cost.

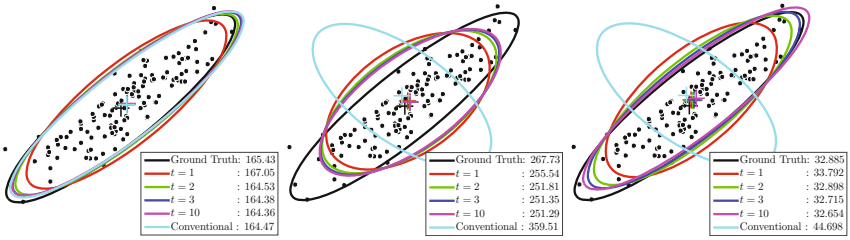

Fig. 3. Visualization of the covariance (ellipse) and center (cross) of the estimated whitening transformation at iteration t and the conventional estimate. The example is performed using the set of 2D points of Fig. 2. The ground truth distribution that created the data points is shown in black. The conventional estimate is shown in cyan. We show the effective estimate of the t^{th} iteration. The two approaches are compared without an outlier in (a) or with an outlier using ℓ_1 in (b) or Cauchy cost function in (c). The outlier is placed at a distance equal to 10 times the maximum inlier distance. The outlier is not plotted to keep the scale of the figure reasonable. The ℓ_1 (or Cauchy) cost is shown in the legend. (Color figure online)

4 Examples on Synthetic Data

We compare the proposed and the conventional whitening approaches on synthetic 2D data in order to demonstrate the robustness of our method to outliers. We sample a set of 2D points from a normal distribution, which is shown in Fig. 2(a) and then add an outlier and show the result in Fig. 2(b). In the absence of outliers, both methods provide a similar estimation as shown in Fig. 3. It is also shown how the iterative approach reduces the cost at each iteration. With the presence of an outlier, the estimation of the conventional approach is largely affected, while the robust method gives a much better estimation, as shown in Fig. 3. Using the Cauchy cost function the estimated covariance is very close to

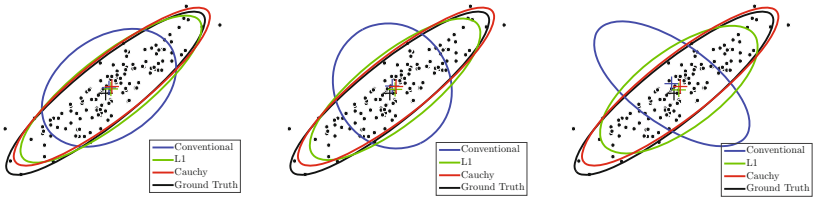

Fig. 4. Visualization of the covariance (ellipse) and center (cross) of the estimated whitening transformation using the conventional approach and ours. The example is performed using the set of 2D points of Fig. 2. The two approaches are compared for the case of an outlier placed at distance equal to 3 (a), 5 (b) and 10 (c) times the maximum inlier distance. The outlier is not shown to keep the resolution high.

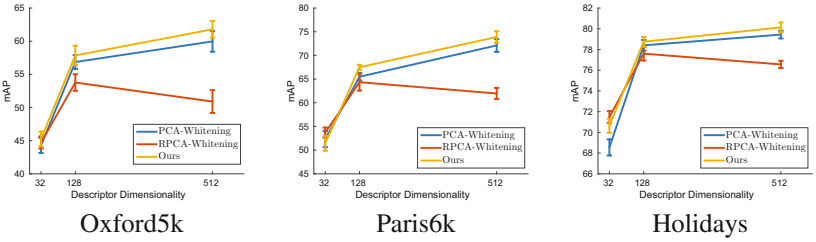

Fig. 5. Retrieval performance comparison using mAP on 3 common benchmarks. Comparison of the conventional PCA whitening, RPCA whitening and our approach for descriptors of varying dimensionality. The training set contains a small subset of 512 vectors randomly selected. The experiment is performed 10 times and mean performance is reported while standard deviation is shown on the curves. Descriptors extracted using VGG.

that of the ground truth. The weights assigned to each point with the robust approach are visualized in Fig. 2 and show how the outlier is discarded in the final estimation. Finally, in Fig. 4, we compare the conventional way with our approach for outlier of increasing distance.

5 Experiments

In this section, the robust whitening is applied to real-application data. In particular, we test on SPOC [4] descriptors, which are CNN-based image descriptors constructed via sum pooling of network activations in the internal convolutional layers. We evaluate on 3 popular retrieval benchmarks, namely Oxford5k, Paris6k and Holidays (the upright version), and use around 25 k training images to learn the whitening. We use VGG network [21] to extract the descriptors and, in contrast to the work of Babenko and Lempitsky [4], we do not ℓ_2-normalize the input vectors. The final ranking is obtained using Euclidean distance between the query and the database vectors. Evaluation is performed by measuring mean Average Precision (mAP). As in the case of conventional whitening, the dimension reduction is performed by preserving those dimensions that have the highest variance. This is done by finding an eigenvalue decomposition of the estimated covariance and ordering the eigenvectors according to decreasing eigenvalue.

There are many approaches performing robust PCA [7,24,25] by assuming that the data matrix can be decomposed into the sum of a low rank matrix and a sparse matrix corresponding to the outliers. We employ the robust PCA (RPCA) method by Candès *et al.* [7] to perform a comparison. The low rank matrix is recovered and PCA whitening is learned on this.

We present results in Table 1, where the robust approach offers a consistent improvement over the conventional PCA whitening [4]. Especially in the case where the whitening is learned on few training vectors, the improvement is larger as outliers will heavily influence the conventional whitening, as shown in Fig. 5.

Table 1. Retrieval performance comparison using mAP on 3 common benchmarks. Comparison of retrieval using the initial sum-pooled CNN activations, post-processing using the baselines and our methods for unsupervised and supervised whitening. Results for descriptors of varying dimensionality. The full training set is used. Descriptors extracted using VGG. S: indicates the use of supervision.

Dataset		Oxford5k			Paris6k			Holidays		
Method	S	32D	128D	512D	32D	128D	512D	32D	128D	512D
Raw		–	–	51.4	–	–	61.6	–	–	78.8
PCA whitening		44.7	56.6	66.7	53.4	67.0	77.1	69.6	78.4	80.6
RPCA whitening		44.0	52.4	55.6	**55.9**	61.1	65.1	70.5	75.8	77.4
Ours		**45.8**	**58.5**	**67.7**	50.0	**68.3**	**78.4**	**70.7**	**78.8**	**81.8**
LDP	×	39.4	59.9	68.8	56.1	70.2	76.6	67.5	77.7	80.8
Ours	×	**49.9**	**62.3**	**70.3**	**57.6**	**72.0**	**78.0**	**69.0**	**78.6**	**82.1**

Our approach is also better than RPCA whitening for large dimensionalities. It seems that RPCA underestimates the rank of the matrix and does not offer any further improvements for large dimensions.

6 Discussion

The applicability of the proposed method goes beyond robust whitening. Consider, for example, the task of affine-invariant descriptors of local features, such as MSERs [18]. A common approach is to transform the detected feature into a canonical frame prior to computing a robust descriptor based on the gradient map of the normalized patch (SIFT [17]). To remove the effect of an affine transformation, a centre of gravity and centered second-order moment (covariance matrix) are used. It can be shown that both the centre of gravity and the covariance matrix are affine-covariants, *i.e.* if the input point set is transformed by an affine transformation A, they transform with the same transformation A.

The proposed method searches μ and P by minimization over all possible affine transformations with a fixed determinant. In turn, μ is fully affine covariant and P is affine covariant up to an unknown scale (and rotation, $P^\top P$ cancels the rotation). To the best of our knowledge, this type of robust-to-outliers covariants have not been used.

7 Conclusions

We cast the problem of data whitening as minimization of robust cost functions. In this fashion we iteratively estimate a whitening transformation that is robust to the presence of outliers. With the use of synthetic data, we show that our estimation is almost unaffected even with extreme cases of outliers, while it also offers improvements when whitening CNN descriptors for image retrieval.

References

1. Aftab, K., Hartley, R.: Convergence of iteratively re-weighted least squares to robust M-estimators. In: IEEE Winter Conference on Applications of Computer Vision (2015)
2. Arandjelović, R., Zisserman, A.: Three things everyone should know to improve object retrieval. In: CVPR (2012)
3. Arandjelovic, R., Zisserman, A.: All about VLAD. In: CVPR (2013)
4. Babenko, A., Lempitsky, V.: Aggregating local deep features for image retrieval. In: ICCV (2015)
5. Bo, L., Sminchisescu, C.: Efficient match kernel between sets of features for visual recognition. In: NIPS (2009)
6. Cai, H., Mikolajczyk, K., Matas, J.: Learning linear discriminant projections for dimensionality reduction of image descriptors. IEEE Trans. PAMI **33**(2), 338–352 (2011)
7. Candès, E.J., Li, X., Ma, Y., Wright, J.: Robust principal component analysis? J. ACM **58**(3), 11 (2011)
8. Carreira, J., Caseiro, R., Batista, J., Sminchisescu, C.: Semantic segmentation with second-order pooling. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7578, pp. 430–443. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33786-4_32
9. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR (2005)
10. De la Torre, F., Black, M.J.: Robust principal component analysis for computer vision. In: ICCV (2001)
11. Gordo, A., Almazan, J., Revaud, J., Larlus, D.: Deep image retrieval: learning global representations for image search. In: arXiv (2016)
12. Huber, P.J.: Projection pursuit. In: The annals of Statistics (1985)
13. Jégou, H., Chum, O.: Negative evidences and co-occurences in image retrieval: the benefit of PCA and Whitening. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, pp. 774–787. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33709-3_55
14. Jegou, H., Perronnin, F., Douze, M., Sánchez, J., Perez, P., Schmid, C.: Aggregating local image descriptors into compact codes. IEEE Trans. PAMI **34**(9), 1704–1716 (2012)
15. Ke, Y., Sukthankar, R.: PCA-SIFT: a more distinctive representation for local image descriptors. In: CVPR (2004)
16. Lim, J.J., Zitnick, C.L., Dollár, P.: Sketch tokens: a learned mid-level representation for contour and object detection. In: CVPR (2013)
17. Lowe, D.G.: Object recognition from local scale-invariant features. In: ICCV (1999)
18. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide-baseline stereo from maximally stable extremal regions. Image Vis. Comput. **22**(10), 761–767 (2004)
19. Perronnin, F., Dance, C.: Fisher kernels on visual vocabularies for image categorization. In: CVPR (2007)
20. Poppe, R.: A survey on vision-based human action recognition. In: Image and Vision Computing (2010)
21. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: arXiv (2014)

22. Tolias, G., Sicre, R., Jégou, H.: Particular object retrieval with integral max-pooling of CNN activations. In: arXiv (2015)
23. Turk, M.A., Pentland, A.P.: Face recognition using eigenfaces. In: CVPR (1991)
24. Wright, J., Ganesh, A., Rao, S., Peng, Y., Ma, Y.: Robust principal component analysis: exact recovery of corrupted low-rank matrices via convex optimization. In: NIPS (2009)
25. Xu, H., Caramanis, C., Sanghavi, S.: Robust PCA via outlier pursuit. In: NIPS (2010)