Robust Principal Component Analysis for Generalized Multi-View Models

Frank Nussbaum[1,2] **Joachim Giesen**[1]

[1]Friedrich-Schiller-Universität, Jena, Germany
[2]DLR Institute of Data Science, Jena, Germany

Abstract

It has long been known that principal component analysis (PCA) is not robust with respect to gross data corruption. This has been addressed by robust principal component analysis (RPCA). The first computationally tractable definition of RPCA decomposes a data matrix into a low-rank and a sparse component. The low-rank component represents the principal components, while the sparse component accounts for the data corruption. Previous works consider the corruption of individual entries or whole columns of the data matrix. In contrast, we consider a more general form of data corruption that affects groups of measurements. We show that the decomposition approach remains computationally tractable and allows the exact recovery of the decomposition when only the corrupted data matrix is given. Experiments on synthetic data corroborate our theoretical findings, and experiments on several real-world datasets from different domains demonstrate the wide applicability of our generalized approach.

1 INTRODUCTION

Principal component analysis (PCA) is a classical data dimension reduction technique based on the assumption that given high-dimensional data lies near some low-dimensional subspace, see Pearson [1901]. Formally, assume observed data points $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^m$ that are combined into a data matrix $X \in \mathbb{R}^{m \times n}$. Approximating the data matrix X by a low-rank matrix L can be formulated as an optimization problem

$$\min_{L \in \mathbb{R}^{m \times n}} \|X - L\| \text{ subject to } \operatorname{rank}(L) \leqslant k,$$

where $\|\cdot\|$ is some suitable norm. The classical and still popular choice, see Hotelling [1933], Eckart and Young [1936], uses the Frobenius norm $\|\cdot\|_F$, which renders the optimization problem tractable. The Frobenius norm does not perform well though for *grossly corrupted data*. A single grossly corrupted entry in X can change the estimated low-rank matrix L significantly, that is, the Frobenius norm approach is not robust against data corruption. An obvious remedy is replacing the Frobenius norm by the ℓ_1-norm $\|\cdot\|_1$, but this renders the optimization problem intractable because of the non-convex rank constraint. Alternatively, we can explicitly model a component that captures data corruption. This leads to a decomposition

$$X = L + S$$

of the data matrix into a low-rank component L as before and a matrix S of outliers. The structure of the outlier matrix S depends on the data corruption mechanism and is commonly assumed to be sparse. In practice, low-rank + sparse decompositions can be computed efficiently through the convex problem

$$\min_{L,\, S \in \mathbb{R}^{m \times n}} \|L\|_* + \gamma \|S\|_{1,2} \text{ subject to } X = L + S, \quad (1)$$

where $\gamma > 0$ is a trade-off parameter between the nuclear norm $\|\cdot\|_*$, which promotes low rank on L, and the $\ell_{1,2}$-norm $\|S\|_{1,2} = \sum_g \|s_g\|_2$, which promotes *structured* sparsity on S given a partitioning of the entries in S into groups s_g. The groups are determined by the assumed data corruption mechanism.

In the simplest data corruption mechanism, *individual* entries of the data points can be corrupted. Under this model, the $\ell_{1,2}$-norm reduces to the ℓ_1-norm, that is, the groups s_g consist of single elements. Wright et al. [2009], Candès et al. [2011], and Chandrasekaran et al. [2011] were first to investigate this model. They show that exact recovery using the specialized version of Problem (1) is possible, that is, in many cases the corruptions can be separated from the data.

Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021), PMLR 161:686–695.

An alternative corruption mechanism that was introduced independently by McCoy and Tropp [2011] and Xu et al. [2010] corrupts *whole* data points, which are referred to as outliers. Here, the groups s_g for the $\ell_{1,2}$-norm are the columns of the data matrix X. Xu et al. [2010] show that exact recovery is also possible in the column-sparse scenario.

In this work, we study a more general data corruption mechanism, where the data points are partitioned as

$$x^{(i)} = (x_1^{(i)}, \dots, x_d^{(i)}) \in \mathbb{R}^{m_1} \times \dots \times \mathbb{R}^{m_d} = \mathbb{R}^m,$$

that is, they form d groups. We assume that for each data point, the groups can be individually corrupted. Hence, the groups in Problem (1) are given by

$$S = \left(\begin{array}{c|c|c|c} s_{11} & s_{12} & \cdots & s_{1n} \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline s_{d1} & s_{d2} & \cdots & s_{dn} \end{array}\right) \in \mathbb{R}^{m \times n},$$

where $s_{ij} \in \mathbb{R}^{m_i}$. In Section 2, we show that exact recovery is still possible for our more general data corruption mechanism. This mechanism has a natural interpretation in terms of generalized multi-view models Sun [2013], Zhao et al. [2017], Zhang et al. [2019], where each data point is obtained by measurements from different sensors, and every sensor can measure several variables. Sensor failures in this model result in corrupted measurements for the group of variables measured by the failing sensor, but only for the data points that were measured while the sensor was not working correctly. Of course, data corruption and sensor failures are an abstraction for what can also be anomalies or outliers in applications, see Figure 1 for a real-world example.

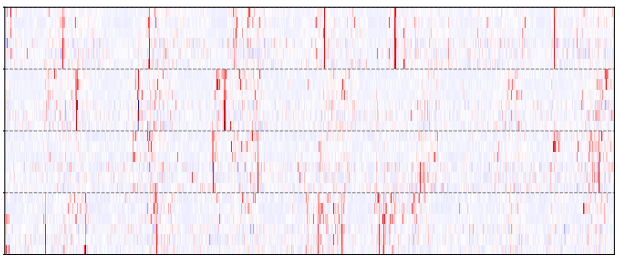

Figure 1: The electrical load profiles of four households from one week. Power consumption for each household is measured in terms of six quantities per time step, which form groups of six elements. In the plot of the data, these groups respectively span six rows, whereas the columns represent time steps. Note that for this data, we expect data corruption (outliers) in the form of short-term usage of electrical devices. More details can be found in Section 3.

Observe that the approach in Wright et al. [2009], Candès et al. [2011], Chandrasekaran et al. [2011] corresponds to the special case where each sensor just measures a single variable, and the approach in McCoy and Tropp [2011], Xu et al. [2010] corresponds to the special case where a single sensor measures all variables.

Of course, it is possible to think of data corruption mechanisms that correspond to even more general group structures, for example, rectangular groups formed by sub-matrices of the data matrix. The latter case does not pose any extra technical challenges. Hence, here we keep the exposition simple and stick with generalized multi-view models. These models are flexible and suitable for many real-world applications. We provide some examples in Section 3, namely, the identification of periods of large power consumption from electrical grid data, the reconstruction of RGB images, and the detection of weather anomalies using wave hindcast data.

Some of the applications have data in the form of tensors. Therefore, we briefly discuss some additional important related work that concerns robust tensor principle component analysis (RTPCA). The most closely related works follow the convex optimization approach. Their main modeling effort lies in the generalization of low rank and the nuclear norm to tensors. For example, Huang et al. [2014] propose RTPCA using the sum of nuclear norms (SNN), which is based on Tucker rank. For 3D tensors, Zhang et al. [2014], Lu et al. [2016, 2019], Zhou and Feng [2017] use a nuclear norm based on t-SVD and tensor tubal rank. Most works, including Huang et al. [2014], Lu et al. [2016, 2019], assume the simple data corruption mechanism that corrupts individual entries of the data tensor. Zhang et al. [2014], Zhou and Feng [2017] consider outliers distributed along slices of 3D tensors. The latter data corruption mechanism is a special case of our multi-view models when all groups have the same size (and the data matrix is viewed as a flattened version of a tensor). However, our general multi-view models allow for different group sizes, which gives them additional flexibility and distinguishes them from all existing RTPCA models.

2 EXACT RECOVERY

In this section, we prove exact recovery for generalized multi-view models. For that, we assume a data matrix X with underlying *true* decomposition $X = L^\star + S^\star$ into a low-rank matrix $L^\star$ and a group-sparse matrix $S^\star$. We investigate under which conditions the pair $(L^\star, S^\star)$ can be obtained as the guaranteed solution to Problem (1) with suitably chosen γ.

Algebraic varieties and tangent spaces. Low rank and group sparsity are algebraic properties that can be formalized by algebraic matrix varieties. We need them for our analysis, hence here we briefly introduce them. First, the low-rank matrix variety of matrices with rank at most r is given by

$$\mathcal{L}(r) = \{L \in \mathbb{R}^{m\times n} : \operatorname{rank}(L) \leqslant r\}.$$

As for example shown in Shalit et al. [2010], a rank-r matrix L is a smooth point in $\mathcal{L}(r)$ and has tangent space

$$\mathcal{T}(L) = \left\{UX^\top + YV^\top : X \in \mathbb{R}^{n\times r}, Y \in \mathbb{R}^{m\times r}\right\},$$

where $L = UDV^\top$ is the (restricted) singular value decomposition of L with $U \in \mathbb{R}^{m\times r}$, $V \in \mathbb{R}^{n\times r}$, and diagonal $D \in \mathbb{R}^{r\times r}$. Second, the variety of group-structured matrices with at most s non-zero groups is given by

$$\mathcal{S}(s) = \{S \in \mathbb{R}^{m\times n} : |\operatorname{gsupp}(S)| \leqslant s\},$$

where

$$\operatorname{gsupp}(S) = \left\{(i,j) : 1 \leqslant i \leqslant d, 1 \leqslant j \leqslant n, s_{ij} \not\equiv 0\right\}$$

is the *group support* of S. Remember that s_{ij} is the sub-vector of the j-th column of S that corresponds to the i-th group of variables. A matrix $S \in \mathcal{S}(s)$ with $|\operatorname{gsupp}(S)| = s$ is a smooth point in $\mathcal{S}(s)$ with tangent space

$$\mathcal{Q}(S) = \{A \in \mathbb{R}^{m\times n} : \operatorname{gsupp}(A) \subseteq \operatorname{gsupp}(S)\}.$$

An intuitive version of Problem (1). Instead of directly analyzing Problem (1), it turns out to be useful to first consider the non-convex feasibility problem of finding L and S such that

$$L \in \mathcal{L}(\operatorname{rank}(L^\star)),\quad S \in \mathcal{S}(|\operatorname{gsupp}(S^\star)|),\quad \text{and } X = L + S. \tag{2}$$

Netrapalli et al. [2014] tried to directly solve a non-convex problem similar to Problem (2) for the non-group case, where individual entries can be corrupted. They assumed a priori estimates of rank and sparsity. However, the true varieties in Problem (2) are unknown in practice. Yet studying Problem (2) leads to necessary conditions for successful recovery.

A first observation is that clearly, the true components $(L^\star, S^\star)$ solve Problem (2). We want this solution to be (locally) unique since we are interested in unique recovery. Any other *nearby* feasible decomposition for Problem (2) must satisfy

$$(L^\star - \Delta, S^\star + \Delta) \in \mathcal{L}(\operatorname{rank}(L^\star)) \times \mathcal{S}(|\operatorname{gsupp}(S^\star)|)$$

for some *small* matrix Δ. Here, if $S^\star + \Delta \in \mathcal{S}(|\operatorname{gsupp}(S^\star)|)$ for small Δ, then it must hold $\Delta \in \mathcal{Q}(S^\star)$. In contrast, $L^\star - \Delta \in \mathcal{L}(\operatorname{rank}(L^\star))$ for small Δ only implies that Δ is contained in a tangent space to the low-rank matrix variety that is close to $\mathcal{T}(L^\star)$. This is due to the local curvature of the low-rank matrix variety. Nevertheless, for local uniqueness of the solution to Problem (2), it suffices to only consider the tangent spaces $\mathcal{T}(L^\star)$ and $\mathcal{Q}(S^\star)$:

Lemma 1. *If the tangent spaces $\mathcal{T}(L^\star)$ and $\mathcal{Q}(S^\star)$ are* transverse*, that is, if*

$$\mathcal{T}(L^\star) \cap \mathcal{Q}(S^\star) = \{0\},$$

then the solution $(L^\star, S^\star)$ to Problem (2) is locally unique.

We prove this lemma in the supplementary material. Next, we discuss when tangent spaces are transverse.

Sufficient condition for transversality. One basic scenario with non-transverse tangent spaces $\mathcal{T}(L^\star)$ and $\mathcal{Q}(S^\star)$ can occur when at least one of the components $L^\star$ or $S^\star$ is simultaneously low rank and group sparse. Therefore, to avoid confusion of the components, we constrain the matrices $L^\star$ and $S^\star$ such that they cannot be both group sparse and low rank at the same time.

For that, we first define the *maximum group degree* $\operatorname{gdeg}_{\max}(S^\star)$ as the maximum number of non-zero groups that appear in a row or column of $S^\star$. If $\operatorname{gdeg}_{\max}(S^\star)$ is small, then the non-zero groups are *not* concentrated in just a few rows and columns, which means that the matrix $S^\star$ is likely not low rank. Next, to ensure that $L^\star$ is not group sparse, we want its entries to be spread-out. This is the case if the row and column spaces of $L^\star$ are *incoherent*, that is, if

$$\operatorname{coh}(L^\star) = \max\{\operatorname{coh}(\operatorname{colspace}(L^\star)), \operatorname{coh}(\operatorname{rowspace}(L^\star))\}$$

is small. Here, the incoherence of a subspace $V \subseteq \mathbb{R}^n$ is defined as $\operatorname{coh}(V) = \max_i \|P_V e_i\|_2$, that is, as the maximum length of a projected standard-basis vector e_i of $\mathbb{R}^n$. Incoherence thus measures how well the subspace is aligned with the standard coordinate axes. A high value indicates that the subspace is well-aligned. Having introduced these notions, it turns out that bounding the product $\operatorname{coh}(L^\star)\operatorname{gdeg}_{\max}(S^\star)$ implies *transversality* of the tangent spaces.

Lemma 2. *Define $\eta = \max_{i=1}^d m_i$ to be the maximum number of variables that a group spans. Let*

$$\operatorname{coh}(L^\star)\operatorname{gdeg}_{\max}(S^\star) < 1/2\eta^{-3/4}.$$

Then, it holds that $\mathcal{T}(L^\star) \cap \mathcal{Q}(S^\star) = \{0\}$.

The proof of Lemma 2 can be found in the supplementary material. The insights from analyzing Problem (2) can be used to prove similar results for the convex Problem (1). Indeed, an only slightly stronger upper bound on the product $\operatorname{coh}(L^\star)\operatorname{gdeg}_{\max}(S^\star)$ implies exact recovery of $(L^\star, S^\star)$ by Problem (1) for a range of values of γ.

Theorem 1. *Let* $X = L^\star + S^\star$ *and suppose that*

$$\operatorname{coh}(L^\star)\operatorname{gdeg}_{\max}(S^\star) < 1/12\,\eta^{-3/4}.$$

Then, the range $(\gamma_{\min}, \gamma_{\max})$ *is non-empty, where*

$$\gamma_{\min} = \frac{2\eta^{1/2}\ \operatorname{coh}(L^\star)}{1 - 8\eta^{3/4}\operatorname{coh}(L^\star)\operatorname{gdeg}_{\max}(S^\star)}$$
$$\gamma_{\max} = \frac{1 - 6\eta^{3/4}\operatorname{coh}(L^\star)\operatorname{gdeg}_{\max}(S^\star)}{\eta^{1/4}\operatorname{gdeg}_{\max}(S^\star)}$$

Moreover, for any γ *in that range, Problem* (1) *with regularization parameter* γ *has the* unique *solution* $(L^\star, S^\star)$.

This result generalizes the recovery results in Chandrasekaran et al. [2011] by including the dependency on η. Theorem 1 can in fact be strengthened by working with more technical notions instead of incoherence and maximum group degree. These technical notions are norm-compatibility constants that measure how the $\ell_{\infty,2}$- and spectral norms compare for elements from different tangent spaces:

$$\mu(\mathcal{Q}(S)) = \max_{M \in \mathcal{Q}(S),\, \|M\|_{\infty,2}=1} \|M\| \quad \text{and}$$
$$\xi(\mathcal{T}(L)) = \max_{M \in \mathcal{T}(L),\, \|M\|=1} \|M\|_{\infty,2}.$$

Here, $\mathcal{Q}(S)$ is the tangent space at a point S to the group-sparse matrix variety $\mathcal{S}(|\operatorname{gsupp}(S)|)$, and $\mathcal{T}(L)$ is the tangent space at a point L to the low-rank matrix variety $\mathcal{L}(\operatorname{rank}(L))$. The following lemma, which we prove in the supplementary material, relates the norm compatibility constants with maximum group degree and incoherence.

Lemma 3. *Let* $S \in \mathcal{S}(|\operatorname{gsupp}(S)|)$ *and* $L \in \mathcal{L}(\operatorname{rank}(L))$. *Then, the following bounds hold:*

$$\operatorname{gdeg}_{\max}(S) \geqslant \eta^{-1/4}\mu(\mathcal{Q}(S)) \quad \textit{and}$$
$$\operatorname{coh}(L) \geqslant 1/2\eta^{-1/2}\ \xi(\mathcal{T}(L)).$$

Due to these lower bounds, the product $\operatorname{coh}(L^\star)\operatorname{gdeg}_{\max}(S^\star)$ in Theorem 1 can only be small if also the product $\mu(\mathcal{Q}(S^\star))\xi(\mathcal{T}(L^\star))$ is small. This is reflected in the assumption of the next theorem that improves Theorem 1.

Theorem 2. *Let* $X = L^\star + S^\star$ *and suppose that*

$$\mu(\mathcal{Q}(S^\star))\xi(\mathcal{T}(L^\star)) < 1/6.$$

Then, the range $(\gamma^\circ_{\min}, \gamma^\circ_{\max})$ *is non-empty, where*

$$\gamma^\circ_{\min} = \frac{\xi(\mathcal{T}(L^\star))}{1 - 4\mu(\mathcal{Q}(S^\star))\xi(\mathcal{T}(L^\star))}$$
$$\gamma^\circ_{\max} = \frac{1 - 3\mu(\mathcal{Q}(S^\star))\xi(\mathcal{T}(L^\star))}{\mu(\mathcal{Q}(S^\star))}$$

Moreover, for any γ *in that range, Problem* (1) *with regularization parameter* γ *has the* unique *solution* $(L^\star, S^\star)$.

We prove Theorem 2 in the supplementary material. Theorem 1 can be proven as a simple consequence of Theorem 2 using the bounds from Lemma 3.

It should be noted that in real-word situations, Theorem 1 and Theorem 2 provide little guidance for the choice of γ. This is because the true maximum group degree and incoherence are unknown, let alone the norm compatibility constants. This leaves the choice of γ up to the user. In the experimental section, we investigate two heuristics for selecting the parameter γ. Since we also intend to experiment with synthetic data, we need to generate *random* low-rank + group-sparse decompositions. Therefore, we introduce a random decomposition model and provide a theoretical result that concerns the recovery of random decompositions drawn from this model.

Random decompositions. As in Candès and Recht [2009], we assume that a rank-r-matrix $L^\star$ is drawn from the *random orthogonal model*, that is, by setting $L^\star = UV^\top/\sqrt{mn}$, where $U \in \mathbb{R}^{m\times r}$ and $V \in \mathbb{R}^{n\times r}$ are drawn at random with independent standard Gaussian entries. The column spaces of U and V are incoherent with high probability. Indeed, by [Candès and Recht, 2009, Lemma 2.2], there exists a constant c such that it holds

$$\begin{aligned}\operatorname{coh}(L^\star) &= \max\{\operatorname{coh}(\operatorname{colspace}(U)), \operatorname{coh}(\operatorname{colspace}(V))\} \\ &\leqslant c \max\left\{\frac{\max(r, \log m)}{m}, \frac{\max(r, \log n)}{n}\right\}\end{aligned} \tag{3}$$

with high probability, that is, with a probability that converges to one as m and n grow to infinity.

Next, we sample $S^\star$ as follows: First, the group support $\operatorname{gsupp}(S^\star)$ is sampled at random using independent Bernoulli variables, where each group is non-zero with probability p. Note that this type of sampling is characteristic for $G(n, p)$ random graph models, see for example Bollobás [2001]. As in Candès et al. [2011], we sample the entries of the groups that belong to the

support uniformly at random from $\{-1, 1\}$. Under this random group sparsity model, the maximum group degree is independent from the precise values of the non-zero entries. Specifically, the following holds:

Lemma 4. *If $S^\star$ is drawn from the random group sparsity model, then the maximum group degree satisfies with high probability (that converges to one as n and d grow to infinity) that*

$$\mathrm{gdeg}_{\max}(S^\star) \leqslant 2ap + 3\sqrt{ap},$$

where $a = \max\{n, d\}$.

Proof. To bound the maximum group degree, we must bound the number of non-zero groups in each row and column. We bound the number of non-zero groups for a single row first. The result then follows from applying a union bound.

The number of non-zero groups in a fixed row is a binomially-distributed random variable $Z \sim \mathrm{Bin}(n, p)$. A consequence of Talagrand's inequality is that for $0 \leqslant t \leqslant np = \mathbb{E}Z$ it holds

$$\mathbb{P}(Z \geqslant np + t + 3\sqrt{np}) \leqslant \exp\left(-t^2/(16np)\right),$$

see Habib et al. [2013]. We set $t = np$ and obtain

$$\mathbb{P}(Z \geqslant 2np + 3\sqrt{np}) \leqslant \exp\left(-np/16\right).$$

Similarly, we have for the columns that

$$\mathbb{P}(Z \geqslant 2dp + 3\sqrt{dp}) \leqslant \exp\left(-dp/16\right).$$

Hence, with $a = \max(n, d)$ and by the union bound, the probability that any row or column has more than $2ap + 3\sqrt{ap}$ non-zero groups is at most $m\exp\left(-np/16\right) + n\exp\left(-dp/16\right)$, which is small for (comparably) large n and d. □

The following corollary shows that if the group-selection probability p is not too high, then random decompositions can be recovered exactly with high probability.

Corollary 1. *Let $(L^\star, S^\star)$ be sampled from the random decomposition model with sufficiently large n and d. Let*

$$p < \frac{\left(\sqrt{9 + 2/3\eta^{-3/4}/\kappa} - 3\right)^2}{16a},$$

where

$$\kappa = c\max\left\{\frac{\max(r, \log m)}{m}, \frac{\max(r, \log n)}{n}\right\}$$

is as in Inequality (3). Then, the assumption of Theorem 1 holds with high probability. Hence, with high probability, the components $(L^\star, S^\star)$ are the guaranteed solution to Problem (1) with input $X = L^\star + S^\star$ and $\gamma \in (\gamma_{\min}, \gamma_{\max})$.

Proof. We show that the assumption

$$\mathrm{coh}(L^\star)\,\mathrm{gdeg}_{\max}(S^\star) < 1/12\,\eta^{-3/4}$$

of Theorem 1 holds with high probability. Using the upper bound on the maximum group degree from Lemma 4 and that by Inequality (3) the incoherence satisfies $\mathrm{coh}(L^\star) \leqslant \kappa$ with high probability, it suffices to show that

$$(2ap + 3\sqrt{ap})\,\kappa < 1/12\,\eta^{-3/4}.$$

This is a quadratic inequality in $\sqrt{p}$. Solving it yields

$$\sqrt{p} < \frac{-3}{4\sqrt{a}} + \sqrt{\frac{9}{16a} + \frac{1}{24a}\eta^{-3/4}\kappa^{-1}}.$$

It can be checked that this is equivalent to the assumption of Corollary 1 by taking squares. This finishes the proof after applying Theorem 1. □

Note that the right-hand side of the inequality in Corollary 1 becomes small if r is large. Hence, for large r, the group-selection probability p is required to be small in order to guarantee exact recovery with high probability. Moreover, if r and p are both small, then exact recovery should be easy.

3 EXPERIMENTS

In this section, we experiment with synthetic and real-world data. We adapt the alternating direction method of multipliers (ADMM) Boyd et al. [2011] for solving Problem (1), see the supplementary material. To accelerate our solver, we use fast randomized singular value thresholding based on Halko et al. [2011].

3.1 SYNTHETIC DATA

In our first experiment, we intent to experimentally verify the theory from the previous section. For that, we generate synthetic data in the form of random pairs $(L^\star, S^\star)$ that we sample according to the random decomposition model that we introduced before. For each random decomposition $(L^\star, S^\star)$, we check if Problem (1) can be used to exactly recover $(L^\star, S^\star)$, using only the compound matrix $L^\star + S^\star$ as input. Here, our main goal is to vary the rank r of $L^\star$ and the group-selection probability p for $S^\star$, where as a consequence of Corollary 1, we expect that successful recovery is more likely possible if $(L^\star, S^\star)$ is sampled with not too large r and p.

More specifically for this experiment, we fix $n = 500$ and the group structure $x = (x_1, \ldots, x_{100}) \in \mathbb{R}^{500}$, where each group $x_i \in \mathbb{R}^5$ consists of five features.

Hence, $X = L^\star + S^\star \in \mathbb{R}^{500\times 500}$. Then, for selected pairs (r, p), we respectively create 10 different random decompositions to average out sampling effects. We try to recover these decompositions by solving instances of Problem (1). However, we still need to choose a suitable regularization parameter γ for each problem. In the following, we compare the rates of successful recovery of two heuristics for choosing γ.

For the first heuristic, observe that according to Theorem 1 exact recovery is possible for a *range* of values for γ. Hence, if successful recovery is possible for a problem, then we expect that there exists an interval of regularization parameters that yield the correct solution. In particular, the solution is the same for all γ from this interval–we say that the solution is *stable* in this interval. As in Chandrasekaran et al. [2011], we use this fact to search for an interval of values for γ, where the solution to Problem (1) is stable (and both components are non-zero). If the search for such an interval is successful, then we check if the solution, which is the same for all γ from the interval, has the correct algebraic properties. If this is the case, we consider the recovery for the given problem as successful. Otherwise we declare failure.

For convenience, we rewrite the objective of Problem (1) as $(1-\alpha)\|L\|_* + \alpha\|S\|_{1,2}$ and denote its solution by (L_α, S_α), where α is in the *compact* interval $[0, 1]$. Then, we equivalently search for an interval of values for α, where the solution does not change. For that, we track how the solution changes by calculating the differences

$$diff_\alpha = \|L_{\alpha-\delta} - L_\alpha\|_F + \|S_{\alpha-\delta} - S_\alpha\|_F$$

along the solution path obtained from a grid search with step size $\delta = 10^{-2}$, see Figure 2.

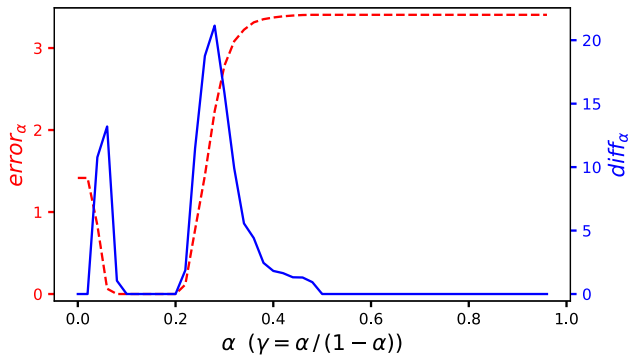

Figure 2: Search for a stable solution to Problem (1). The blue line shows the change of the solution at each step of the grid search. The red line shows the recovery error. For roughly $\alpha \in [0.1, 0.2]$, the solution is stable with almost zero recovery error.

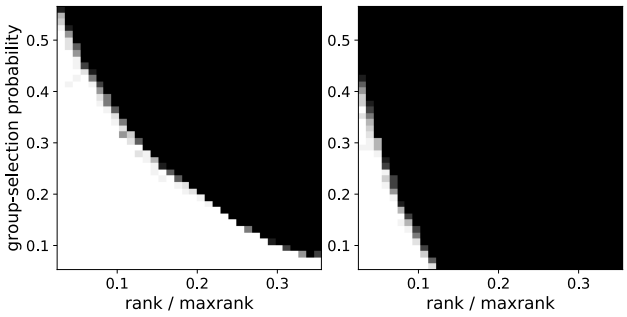

Figure 3: Recovery results for varying rank r (displayed as fractions $r/\max(m, n)$ of the maximum possible rank) and varying group-selection probability p. Trials were repeated 10 times for selected pairs (r, p). Empirical success probabilities are encoded as grey values, where white indicates a probability of 1 and black a probability of 0. The left plot shows the results when γ is selected based on a search for a stable solution. The right plot shows the results that correspond to the ad-hoc choice $\gamma = 1/\sqrt{\max(m, n)}$.

The change of the solution follows a typical pattern, which can also be seen in Figure 2. There are three intervals, where the solution is stable: First, for very small values of α, there is little group-sparse regularization, hence the solution has a zero low-rank component. Likewise, for too large values of α, the solution always has a zero group-sparse component. The third interval with a stable solution in the middle is the one that we are looking for. In the example shown in Figure 2, the recovery error

$$error_\alpha = \|L_\alpha - L^\star\|_F + \|S_\alpha - S^\star\|_F$$

is close to zero for all values in this interval. Note that the recovery error is unknown in practice.

The search for γ (or equivalently α) as outlined above requires solving several instances of Problem (1). Therefore, as a second heuristic, we also compare the rate of successful recovery to the rate when the ad-hoc choice $\gamma = 1/\sqrt{\max(m, n)}$ is used instead of searching. This value was suggested for learning RPCA decompositions under entry-wise data corruption, see Candès et al. [2011].

The results of the experiment, see Figure 3, support the theory and effectively demonstrate that exact recovery is possible. Moreover, they confirm that for smaller r and p, that is, for smaller ranks and group-selection probabilities, successful recovery becomes easier.

The results also show that it may pay off to perform the search for an interval, where the solution is stable. This is because decompositions with much greater ranks and group-selection probabilities can still be recovered

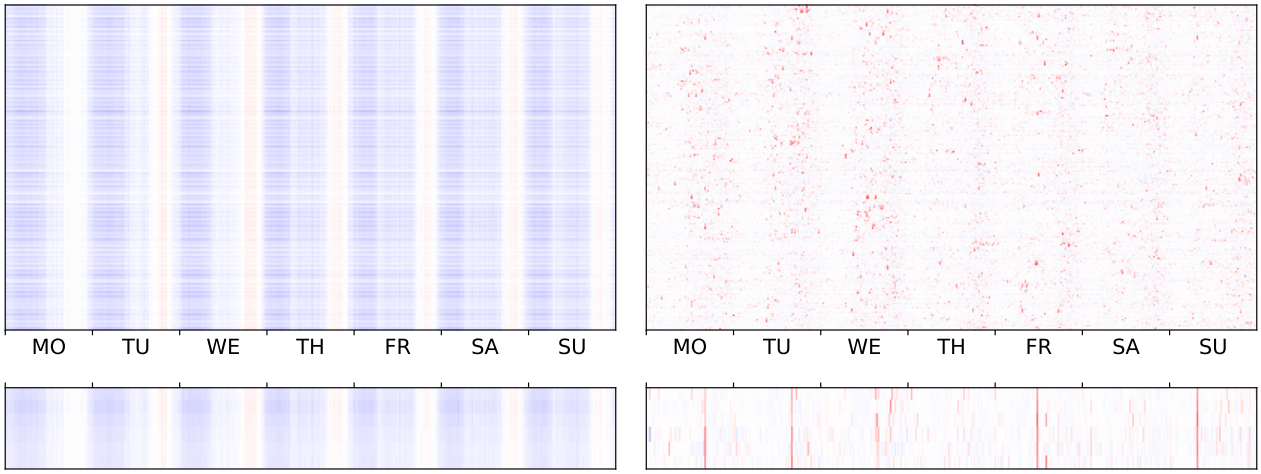

Figure 4: Decomposition of the electrical load profiles of 74 households over the course of one week. The left plot shows a typical repeated low-rank day-and-night pattern. The right plot shows the outlier component that captures periods of large loads, when some electrical devices consumed power. On the bottom, the decomposition for a single household is shown.

successfully, though using the ad-hoc choice fails. On the other hand, the ad-choice can also be tuned by hand if there is a priori knowledge about the solution. For example, if the outlier matrix is very sparse, then a larger value of γ can be used.

3.2 REAL-WORLD DATA

In the following, our goal is to demonstrate the wide applicability of robust principal component analysis for generalized multi-view models. For that, we briefly discuss three real-world applications.

Identification of power consumption. In households, aside from the base load, power consumption usually takes place infrequently and during a limited period of time when electrical devices are turned on. Thus, momentary power consumption in households has characteristic features of outliers. Hence, we first show that our model can be used to identify periods of large power consumption from electrical grid data.

Specifically, we use a dataset that contains the electrical load profiles of 74 representative German residential buildings from the year 2010. The dataset, which was obtained from Tjaden et al. [2015], constitutes a time series with a temporal resolution of one second. For illustrative purposes, we restrict the dataset to the first week. For each residential building, the electrical load profiles consist of six quantities that correspond to three phases, respectively, of active and reactive power. Hence, each of the 74 residential buildings entails a group of six elements, and thus at each time step a 444-dimensional vector is observed. In total, the data matrix is of size $444 \times 10\,080$, including one observation per second of the week. Note that sample data from the first four households is shown in Figure 1.

The solution to Problem (1) is stable around $\gamma = 10^{-2}$. Figure 4 shows the corresponding decomposition. There is a noticeable general pattern of electrical load profiles that is explained by the alternation of day and night: During sleeping hours there are few devices that consume power. However, during day-time hours there generally is increased activity, with the most electrical power being consumed in the evening hours. The low-rank component of the decomposition in Figure 4 captures the repeated general pattern. Meanwhile the group-sparse component identifies periods of larger electrical loads, caused by electrical devices that momentarily consumed power.

Reconstruction of RGB images (multi-view data). Here, we briefly show that robust PCA for generalized multi-view models can be used to improve RGB images. For that, we apply our robust PCA model on a multi-view dataset that consists of images from the *Amsterdam Library of Object Images* Geusebroek et al. [2005], which is equipped with additional views from Schubert and Zimek [2019]. In the dataset, the data points are RGB images of the same object under 36 different light conditions. Each image has 144×192 pixels, where each pixel constitutes a different view of the image. The first additional view for each image consists of the first 13 Haralick features (radius 1 pixel), see Haralick [1979], and the second additional view is

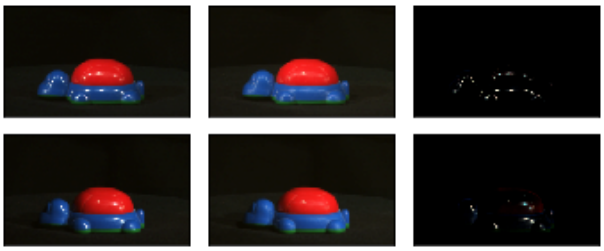

Figure 5: Robust recovery of RGB images. From left to right, the original images, the reconstructed low-rank images, and the outlier components are shown. Overexposures have been removed in the low-rank component and appear in the outlier component.

a standard RGB color histogram with 8 uniform bins. The whole data matrix has dimensions $82\,965 \times 36$. Exemplary results of applying robust PCA with $\gamma = 10^{-2}$ for a typical object of the Amsterdam Library of Object Images are shown in Figure 5. In the low-rank component, spotlights have been reduced. Additional results from datasets with other objects from the Amsterdam Library of Object Images can be found in the supplementary material.

Detection of weather anomalies. The wave hindcast dataset *coastdat1* Helmholtz Centre for Materials and Coastal Research [2012] contains a time series of wave conditions in the southern North Sea. We use data for the year 2007 with a resolution of one hour. The covered area is 51.0N to 56.5N and $-3.0W$ to 10.5E, using a grid size of approximately 0.05 degrees latitude and 0.10 degrees longitude. At each grid point, the sea state is described by the variables *significant weight height* (*hs*) and *mean wave period* (*mp*), which are derived from 2D wave spectra Groll and Weisse [2016].

The sea state at each grid position naturally defines a group of two parameters. Hence, to apply robust PCA for these groups, we change the data representation for a single time step from grid to a vector that contains the groups from all 6 324 sea-side grid positions. Hence, the data matrix has overall size $12\,648 \times 8\,760$, where each column corresponds one hour of the year.

The resulting decomposition for $\gamma = 10^{-3}$ can be found in Figure 6. Here, we only show the decomposition for selected time steps, and instead of the columns of the data matrix we show the covered area for the *mp* feature. The corresponding decompositions for the *hs* feature can be found in the supplementary material. We picked November, 9th as a special date since at this time there was a cyclone that caused severe floods, that is, a strong weather anomaly. This is reflected in the outlier component in Figure 6, which highlights

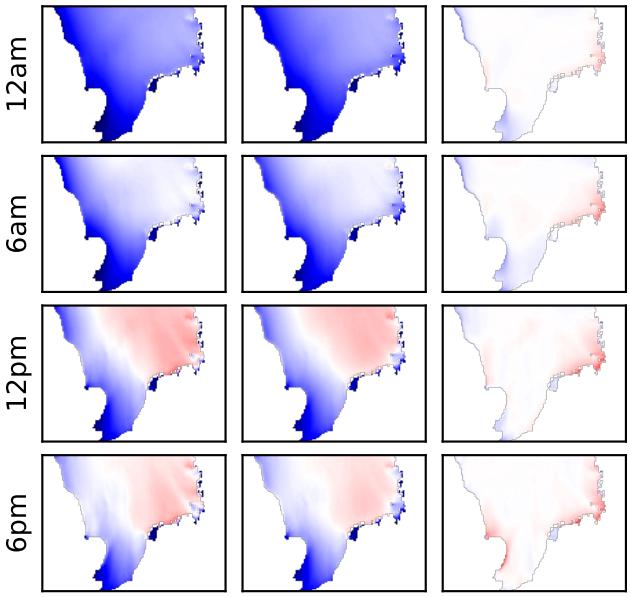

Figure 6: Wave hindcast data: The *mp* feature is shown from four time steps of November 9th, 2007 when cyclone Tilo caused severe North Sea floods (storm surges). From left to right, the columns show the original data, the low-rank components, and the outlier components. In the outlier components, the coastal lines show increased energy (red).

areas, where the storm was particularly strong. In the supplementary material, we also show the decomposition for a normal day without weather anomalies. This experiment shows that generalized multi-view RPCA models can also be used to detect anomalies.

4 CONCLUSION

In this work, we introduced robust principal component analysis for generalized multi-view models, where observations are structured in groups of measurements. A theoretically well-founded convex optimization problem can be used to separate principal components from groups of outliers. We empirically evaluated the rates of successful recovery for different decompositions using synthetic data. We presented a variety of real-world applications with naturally arising groups. The learned decompositions yield insights into the data, such as, general patterns and anomalies.

Acknowledgements

This work was supported by the German Science Foundation (DFG) grant (GI-711/5-1) within the priority program (SPP 1736) "Algorithms for Big Data". We thank Yanira Guanche for providing us with data.

References

Béla Bollobás. *Random graphs*. Number 73. Cambridge University Press, 2001.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717, 2009.

Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

Venkat Chandrasekaran, Sujay Sanghavi, Pablo A. Parrilo, and Alan S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Jan-Mark Geusebroek, Gertjan J. Burghouts, and Arnold W. M. Smeulders. The Amsterdam library of object images. *International Journal of Computer Vision*, 61(1):103–112, 2005.

Nikolaus Groll and Ralf Weisse. coastdat-2 North Sea wave hindcast for the period 1949-2014 performed with the wave model wam. *World Date Center for Climate (WDCC) at DKRZ*, 2016.

Michel Habib, Colin McDiarmid, Jorge Ramirez-Alfonsin, and Bruce Reed. *Probabilistic methods for algorithmic discrete mathematics*, volume 16. Springer Science & Business Media, 2013.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.

Robert M Haralick. Statistical and structural approaches to texture. *Proceedings of the IEEE*, 67(5): 786–804, 1979.

Helmholtz Centre for Materials and Coastal Research. coastdat-1 waves north sea wave spectra hindcast (1948-2007), 2012. Geesthacht.

Roger A. Horn and Charles R. Johnson. *Matrix analysis*. Cambridge university press, 2012.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441, 1933.

Bo Huang, Cun Mu, Donald Goldfarb, and John Wright. Provable low-rank tensor recovery. *Optimization-Online*, 4252(2):455–500, 2014.

Canyi Lu, Jiashi Feng, Yudong Chen, Wei Liu, Zhouchen Lin, and Shuicheng Yan. Tensor robust principal component analysis: Exact recovery of corrupted low-rank tensors via convex optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5249–5257, 2016.

Canyi Lu, Jiashi Feng, Yudong Chen, Wei Liu, Zhouchen Lin, and Shuicheng Yan. Tensor robust principal component analysis with a new tensor nuclear norm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

Michael McCoy and Joel A. Tropp. Two proposals for robust PCA using semidefinite programming. *Electronic Journal of Statistics*, 5:1935–7524, 2011.

Praneeth Netrapalli, UN Niranjan, Sujay Sanghavi, Animashree Anandkumar, and Prateek Jain. Non-convex robust PCA. In *Advances in Neural Information Processing Systems*, pages 1107–1115, 2014.

Frank Nussbaum and Joachim Giesen. Pairwise sparse + low-rank models for variables of mixed type. *Journal of Multivariate Analysis*, 178:104601, 2020. ISSN 0047-259X. doi: https://doi.org/10.1016/j.jmva.2020.104601. URL `http://www.sciencedirect.com/science/article/pii/S0047259X19303756`.

Karl Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2 (11):559–572, 1901.

Erich Schubert and Arthur Zimek. ELKI: A large open-source library for data analysis - ELKI release 0.7.5 "heidelberg". *CoRR*, abs/1902.03616, 2019. URL `http://arxiv.org/abs/1902.03616`.

Jssai Schur. Bemerkungen zur Theorie der beschränkten Bilinearformen mit unendlich vielen Veränderlichen. *Journal für die reine und Angewandte Mathematik*, 140:1–28, 1911.

Uri Shalit, Daphna Weinshall, and Gal Chechik. Online learning in the manifold of low-rank matrices. In *Advances in Neural Information Processing Systems*, pages 2128–2136, 2010.

Shiliang Sun. A survey of multi-view machine learning. *Neural Computing and Applications*, 23(7-8):2031–2038, 2013.

Tjarko Tjaden, Joseph Bergner, Johannes Weniger, and Volker Quaschning. Representative electrical load profiles of residential buildings in germany with a temporal resolution of one second. *ResearchGate: Berlin, Germany*, 2015.

Martin J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Information Theory*, 55(5):2183–2202, 2009.

Alistair Watson, G. Characterization of the subdifferential of some matrix norms. *Linear algebra and its applications*, 170:33–45, 1992.

John Wright, Arvind Ganesh, Shankar Rao, Yigang Peng, and Yi Ma. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In *Advances in neural information processing systems*, pages 2080–2088, 2009.

Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Robust PCA via outlier pursuit. In *Advances in Neural Information Processing Systems*, pages 2496–2504, 2010.

Rui Zhang, Feiping Nie, Xuelong Li, and Xian Wei. Feature selection with multi-view data: A survey. *Information Fusion*, 50:158–167, 2019.

Zemin Zhang, Gregory Ely, Shuchin Aeron, Ning Hao, and Misha Kilmer. Novel methods for multilinear data completion and de-noising based on tensor-SVD. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3842–3849, 2014.

Jing Zhao, Xijiong Xie, Xin Xu, and Shiliang Sun. Multi-view learning overview: Recent progress and new challenges. *Information Fusion*, 38:43–54, 2017.

Pan Zhou and Jiashi Feng. Outlier-robust tensor PCA. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2263–2271, 2017.