

robust principal component analysis (robust PCA)

Author and editors

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland

Editors: Konstantina Olioumtsevs, Ekkehard Schnoor, Aalto University, Espoo, Finland

*Corresponding author: alex.jung@aalto.fi

Abstract

Robust principal component analysis (robust PCA) decomposes a data matrix into the sum of a matrix of low rank and a matrix with few nonzero entries. A small number of grossly corrupted entries then lands in the second part instead of tilting the principal directions, which the squared error loss of principal component analysis (PCA) lets a single such entry do. Principal component pursuit computes the decomposition by minimizing the sum of the singular values of the first part plus a weighted sum of the absolute entries of the second. That is a convex problem whose solution is the true decomposition under conditions on the low-rank part and on the positions of the corrupted entries. The alternating direction method of multipliers (ADMM) solves it by alternating singular value thresholding and soft thresholding. Uses are the separation of a static background from moving objects in video and the detection of faulty measurements.

Definition

The hourly air temperature at Krems an der Donau in 2024, arranged as a matrix with one row per day and one column per hour, has 366 rows and 24 columns, and every row is the daily cycle on top of that day's level. Three singular values carry 99.5% of its energy: the matrix is nearly of rank three. Now let 5% of the entries be sensor faults, each off by 20 to 40 degrees. Principal component analysis (PCA) of the corrupted matrix, its best rank-three approximation by a truncated singular value decomposition (SVD), ends up 19% away from the clean matrix, since the squared error loss lets the faults tilt the directions. Robust PCA decomposes the corrupted matrix into a low-rank part and a part with few nonzero entries: the first is 6% away from the clean matrix, and the entries of the second at the scale of a fault are exactly the faults (Fig. 1).

Let $\mathbf{X} \in \mathbb{R}^{m \times d}$ be the data matrix whose rows are the feature vectors of m data points. PCA finds the matrix of rank d' nearest to $\mathbf{X}$ in the sum of squared entries, which is the truncated SVD (Eckart and Young, 1936). A single entry of $\mathbf{X}$ that is off by a large amount changes every singular vector, since the squared

error grows with the square of that amount. Robust PCA assumes instead that

$$\mathbf{X} = \mathbf{L} + \mathbf{S}, \quad (1)$$

with a matrix $\mathbf{L}$ of low rank and a matrix $\mathbf{S}$ with few nonzero entries of arbitrary size, and recovers both. Minimizing the rank of $\mathbf{L}$ plus the number of nonzero entries of $\mathbf{S}$ subject to (1) is intractable, and principal component pursuit replaces each count by its convex surrogate (Candès et al., 2011):

$$\min_{\mathbf{L}, \mathbf{S} \in \mathbb{R}^{m \times d}} \|\mathbf{L}\|_* + \alpha \|\mathbf{S}\|_1 \quad \text{subject to } \mathbf{L} + \mathbf{S} = \mathbf{X}, \quad (2)$$

with the nuclear norm $\|\mathbf{L}\|_*$, the sum of the singular values of $\mathbf{L}$, in place of the rank, and the ℓ_1 norm $\|\mathbf{S}\|_1$, the sum of the absolute entries, in place of their count, as for the least absolute shrinkage and selection operator (Lasso). Both are penalty terms, the loss-penalization route of regularization. The weight $\alpha = 1/\sqrt{\max\{m, d\}}$ needs no tuning: if the singular vectors of $\mathbf{L}$ are spread over all rows and columns rather than concentrated on a few, and the nonzero entries of $\mathbf{S}$ sit at positions chosen uniformly at random and are not too many, then the minimizer of (2) is the true pair $(\mathbf{L}, \mathbf{S})$ with high probability (Candès et al., 2011, Thm. 1.1). On the temperature matrix the first condition fails only mildly, since the rows are not exactly of rank three, and the pursuit then assigns the small remainder of the daily curves to $\mathbf{S}$ as well; the faults are still the entries of $\mathbf{S}$ at their own scale.

The alternating direction method of multipliers (ADMM) solves (2) by alternating two proximal operators (Candès et al., 2011, Sect. 5): the proximal operator of the nuclear norm, which computes an SVD and shrinks every singular value toward zero by a threshold, and the proximal operator of the ℓ_1 norm, which shrinks every entry toward zero by a threshold, followed by an update of the multiplier of the constraint. Each iteration applies the same operator to the triple of $\mathbf{L}$, $\mathbf{S}$ and the multiplier, so the method is a fixed-point iteration, whose fixed points are the saddle points of the augmented Lagrangian and therefore the solutions of (2); the operator is averaged, which is what guarantees convergence for a convex problem (Boyd et al., 2011, Sect. 3.2). On the temperature matrix the recovered $\mathbf{L}$ has rank nine out of 24 and the pursuit takes a few thousand iterations, each an SVD of a 366×24 matrix.

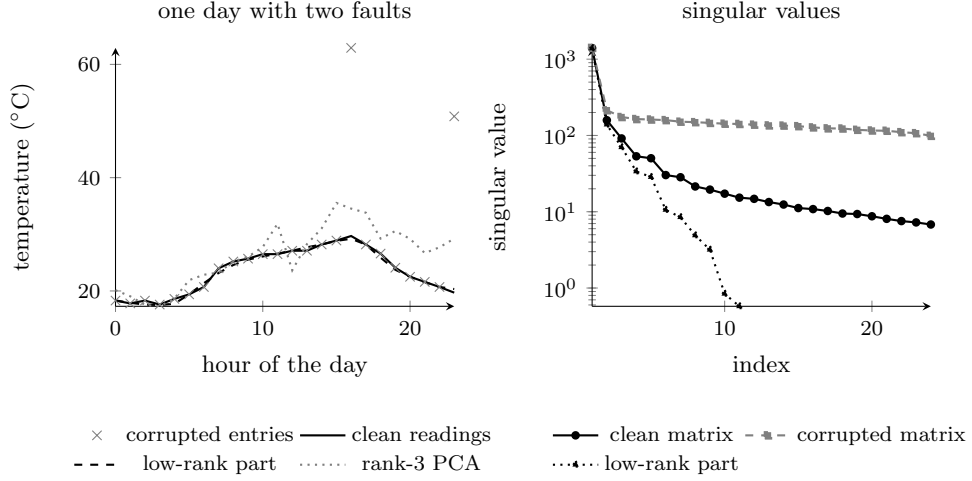


Figure 1: Robust PCA on the 366×24 matrix of hourly temperatures at Krems an der Donau in 2024 with 5% of the entries corrupted by 20 to 40 degrees. Left: one day with two faults, its clean readings, the row of the recovered low-rank part, and the row of the rank-three PCA approximation of the corrupted matrix, which the faults pull off the curve. Right: the singular values of the clean, the corrupted and the recovered matrix; the faults lift the tail of the corrupted one. Data generated by `pythondemos/rpca.py`

The sparse part is as useful as the low-rank part. In a video from a fixed camera, with one frame per column, the static background is the low-rank part and the moving people are the sparse part, so robust PCA separates the two without a model of either (Candès et al., 2011, Sect. 1). In a matrix of sensor readings, the sparse part flags the faulty entries, which is anomaly detection with the low-rank structure as the model of normal behavior. If entries of $\mathbf{X}$ are missing rather than corrupted, the same nuclear-norm criterion with the constraint restricted to the observed entries recovers $\mathbf{L}$, which is the matrix-completion problem (Candès et al., 2011, Sect. 1.6).

Cross References

PCA, sparse PCA, PPCA, SVD, rank, ADMM, proximal operator, fixed-point iteration, Lasso, regularization, outlier, anomaly detection, robustness, missing data

References

1. Candès et al. (2011). Robust Principal Component Analysis?. Journal of the ACM, 58(3), pp. 11:1–11:37.

2. Boyd et al. (2011). Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Found. Trends Mach. Learn.*, 3(1), pp. 1–122.
3. Eckart and Young (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3), pp. 211–218.