Draft Outline:

Introduction:

- Motivation and background for Robust PCA

- Problem Formulation

Theory:

- Analysis of the proof of the main result (probabilistic guarantees for reliable recovery)

- Survey related work and variants / extensions of the Robust PCA problem

Algorithms:

- Review and implementation of different algorithms to solve Robust PCA

- Comparison of complexity, convergence rate, scalability etc of different algorithms

- Exploration of algorithms for large-scale problems (including in particular fast SVD and partial SVD methods)

Application:

- Apply to voice signal recognition and neural activity identification (Brain-Machine interface project)

- Apply to classical image processing

Discussion and improvement

- Practical issues when using the Robust PCA framework (validity of assumptions for realistic problems, scalability, etc.)

- Identification of promising fields of application outside vision

- Exploiting additional application-driven information on the low-rank and noise components to improve guarantees. How to adapt existing algorithms to solve the resulting formulation?

- Stable Robust PCA problem: How to use other models for the non-sparse noise component?

Introduction and motivation:

In todays information age, we are always surrounded by big data sets. Due to the nature of many problems, many data are redundant and thus can be modeled as the low rank component of an observed matrix. However, the data sets are usually corrupted by some sparse random data, which is not related to the compact information that we want to obtain. This makes it hard to recover the low rank component of interest from the observed matrix. This category of problems arises in many fields like computer vision, signal processing, web data analysis, etc. To address this problem, the Robust PCA framework has been introduced, which is a heuristic to recover low rank matrix corrupted by sparse noise through convex optimization. Under suitable assumptions exact recovery of the individual components can be achieved with high probability. Therefore Robust PCA is a potentially powerful tool for obtaining compact information from redundant data sets.

Given an observed matrix M, Robust PCA solves the following convex optimization problem.

$$p^* = \min_{L,S} \|L\|_* + \lambda \|S\|_1$$
$$\text{s.t.} \quad M = L + S$$

We note that this formulation agrees with our intuition that nuclear norm encourage low rank and ℓ_1-norm encourages sparsity. If the low rank component is not too sparse and under some assumptions on the sparsity pattern of the noise component the problem recovers the individual components with high probability as the dimension of the matrix grows.

Theory: We first present the main results regarding Robust PCA and then focus our discussion on the probabilistic guarantee for recovering the low rank component. Then, we survey related topics on recovering low rank matrices from corrupted data, including generalizations to additional non-sparse noise components.

Algorithm: Since Robust PCA is often targeted at large data sets, the theoretical guarantees themselves are only useful if we have fast algorithms to solve the problem. Since the problem is formulated as a convex problem, we survey various convex optimization algorithms for solving this problem. In particular, we compare their effectiveness in solving the problem regarding complexity and theoretical and empirical convergence rates.

Applications: We apply Robust PCA to some signal recovery problems appearing in vision, biological applications and voice recognition. We discuss the obtained results and comment on the suitability of applying the Robust PCA framework to these problems.

Discussion and Improvements: The Robust PCA framework makes a number of assumptions that might be limiting applicability to some practical problems. On the other hand, domain knowledge might provide additional information about the nature of the low-rank and noise components that can be used to improve recovery guarantees or computational efficiency. We discuss a number of possible extensions that make use of this additional information, both in terms of theory and algorithms.