Robust Principal Component Analysis

Reported by Apexa Joshi - 1644001
Seas,Ahmedabad University
Mtech 2nd sem

***Abstract*—In this era of Big Data, there are datasets of high dimension that can have a fraction of missing entries and also lots of entries are corrupted and the problem is to fill those missing entries and also to correct those erroneous data. Suppose we have a data matrix,which is the superposition of a low-rank component and a sparse component and we want to recover each component individually this can be done using convex optimization techniques which are discussed in the paper.**

Keywords:Robust PCA, Low rank, Sparse Matrix,,Convex Optimization, Principal Component pursuit, Matrix Norms, Augmented Lagrange Multiplier.

I. Introduction

PCA is one of the most widely used statistical tool for data analysis and dimensionality reduction but the problem is that it is very sensitive to outliers.A single error point can drastically change the PCA component since classical PCA gives the direction of maximum variance. Thus, a single grossly corrupted entry in data could render the estimated matrix arbitrarily far from the true matrix.Gross errors occur in many applications such as Image Processing, Web data Analysis,Bioinformatics,sensor failures etc. hence it is important to make PCA robust. If we are given a large data matrix and we know that it can be decomposed as M = L 0 + S 0 where L 0 has low rank and S 0 is sparse.We have no information about the low- dimensional column and row space of L 0 .We also have no information of the location of nonzero entries in the sparse matrix and our goal is to recover the low-rank and the sparse components accurately. One of the applications for this could be video surveillance where the low rank component would be the background which would be static while the sparse component would contain some activity on the foreground.

II. THEORETICAL ASPECTS

This problem of separating the low rank and the sparse component from the data matrix can be solved using convex optimization technique Principal Component Pursuit. Recovering using PCP :

M= $L_0 + S_0$
where L 0 is unknown(rank is unknown)
S 0 is unknown(No. of nonzero entries,magnitudes,locations)

$$\textit{minimize} \;\; \parallel L \parallel + \parallel S \parallel$$

$$\textit{subject to } L\text{+ } S\text{= } M$$

A. Robust Principal Component Analysis

PCA is unequivocally the best statistical tool for data analysis and dimensionality reduction. However its brittleness to small corrupted data in data matrix puts its validity in jeopardy, as this small corruption could render the estimated $\hat{\text{L}}$ arbitrarily far from true L_0. Such problems are ubiquitous in modern applications such as image processing, web data analysis and many more.The problems mentioned above are the idealized version of robust PCA, where we recover low rank matrix from highly corrupted data matrix M such that $M = L_0 + S_0$ where unlike classical PCA S_0 can have arbitrarily large magnitude. and their support is assumed to be sparse.

III. ALGORITHM

The theorem above shows that incoherent low-rank matrices can be recovered from non-vanishing fractions of gross errors in polynomial time. For small problem sizes, PCP can be performed sing off-the-shelf tools such as interior pint methods. Despite the superior converge rates, interior point methods are typically limited to small problems, say n ¡ 100, due to On6 complexity of computing a step direction.

A. Principal component pursuit by alternating directions

initialize : $S_0 = Y_0 = 0; \mu > 0$ while not converged do
compute $L_{h+1} = D_{\frac{1}{\mu}}(M - S_h + \mu^{-1}Y_h)$;
compute$S_{k+1} = S_{\frac{\lambda}{\mu}}(M - L_{h+1} + \mu^{-1}Y_h)$;
compute$Y_{h+1} = Y_h + \mu(M - L_{h+1} - S_{h+1})$;
end while
output : L,S.

IV. APPLICATIONS

There are many important applications in which the data can be modeled as a low rank and sparse matrix.

A. Video Surveillance

Suppose we are given a surveillance frames of a video, we need to identify activities that stand out from background. If we stack video as a matrix, then the low rank matrix corresponds to the stationary background and sparse matrix captures moving objects from the video.

B. Face Recognition

It is well known that images of a convex, Lambertian surface un- der varying illuminations span a low-dimensional subspace. This fact has been a main reason why low-dimensional models are mostly effective for imagery data. In particular, images of a humans face can be well-approximated by a low-dimensional subspace. Being able to correctly retrieve this subspace is crucial in many applications such as face recognition and alignment. However, realistic face images often suffer from self-shadowing, specularities, or saturations in brightness, which make this a difficult task and subsequently compromise the recognition performance.

C. Ranking and Collaborative Filtering

The problem of anticipating user tastes is gaining increasing importance in online commerce and advertisement. Companies now routinely collect user rankings for various products, for example, movies, books, games, or web tools, among which the Netflix Prize for movie ranking is the best known [Netflix, Inc.]. The problem is to use incomplete rankings provided by the users on some of the products to predict the preference of any given user on any of the products. This problem is typically cast as a low-rank matrix completion problem. However, as the data collection process often lacks control or is sometimes even ad hoca small portion of the available rankings could be noisy and even tampered with. The problem is more challenging since we need to simultaneously complete the matrix and correct the errors.

Similar problems also arise in many other applications such as Ranking and collaborative Filtering, Latent Semantic Indexing, graphical model learning, linear system identification, and coherence decomposition in optical systems

V. Main Result

A. Theorem 1.1

Suppose that the support set Ω of S_0 is uniformly distributed among all sets of cardinality m, and that sgn($[S_0]_{ij}$) = Σ_{ij} for all (i, j) $\in$ Ω. Then, there is a numerical constant c such that with probability at least 1 - cn^{-10} (over the choice of support of S_0), Principal Component Pursuit (1) with λ = 1/$\sqrt{n}$ is exact, that is, $\hat{L}$ = L_0 and $\hat{S}$ = S_0, provided that

$$rank(L_0) \leq \rho_r n \mu^{-1}(logn)^{-2} \, and \, m \leq \rho_s n^2$$

In this equation, ρ_r and ρ_s are positive numerical constants. In the general rectangular case, where L_0 is $n_1 x n_2$, PCP with λ = $\frac{1}{\sqrt{n_{(1)}}}$ succeeds with probability at least $1 - cn_{(1)}^{-10}$, provided that rank(L_0) $\leq$ $\rho_r n_{(2)} \mu^{-1}(logn_{(1)})^{-2}$ and $m \leq \rho_s n_1 n_2$.

Matrix L_0 whose principal components are spreaded can be recovered with probability almost onefrom arbitary and completely unknown corruption patterns. It also works for higher ranks like n/$log(n)^2$ when μ is not large. Minimizing

$$||L||_* + \frac{1}{\sqrt{n_{(1)}}}||S||_1$$

where,

$$n_{(1)} = max(n_1, n_2)$$

under the assumption of theorem, this always gives correct answer. Here we chose λ = $\frac{1}{\sqrt{n_{(1)}}}$ but it is not clear why that has happened. It has been due to mathematical analysis why we are taking that value.

VI. Conclusion

We can conclude from the above explanation and the results that one can disentangle the low-rank and the sparse components exactly by convex programming, this provably works under quite broad conditions. Also the above method can be used for matrix completion and matrix recovery from sparse errors and this also works in the case when there are both incomplete and corrupted entries.

References

[1] Cand‘es, E. J., Li, X., Ma, Y., Wright, J. (2011). Robust principal component analysis?. Journal of the ACM (JACM), 58(3), 11. Chicago

[2] Chandrasekaran V., Sanghavi, S., Parrilo, P.,AND Willsky, A. 2009. Rank-sparsity incoherence for matrix decomposition. Siam J. Optim., to appear http://arxiv.org/abs/0906.2220.

[3] E. Candes, X. Li, Y. Ma, and J. Wright. Robust principal component analysis. 2011. https://github.com/dlaptev/RobustPCA.

[4] Analysis of Robust PCA via Local Incoherence. Huishuai Zhang, Yi Zhou, Yingbin Liang.