Robust and Effective Factorization Machines

Anoymous

Abstract

Factorization Machine (FM) (Rendle 2010; 2012) is a generic framework which involves a weight matrix $\mathbf{Z} \in \mathbb{R}^{d\times d}$ in the formulation to take into account the pairwise interactions between features. Due to its capability of efficiently modeling complex non-linear data, FM has achieved great success in various of classification and regression tasks. The key advantage of FM is that it learns the low-rank feature interaction matrix in a factorized from, $\mathbf{Z} = \mathbf{V}\mathbf{V}^\top$, where $\mathbf{V} \in \mathbb{R}^{d\times k}$ and $k \ll d$. This could avoid overfitting, learn correlation among samples and allow to compute prediction efficiently. Although the factorized interaction allow us to reduce computation cost from $O(d^2)$ to $O(kd)$ and the computation of gradients is also easier, it can only find local optima because of its non-convex formulation. Moreover, the predictive accuracy of FM is sensitive to the rank parameter k, since a small deviation of k from optimal value may have enormous effect on the performance. But it is often difficult to choose or tune the best k from large range.

Introduction

Factorization Machine (FM) (Rendle 2010; 2012) is a generic framework which involves a weight matrix $\mathbf{Z} \in \mathbb{R}^{d\times d}$ in the formulation to take into account the pairwise interactions between features. Due to its capability of efficiently modeling complex non-linear data, FM has achieved great success in various of classification and regression tasks. The key advantage of FM is that it learns the low-rank feature interaction matrix in a factorized from, $\mathbf{Z} = \mathbf{V}\mathbf{V}^\top$, where $\mathbf{V} \in \mathbb{R}^{d\times k}$ and $k \ll d$. This strategy make it avoid overfitting, learn correlation among samples and compute prediction efficiently. Although the factorized interaction allow us to reduce computation cost from $O(d^2)$ to $O(kd)$ and the computation of gradients is also easier, it can only find local optima because of its non-convex formulation. Moreover, the predictive accuracy of FM is sensitive to the rank parameter k, since a small deviation of k from optimal value may have enormous effect on the performance. But it is often difficult to choose or tune the best k from large range.

To solve this problem, some authors (Blondel, Fujino, and Ueda 2015; Yamada et al. 2015) proposed convex variants of FM model by relaxing the low-rank structure of feature interaction matrix $\mathbf{Z}$ with trace norm regularization $\|\mathbf{Z}\|_*$, which is defined as the sum of all singular values of $\mathbf{Z}$. Following the recent work about large scale optimization with trace norm constraint (Shalev-shwartz, Gonen, and Shamir 2011), Blondel et al. presented an efficient greedy coordinate descent algorithm which enjoys global convergence. Yamada et al. formulated the objective as a semidefinite programming problem and derived an efficient optimization procedure with Hazan's algorithm (Hazan 2008).

Unfortunately, although the trace norm is capable of including low-rank and sparse structures (Tao and Yuan 2011) and learning potential correlations from sample, such convex formulation is based on the strong assumptions which may not hold in real-world applications and the approximation error between rank minimization and trace norm constraint often can not be neglected (Sun, Xiang, and Ye 2013). For example, with the change of non-zero singular value, the trace norm value will change together but the rank value will remain the same.

Like most supervised learning methods, existing FM method train model with large amounts of labeled data, where outliers will be included unavoidably. These incorrectly labeled data is significantly different from normal data and could mislead the model training task, such that the learned model are not optimal and the prediction performance is reduced. Hence it is necessary to learn FM model resistant to outliers.

To address these challenging problems, we propose a novel robust and effective Factorization Machine model and derive an effective optimization algorithm with rigorous theoretical guarantees. To make FM model immune to the threat of outliers, we replace the squared loss in existing FM model with a novel capped ϵ-intensive loss $\ell_{\epsilon_1,\epsilon_2} = \min\{\max\{|u| - \epsilon_1, 0\}, \epsilon_2\}$, where ϵ_1 and ϵ_2 are threshold values. It is obvious that this loss make FM model robust to light and heavy outliers since it could eliminate outliers which often have large residual ϵ_2. Furthermore, the error $|u|$ would become unimportant so long as it falls below an insensitivity parameter ϵ_1. To narrow the gap between rank minimization and trace norm, we develop a novel capped squared trace norm as low-rank regularization. It is defined as $R_{\epsilon_3}(\mathbf{Z}) = \sum_s \min\{\lambda_s^2, \epsilon_3\}$ where ϵ_3 is a threshold value and λ_s is the enginvalues of interaction matrix $\mathbf{Z}$. In this term, we can approximate the rank function by

$rank(\mathbf{Z}) \approx \sum_s \min\{1, \frac{\lambda_s^2}{\epsilon_3}\}$. Noted that if all the squared singular values of $\mathbf{Z}$ are greater than ϵ_3, then the approximation will become equality which leads to a better approximation compared with traditional trace norm. Moreover, the capped squared trace norm penalizes the squared singular values that are less than ϵ_3, which weakens the effect of non-relevant feature interaction, such that the FM model is more robust and stable in real world applications. The novel loss function $\ell_{\epsilon_1,\epsilon_2}$ and regularization $R_{\epsilon_3}(\mathbf{Z})$ involves the non-convexity and non-smoothness of the objective, which make it challenging to optimize. To tackle this issue, we introduce a new efficient two-stage algorithm with rigorously proved local optimum convergence. Extensive experiments on both toy data and benchmark datasets show our proposed new robust FM models correctness and effectiveness.

Problem Formulation

Factorization Machine (FM) is an increasingly popular method for efficiently utilize second-order feature interaction in classification or regression task. The prediction equation of a standard 2-order FMs mode is:

$$\hat{y}(\mathbf{x}|\mathbf{w}, \mathbf{V}) = \mathbf{w}^\top \mathbf{x} + \sum_{j=1}^{d} \sum_{j'=j+1}^{d} (\mathbf{V}\mathbf{V}^\top)_{jj'} x_j x_{j'},$$

where d is the dimensionality of feature vector $\mathbf{x} \in \mathbb{R}^d$, $k \ll p$ is a hyper-parameter that denotes the dimensionality of latent factors and $\mathbf{V} \in \mathbb{R}^{d\times k}, \mathbf{w} \in \mathbb{R}^d$ are the model parameters to be estimated. Here we consider a generalized prediction function:

$$\hat{y}(\mathbf{x}|\mathbf{w}, \mathbf{Z}) = \mathbf{w}^\top \mathbf{x} + \sum_{j=1}^{d} \sum_{j'=1}^{d} z_{jj'} x_j x_{j'} = \mathbf{w}^\top \mathbf{x} + \langle \mathbf{Z}, \mathbf{x}\mathbf{x}^\top \rangle,$$

where $z_{jj'}$ is the elements of the low-rank sysmmetric positive semi-definite matrix $\mathbf{Z} \in \mathbb{S}^{d\times d}$, i.e., $rank(\mathbf{Z}) \ll d$. Given a training set $[\mathbf{x}_1, \ldots, \mathbf{x}_n]^\top \in \mathbb{R}^{n\times d}$ and corresponding targets $[y_1, \ldots, y_n]^\top \in \mathbb{R}^n$, model parameters can be learned by using the principle of empirical risk minimization and solving the following non-convex problem:

$$\min_{\mathbf{w}\in\mathbb{R}^d, \mathbf{Z}\in\mathbb{S}_+^{d\times d}} \sum_{i=1}^{n} \ell(y_i, \hat{y}(\mathbf{x}_i|\mathbf{w}, \mathbf{Z})) + \frac{\alpha}{2} R(\mathbf{w}) + \frac{\beta}{2} R(\mathbf{Z}),$$

where $R(\mathbf{w}), R(\mathbf{Z})$ are the regularization terms and ℓ is a loss function incurred.

In the existing FM model, it usually apply ℓ_2-norm loss for regression task, ℓ_2-norm regularization for linear term and Frobenius norm regularization for quadratic term. To endow FM model with robustness capability, a natural way is to use absolute loss function instead . However, if the extreme odd points incur very large residuals, they will still have significantly negative effects on the performance of FM model. To further better the performance of FM model, we propose to adopt the capped ϵ-intensive loss function $\ell_{\epsilon_1,\epsilon_2}$ for robust prediction. In recent research work (Zhang 2009; Gao et al. 2015; Jiang, Nie, and Huang 2015), the capped ℓ_1-norm loss was successfully used to approximate the ℓ_0 norm. The FM model with our novel capped ϵ-intensive loss could be expressed as:

$$\ell_{capped}^{FM}(y_i, \hat{y}(\mathbf{x}_i|\mathbf{w}, \mathbf{Z})) = \min\{error, \epsilon_2\}$$
$$= \min\{\max\{|y_i - \mathbf{w}^\top \mathbf{x}_i - \langle \mathbf{Z}, \mathbf{x}_i\mathbf{x}_i^\top \rangle| - \epsilon_1, 0\}, \epsilon_2\}. \quad (1)$$

In this term, if the error of a sample is larger than ϵ_2, we consider this sample as extreme outlier and its error is also capped as ϵ_2 such that its effect to the whole FM model is fixed. In addition, we ignore this sample if its residual falls below the insensitivity threshold ϵ_1. For other normal samples, our objective will minimize $error = |y_i - \mathbf{w}^\top \mathbf{x}_i - \langle \mathbf{Z}, \mathbf{x}_i\mathbf{x}_i^\top \rangle| - \epsilon_1$, which is similar with the ℓ_1-norm loss. Therefore, the proposed capped ℓ_1-norm is a more robust loss function than traditional ℓ_2-norm loss in FM model. It is worth noting that if ϵ_1 is set as 0 and ϵ_2 is set as ∞, our capped ϵ-intensive loss degenerates to the absolute loss.

Recently, trace norm regularization has been utilized as the convex relaxation of the rank minimization so as to approximate the low-rank structure of feature interaction matrix in FM model(Blondel, Fujino, and Ueda 2015; Yamada et al. 2015). However, as we mention before, if the non-zero singular values of matrix $\mathbf{Z}$ changes, the trace norm of $\mathbf{Z}$ will change simultaneously, but the rank of $\mathbf{Z}$ stays the same. Thus, there is still a gap between trace norm and rank minimization. To make much tighter approximation and more robust model, we propose a novel capped squared trace norm to uncover the low rank structure of feature interaction matrix $\mathbf{Z}$, which is defined as:

$$R_{\epsilon 3}(\mathbf{Z}) = \sum_s \min\{\lambda_s^2, \epsilon_3\}, \quad (2)$$

where λ_s is the eigenvalue of matrix $\mathbf{Z}$ and ϵ_3 is a threshold value. We can approximate the rank function by $rank(\mathbf{Z}) \approx \sum_s \min\{1, \frac{\lambda_s^2}{\epsilon_3}\}$. The smaller ϵ_3, the more accurate the approximation would be. Thus, by carefully choosing ϵ_3, we can recovery matrix $\mathbf{Z}$ more accurately than the trace norm.

Variational trace norms have been developed for better approximating the rank minimization in rent years. The authors in (Law, Thome, and Cord 2014; Huo et al. 2017) considered to minimize the k-smallest singular values. Although it can avoid the effect of large singular values and achieve a better approximation, this method suffers from the tedious best rank parameter selection process in the same way as the traditional trace norm. The authors in (Sun, Xiang, and Ye 2013; Huo, Nie, and Huang 2016) proposed to minimize the sum of singular values which are smaller than a threshold value. However, minimizing the sum of capped singular values would lead to a sparse solution, that is, some small singular value will become zero while others may get large values. Our novel capped squared trace norm could solve the above problems, since it avoids the cumbersome rank parameter selection process and minimizes the sum of capped squared singular values whose solution will be shrunk near to zero.

With the aforementioned capped norms, we propose to solve the following non-convex and non-smooth formulation:

$$\min_{\mathbf{w}\in\mathbb{R}^d, \mathbf{Z}\in\mathbb{S}_+^{d\times d}} \sum_{i=1}^{n} \ell_{capped}^{FM}(y_i, \hat{y}(\mathbf{x}_i|\mathbf{w}, \mathbf{Z})) + \frac{\alpha}{2}\|\mathbf{w}\| + \frac{\beta}{2} R_{\epsilon_3}(\mathbf{Z}), \quad (3)$$

where we use ℓ_2-norm regularization for linear term to avoid overfitting. Clearly, the new objective function in problem (3) is not convex and not smooth due to the definition of capped ϵ-intensive loss function (1) and capped squared trace norm regularization (2), which are difficult to solve. In next section, we will propose an efficient optimization algorithm to solve it.

Optimization Algorithm

In previous work (Sun, Xiang, and Ye 2013; Gong, Ye, and Zhang 2012), they utilize ADMM to optimize the capped norm based objectives, but it is difficult to directly applied to our problem. In this section, we first use the re-weighted method to repeatedly transform the original objective to a convex relaxation, then use the proximal gradient method to solve this new subproblem. Specifically, according to the multi-stage convex relaxation techniques (Zhang 2009; 2010; Nie et al. 2010; Nie, Yuan, and Huang 2014) and alternating minimization optimization procedure, we could transform the original objective to the following formulation:

$$\min_{\mathbf{w}\in\mathbb{R}^d,\mathbf{Z}\in\mathbb{S}_+^{d\times d}} \sum_{i=1}^{n} e_i(\max\{|y_i - \mathbf{w}^\top\mathbf{x}_i - \langle\mathbf{Z},\mathbf{x}_i\mathbf{x}_i^\top\rangle| - \epsilon_1, 0\})^2 + \frac{\alpha}{2}\|\mathbf{w}\| + \sum_s \min\{\lambda_s^2, \epsilon_3\}, \quad (4)$$

$$e_i = \begin{cases} \frac{1}{2error}, & 0 < error \le \epsilon_2; \\ 0, & otherwise, \end{cases}$$

where $error = |y_i - \mathbf{w}^\top\mathbf{x}_i - \langle\mathbf{Z},\mathbf{x}_i\mathbf{x}_i^\top\rangle| - \epsilon_1$ and this new objective funcion can be solved via the iterative re-weighted optimization strategy. From the formulation (4), we can observe that it is similar with ℓ_2 loss in original FM model except the weights e_i for each sample. From the updating rule of weights e_i, we can find that the samples with lower residuals have higher weights, which is consistent with robustness of FM model. Thus, we adopt the same strategy by using a two-block coordinate descent algorithm. That is, we alternate between minimizing with respect to $\mathbf{w}$ and $\mathbf{Z}$ until convergence. When the algorithm terminates, it returns $\mathbf{w}$ and $\mathbf{Z}$. In the next section, we will prove that our objective function will converge.

With feature interaction matrix $\mathbf{Z}$ fixed, we need to optimize

$$\min_{\mathbf{w}\in\mathbb{R}^d} \sum_{i=1}^{n} e_i(\max\{|y_i - \mathbf{w}^\top\mathbf{x}_i - \langle\mathbf{Z},\mathbf{x}_i\mathbf{x}_i^\top\rangle| - \epsilon_1, 0\})^2 + \frac{\alpha}{2}\|\mathbf{w}\|_2^2, \quad (5)$$

which is similar with the standard linear model objective except the constant bias term incurred by $\mathbf{Z}$ and ϵ-intensive loss. Therefore, we can solve objective (5) by standard methods, like gradient descent (Rendle 2010).

With linear term parameter $\mathbf{x}$ fixed, we can reformulate the objective (4) as the following optimization problem:

$$\min_{\mathbf{Z}\in\mathbb{S}_+^{d\times d}} \sum_{i=1}^{n} e_i(\max\{|y_i - \mathbf{w}^\top\mathbf{x}_i - \langle\mathbf{Z},\mathbf{x}_i\mathbf{x}_i^\top\rangle| - \epsilon_1, 0\})^2 + \frac{\beta}{2}\sum_s \min\{\lambda_s^2, \epsilon_2\}, \quad (6)$$

However, the capped trace norm is non-convex and non-smooth which make it difficult for direct optimization. Similarly, We follow the same multi-stage convex relaxation technique, and convert this objective (6) to a sub-problem.

Specifically, define a singluar value decomposition of symmetric matrix $\mathbf{Z}$ as $\mathbf{Z} = \mathbf{P}\mathbf{\Sigma}\mathbf{P}^\top = \sum_s \lambda_s\mathbf{p}_s\mathbf{p}_s^\top$, where $\mathbf{P}$ is an orthogonal matrix with columns $\mathbf{p}_s \in \mathbb{R}^d$ and $\mathbf{\Sigma} = diag(\boldsymbol{\lambda})$ is diagonal singular matrix in ascending order . Denote the set of index whose singular value is smaller than ϵ_3 as $M = \{s|\lambda_s \le \epsilon_3\}$, and their corresponding eignvectors as $\mathbf{P}_M$. It is easy to have $Tr(\mathbf{P}_M^\top\mathbf{Z}\mathbf{Z}^\top\mathbf{P}_M) = \sum_{s\in M}\lambda_s^2$. The sub-problem of the objective (6) is formulated as:

$$\min_{\mathbf{Z}\in\mathbb{S}^{d\times d}} \sum_{i=1}^{n} e_i \max\{|y_i - \mathbf{w}^\top\mathbf{x}_i - \langle\mathbf{Z},\mathbf{x}_i\mathbf{x}_i^\top\rangle| - \epsilon_1, 0\})^2 + \frac{\beta}{2}\sum_s Tr(\mathbf{P}_M^\top\mathbf{Z}\mathbf{Z}^\top\mathbf{P}_M). \quad (7)$$

With matrix parameter $\mathbf{P}_M$ fixed, this sub-problem is convex. We could use proximal gradient descent to optimize it iteratively. At each round, let the residual $r_i = y_i - \mathbf{w}^\top\mathbf{x}_i - \langle\mathbf{Z},\mathbf{x}_i\mathbf{x}_i^\top\rangle$, then the subgradient with respect to $\mathbf{Z}$ is:

$$\nabla_{\mathbf{Z}} = \sum_{i=1}^{n} e_i\pi_i\mathbf{x}_i\mathbf{x}_i^\top + \beta\mathbf{P}_M\mathbf{P}_M^\top\mathbf{Z}. \quad (8)$$

$$\pi_i = \begin{cases} -r_i + \epsilon_1, & r_i \ge \epsilon_1; \\ -r_i - \epsilon_1, & -r_i \ge \epsilon_1; \\ 0 & otherwise \end{cases}.$$

After each step, $\mathbf{Z}$ is projected onto the positive semidefinite cone with

$$\mathbf{Z} = \Pi_{\mathbb{S}^{d\times d}}(\mathbf{Z} - \eta\nabla_{\mathbf{Z}}), \quad (9)$$

where η is the step size. Our optimization algorithm is summarized in 1.

Algorithm 1 Optimization Algorithm for Robust and Effective Factorization Machine Model

Input: Training data$\{(x_n, y_n\}_{n=1}^N$, parameters $\alpha, \beta, \epsilon_2, \epsilon_3$
Initialization: $e_i = 1$ for $i = 1, 2, \ldots, n$
while not converge **do**
  Update $\mathbf{w}$ according to (5)
  Update $\mathbf{Z}$ and $\mathbf{P}_M\mathbf{P}_M^\top$ according to (7)
  Compute $e_i = 1$ for $i = 1, 2, \ldots, n$ according to (4)
end while
Output: model parameter w, $\mathbf{Z}$

Noted that we only need to compute $\mathbf{P}_M\mathbf{P}_M^\top$ instead of $\mathbf{P}_M$ and $\mathbf{P}_M^\top$. Suppose $\mathbf{P} = [\mathbf{P}_M, \mathbf{P}_{\bar{M}}]$, where $\bar{M} = \{s|\lambda_s > \epsilon_3\}$ and $\mathbf{P}_{\bar{M}}$ are singular values whose corresponding singular vector larger than ϵ_3. It is easy to see that $\mathbf{P}_M\mathbf{P}_M^\top = I - \mathbf{P}_{\bar{M}}\mathbf{P}_{\bar{M}}^\top$. Since $\mathbf{Z}$ is a low-rank matrix, if we set ϵ_3 appropriately, the set $\bar{M}$ could be quite small. Thus it could be quite efficient to compute $\mathbf{P}_{\bar{M}}\mathbf{P}_{\bar{M}}^\top$ with truncated SVD algorithm first and then update $\mathbf{P}_M\mathbf{P}_M^\top$ accordingly.

However, the above batch learning method require to compute the whole training set at each round, which prevents it from processing big data due to possible mem-

ory bottleneck. Therefore, we propose an efficient stochastic proximal gradient method to tackle problem (7). Specifically, we randomly choose a mini-batch $I \subset \{1, \ldots, n\}$ of size b and update $\mathbf{Z}$ according to it at each iteration. Its gradient is:

$$\nabla_{\mathbf{Z},I} = \sum_{i=1}^{b} e_{I_i} \pi_{I_i} \mathbf{x}_{I_i} \mathbf{x}_{I_i}^\top + \beta \mathbf{P}_M \mathbf{P}_M^\top \mathbf{Z}.$$

We introduce an efficient way to compute the updating rule (9). The key idea is to utilize the incremental SVD (Brand 2006) to incrementally calculate the SVD of $\mathbf{Z} - \eta \nabla_{\mathbf{Z},I}$. Let the symmetric and low rank matrix $\mathbf{Z}$ has rank k and its economy SVD is $\mathbf{Z} = \mathbf{P}_k \Sigma_k \mathbf{P}_k^\top$, where $\mathbf{P}_k \in \mathbb{R}^{d \times k}$ and $\Sigma_k \in \mathbb{R}^{k \times k}$. It is easy to see that matrix $\nabla_{\mathbf{Z},I}$ is symmetric and low rank, and thus could be represented as $\mathbf{A}\mathbf{A}^\top$ where $\mathbf{A} \in \mathbb{R}^{d \times c}$. Thus, the SVD of $\mathbf{Z} - \eta \mathbf{A}\mathbf{A}^\top$ can be computed in $O(d(r+c)^2 + (r+c)^3)$ time with $O(d(r+c))$ memory (Brand 2006; Zhang and Zhang 2017).

Let $\mathbf{U}$ be an orthogonal basis of the column space of $(I - \mathbf{P}_k \mathbf{P}_k^\top)\mathbf{A}$, and set $R_{\mathbf{A}} = \mathbf{U}^\top (I - \mathbf{P}_k \mathbf{P}_k^\top)\mathbf{A}$. Note that $cols(\mathbf{U}) = rows(R_{\mathbf{A}}) = rank((I - \mathbf{P}_k \mathbf{P}_k^\top)\mathbf{A}) \leq c$, and may be zero if $\mathbf{A}$ lies in the column space of $\mathbf{P}_k$. Then, we have

$$[\mathbf{P}_k \quad \mathbf{A}] = [\mathbf{P}_k \quad \mathbf{U}] \begin{bmatrix} I & \mathbf{P}_k^\top \mathbf{A} \\ 0 & R_{\mathbf{A}} \end{bmatrix}$$

Then, we can easily get

$$\mathbf{Z} - \mathbf{A}\mathbf{A}^\top = [\mathbf{P}_k \quad \mathbf{U}]\, \mathbf{K}\, [\mathbf{P}_k \quad \mathbf{U}]^\top,$$

where

$$\begin{aligned} \mathbf{K} &= \begin{bmatrix} I & \mathbf{P}_k^\top \mathbf{A} \\ 0 & R_{\mathbf{A}} \end{bmatrix} \begin{bmatrix} \Sigma_k & 0 \\ 0 & -I \end{bmatrix} \begin{bmatrix} I & \mathbf{P}_k^\top \mathbf{A} \\ 0 & R_{\mathbf{A}} \end{bmatrix}^\top \\ &= \begin{bmatrix} \Sigma_k & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} \mathbf{P}_k^\top \mathbf{A} \\ R_{\mathbf{A}} \end{bmatrix} \begin{bmatrix} \mathbf{P}_k^\top \mathbf{A} \\ R_{\mathbf{A}} \end{bmatrix}^\top \in \mathbb{R}^{(r+u)\times(r+u)} \end{aligned}$$

and $u = cols(\mathbf{U})$. Let the SVD of $\mathbf{K}$ be $\mathbf{K} = \hat{\mathbf{P}} \hat{\Sigma} \hat{\mathbf{P}}^\top$. Then, the SVD of $\mathbf{Z} - \mathbf{A}\mathbf{A}^\top$ is given by

$$\mathbf{Z} - \mathbf{A}\mathbf{A}^\top = ([\mathbf{P}_k \quad \mathbf{U}]\, \hat{\mathbf{P}}) \hat{\Sigma} ([\mathbf{P}_k \quad \mathbf{U}]\, \hat{\mathbf{P}})^\top.$$

Computational Complexity

Convergence Analysis

Lemma 1. *According to (Theobald 1975), any two hermitian matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$ satisfy the inequality ($\lambda_i(\mathbf{A}), \lambda_i(\mathbf{B})$ are singular values sorted in the same order)*

$$\sum_{i=1}^{n} \lambda_i(\mathbf{A}) \lambda_{n-i+1}(\mathbf{B}) \leq Tr(\mathbf{A}^\top \mathbf{B}) \leq \sum_{i=1}^{n} \lambda_i(\mathbf{A}) \lambda_i(\mathbf{B})$$

Lemma 2. *Let $\mathbf{Z} = \mathbf{P}\mathbf{\Sigma}\mathbf{P}^\top$, $\mathbf{\Sigma} = diag(\boldsymbol{\lambda})$ is diagonal singular matrix in ascending order and $M = \{s | \lambda_s \leq \epsilon_3\}$ is the set of index whose singular values are smaller than ϵ_3. $\hat{\mathbf{Z}} = \hat{\mathbf{P}}\hat{\mathbf{\Sigma}}\mathbf{P}^\top$ is the updated matrix after $\mathbf{Z}$ with diagonal singular matrix $\hat{\mathbf{\Sigma}} = diag(\hat{\boldsymbol{\lambda}})$ in ascending order and $\hat{M} = \{s | \hat{\lambda}_s \leq \epsilon_3\}$ is the set of index whose singular values are smaller than ϵ_3. Then we have*

$$\begin{aligned} &\sum_s \min\{\hat{\lambda}_s^2, \epsilon_2\} - Tr(\mathbf{P}_M^\top \hat{\mathbf{Z}} \hat{\mathbf{Z}}^\top \mathbf{P}_M) \\ &\leq \sum_s \min\{\lambda_s^2, \epsilon_2\} - Tr(\mathbf{P}_M^\top \mathbf{Z}\mathbf{Z}^\top \mathbf{P}_M) \end{aligned} \tag{10}$$

Proof. According to the definition of $\mathbf{P}_M, \mathbf{Z}$, it is apparent that $Tr(\mathbf{P}_M^\top \mathbf{Z}\mathbf{Z}^\top \mathbf{P}_M) = Tr(\mathbf{P}_M \mathbf{P}_M^\top \mathbf{P}\mathbf{\Sigma}^2\mathbf{P}^\top) = \sum_{s \in M} \lambda_s^2$. The RHS of inequality (10) is equivalent to

$$\sum_{s \in M} \lambda_s^2 + \sum_{s \notin M} \epsilon_3 - \sum_{s \in M} \lambda_s^2 = \sum_{s \notin M} \epsilon_3. \tag{11}$$

Similarly, with the definition of $\hat{P}, \hat{M}$ and Lemma 1, we know that $Tr(\mathbf{P}_M^\top \hat{\mathbf{Z}}\hat{\mathbf{Z}}^\top \mathbf{P}_M) = Tr(\mathbf{P}_M \mathbf{P}_M^\top \hat{\mathbf{P}}\hat{\mathbf{\Sigma}}^2\hat{\mathbf{P}}^\top) \geq \sum_{s \in M} \hat{\lambda}_s^2$. Since $\hat{M}$ denotes the total eigenvalues of $\hat{\mathbf{Z}}$ that are smaller than ϵ_3, no matter how the index set $\hat{M}$ varies from M, we could obtain that $\sum_{s \in \hat{M}} \hat{\lambda}_s^2 + \sum_{s \notin \hat{M}} \epsilon_3 \leq \sum_{s \in M} \hat{\lambda}_s^2 + \sum_{s \notin M} \epsilon_3$. Therefore, from the LHS of inequality (10), we have

$$\begin{aligned} &\sum_{s \in \hat{M}} \hat{\lambda}_s^2 + \sum_{s \notin \hat{M}} \epsilon_3 - Tr(\mathbf{P}_M^\top \hat{\mathbf{Z}}\hat{\mathbf{Z}}^\top \mathbf{P}_M) \\ &\leq \sum_{s \in M} \hat{\lambda}_s^2 + \sum_{s \notin M} \epsilon_3 - \sum_{s \in M} \hat{\lambda}_s^2 = \sum_{s \notin M} \epsilon_3 \end{aligned} \tag{12}$$

Combining inequality (11) and (12) completes the proof. □

Experimental Results

References

Blondel, M.; Fujino, A.; and Ueda, N. 2015. Convex factorization machines. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 19–35. Springer.

Brand, M. 2006. Fast low-rank modifications of the thin singular value decomposition. *Linear algebra and its applications* 415(1):20–30.

Gao, H.; Nie, F.; Cai, W.; and Huang, H. 2015. Robust capped norm nonnegative matrix factorization: Capped norm nmf. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 871–880. ACM.

Gong, P.; Ye, J.; and Zhang, C.-s. 2012. Multi-stage multi-task feature learning. In *Advances in Neural Information Processing Systems*, 1988–1996.

Hazan, E. 2008. Sparse approximate solutions to semidefinite programs. In *Latin American Symposium on Theoretical Informatics*, 306–316. Springer.

Huo, Z.; Gao, S.; Cai, W.; and Huang, H. 2017. Video recovery via learning variation and consistency of images. In *AAAI*, 4082–4088.

Huo, Z.; Nie, F.; and Huang, H. 2016. Robust and effective metric learning using capped trace norm: Metric learning via capped trace norm. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1605–1614. ACM.

Jiang, W.; Nie, F.; and Huang, H. 2015. Robust dictionary learning with capped l1-norm. In *IJCAI*, 3590–3596.

Law, M. T.; Thome, N.; and Cord, M. 2014. Fantope regularization in metric learning. In *Proceedings of the IEEE*

Conference on Computer Vision and Pattern Recognition, 1051–1058.

Nie, F.; Huang, H.; Cai, X.; and Ding, C. H. 2010. Efficient and robust feature selection via joint 2, 1-norms minimization. In *Advances in neural information processing systems*, 1813–1821.

Nie, F.; Yuan, J.; and Huang, H. 2014. Optimal mean robust principal component analysis. In *Proceedings of the 31st international conference on machine learning (ICML-14)*, 1062–1070.

Rendle, S. 2010. Factorization machines. *Proceedings - IEEE International Conference on Data Mining, ICDM* 995–1000.

Rendle, S. 2012. Factorization machines with libfm. *ACM Transactions on Intelligent Systems and Technology (TIST)* 3(3):57.

Shalev-shwartz, S.; Gonen, A.; and Shamir, O. 2011. Large-scale convex minimization with a low-rank constraint. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 329–336.

Sun, Q.; Xiang, S.; and Ye, J. 2013. Robust principal component analysis via capped norms. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 311–319. ACM.

Tao, M., and Yuan, X. 2011. Recovering low-rank and sparse components of matrices from incomplete and noisy observations. *SIAM Journal on Optimization* 21(1):57–81.

Theobald, C. 1975. An inequality for the trace of the product of two symmetric matrices. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 77, 265–267. Cambridge Univ Press.

Yamada, M.; Lian, W.; Goyal, A.; Chen, J.; Wimalawarne, K.; Khan, S. A.; Kaski, S.; Mamitsuka, H.; and Chang, Y. 2015. Convex factorization machine for regression. *arXiv preprint arXiv:1507.01073*.

Zhang, J., and Zhang, L. 2017. Efficient stochastic optimization for low-rank distance metric learning. In *AAAI*, 933–940.

Zhang, T. 2009. Multi-stage convex relaxation for learning with sparse regularization. In *Advances in Neural Information Processing Systems*, 1929–1936.

Zhang, T. 2010. Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research* 11(Mar):1081–1107.