EE-556 Homework 4

Edoardo Debenedetti

January 5, 2020

1 Theory: Projection free convex low-rank matrix optimization

1.1 Projection onto the nuclear norm ball

1.1.a Projection proof

Proof. Let $\mathbf{Z} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ be the SVD of $\mathbf{Z} \in \mathbb{R}^{p\times m}$, and $\mathcal{X} = \{\mathbf{X} : \|\mathbf{X}\|_* \leq \kappa\}$. We can write the projection of $\mathbf{Z}$ onto $\mathcal{X}$ as:

$$\mathrm{proj}_{\mathcal{X}}(\mathbf{Z}) = \arg\min_{\mathbf{X}\in\mathcal{X}} \|\mathbf{X} - \mathbf{Z}\|_F \tag{1}$$

Let us first look for the projection, onto $\mathcal{X}$ of $\mathbf{\Sigma}$:

$$\mathrm{proj}_{\mathcal{X}}(\mathbf{\Sigma_Z}) = \arg\min_{\mathbf{\Sigma_X}\in\mathcal{X}} \|\mathbf{\Sigma_X} - \mathbf{\Sigma_Z}\|_F \tag{2}$$

The nuclear norm $\|\mathbf{\Sigma}\|_*$ is given by the sum of its singular values. However, as Σ is diagonal, its singular values are given by the diagonal itself, and their sum is the sum of the diagonal elements. Moreover, its diagonal elements are the only non-zero elements, and then, when computing its $\ell1$-norm $\|\mathbf{\Sigma}\|_1$, we sum only the absolute value of the diagonal. Moreover, as the entries of $\mathbf{\Sigma}$ are singular values, they are all non-negative and they are equal to their absolute value. Hence, $\|\mathbf{\Sigma}\|_* = \|\mathbf{\Sigma}\|_1$, and, when considering $\mathbf{\Sigma}$, $\mathcal{X} = \{\mathbf{X} : \|\mathbf{X}\|_* \leq \kappa\} = \{\mathbf{X} : \|\mathbf{X}\|_1 \leq \kappa\}$. Hence,

$$\mathrm{proj}_{\mathcal{X}}(\mathbf{\Sigma_Z}) = \arg\min_{\mathbf{\Sigma_X}\in\mathcal{X}} \|\mathbf{\Sigma_X} - \mathbf{\Sigma_Z}\|_F = \mathbf{\Sigma_Z^{\ell1}} \tag{3}$$

where $\mathbf{\Sigma_Z^{\ell1}}$ is the projection of $\mathbf{\Sigma_Z}$ onto the $\ell1$-norm ball of radius κ. Moreover, we can state that

$$\min_{\mathbf{\Sigma_X}\in\mathcal{X}} \|\mathbf{\Sigma_X} - \mathbf{\Sigma_Z}\|_F = \|\mathbf{\Sigma_Z^{\ell1}} - \mathbf{\Sigma_Z}\|_F \tag{4}$$

Now let us consider, again, $\|\mathbf{X}-\mathbf{Z}\|_F$. Thanks to Mirsky's inequality we know that $\|\mathbf{X}-\mathbf{Z}\|_F \geq \|\mathbf{\Sigma_X} - \mathbf{\Sigma_Z}\|_F$, and then, thanks to Equation (4), we can state that $\|\mathbf{X} - \mathbf{Z}\|_F \geq \|\mathbf{\Sigma_Z^{\ell1}} - \mathbf{\Sigma_Z}\|_F$, which means that $\|\mathbf{\Sigma_Z^{\ell1}} - \mathbf{\Sigma_Z}\|_F$ is a lower bound for the projection argument.

If we set $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}_{\mathbf{Z}}^{\ell 1}\mathbf{V}^T$, we then have

$$\|\mathbf{U}\mathbf{\Sigma}_{\mathbf{Z}}^{\ell 1}\mathbf{V}^T - \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\| = \tag{5}$$
$$= \|\mathbf{U}(\mathbf{\Sigma}_{\mathbf{Z}}^{\ell 1}\mathbf{V}^T - \mathbf{\Sigma}\mathbf{V}^T)\| = \tag{6}$$
$$= \|\mathbf{U}(\mathbf{\Sigma}_{\mathbf{Z}}^{\ell 1} - \mathbf{\Sigma})\mathbf{V}^T\| = \tag{7}$$
$$= \|\mathbf{\Sigma}_{\mathbf{Z}}^{\ell 1} - \mathbf{\Sigma}\| \tag{8}$$

It is worth noting that between Equation (7) and Equation (8) we exploited the fact that $\mathbf{U}$ and $\mathbf{V}$ are unitary matrices and hence orthogonal, and that Frobenius norm is invariant of orthogonal matrices.

We can now note that if we set $\mathbf{X}$ as we did above, we get a norm which is equal to a lower-bound, and hence is a minimum, which means that

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}_{\mathbf{Z}}^{\ell 1}\mathbf{V}^T \in \arg\min_{\mathbf{X}\in\mathcal{X}} \|\mathbf{X} - \mathbf{Z}\|_F = \text{proj}_{\mathcal{X}}(\mathbf{Z}) \tag{9}$$

$\square$

1.1.b Implementation

We can observe from Table 1 that the average duration with 1M entries matrix is exponentially larger than that with 100k entries. This is coherent with SVD decomposition complexity, which is exponential: $\mathcal{O}(\min(m^2p, mp^2))$.

Table 1: Duration of matrix projections onto the nuclear norm ball with $\kappa = 5000$.

Measure	1	2	3	4	5	Average
100k (s)	0.7198	0.6789	0.6661	0.677	0.6504	0.67844
1M (s)	37.6	35.4	35.42	35.27	35.25	35.788

1.2 LMO of nuclear norm

1.2.a LMO proof

Given that

$$\underset{\mathcal{X}}{\text{lmo}}(\mathbf{Z}) = \arg\min_{\mathbf{X}\in\mathcal{X}} \langle \mathbf{X}, \mathbf{Z} \rangle \tag{10}$$

where $\langle \mathbf{X}, \mathbf{Z} \rangle = \text{Tr}(\mathbf{Z}^T\mathbf{X})$, we want to show that

$$-\kappa \mathbf{u}\mathbf{v}^T \in \underset{\mathcal{X}}{\text{lmo}}(\mathbf{Z}) \tag{11}$$

As $\kappa \mathbf{u}\mathbf{v}^T \in \mathcal{X}$ by definition, we just need to show that $\langle \mathbf{X}, \mathbf{Z} \rangle \geq \langle -\kappa \mathbf{u}\mathbf{v}^T, \mathbf{Z} \rangle, \forall \mathbf{X} \in \mathcal{X}$.

Proof. First, let us work on $\langle -\kappa \mathbf{u}\mathbf{v}^T, \mathbf{Z}\rangle$:

$$\langle \kappa \mathbf{u}\mathbf{v}^T, \mathbf{Z}\rangle = \mathrm{Tr}(\mathbf{Z}^T(-\kappa \mathbf{u}\mathbf{v}^T)) = \tag{12}$$

$$= -\kappa\, \mathrm{Tr}(\mathbf{Z}^T \mathbf{u}\mathbf{v}^T) \tag{13}$$

Thanks to variational characterization of singular vectors, $\mathbf{Z}^T\mathbf{u} = \sigma\mathbf{v}$, where σ is the singular value corresponding to the singular vectors $\mathbf{u}$ and $\mathbf{v}$, i.e. the largest singular value. Hence we get

$$\langle \kappa \mathbf{u}\mathbf{v}^T, \mathbf{Z}\rangle = -\kappa\, \mathrm{Tr}(\sigma \mathbf{v}\mathbf{v}^T) = \tag{14}$$

$$= -\kappa\sigma\, \mathrm{Tr}(\mathbf{v}\mathbf{v}^T) = -\kappa\sigma \tag{15}$$

where we leveraged the fact that $\mathbf{v}$ is a column of the unitary matrix $\mathbf{V}$ and then its inner product with itself $\langle \mathbf{v}, \mathbf{v}\rangle = \mathrm{Tr}(\mathbf{v}^T\mathbf{v}) = \mathrm{Tr}(\mathbf{v}\mathbf{v}^T) = 1$.

If we now take in consideration $\langle \mathbf{X}, \mathbf{Z}\rangle$, we know that, thanks to Hölder's inequality

$$|\langle \mathbf{X}, \mathbf{Z}\rangle| \leq \|\mathbf{X}\|_q \cdot \|\mathbf{Z}\|_r, \quad r > 1,\ \frac{1}{q} + \frac{1}{r} = 1 \tag{16}$$

If we choose the nuclear norm for $\mathbf{X}$ and the infinity norm for $\mathbf{Z}$, then $q = 1$ and $r = \infty$, and $\frac{1}{q} + \frac{1}{r} = 1$. Then we have that

$$|\langle \mathbf{X}, \mathbf{Z}\rangle| \leq \|\mathbf{X}\|_* \cdot \|\mathbf{Z}\|_\infty \tag{17}$$

Removing the absolute value we obtain $-\|\mathbf{X}\|_* \cdot \|\mathbf{Z}\|_\infty \leq \langle \mathbf{X}, \mathbf{Z}\rangle \leq \|\mathbf{X}\|_* \cdot \|\mathbf{Z}\|_\infty$, which leads to

$$\langle \mathbf{X}, \mathbf{Z}\rangle \geq -\|\mathbf{X}\|_* \cdot \|\mathbf{Z}\|_\infty \tag{18}$$

Since $\mathbf{X} \in \mathcal{X}$, then $\|\mathbf{X}\|_* \leq \kappa$, which leads to $-\|\mathbf{X}\|_* \geq -\kappa$. Hence, $-\|\mathbf{X}\|_* \cdot \|\mathbf{Z}\|_\infty \geq -\kappa\|\mathbf{Z}\|_\infty$, and we obtain

$$\langle \mathbf{X}, \mathbf{Z}\rangle \geq -\|\mathbf{X}\|_* \cdot \|\mathbf{Z}\|_\infty \geq -\kappa\|\mathbf{Z}\|_\infty \tag{19}$$

Since $\|\mathbf{Z}\|_\infty = \sigma$, where σ is the largest singular value of $\mathbf{Z}$, we show that

$$\langle \mathbf{X}, \mathbf{Z}\rangle \geq -\kappa\sigma \tag{20}$$

Finally combining Equation (15) and Equation (20), we obtain

$$\langle \mathbf{X}, \mathbf{Z}\rangle \geq -\kappa\sigma = \langle \kappa \mathbf{u}\mathbf{v}^T, \mathbf{Z}\rangle \tag{21}$$

Hence

$$\langle \mathbf{X}, \mathbf{Z}\rangle \geq \langle \kappa \mathbf{u}\mathbf{v}^T, \mathbf{Z}\rangle \tag{22}$$

□

1.2.b Implementation

We can see from Table 2 that the duration with a 1M entries matrix is less than 10 times larger than the one with 100k entries, which is a significantly smaller ratio than that of projection.

Table 2: Duration of LMO computation with $\kappa = 5000$.

Measure	1	2	3	4	5	Average
100k	0.1124	0.4546	0.0401	0.0539	0.0369	0.13958
1M	0.7298	1.227	1.14	1.001	0.2803	0.87562

2 Hands-on experiment 1: Crime Scene Investigation with Blind Deconvolution

2.1 Frank-Wolfe for Blind Deconvolution

2.1.a Objective function Lipschitz-smoothness

Proof. We first compute the gradient of $f(\mathbf{X}) = \frac{1}{2}\|\mathbf{A}(\mathbf{X}) - \mathbf{b}\|_2^2$, where $\mathbf{A}$ is a linear operator and therefore can be treated as a matrix. The gradient is

$$\nabla f(\mathbf{X}) = \nabla \frac{1}{2}\|\mathbf{A}(\mathbf{X}) - \mathbf{b}\|_2^2 = \mathbf{A}^T(\mathbf{A}(\mathbf{X}) - \mathbf{b}) \tag{23}$$

Then, in order to prove the Lipschitz smoothness of the gradient, we need to find a bounded L such that

$$\|\nabla f(\mathbf{X}_1) - \nabla f(\mathbf{X}_2)\|_2 \leq L\|\mathbf{X}_1 - \mathbf{X}_2\|_2 \tag{24}$$

Let us work on the first term of the above inequality:

$$\|\nabla f(\mathbf{X}_1) - \nabla f(\mathbf{X}_2)\|_2 = \|\mathbf{A}^T(\mathbf{A}(\mathbf{X}_1) - \mathbf{b}) - \mathbf{A}^T(\mathbf{A}(\mathbf{X}_2) - \mathbf{b})\|_2 = \tag{25}$$

$$= \|\mathbf{A}^T(\mathbf{A}(\mathbf{X}_1) - \mathbf{b} - \mathbf{A}(\mathbf{X}_2) + \mathbf{b})\|_2 = \tag{26}$$

$$= \|\mathbf{A}^T\mathbf{A}(\mathbf{X}_1 - \mathbf{X}_2)\|_2 \leq \|\mathbf{A}^T\mathbf{A}\|_{2\to 2} \cdot \|\mathbf{X}_1 - \mathbf{X}_2\|_2 \tag{27}$$

Where we used Cauchy-Schwartz inequality. As $\|\mathbf{A}^T\mathbf{A}\|_{2\to 2}$ is bounded, Equation (24) is satisfied. □

2.2 Implementation

As can be seen from Figure 1, the original plate number is *J209 LTL*. The best result is obtained after just 30 iterations using $\kappa = 100$, and a 17×17 support. Using a smaller support leads to a too dark and indistinguishable picture, while a larger support can not run due to memory issues since with a larger support the matrices to work with become larger as well. Instead, using a too small κ (i.e. 1 or 10) leads to blurred and indistinguishable results, while large κs (i.e. 1k, 10k and 100k) lead to good results after more iterations. Interestingly, the quality of deblurring degenerates after 100 iterations and the plate gets less readable.

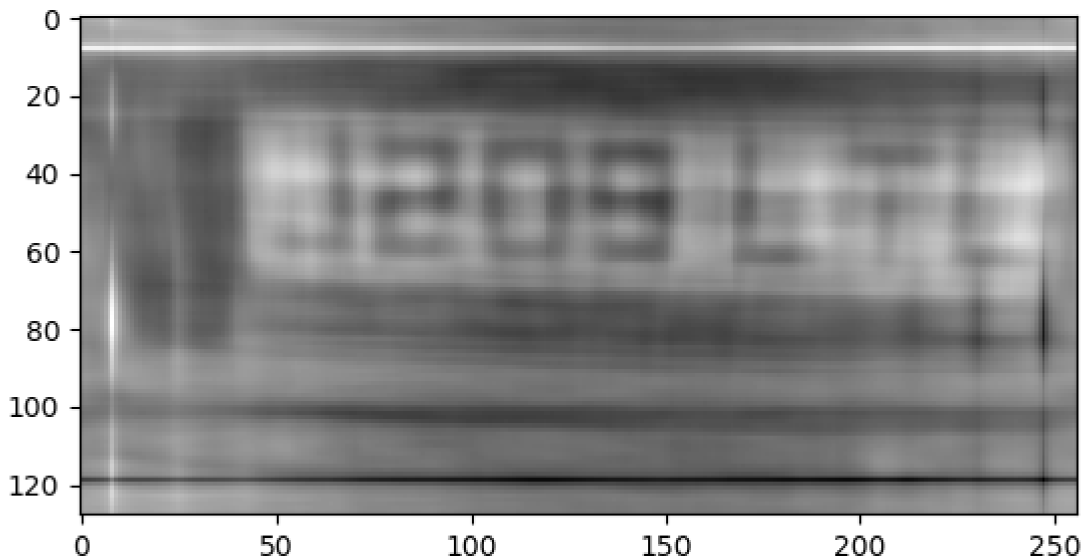

Figure 1: Plate deblurring after 30 iterations using $\kappa = 100$

3 Hands-on experiment 2: k-means Clustering by Semidefinite Programming

3.1 Conditional Gradient Method for Clustering Fashion-MNIST data

3.1.a Domain convexity

Proof. Given the domain $\mathcal{X} = \{\mathbf{X} : \mathrm{Tr}(\mathbf{X}) \leq \kappa, \mathbf{X} \in \mathbb{C}^{p\times p}, \mathbf{X} \succeq 0\}$, in order to prove that it is a convex set, we need to prove that, given $\mathbf{X}_1, \mathbf{X}_2 \in \mathcal{X}$ and $\alpha \in [0, 1]$

$$\mathbf{X}_3 = \alpha\mathbf{X}_1 + (1-\alpha)\mathbf{X}_2 \in \mathcal{X} \tag{28}$$

First we can trivially prove that $\mathbf{X}_3 \in \mathbb{C}^{p\times p}$ as it is the sum of two rescaled matrices $\mathbf{X}_1, \mathbf{X}_2 \in \mathbb{C}^{p\times p}$ and the sum of two matrices preserves the matrices dimensions. Moreover, since $\alpha \in [0, 1]$, also $(1-\alpha) \in [0, 1]$ and, more specifically, both $\alpha \geq 0$ and $1-\alpha \geq 0$, then the scaled $\alpha\mathbf{X}_1 \succeq 0$ and $(1-\alpha)\mathbf{X}_2 \succeq 0$ and their sum $\mathbf{X}_3 \succeq 0$ as well, as it is the sum of two positive semi-definite matrices. Now, we need only to prove that $\mathrm{Tr}(\mathbf{X}_3) \leq \kappa$:

$$\mathrm{Tr}(\mathbf{X}_3) = \mathrm{Tr}(\alpha \mathbf{X}_1 + (1-\alpha)\mathbf{X}_2) = \alpha \, \mathrm{Tr}(\mathbf{X}_1) + (1-\alpha) \, \mathrm{Tr}(\mathbf{X}_2)) \tag{29}$$

As $\mathbf{X}_1, \mathbf{X}_2 \in \mathcal{X}$, both $\mathrm{Tr}(\mathbf{X}_1) \leq \kappa$ and $\mathrm{Tr}(\mathbf{X}_2) \leq \kappa$. Considering the upper-bound where both $\mathrm{Tr}(\mathbf{X}_1) = \kappa$ and $\mathrm{Tr}(\mathbf{X}_2) = \kappa$, we obtain

$$\alpha\kappa + (1-\alpha)\kappa = \alpha\kappa + \kappa - \alpha\kappa = \kappa \in \sup \mathrm{Tr}(\mathbf{X}_3) \tag{30}$$

We have then shown that

$$\mathrm{Tr}(\mathbf{X}_3) \leq \kappa \tag{31}$$

Which satisfies the last requirement we needed to prove. □

3.1.b Penalized objective and gradient

As the constraints we need to satisfy are

1. $\mathbf{X}\mathbf{1} = \mathbf{1}$
2. $\mathbf{X}^T\mathbf{1} = \mathbf{1}$
3. $\mathbf{X} \geq 0$

Given that the quadratic penalty function is given by

$$\mathrm{QP}_{\mathcal{Y}}(x) = \mathrm{dist}^2(Tx, \mathcal{Y}) = \min_{y \in \mathcal{Y}} \|y - Tx\|^2 \tag{32}$$

we can express them, analogously as done in lecture 11, as respectively

1. $\frac{1}{2\beta}\|A_1(x) - b_1\|^2$, where $A_1(x) = \mathbf{X}\mathbf{1}$, and $b_1 = \mathbf{1}$, as it is the only allowed value in $\mathcal{Y}$
2. $\frac{1}{2\beta}\|A_2(x) - b_2\|^2$, where $A_2(x) = \mathbf{X}^T\mathbf{1}$, and $b_2 = \mathbf{1}$, as it is the only allowed value in $\mathcal{Y}$
3. $\frac{1}{2\beta}\mathrm{dist}^2(x, \mathcal{K})$

Hence, we can express the penalized function $f_\beta(x)$ as

$$f_\beta(x) = f(x) + \frac{1}{2\beta}\|A_1(x) - b_1\|^2 + \frac{1}{2\beta}\|A_2(x) - b_2\|^2 + \frac{1}{2\beta}\mathrm{dist}^2(x, \mathcal{K}) \tag{33}$$

Moreover, we can show that $\mathrm{dist}^2(x, \mathcal{K}) = (x - \mathrm{proj}_{\mathcal{K}}(x))^2$, as $\mathrm{proj}_{\mathcal{K}} \in \arg\min_{y \in \mathcal{Y}} \|y - x\|$ by definition of projection. Hence, we can express $f_\beta(x)$ as

$$f_\beta(x) = f(x) + \frac{1}{2\beta}\|A_1(x) - b_1\|^2 + \frac{1}{2\beta}\|A_2(x) - b_2\|^2 + \frac{1}{2\beta}(x - \mathrm{proj}_{\mathcal{K}}(x))^2 \tag{34}$$

We can now compute the gradient $\nabla f_\beta(x)$

$$\nabla f_\beta(x) = \nabla f(x) + \frac{1}{\beta}(A_1^T(A_1(x) - b_1) + A_2^T(A_2 x) - b_2) + (x - \mathrm{proj}_{\mathcal{K}}(x)) \tag{35}$$

In order to accomplish this, it is worth noting that we leveraged the fact that both A_1 and A_2 are linear operators, and Danskin's theorem to derive the gradient of the last addendum.

3.1.c v_k derivation

In order to derive v_k we are still missing the gradient of $f(x)$ and an explicit form for $\text{proj}_{\mathcal{K}}(x)$. The former is

$$\nabla\langle \mathbf{C}, \mathbf{X}\rangle = \nabla \operatorname{Tr}(\mathbf{C}^T\mathbf{X}) = \nabla \operatorname{Tr}(\mathbf{X}^T\mathbf{C}) = \mathbf{C} \tag{36}$$

In order to derive an explicit form for $\text{proj}_{\mathcal{K}}(x)$, we should first note that the set $\mathcal{K}$ operates element-wise, then also the projection operates element-wise. We can then define a set κ of elements $\mathbf{Z}_{ij}$ of $\mathbf{Z}$, such that $\mathbf{Z}_{ij} \geq 0$. Then, given $\text{proj}_{\kappa}(\mathbf{Z}_{ij}) = \arg\min_{\mathbf{X}_{ij}} |\mathbf{X}_{ij} - \mathbf{Z}_{ij}|$ we have two cases:

1. $\mathbf{Z}_{ij} \geq 0$, which means that $\mathbf{Z}_{ij} \in \kappa$, hence $\text{proj}_{\kappa}(\mathbf{Z}_{ij}) = \mathbf{Z}_{ij}$
2. $\mathbf{Z}_{ij} < 0$

For the second case, given $\arg\min_{\mathbf{X}_{ij}} |\mathbf{X}_{ij} - \mathbf{Z}_{ij}|$, we get that $\min_{\mathbf{X}_{ij}} |\mathbf{X}_{ij} - \mathbf{Z}_{ij}| = \mathbf{Z}_{ij}$, that is obtained with $\mathbf{X}_{ij} = 0$. Hence, merging the results obtained in the two cases,

$$\text{proj}_{\mathcal{K}}(x) = \max(0, x) \tag{37}$$

where max is applied element-wise.

3.1.d Implementation

No specific penalizer parameter The k-means value before the clustering is 150.9680. We first run the experiment using the same penalizer coefficient for all the constraints. With this setting we obtain a final k-means value of 98.5119, and what is plotted in Figure 2, Figure 3 and Figure 4. The obtained misclassification rate is 0.1600, while the final objective value is -286.55, which is lower than the provided 57.05339187. This can be due to the fact that $\mathbf{X}$ can have several negative values (the penalization notwithstanding) such that $\langle \mathbf{C}, \mathbf{X}\rangle < 0$, which leads to a negative f.

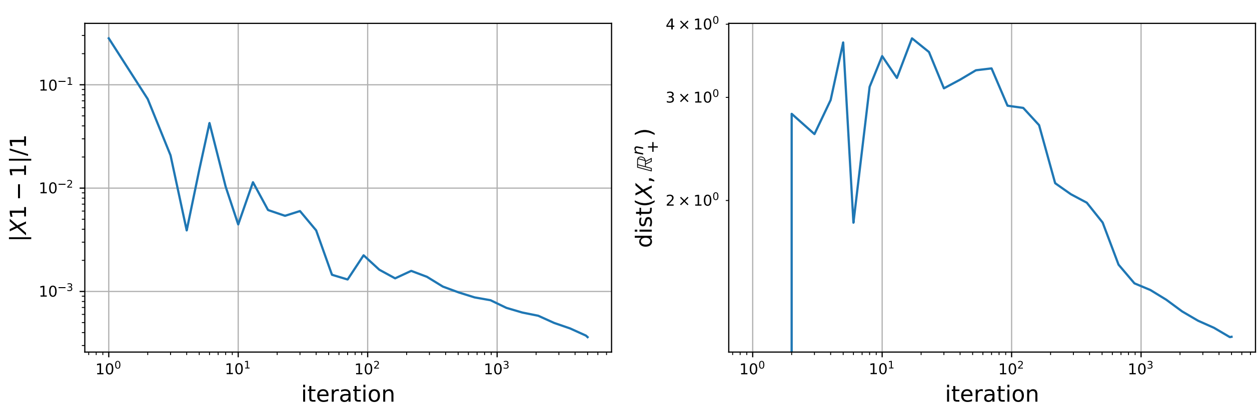

Figure 2: Convergence for k-means clustering, using no specific β coefficient for the $\mathcal{K}$ quadratic penalty function.

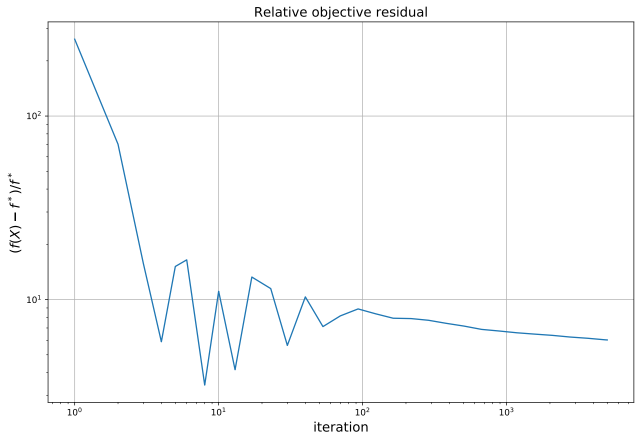

Figure 3: Relative objective residual for k-means clustering, using no specific β coefficient for the $\mathcal{K}$ quadratic penalty function.

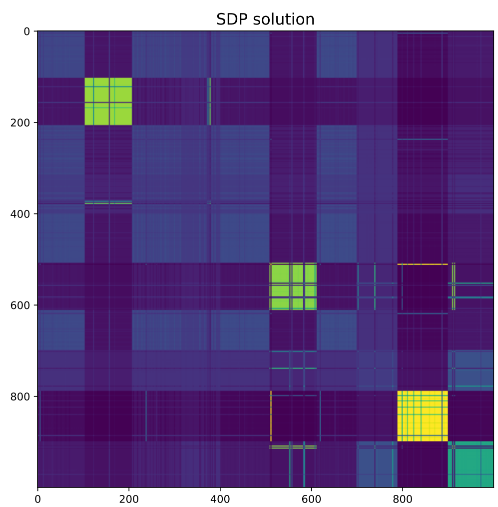

Figure 4: SDP solution for k-means clustering, using no specific β coefficient for the $\mathcal{K}$ quadratic penalty function.

Specific penalizer parameter Next, an experiment is run using a specific $\beta = 1000$ as coefficient for the $\mathcal{K}$ quadratic penalty function. With this setting we obtain a final k-means value of 48.6801, and what is plotted in Figure 5, Figure 6 and Figure 7. The obtained misclassification rate is 0.1330, while the final objective value is -128.41, which is lower than the provided 57.05339187, but larger than the value obtained using no different penalization parameter. This can be thanks to the stronger enforcement of the positivity of the entries of the matrix $\mathbf{X}$, which could have led to a less negative result. It is also worth noting that the convergence is less oscillating in this setting than in the previous one, and that the SDP solution in Figure 7 has clearer cluster separations between classes than those in Figure 4.

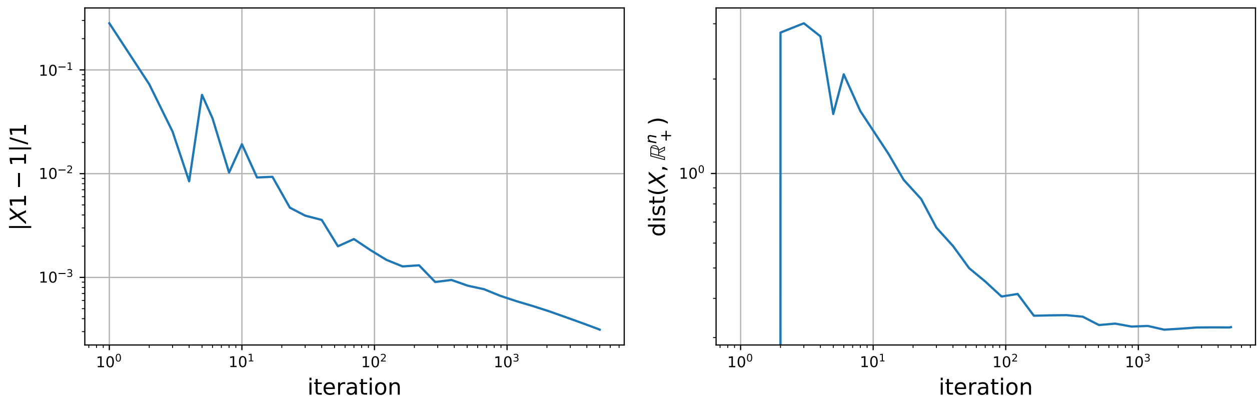

Figure 5: Convergence for k-means clustering, using $\beta = 1000$ as coefficient for the $\mathcal{K}$ quadratic penalty function.

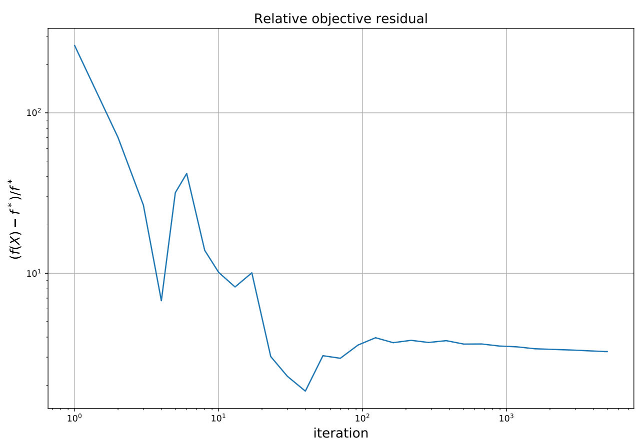

Figure 6: Relative objective residual for k-means clustering, using $\beta = 1000$ as coefficient for the $\mathcal{K}$ quadratic penalty function.

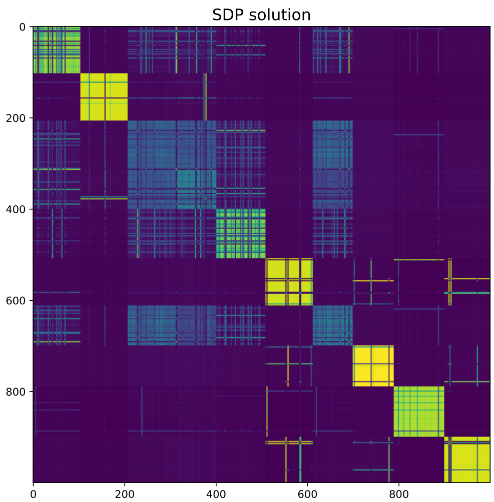

Figure 7: SDP solution for k-means clustering, using $\beta = 1000$ as coefficient for the $\mathcal{K}$ quadratic penalty function.