Revisiting clustering as matrix factorisation on the Stiefel manifold

Stephane Chretien, Benjamin Guedj

Submitted vol XX:1–18, 2019

Revisiting clustering as matrix factorisation on the Stiefel manifold

Stéphane Chrétien STEPHANE.CHRETIEN@NPL.CO.UK
Hampton Road, Teddington, TW11 0LW, UK

Benjamin Guedj BENJAMIN.GUEDJ@INRIA.FR
Inria and University College London

Abstract

This paper studies clustering for possibly high dimensional data (*e.g.* images, time series, gene expression data, and many other settings), and rephrase it as low rank matrix estimation in the PAC-Bayesian framework. Our approach leverages the well known Burer-Monteiro factorisation strategy from large scale optimisation, in the context of low rank estimation. Moreover, our Burer-Monteiro factors are shown to lie on a Stiefel manifold. We propose a new generalized Bayesian estimator for this problem and prove novel prediction bounds for clustering. We also devise a componentwise Langevin sampler on the Stiefel manifold to compute this estimator.

Keywords: Clustering, concentration inequalities, non-negative matrix factorisation, Gaussian mixtures, PAC-Bayes, optimisation on manifolds.

1. Introduction

Clustering, *i.e.,* unsupervised classification, is a central problem in machine learning and has attracted great attention since the origins of statistics, via model-based learning, but recently regained a lot of interest from theoreticians, due to its similarities with community detection (Arias-Castro and Verzelen, 2014; Verzelen and Arias-Castro, 2015). On the application side, clustering is pervasive in data science, and has become a basic tool in computer science, bio-informatics, finance, metrology, to name but a few.

1.1. Historical background

The problem of identifying clusters in a data set can be addressed using an wide variety of tools. Two main approaches can be delineated, namely the model-based approach and the non-model based approach. Techniques such as hierarchical clustering (Hastie et al., 2009), minimum spanning tree-based approaches (Blum et al., 2016), K-means algorithms (Hastie et al., 2009), belong to the non-model based family of methods. Model-based techniques mostly rely on mixture modelling (McLachlan and Peel, 2004) and often offer better interpretability whilst being easily amenable to uncertainty quantification analysis. The EM algorithm (Dempster et al., 1977; McLachlan and Peel, 2004) is often the algorithm of choice in the frequentist approach while many Monte Carlo Markov Chain techniques have been proposed for estimation in the Bayesian setting.

In recent years, the clustering problem has revived a surge of interest in a different setting, namely community detection in random graphs. Tools from spectral graph theory and convex optimisation, combined with recent breakthrough from random matrix theory where put to work in devising efficient clustering methods that operate in polynomial time. The celebrated example of Max-Cut, a well known NP-hard combinatorial optimisation problem strongly related to bi-

clustering and with a tight Semi-Definite Programming (SDP) relaxation discovered by Goemans and Williamson (1995), is an example among the many successes of the convex optimisation approach to addressing machine learning problems. SDP is the class of optimisation problems that consist in minimising a linear function over the sets of Positive Semi-Definite matrices that satisfy a set of linear (in)equalities. Goemans and Williamson (1995) subsequently triggered a long lasting trend of research in convex relaxation with many application in data science, and recent results proposing tighter relaxations to the clustering problem can be found in Guédon and Vershynin (2016), Chrétien et al. (2016), Giraud and Verzelen (2018). Some of these methods even apply to any kind of data set endowed with a relevant affinity measure computed from the pairwise distances between the data points, and share the common feature of using low-rank matrix representations of the clustering problem. The theoretical tools behind the analysing of the performance of these convex optimisation-based methods are also quite fascinating and range from random matrix theory (Bandeira, 2018; Vershynin, 2018), concentration inequalities for quadratic forms of random vectors (Rudelson and Vershynin, 2013) and optimisation theory (optimality conditions, see Royer, 2017), localisation arguments in statistical learning theory (Giraud and Verzelen, 2018), Grothendieck's inequality (Guédon and Vershynin, 2016; Montanari and Sen, 2015), to name but a few.

The main drawback, however, of the current lines of approach to the performance analysis of these powerful convex *SDP* and *spectral relaxations* is that they all depend on the separation between clusters, *i.e.,* the minimum distance between two points from different clusters, a crucial parameter in the aforecited analyses. In real data sets however, sufficient inter-cluster separation rarely holds and overlaps between clusters are the common situation. This leaves open the difficult problem of finding an alternative theoretical route for controlling the estimation error. On the computational side, the sample size is also a problem for SDP relaxations for which off-the-shelf software does not scale to big data. A remedy to this problem is to use the Burer-Monteiro factorisation consisting in solving in U where $X = UU^t$ is the variable of the SDP at hand (Burer and Monteiro, 2003). The Burer-Monteiro factorisation results in a non-convex optimisation problem whose local minimisers are global minimisers when the number of columns of U is sufficiently large (Burer and Monteiro, 2005; Boumal et al., 2016). In practice however, the rank of the sought matrix is simply equal to the number of clusters, and whether such small priors on the rank of the Burer-Monteiro factorisation are compatible with the local/global equivalence of the minimisers in general remains an open question to this day. A final source of frustration in our list, is that there does not seem to exist any method for quantifying the uncertainty of the results in these convex optimisation-based approaches to clustering.

In the present paper, we propose a generalized Bayesian approach to clustering which hinges on low rank estimation of a clustering matrix. We then leverage arguments from the PAC-Bayesian theory for controlling the error which does not use any prior estimate of separation. Our approach is based on the estimation of a normalised version T^* of the adjacency matrix of the clustering, which can be factorised into $T^* = U^*U^{*^t}$, where U^* has orthonormal, non-negative columns. Leveraging this structure leads to sampling on the intersection of the Stiefel manifold Edelman et al. (1998) and the non-negative orthant, which is another surprising manifestation of the power of non-negative matrix factorisation (NMF) in clustering problems. Solving this factorised version in the PAC-Bayesian setting is the sampling counterpart of the Burer-Monteiro approach to the numerical solution of high dimensional SDP. The PAC-Bayesian approach (initiated by Shawe-Taylor and Williamson, 1997; McAllester, 1998, 1999; Catoni, 2004, 2007; see Guedj, 2019, for a

recent survey) moreover makes no prior use of the separation and at the same time makes it possible to obtain state-of-the-art risk bounds.

1.2. Our contribution

The main goal of the present paper is to study the clustering problem from a low rank Stiefel matrix, i.e. matrices with orthonormal columns, view point, and present a PAC-Bayesian analysis of the related statistical estimation problem. Our approach is in particular inspired by recent work on low rank approximation for k-means (Boutsidis et al., 2009; Cohen et al., 2015), where the representation of clustering using the matrix T^* is explicitly stated (although no algorithm is provided), and PAC-Bayesian bounds for Non-Negative Matrix factorisation (as introduced by Alquier and Guedj, 2017, although they do not establish the link between NMF and clustering). To the best of our knowledge, the representation in Boutsidis et al. (2009) using the matrix T^* has never been studied from a statistical learning perspective.

We present our main result (Theorem 1, which states an inequality holding in expectation on the prediction performance) in Section 2 and its proof in Section 3. Our second main result is Theorem 2, which specifies the results of Theorem 1 in the case where we assume that the family of means is incoherent. Section 4 is devoted to our algorithm (an alternating Langevin sampler which relies on computing gradients on the Stiefel manifold), and the paper closes with a discussion and comments on future work in 5. Additional proofs are gathered in Appendix A.

1.3. Notation

The notation used in the present paper is fairly standard. The canonical scalar product in $\mathbb{R}^d$ will be denoted by $\langle \cdot, \cdot \rangle$, the ℓ_p norms by $\| \cdots \|_p$. For matrices in $\mathbb{R}^{d \times n}$, the operator norm will be denoted by $\| \cdot \|$ and the Frobenius norm by $\| \cdot \|_F$. The Stiefel manifold of order (n, R), i.e. the set of matrices in $\mathbb{R}^{n \times R}$ with orthonormal columns, will be denoted by $\mathbb{O}_{n,R}$, and $\mathbb{O}_{n,R,+}$ will denote the subset of the Stiefel manifold $\mathbb{O}_{n,R}$ consisting of componentwise nonnegative matrices. The matrices in $\mathbb{O}_{n,R}$ will sometimes be identified with matrices in $\mathbb{R}^{n \times n}$ where the first R columns form an orthonormal family and the remaining $n - R$ columns are set to zero. The gradient operator acting on differentiable multivariate functions will be denoted by ∇.

2. Non-negative factorisation of the Stiefel manifold

This section is devoted to the presentation of our framework and our main theoretical result.

2.1. Model

Let data points $x_1, \ldots, x_n$ be vectors in $\mathbb{R}^d$ and let X denote the matrix

$$X = [x_1, \ldots, x_n] .$$

Let $\mu_1, \ldots, \mu_K$ be $K \in \mathbb{N} \backslash \{0\}$ vectors in $\mathbb{R}^d$. We will say that x_i belongs to cluster $k \in \{1, \ldots, K\}$ if $x_i = \mu_k + E_i$ for some centered random vector $E_i \in \mathbb{R}^d$. For each $i = 1, \ldots, n$, we will denote by k_i the label of the cluster to which x_i belongs. For each k, we will denote by I_k the index set of the points which belong to cluster k and n_k its cardinality. Now, we can decompose X as $X = M + E$ with

$$M = [\mu_{k_1}, \ldots, \mu_{k_n}] ,$$

$$E = [\varepsilon_1, \dots, \varepsilon_n]$$

More assumptions about the noise matrix E will be introduced and their consequences on the performance of our clustering method will be studied in Theorem 2.

2.2. Ideal solution

If we try to estimate the columns of M, one simple way is to use a convex combination of the x_i's for each of them. In other words, one might try to approximate X by XT^* where T^* is a $\mathbb{R}^{n\times n}$ matrix. One simple way to proceed is to set T as the matrix which computes the cluster means, given by

$$T^*_{i,j} = \begin{cases} \frac{1}{n_k} & \text{if } x_i \text{ and } x_j \text{ belong to the same cluster } k, \textit{i.e.}, k_i = k_j \\ 0 & \text{otherwise.} \end{cases}$$

Thus, each column $i = 1, \dots, n$ of XT^* is simply the mean over cluster k_i. This type of solution is well-motivated by the fact that the mean is the least-squares solution of the approximation of μ_k by the observation points. The matrix T^* defined as above enjoys the following desirable properties: (i) its rank is exactly the number of clusters (ii) it is nonnegative (iii) the columns corresponding to different clusters are orthogonal.

One important fact to notice is that the eigenvalue decomposition of T^* is explicit and given by

$$T^* = U^* U^{*t} \tag{2.1}$$

with

$$U^* = \left[\frac{1}{\sqrt{n_1}} 1_{I_1}, \dots, \frac{1}{\sqrt{n_K}} 1_{I_K} \right], \tag{2.2}$$

and therefore, all the eigenvalues of T^* are equal to one.

Based on this decomposition, we can now focus on estimating U^* rather than T^*, the reason being that working on estimating U^* with $\hat{U} \geq 0$ will automatically enforce positive semi-definiteness of $\hat{T}$ (the estimator of T^*) and non-negativity of its components. Moreover, enforcing the orthogonality of the columns of $\hat{U}$, combined with the non-negativity of its components, will enforce the columns of $\hat{U}$ to have disjoint supports.

Adopting a generalized Bayesian strategy (inspired by Alquier and Guedj, 2017), we will then define a prior distribution on $\hat{U}$ and study the main properties of the resulting (generalized) posterior distribution.

2.3. The latent variable model

In order to perform an accurate estimation, we need to devise meaningful priors which will account for the main constraints our estimator should satisfy, namely (i) nonnegativity of the entries (ii) orthogonality of the columns (iii) the columns have unit ℓ_2 norm (iv) group sparsity of the columns.

In order to simplify this task, we will introduce a *latent (matrix) variable* O with uniform distribution on the orthogonal group, and *build priors on* U that will promote group sparsity of the columns and non-negativity (component-wise).

Prior on (U, O)**.** Let $\mathcal{U}_R$ denote the set of matrices of the form

$$U = \begin{bmatrix} U_{1,1} & \cdots & U_{1,R} & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ U_{n,1} & \cdots & U_{n,R} & 0 & \cdots & 0 \end{bmatrix}.$$

Let $\mathbb{O}_{n,R}$ denote the Stiefel manifold, *i.e.,* the manifold of all matrices with R orthonormal columns in $\mathbb{R}^n$. The prior on $(U, O) \in \mathcal{U}_R \times \mathbb{O}_{n,R}$ is given by

$$\pi_{U,O}(U, O) = \pi_{U|O}(U)\, \pi_O(O)$$

with

$$\pi_{U_{i,r}|O}(U) = \frac{1}{\sqrt{2\pi}\mu} \exp\left(-\frac{\|U_{i,r} - |O_{i,r}|\|_F^2}{2\mu^2}\right), \quad i = 1, \ldots, n, \quad r = 1, \ldots, R,$$

with R being a fixed integer and π_O being the uniform distribution on the Stiefel manifold $\mathbb{O}_{n,R}$.

2.4. Generalized posterior and estimator

Following the approach of Alquier and Guedj (2017), we use a loss term (instead of a likelihood, hence the term "generalized Bayes", see Guedj, 2019, for a survey) given by

$$L_\lambda(U) = \exp\left(-\frac{\lambda}{2}\, \|X - XUU^t\|_F^2\right)$$

for some fixed positive parameters λ and μ. The resulting generalized posterior (also known as a *Gibbs measure*) is defined as

$$\rho(U, O) = \frac{1}{Z_\lambda}\, L_\lambda(U)\, \pi_{U|O}(U)\, \pi_O(O),$$

where Z_λ denotes the normalisation constant $Z_\lambda = \int\, L_\lambda(U)\, \pi_{U|O}(U)\, \pi_O(O)\, dU$. Finally we let $\hat{U}_\lambda$ denote the posterior mean of U, i.e.

$$\hat{U}_\lambda = \int U\, L_\lambda(U)\, \pi_{U|O}(U) \pi_O(O) dU dO.$$

2.5. A PAC-Bayesian-flavored error bound

Our main result is the following theorem.

Theorem 1 *Let* $\nu_{\min}$ *and* $\nu_{\max}$ *be such that*

$$\nu_{\min} \leq \min_{\tilde{U} \in \mathbb{O}_{n,R,+},\ M\tilde{U}\tilde{U}^t = M} \|E(I - \tilde{U}\tilde{U}^t)\|_F \quad \textit{and} \quad \nu_{\max} \geq \max_{\tilde{U} \in \mathbb{O}_{n,R,+},\ M\tilde{U}\tilde{U}^t = M} \|E(I - \tilde{U}\tilde{U}^t)\|_F.$$

Then, for all $\varepsilon > 0$*, and for all* $c_O > 0$ *and* $c_U > c_O$ *such that*

$$c_U\, (2 + c_U) \leq \varepsilon \frac{\nu_{\min}}{\|M\| + \|E\|}, \tag{2.3}$$

for any $R \in \{1, \dots, n\}$ and for c and ρ sufficiently small universal constants, we have

$$\mathbb{E}\left[\left\|M\left(T^* - \hat{U}_\lambda \hat{U}_\lambda^t\right)\right\|_F\right] \leq (1+\varepsilon) \min_{\tilde{U} \in \mathbb{O}_{n,R,+},\ M\tilde{U}\tilde{U}^t = M} \|M(T^* - \tilde{U}\tilde{U}^t)\|_F \tag{2.4}$$

$$+ \sqrt{\frac{1}{\exp\left(\frac{\left(\sqrt{c_U^2 - c_O^2} - \sqrt{nR}\right)^2}{2}\right) - 1}} + \sqrt{\left(nR - \frac{1}{2}(R^2+R)\right)\log(\rho^{-1}) + \log(c^{-1})}$$

$$+ (2+\varepsilon)\nu_{\max}.$$

This theorem gives a prediction bound on the difference between the true and the estimated cluster matrices filtered by the matrix of means. Up to our knowledge, this is the first oracle bound for clustering using a generalized Bayesian NMF. Note that the oracle bound is not sharp as the leading constant is $1 + \varepsilon > 1$, however ε may be chosen arbitrarily close to 0.

Note also that the claim that this result is PAC-Bayesian-flavored comes from the fact that the prediction machinery is largely inspired by Alquier and Guedj (2017), and the scheme of proof builds upon the PAC-Bayesian bound from Dalalyan and Tsybakov (2008). Hence we kept that PAC-Bayesian filiation, even though the bound holds in expectation.

The following Theorem gives a more precise bound in the case where the noise E is assumed to be iid Gaussian.

Theorem 2 *Assume that the dimension is larger that the number of clusters, i.e. $d > K$. In addition to the assumptions of Theorem 1, assume that E is iid Gaussian with minimum (resp. maximum) one-dimensional variance $\sigma_{\min}^2$ (resp. $\sigma_{\max}^2$) and assume also that the μ_k have Euclidean norm less that 1 and pairwise scalar products less than μ in absolute value. Then, as long as $\mu < 1/(K-1)$, for all $\varepsilon > 0$, and for all $c_O > 0$ and $c_U > c_O$ such that*

$$c_U\,(2 + c_U) \leq \varepsilon \frac{\nu_{\min}}{\sqrt{\left(\max_{k=1}^K\ n_k\right)\ \mu(K-1) + 1} + \sigma_{\max}\left(\sqrt{n} + 2\sqrt{d}\right)}, \tag{2.5}$$

with probability at least

$$1 - \exp(-d) - \left(\frac{c}{\varepsilon}\right)^{nR - R(R+1)/2}\left(\frac{2}{\sqrt{\pi n(n-R)}}\left(t_{\min}\, e/2\right)^{n(d-R)/4} + \exp(-t_{\max})\right) - \exp(-nu^2/8),$$

we have

$$\sum_{k=1}^K \sum_{i_k \in I_k}\left(\sum_{i'_k \in I_k} T^*_{\pi, i'_k, i_k} - \hat{U}_{\lambda,\pi,i'_k}\hat{U}^t_{\lambda,\pi,i_k}\right)^2$$

$$\leq \frac{(1+\varepsilon)\ \sqrt{1+\mu(K-1)}}{1-\mu(K-1)} \min_{\tilde{U} \in \mathbb{O}_{n,R,+},\ M\tilde{U}\tilde{U}^t = M} \|M(T^* - \tilde{U}\tilde{U}^t)\|_F$$

$$+ \sqrt{\frac{1}{\exp\left(\frac{\left(\sqrt{c_U^2 - c_O^2} - \sqrt{dR}\right)^2}{2}\right) - 1}} + \sqrt{(dR - \frac{1}{2}(R^2+R))\log(\rho^{-1}) + \log(c^{-1})} + (2+\varepsilon)\nu_{\max}.$$

with

$$t_{\min} = \frac{\left(\frac{\nu_{\min}}{\sigma_{\min}} + 4\varepsilon\sqrt{nd+u}\right)^2}{n(n-R)} \quad \text{and} \quad t_{\max} = \left(\frac{\nu_{\max}}{\sigma_{\max}} - 4\varepsilon\sqrt{nd+u}\right)^2 - \sqrt{n(n-R)},$$

This theorem shows that a bound on the difference of the cluster matrices can be obtained when the matrix of means is sufficiently incoherent. Notice that this bound is not exactly component-wise, but considers a sum over clusters, which is perfectly relevant because the matrix M does not distinguish between points in the same cluster. As expected, the smaller the coherence μ, the better the oracle bound.

The proof of Theorem 2 is deferred to Appendix A.

3. Proof of Theorem 1

We break down the proof in the following successive elementary steps.

3.1. Initial PAC-Bayesian bound

Theorem 3.1 (adapted from Dalalyan and Tsybakov, 2008) *For $\lambda \leq 1/4$,*

$$\mathbb{E}\left[\|X - X\hat{U}_\lambda\hat{U}_\lambda^t\|_F^2\right] \leq \inf_\rho \left\{\int \|X - X\hat{U}_\lambda\hat{U}_\lambda^t\|_F^2 \, \rho(U)\, dU\right\} + \frac{KL(\rho,\pi)}{\lambda} \tag{3.6}$$

where the infimum is taken over all probability measures ρ which are absolutely continuous with respect to π. Here KL denotes the Kullback-Leibler divergence.

3.2. Bounding the integral part

In order to bound the integral part in the bound given in (3.6), we define for any R and any matrix $U^0 \in \mathcal{U}_R \cap \mathbb{O}_{n,R,+}$ for any $c_U,\ c_O \in (0,1]$, the measure

$$\rho_{R,U^0,c_U,c_O}(U,O) = \frac{1_{\|U-U^0\|_F \leq c_U,\ \||O|-U^0\|_F \leq c_O}\ \ \pi_{U,O}(U,O)}{\pi_{U,O}\left(\|U-U^0\|_F \leq c_U,\ \||O|-U^0\|_F \leq c_O\right)}.$$

Define $c = (c_U, c_O)$. Using these distributions we will be able to prove the following bound.

Lemma 3.2 *We have*

$$\int \|X - XUU^t\|_F^2 \rho_{R,U^0,c_U,c_O}(U,O)\, dUdO \leq \left(\|X - XU^0U^{0^t}\|_F + c_U\,(2+c_U)\ (\|M\| + \|E\|)\right)^2.$$

Proof Note that

$$\int \|X - XUU^t\|_F^2 \rho_{R,U^0,c_U,c_O}(U,O)\, dUdO = \int \Big(\|X - XU^0U^{0^t}\|_F^2 + 2\langle X(UU^t - U^0U^{0^t}, X - XU^0U^{0^t}\rangle$$
$$+ \|XUU^t - XU^0U^{0^t}\|_F^2\Big)\, \rho_{R,U^0,c_U,c_O}(U,O)\, dUdO$$

and thus,

$$\begin{aligned}\int \|X - XUU^t\|_F^2 \quad \rho_{R,U^0,c_U,c_O}(U,O)\, dUdO \\ = \int \Big(\|X - XU^0U^{0^t}\|_F^2 + 2\langle X(UU^t - U^0U^{0^t}), X - XU^0U^{0^t}\rangle \\ + \|XUU^t - XU^0U^{0^t}\|_F^2 \Big)\, \rho_{R,U^0,c_U,c_O}(U,O)\, dUdO.\end{aligned}$$

By the Cauchy-Schwartz inequality, we get

$$\begin{aligned}\int \|X - XUU^t\|_F^2 \quad \rho_{R,U^0,c_U,c_O}(U,O)\, dUdO = \|X - XU^0U^{0^t}\|_F^2 \\ + \|X - XU^0U^{0^t}\|_F \int 2\, \|X(UU^t - U^0U^{0^t})\|_F\, \rho_{R,U^0,c_U,c_O}(U,O)\, dUdO \\ + \int \|X(UU^t - U^0U^{0^t})\|_F^2\, \rho_{R,U^0,c_U,c_O}(U,O)\, dUdO\end{aligned}$$

Note further that since (U,O) must belong to the support of ρ_{R,U^0,c_U,c_O}, we have

$$\begin{aligned}\|X(UU^t - U^0U^{0^t})\|_F &= \left\|X\left(U(U-U^0)^t + (U-U^0)U^{0^t}\right)\right\|_F \\ &\leq \left\|XU(U-U^0)^t\right\|_F + \left\|X(U-U^0)U^{0^t}\right\|_F.\end{aligned}$$

Now, on the one hand, we have

$$\begin{aligned}\left\|XU(U-U^0)^t\right\|_F^2 &\leq \|XU\|^2\, \|U-U^0\|_F^2 \\ &\leq \left(\|MU^0\| + \|EU^0\| + (\|M\| + \|E\|)\|U-U^0\|\right)^2\, \|U-U^0\|_F^2\end{aligned}$$

and since the columns of U^0 are orthonormal, we have $\|U^0\| = 1$, which gives

$$\left\|XU(U-U^0)^t\right\|_F \leq (\|M\| + \|E\|)\left(1 + \|U-U^0\|_F\right)\, \|U-U^0\|_F.$$

On the other hand,

$$\left\|X(U-U^0)U^{0^t}\right\|_F^2 \leq \|X\|^2 \left\|(U-U^0)U^{0^t}\right\|_F^2 \leq \|X\|^2\, \|U^0\|^2 \left\|(U-U^0)\right\|_F^2$$

and using again that $\|U^0\| = 1$,

$$\left\|X(U-U^0)U^{0^t}\right\|_F \leq \|X\|\, \left\|(U-U^0)\right\|_F.$$

From this, we easily deduce that

$$\begin{aligned}\int \|X - X(UU^t)\|_F\, \rho_{R,U^0,c_U,c_O}(U,O)\, dUdO &\leq (\|M\| + \|E\|)\left((1+c_U)\, c_U + \, c_U\right), \\ &\leq c_U\,(2+c_U)\, \left(\|M\| + \|E\|\right),\end{aligned}$$

and

$$\int \|X - X(UU^t)\|_F^2\, \rho_{R,U^0,c_U,c_O}(U,O)\, dUdO \leq c_U^2\,(2+c_U)^2\, \left(\|M\| + \|E\|\right)^2$$

which completes the proof. ∎

3.3. Upper bound on the Kullback-Leibler divergence

Lemma 3.3 *We have*

$$KL\left(\rho_{R,U^0,c},\pi\right) \leq \frac{1}{\exp\left(\frac{\left(\frac{1}{\mu}\sqrt{c_U^2-c_O^2}-\sqrt{nR}\right)^2}{2}\right)-1} + \left(nR-\frac{1}{2}(R^2+R)\right)\log(\rho^{-1})+\log(c^{-1}),$$

for some $\rho, c > 0$ sufficiently small.

Proof By definition,

$$\begin{aligned} KL(\rho_{R,U^0,c_U,c_O},\pi) &= \int \rho_{R,U^0,c_U,c_O}(U,O)\log\left(\frac{\rho_{R,U^0,c_U,c_O}(U,O)}{\pi_{U,O}(U,O)}\right)\,dUdO, \\ &= \log\left(\frac{1}{\int 1_{\|U-U^0\|_F\leq c_U,\ \||O|-U^0\|_F\leq c_O}\,\pi_{U,O}(U,O)\,dUdO}\right)\,dUdO. \end{aligned}$$

We then have

$$\begin{aligned} &\pi_{U,O}\left(\|U-U^0\|_F^2\leq c_U^2,\quad \||O|-U^0\|_F^2\leq c_O^2\right) \\ &\geq \pi_{U,O}\left(\|U-|O|\|_F^2\leq c_U^2-c_0^2,\quad \||O|-U^0\|_F^2\leq c_O^2\right) \\ &= \int_{\||O|-U^0\|_F^2\leq c_O^2}\left(\int 1_{\|U-|O|\|_F^2\leq c_U^2-c_O^2}\quad \pi_{U|O}(U)dU\right)\pi_O(O)dO \\ &= \int_{\||O|-U^0\|_F^2\leq c_O^2}\pi_{U|O}\left(\sum_{i=1}^d\sum_{r=1}^R(U_{i,r}-|O_{i,r}|)^2\leq c_U^2-c_O^2\right)\pi_O(O)dO. \end{aligned}$$

As long as $c_U^2 \geq dR$, the inner probability can be bounded as follows (see equation 7.50 in Massart, 2007):

$$\pi_{U|O}\left(\sqrt{\sum_{i=1}^d\sum_{r=1}^R(U_{i,r}-|O_{i,r}|)^2}\leq\sqrt{c_U^2-c_O^2}\right)\geq 1-\exp\left(-\frac{\left(\frac{1}{\mu}\sqrt{c_U^2-c_O^2}-\sqrt{nR}\right)^2}{2}\right).$$

From this last inequality, we get

$$\begin{aligned} &\pi_{U,O}\left(\|U_{i,r}-U_{i,r}^0\|_F^2\leq c_U^2,\quad \||O|-U^0\|_F^2\leq c_O^2\right) \\ &\qquad\geq\left(1-\exp\left(-\frac{\left(\frac{1}{\mu}\sqrt{c_U^2-c_O^2}-\sqrt{nR}\right)^2}{2}\right)\right)\pi_O\left(\||O|-U^0\|_F^2\leq c_O^2\right). \end{aligned} \tag{3.7}$$

We now use the elementary inequality $\log(1+x)\geq x/(1+x)$ for $x\in(-1,+\infty)$, and a lower bound on $\pi_O\left(\||O|-U^0\|_F^2\leq c_O^2\right)$ (given by Lee and Ruymgaart, 1996). Therefore

$$\pi_O\left(\||O|-U^0\|_F^2\leq c_O^2\right)\geq c\,\rho^{nR-\frac{1}{2}(R^2+R)}$$

for c and ρ sufficiently small. From these two inequalities, we end the proof:

$$\log\left(\pi_{U|O}\left(\sqrt{\sum_{i=1}^{d}\sum_{r=1}^{R}(U_{i,r}-|O_{i,r}|)^2}\leq\sqrt{c_U^2-c_O^2}\right)\right)\geq\frac{-1}{\exp\left(\frac{\left(\frac{1}{\mu}\sqrt{c_U^2-c_O^2}-\sqrt{dR}\right)^2}{2}\right)-1}$$
$$+\left(nR-\frac{1}{2}(R^2+R)\right)\log(\rho)+\log(c).$$

■

3.4. Combining the lemmæ

Combining the results of the previous lemmæ, we get the following proposition.

Proposition 3.4 *Let* $\nu=\min_{\tilde{U}\in\mathbb{O}_{n,R,+},\ M\tilde{U}\tilde{U}^t=M}\ \|E(I-\tilde{U}\tilde{U}^t)\|_F$. *Then, for any* c_U *such that* $c_U\,(2+c_U)\leq\varepsilon\frac{\nu}{(\|M\|+\|E\|)}$, *we have*

$$\mathbb{E}\left[\left\|X-X\hat{U}_\lambda\hat{U}_\lambda^t\right\|_F^2\right]\leq(1+\varepsilon)^2\|X-XU^0U^{0^t}\|_F^2$$
$$+\frac{1}{\exp\left(\frac{\left(\frac{1}{\mu}\sqrt{c_U^2-c_O^2}-\sqrt{nR}\right)^2}{2}\right)-1}+(nR-\frac{1}{2}(R^2+R))\log(\rho^{-1})+\log(c^{-1}).$$

Proof Let $\tilde{U}$ be a minimiser in the numerator of the right hand-side in (2.3). Since

$$\|X-X\tilde{U}\tilde{U}^t\|_F=\|M+E-M\tilde{U}\tilde{U}^t-E\tilde{U}\tilde{U}^t\|_F$$

and since $M=M\tilde{U}\tilde{U}^t$, we get $\|X-X\tilde{U}\tilde{U}^t\|_F=\|E(I-\tilde{U}\tilde{U}^t)\|_F$. As

$$\|E(I-\tilde{U}\tilde{U}^t)\|_F\leq\|E(I-U^0U^{0^t})\|_F$$

for all $U^0\in\mathbb{O}_{n,r,+}\cup\{U\mid MUU^t=M\}$, the claim then follows from combining the results of Lemma 3.2 and Lemma 3.3 above, and taking c_U such that (2.3) holds. ■

3.5. Assembling the elements

We have that

$$\left\|X-X\hat{U}_\lambda\hat{U}_\lambda^t\right\|_F^2\geq\left(\left\|M-M\hat{U}_\lambda\hat{U}_\lambda^t\right\|_F-\nu_{\max}\right)^2=\left(\left\|M\left(U^0U^{0^t}-\hat{U}_\lambda\hat{U}_\lambda^t\right)\right\|_F-\nu_{\max}\right)^2$$

Using Jensen's inequality gives

$$\mathbb{E}\left[\left\|X-X\hat{U}_\lambda\hat{U}_\lambda^t\right\|_F^2\right]\geq\mathbb{E}\left[\left\|X-X\hat{U}_\lambda\hat{U}_\lambda^t\right\|_F\right]^2\geq\mathbb{E}\left[\left\|M\left(T^*-\hat{U}_\lambda\hat{U}_\lambda^t\right)\right\|_F-\nu_{\max}\right]^2$$

which, combined with Proposition 3.4 gives

$$\mathbb{E}\left[\left\|M\left(T^* - \hat{U}_\lambda \hat{U}_\lambda^t\right)\right\|_F\right] \leq (1+\varepsilon)\|X - XU^0U^{0^t}\|_F$$

$$+ \sqrt{\frac{1}{\exp\left(\frac{\left(\frac{1}{\mu}\sqrt{c_U^2 - c_O^2} - \sqrt{nR}\right)^2}{2}\right) - 1}} + \sqrt{(nR - \frac{1}{2}(R^2+R))\log(\rho^{-1}) + \log(c^{-1})} + \nu_{\max}.$$

Given that

$$\|X - XU^0U^{0^t}\|_F \leq \|M - MU^0U^{0^t}\|_F + \nu_{\max}$$

this completes the proof of Theorem 1, since U^0 is any matrix satisfying the constraints.

4. A Langevin sampler

In this section, we present a Langevin sampler for our estimator $\hat{U}_\lambda, \hat{O}_\lambda$ *. Langevin-type samplers were first proposed by Grenander (1983), Grenander and Miller (1994), Roberts and Tweedie (1996), and have attracted a lot a attention lately in the statistical learning community (Dalalyan, 2017; Durmus and Moulines, 2017; Brosse et al., 2018).

4.1. Computing the gradient on the Stiefel manifold

We start with some preliminary material about gradient computation on the Stiefel manifold from Edelman et al. (1998). The Stiefel manifold can be interpreted as the set of equivalence classes

$$[O] = \left\{ \left[O \begin{bmatrix} I_R & 0 \\ 0 & O' \end{bmatrix}, \quad \text{with} \quad O' \in \mathbb{O}_{d-R} \right] \right\}.$$

As can easily be deduced from this quotient representation of the Steifel manifold, the tangent space to the Stiefel manifold at a point O is

$$T_O(\mathbb{O}_{d,R}) = \left\{ O \begin{bmatrix} A & -B^t \\ B & 0 \end{bmatrix}, \quad \text{with} \quad A \in \mathbb{R}^{R\times R} \text{ skew symmetric} \right\}.$$

The canonical metric at a point O is given by

$$g_c = \text{trace}\left(\Delta^t \left(I - \frac{1}{2}OO^t\right)\Delta\right).$$

For $\Delta \in T_O(\mathbb{O}_{d,R})$, the exponential map is given by $O(t) = Oe^{t\Delta}I_{d,R}$. The gradient at O of a function f defined on the Stiefel manifold $\mathbb{O}_{d,R}$ is given by †

$$\nabla f(O) = f_O - Of_O^t O, \tag{4.8}$$

where $f_O(i,i') = \frac{\partial f}{\partial O_{i,i'}}$ for any $i, i' = 1, \ldots, n$.

∗. Notation-wise, we will identify the Stiefel manifold with the set of matrices whose first R columns form an orthonormal family and the remaining $n - R$ columns are set to zero

†. This formula can be obtained using differentiation along the geodesic defined by the exponential map in the direction Δ, for all $\Delta \in T_O(\mathbb{O}_{d,R})$.

4.2. The alternating Langevin sampler

The alternating Langevin sampler is described as follows. It consists in alternating between perturbed gradient steps in the matrix variable O on the Stiefel manifold and perturbed gradient steps in the matrix variable U.

For clarity of the exposition, we give the formula for the gradient of $\|X - XUU^t\|_F^2$ as a function of U:

$$\nabla \left(\|X - X \cdot {}^t\cdot\|_F^2 \right)_U = \left((X - XUU^t)^t X + X^t (X - XUU^t) \right) U.$$

Following Brosse et al. (2018), we propose the following algorithm.

Algorithm 1 The Langevin sampler

Result: A sample $\hat{U}_\lambda$ of the quasi-posterior distribution
initialise $U^{(0)} = O^{(0)}$
for $\ell = 1$ **do**

$$O^{(\ell+1)} = \exp\Bigg(O^{(\ell)}, -h \Big(\operatorname{sign}(O^{(\ell)}) \odot \Big(U^{(\ell)} - |O^{(\ell)}| \Big) \Big)$$
$$- O^{(\ell)} \Big(\operatorname{sign}(O^{(\ell)}) \odot \Big(U^{(\ell)} - |O^{(\ell)}| \Big) \Big)^t O^{(\ell)} + \sqrt{2\,h}\, Z_O^{(\ell)} \Big) \Bigg)$$

$$U^{(\ell+1)} = U^{(\ell)} - h \Bigg(- \Big((X - XU^{(\ell)}U^{(\ell)^t})^t X + X^t (X - XU^{(\ell)}U^{(\ell)^t}) \Big) U^{(\ell)}$$
$$+ \frac{1}{\mu^2} \Big(U^{(\ell)} - |O^{(\ell+1)}| \Big) \Bigg) + \sqrt{2\,h}\, Z_U^{(\ell)}.$$

end

In this algorithm the exponential function $\exp(O, H)$ at O is given by Edelman et al. (1998, Eq. 2.45) using different notation.

5. Conclusion and future work

We propose a novel way to address the standard clustering problem by recasting it for the first time into constrained NMF framework using a Stiefel manifold prior. Up to our knowledge, this strategy is unprecedented in the statistical learning literature. For this approach, a bound in expectation is derived in a very general setting and then specialised to the case of an incoherent matrix of means. We derive a original componentwise Langevin sampler on the Stiefel manifold to compute our estimator.

This paper opens several lines of research. On a computational aspect, we intend to deploy our algorithm and evalute its performance on large real-life data sets. On a more theoretical side, our plans are to investigate the convergence properties of our Langevin sampler along with its convergence rate.

References

Pierre Alquier and Benjamin Guedj. An oracle inequality for quasi-Bayesian nonnegative matrix factorization. *Mathematical Methods of Statistics*, 26(1):55–67, 2017.

Ery Arias-Castro and Nicolas Verzelen. Community detection in dense random networks. *The Annals of Statistics*, 42(3):940–969, 2014.

Afonso S. Bandeira. Random Laplacian matrices and convex relaxations. *Foundations of Computational Mathematics*, 18(2):345–379, 2018.

Avrim Blum, John Hopcroft, and Ravindran Kannan. Foundations of data science. *Draft book*, 2016.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Nicolas Boumal, Vlad Voroninski, and Afonso Bandeira. The non-convex Burer-Monteiro approach works on smooth semidefinite programs. In *Advances in Neural Information Processing Systems*, pages 2757–2765, 2016.

Christos Boutsidis, Petros Drineas, and Michael W Mahoney. Unsupervised feature selection for the k-means clustering problem. In *Advances in Neural Information Processing Systems*, pages 153–161, 2009.

Nicolas Brosse, Alain Durmus, and Eric Moulines. The promises and pitfalls of stochastic gradient Langevin dynamics. In *Advances in Neural Information Processing Systems*, pages 8278–8288, 2018.

Samuel Burer and Renato D. C. Monteiro. Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming*, 103(3):427–444, 2005.

Samuel Burer and Renato DC Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

Olivier Catoni. *Statistical learning theory and stochastic optimization: Ecole d'Eté de Probabilités de Saint-Flour XXXI-2001*. Springer, 2004.

Olivier Catoni. *PAC-Bayesian Supervised Classification*. Lecture Notes-Monograph Series. IMS, 2007.

Stéphane Chrétien and Sébastien Darses. Sparse recovery with unknown variance: a Lasso-type approach. *IEEE Transactions on Information Theory*, 60(7):3970–3988, 2014.

Stéphane Chrétien, Clément Dombry, and Adrien Faivre. A semi-definite programming approach to low dimensional embedding for unsupervised clustering. *arXiv preprint arXiv:1606.09190*, 2016.

Michael B. Cohen, Sam Elder, Cameron Musco, Christopher Musco, and Madalina Persu. Dimensionality reduction for k-means clustering and low rank approximation. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 163–172. ACM, 2015.

Arnak S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3): 651–676, 2017.

Arnak S. Dalalyan and Alexandre B. Tsybakov. Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Machine Learning*, 72(1-2):39–61, 2008.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, pages 1–38, 1977.

Alain Durmus and Eric Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 2017.

Alan Edelman, Tomás A. Arias, and Steven T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

Christophe Giraud and Nicolas Verzelen. Partial recovery bounds for clustering with the relaxed k means. *arXiv preprint arXiv:1807.07547*, 2018.

Michel X. Goemans and David P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)*, 42 (6):1115–1145, 1995.

Ulf Grenander. Tutorial in pattern theory. *Report, Division of Applied Mathematics*, 1983.

Ulf Grenander and Michael I. Miller. Representations of knowledge in complex systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 549–603, 1994.

Benjamin Guedj. A primer on PAC-Bayesian learning. *arXiv preprint arXiv:1901.05353*, 2019.

Olivier Guédon and Roman Vershynin. Community detection in sparse networks via Grothendieck's inequality. *Probability Theory and Related Fields*, 165(3-4):1025–1049, 2016.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Unsupervised learning. In *The elements of statistical learning*, pages 485–585. Springer, 2009.

Aicke Hinrichs, Joscha Prochno, and Jan Vybiral. Entropy numbers of embeddings of Schatten classes. *Journal of Functional Analysis*, 273(10):3241–3261, 2017.

Jeffrey M. Lee and Frits H. Ruymgaart. Nonparametric curve estimation on Stiefel manifolds. *Journal of Nonparametric Statistics*, 6(1):57–68, 1996. doi: 10.1080/10485259608832663. URL https://doi.org/10.1080/10485259608832663.

Pascal Massart. *Concentration inequalities and model selection*. Springer, 2007.

D. McAllester. Some PAC-Bayesian theorems. In *COLT*, pages 230–234, 1998.

D. McAllester. PAC-Bayesian model averaging. In *COLT*, pages 164–171, 1999.

Geoffrey McLachlan and David Peel. *Finite mixture models*. John Wiley & Sons, 2004.

Andrea Montanari and Subhabrata Sen. Semidefinite programs on sparse random graphs and their application to community detection. *arXiv preprint arXiv:1504.05910*, 2015.

Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

Martin Royer. Adaptive clustering through semidefinite programming. In *Advances in Neural Information Processing Systems*, pages 1795–1803, 2017.

Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.

J. Shawe-Taylor and R. C. Williamson. A PAC analysis of a Bayesian classifier. In *COLT*, pages 2–9, 1997.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

Nicolas Verzelen and Ery Arias-Castro. Community detection in sparse random networks. *The Annals of Applied Probability*, 25(6):3465–3510, 2015.

Appendix A. Proof of Theorem 2

A.1. Control of $\mathbb{P}\left(\min_{U\in\mathbb{O}_{n,R}} \|E(I-UU^t)\|_F \leq \nu_{\min}\right)$

The covering number of the Stiefel manifold in the operator norm was computed in Hinrichs et al. (2017) and is given by

$$N(\mathbb{O}_{n,R}, \|\cdot\|, \varepsilon) \leq \left(\frac{c}{\varepsilon}\right)^{nR-R(R+1)/2}. \tag{A.9}$$

Let $\mathcal{N}_\varepsilon$ denote an ε-net in the operator norm for the Stiefel manifold with cardinality $N(\mathbb{O}_{n,R}, \|\cdot\|, \varepsilon)$. For any $U \in \mathbb{O}_{n,R}$, let $U^\sharp$ denote the closest matrix in $\mathcal{N}_\varepsilon$ to U. Then, we have

$$\begin{aligned}
\|E(I-UU^t)\|_F &= \|E(I-(U^\sharp+(U-U^\sharp))(U^\sharp+(U-U^\sharp))^t\|_F \\
&\geq \|E(I-U^\sharp U^{\sharp t})\|_F - \|EU^\sharp(U-U^\sharp)^t)\|_F - \|E(U-U^\sharp)U^{\sharp t})\|_F \\
&\geq \|E(I-U^\sharp U^{\sharp t})\|_F - \|(U-U^\sharp)\|\|EU^\sharp\|_F - \|E(U-U^\sharp)U^{\sharp t}\|_F
\end{aligned} \tag{A.10}$$

Moreover,

$$\begin{aligned}
\|E(U-U^\sharp)U^{\sharp t}\|_F^2 &= \operatorname{trace}(E(U-U^\sharp)U^{\sharp t}U^\sharp(U-U^\sharp)^tE^t) \\
&= \operatorname{trace}(E(U-U^\sharp)(U-U^\sharp)^tE^t) \\
&= \|E(U-U^\sharp)\|_F^2 \\
&\leq \|U-U^\sharp\|^2\,\|E\|_F^2
\end{aligned}$$

Combining this last inequality with (A.10), we get

$$\begin{aligned}\|E(I-UU^t)\|_F &\geq \|E(I-U^\sharp U^{\sharp t})\|_F - \varepsilon\left(\|EU^\sharp\|_F + \|E\|_F\right)\\ &\leq \|E(I-U^\sharp U^{\sharp t})\|_F - 2\,\varepsilon\|E\|_F\end{aligned}$$

Therefore, we obtain that

$$\begin{aligned}\mathbb{P}\left(\min_{U\in\mathbb{O}_{n,R}}\|E(I-UU^t)\|_F \leq \nu_{\min}\right) &\leq \mathbb{P}\left(\min_{U\in\mathcal{N}_\varepsilon}\|E(I-UU^t)\|_F \leq \nu_{\min}+2\varepsilon\|E\|_F\right)\\ &\leq \mathbb{P}\left(\min_{U\in\mathcal{N}_\varepsilon}\|E(I-U^\sharp U^{\sharp t})\|_F \leq \nu_{\min}+4\varepsilon\|E\|_F\right)\end{aligned}$$

where, for any U in $\mathcal{N}_\varepsilon$, $U^\sharp$ will denote the projection in operator norm of U onto $\mathbb{O}_{n,R}$. Moreover

$$\begin{aligned}\mathbb{P}\left(\min_{U\in\mathbb{O}_{n,R}}\|E(I-UU^t)\|_F \leq \nu_{\min}\right) &\leq \mathbb{P}\left(\min_{U\in\mathcal{N}_\varepsilon}\|E(I-U^\sharp U^{\sharp t})\|_F \leq \nu_{\min}+4\varepsilon\|E\|_F,\ \|E\|_F\leq\eta\right)\\ &\quad+\mathbb{P}\left(\min_{U\in\mathcal{N}_\varepsilon}\|E(I-U^\sharp U^{\sharp t})\|_F \leq v+4\varepsilon\|E\|_F,\ \|E\|_F>\eta\right)\\ &\leq \mathbb{P}\left(\min_{U\in\mathcal{N}_\varepsilon}\|E(I-U^\sharp U^{\sharp t})\|_F \leq \nu_{\min}+4\varepsilon\eta\right)\\ &\quad+\mathbb{P}\left(\|E\|_F>\eta\right)\end{aligned}$$

Since E is i.i.d. Gaussian with minimum one-dimensional variance $\sigma^2_{\min}$, for any $U^\sharp\in\mathbb{O}_{n,R}$, the lower tail of $\sigma_{\min}^{-1}\|E(I-U^\sharp U^{\sharp t})\|_F^2$ is dominated by the lower tail of a $\chi^2(n(n-R))$ distribution and therefore, as recalled in Chrétien and Darses (2014, Lemma B1)

$$\mathbb{P}\left(\|E(I-U^\sharp U^{\sharp t})\|_F \leq \sigma\,\sqrt{t\,n(n-R)}\right) \leq \frac{2}{\sqrt{\pi\,n(n-R)}}\,(t\,e/2)^{n(n-R)/4}. \tag{A.11}$$

Let us now tune t and η. On the one hand, by Boucheron et al. (2013), we have that

$$\mathbb{P}\left(\|E\|_F \geq \sigma\sqrt{dn+u}\right) \leq \exp(-nu^2/8).$$

Thus, we will choose $\eta=\sigma\sqrt{dn+u}$. On the other hand, we will take t such that

$$\sigma\sqrt{tn(d-R)} = \nu_{\min}+4\varepsilon\eta$$

i.e.

$$t = \frac{\left(\frac{\nu_{\min}}{\sigma}+4\varepsilon\sqrt{nd+u}\right)^2}{n(n-R)}$$

Therefore, using the union bound we get

$$\mathbb{P}\left(\min_{U\in\mathcal{N}_\varepsilon}\|E(I-U^\sharp U^{\sharp t})\|_F \leq \nu_{\min}\right) \leq \frac{2\,N(\mathbb{O}_{n,R},\|\cdot\|,\varepsilon)}{\sqrt{\pi\,n(n-R)}}\,(t\,e/2)^{n(n-R)/4}.$$

A.2. Control of $\mathbb{P}\left(\max_{U\in\mathbb{O}_{n,R}} \|E(I-UU^t)\|_F \geq \nu_{\max}\right)$

The same strategy applies, with slight modifications. We consider the same ε-net as in the previous subsection. We first easily get

$$\|E(I-UU^t)\|_F \leq \|E(I-U^\sharp U^{\sharp t})\|_F^2 + 2\,\varepsilon\|E\|_F$$

Therefore,

$$\begin{aligned}\mathbb{P}\left(\max_{U\in\mathbb{O}_{n,R}} \|E(I-UU^t)\|_F \geq \nu_{\max}\right) &\leq \mathbb{P}\left(\max_{U\in\mathcal{N}_\varepsilon} \|E(I-UU^t)\|_F \geq \nu_{\max} - 2\varepsilon\|E\|_F\right)\\ &\leq \mathbb{P}\left(\max_{U\in\mathcal{N}_\varepsilon} \|E(I-U^\sharp U^{\sharp t})\|_F \leq \nu_{\max} - 4\varepsilon\|E\|_F\right)\end{aligned}$$

where, for any U in $\mathcal{N}_\varepsilon$, $U^\sharp$ again denotes the projection in operator norm of U onto $\mathbb{O}_{n,R}$. We then get

$$\begin{aligned}\mathbb{P}\left(\max_{U\in\mathbb{O}_{n,R}} \|E(I-UU^t)\|_F \leq \nu_{\max}\right) \leq \mathbb{P}&\left(\max_{U\in\mathcal{N}_\varepsilon} \|E(I-U^\sharp U^{\sharp t})\|_F \leq \nu_{\max} - 4\varepsilon\eta\right)\\ &+ \mathbb{P}\left(\|E\|_F > \eta\right)\end{aligned}$$

Since E is i.i.d. Gaussian with maximum one-dimensional variance $\sigma_{\max}^2$, the upper tail of $\sigma_{\max}^{-1}\|E(I-U^\sharp U^{\sharp t})\|_F^2$ is dominated by that of a $\chi^2(n(n-R))$ distribution for any $U^\sharp$ in $\mathbb{O}_{n,R}$, and therefore, as recalled in Chrétien and Darses (2014, Lemma B1)

$$\mathbb{P}\left(\|E(I-U^\sharp U^{\sharp t})\|_F \geq \sigma_{\max}\left(\sqrt{n(n-R)} + \sqrt{2t}\right)\right) \leq \exp(-t). \tag{A.12}$$

We can now tune t and η. Recall that for all $u>0$, we have

$$\mathbb{P}\left(\|E\|_F \geq \sigma\sqrt{dn+u}\right) \leq \exp(-nu^2/8).$$

Thus, we will choose as before $\eta = \sigma_{\max}\sqrt{dn+u}$. On the other hand, we will take t such that

$$\sigma_{\max}\left(\sqrt{2t} + \sqrt{n(n-R)}\right) = \nu_{\max} - 4\varepsilon\eta$$

i.e.

$$t = \left(\frac{v}{\sigma_{\max}} - 4\varepsilon\sqrt{nd+u}\right)^2 - \sqrt{n(d-R)}.$$

Therefore, using the union bound we get

$$\mathbb{P}\left(\min_{U\in\mathcal{N}_\varepsilon} \|E(I-U^\sharp U^{\sharp t})\|_F \leq v\right) \leq N(\mathbb{O}_{n,R}, \|\cdot\|, \varepsilon)\ \exp(-t).$$

A.3. Control of $\|E\|$

We will also need to control $\|E\|$. Using Vershynin (2018, Section 4.4), we obtain

$$\|E\| \leq \sigma\left(\sqrt{n} + 2\sqrt{d}\right)$$

with probability at least $1-\exp(-d)$.

A.4. Control of $\|M\|$

Finally, we need to compute $\|M\|$ as a function of the coherence μ. Using the Gershgorin bound, we easily obtain

$$\|M\| = \sqrt{\|M^t M - I\| + 1} \leq \sqrt{\left(\max_{k=1}^{K} n_k\right) \mu(K-1) + 1} \tag{A.13}$$

A.5. End of the proof

Let us now study $\|M(T^* \hat{U}_\lambda \hat{U}_\lambda^t)\|_F^2$. Let

$$\Upsilon = [\mu_1, \ldots, \mu_K].$$

Let $\Upsilon^\dagger$ denote the pseudo inverse of Υ, i.e.

$$\Upsilon^\dagger = (\Upsilon^t \Upsilon)^{-1} \Upsilon^t.$$

In particular, let π denote a permutation which orders the data cluster wise, i.e. all data from cluster 1, followed by all data from cluster 2, …, all data from cluster K and let M_π denote the matrix whose columns are permuted by π. Then,

$$\Upsilon^\dagger M_\pi = \begin{bmatrix} 1_{n_1}^t & 0 & \cdots & \cdots & 0 \\ 0 & 1_{n_2}^t & \cdots & \cdots & 0 \\ \vdots & 0 & & & \vdots \\ \vdots & \vdots & & & \vdots \\ 0 & 0 & \cdots & 0 & 1_{n_K}^t \end{bmatrix}$$

Denote by T_π^* the matrix obtained from T^* after reordering its rows and columns using π, and $\hat{U}_{\lambda,\pi}$ denote the matrix obtained from $\hat{U}_\lambda$ after reordering its row using π.

$$\left\| \Upsilon^\dagger M_\pi \left(T_\pi^* - \hat{U}_{\lambda,\pi} U_{\lambda,\pi} \right) \right\|_F^2 = \sum_{k=1}^{K} \sum_{i_k \in I_k} \left(\sum_{i'_k \in I_k} T^*_{\pi, i'_k, i_k} - \hat{U}_{\lambda,\pi,i'_k} \hat{U}^t_{\lambda,\pi,i_k} \right)^2.$$

On the other hand,

$$\sigma_{\max}\left(\Upsilon^\dagger\right) \leq \sigma_{\min}\left(\Upsilon^t \Upsilon\right)^{-1} \sigma_{\max}\left(\Upsilon^t\right) \geq \frac{\sqrt{1+\mu(K-1)}}{1-\mu(K-1)}$$

and we get that

$$\left\| \Upsilon^\dagger M_\pi \left(T_\pi^* - \hat{U}_{\lambda,\pi} U_{\lambda,\pi} \right) \right\|_F^2 \leq \frac{\sqrt{1+\mu(K-1)}}{1-\mu(K-1)} \left\| M_\pi \left(T_\pi^* - \hat{U}_{\lambda,\pi} U_{\lambda,\pi} \right) \right\|_F^2$$

and the result follows.