SPARSE LOW RANK DECOMPOSITION FOR GRAPH DATA SETS

Clément Mantoux[⋆†] *Stanley Durrleman*[⋆‡] *Fabrizio De Vico Fallani*[⋆‡] *Stéphanie Allassonnière*[§†]

[⋆] Aramis Lab, Inria, Paris, France
[†] CMAP, École polytechnique, Palaiseau, France
[‡] Brain and Spine Institute, Inserm, CNRS, Sorbonne University, Paris, France
[§] Centre de Recherche des Cordeliers, Université Paris Descartes, Paris, France

ABSTRACT

Understanding the variability in a graph data set has been an important research topic in network analysis over the last years, and remains a challenging open problem. These data sets are often difficult to analyze, and one aims at highlighting interpretable common patterns and sample-specific variabilities. In this paper, we characterize graphs as the sum of a sparse low-rank common template and sparse low-rank deviations from it. This structure allows to account for real-world network properties: their adjacency matrices are sparse, and a low-rank reflects their organization into clusters as well as the role of core nodes. We propose a variational approach to estimate the template and the deviations based on combined sparse and low-rank regularizers. We demonstrate the performance of our decomposition model on both simulated and real data sets. For example, analyzing air traffic data, we show that sparsity and low rank lead to interpretable results on the structure of airline traffic.

***Index Terms*—** Network modeling, Sparsity, Low rank, Modularity, Non-parametric modeling

1. INTRODUCTION

Network science and graph theory are at the core of a wide range of applications. When dealing with a data set of networks defined on the same set of nodes, understanding the variability among samples is a crucial issue. In neuroscience, the connections between well-defined brain regions are studied on groups of subjects to better understand the human brain's anatomy and function [1]. Computational social science also requires to analyze the evolution of interactions in a fixed population [2]. In this setup, it is interesting to work directly on the networks' adjacency matrices. Yet, to our knowledge, comparing such matrices remains a difficult problem. Kernel methods can be employed to evaluate distances between networks [3], but many theoretically interesting graph kernels require solving NP-hard problems. Other metrics like the cut norm are used in graph theory, but cannot be computed exactly in polynomial time [4]. Consequently, statistical modeling for networks has so far focused on parametric models like Exponential Random Graph Models [5], which take stock on a small set of graph features to characterize a data set's variability.

Non-parametric modeling for multiple networks has only recently started to draw some attention in the literature. The authors of [6] represent networks having a common set of nodes as samples from an extended graphon model. However, this model does not account for classical properties of real-world networks like edge sparsity and low rank [7]. On the other hand, much work has been devoted to extracting denoised estimates from matrices with an underlying sparse or low-rank structure [8, 9, 10, 11], but these techniques have not been used yet in a statistical framework for graph data sets modeling. The issue we wish to address can be formulated as follows: how can the variability of graph data sets be accounted for in a non-parametric way, while providing interpretable patterns either shared by the whole dataset or specific to each graph? For example, given an airline network, can we identify the baseline traffic and how daily traffic fluctuations are structured ?

In this paper, we model the adjacency matrices of weighted networks sharing a common set of nodes as additive deviations from a template network representing the reference interactions. These template and deviations both have a sparse low-rank structure corrupted by a sparse noise. We estimate this decomposition using variational matrix recovery techniques. We show that taking advantage of the networks' common structure allows to efficiently estimate the uncorrupted data and recover the template patterns as well as the deviations. Finally, we show how this decomposition operates on real data sets.

2. RELATED WORK

The recovery of low-rank patterns has been widely addressed in the literature, in various application domains [12, 13, 14]. Many of them rely on singular value analysis and the use of the nuclear norm, and consider sparse noise [8, 10, 11]. More recently, efforts have been made to denoise matrices with both low-rank and sparse structure. Authors in [9] reconstruct such matrices by combining the nuclear and ℓ_1 penalties. The same idea is used by [15] to reconstruct covariance matrices, with additional theoretical guarantees on the convergence rate. All these methods focus on one single matrix at a time.

When it comes to averaging networks or adjacency matrices, little has been done in the literature so far. Kernel methods allow to compute the mean of several graphs, but the result is not always a graph itself. Stochastic block-models divide the nodes of a graph into a fixed, known number K of clusters with simple interactions [16]. This amounts to approximating the network by a block-wise constant adjacency matrix with rank less than K. The recent work of [6] allows to handle multiple adjacency matrices in a single mathematical object. The authors propose a multi-graphon model which represents networks as one-dimensional latent codes, but does not account for edge sparsity nor low rank. However, aggregating adjacency matrices seems an important step toward developing a coherent complete statistical framework for graph data sets analysis, and is at the core of the model we propose.

The research leading to these results has received funding from the European Research Council (ERC) under grant agreement No 678304, European Union'sHorizon 2020 research and innovation program under grant agreement No 666992 (Euro-POND) and No 826421 (TVB-Cloud). It was also funded by in part by the program "Investissements d'avenir" ANR-10-IAIHU-06 and the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute).

3. MODEL SETUP

We denote by $\|\cdot\|_F$, $\|\cdot\|_1$ and $\|\cdot\|_*$ respectively the Frobenius, the ℓ_1 and the nuclear norm (sum of singular values) over the set of adjacency matrices with a given number of nodes m.

We are interested in data sets of weighted, possibly non-symmetric adjacency matrices $A_1, ..., A_n \in \mathbb{R}_+^{m\times m}$. We model each A_i as deriving from a common template matrix T. In the example of the airline network in the introduction, T can be the typical daily traffic load and $A_i - T$ the daily traffic difference due to sample-specific circumstances and random fluctuations. The difference between A_i and T is decomposed into a deviation V_i with simple structure and a residual noise ε_i. This decomposition writes:

$$A_i = T + V_i + \varepsilon_i\,. \tag{1}$$

The template T is assumed to be sparse and have a low-rank structure. Although the network's perturbation V_i is not an adjacency matrix in itself, we also propose to consider it sparse and low-rank. A low-rank structure is the simplest way to account for global changes in the behavior of groups of individuals. For instance, if a matrix A_i differs from the template only in the connections of one given node with its neighbors, the resulting deviation has rank 2. Similarly, an identical change in a group of nodes translates into a very low rank term. Finally, since A_i is sparse the noise ε_i is also sparse.

We want to *simultaneously* remove the noise and separate the template from the deviations. To that end, we propose a variational formulation for the decomposition of Eq. (1). Our approach follows the idea of [9]. In order to denoise an adjacency matrix A, the authors in [9] solve a convex optimization problem to estimate an underlying sparse and low-rank ground truth:

$$S \in \operatorname*{argmin}_S \left\{\ell(S,A) + \lambda\|S\|_* + \rho\|S\|_1\right\}\,,$$

where ℓ is a convex loss. This formulation takes stock on both the sparsity-inducing property of the ℓ_1 penalization and the nuclear norm regularizer, often considered as the convex relaxation of the matrix rank [17]. [9] showed that using both penalties simultaneously improves the matrix recovery in terms of support and Root Mean Square Error (RMSE). Later, a better penalty was proposed by [18] to account for both the sparsity and low rank in a single regularizer, but the related estimation procedure turns out to be very computationally demanding, which led us to keep to the method of [9].

As a generalization of this previous framework, we propose to estimate the template T and the deviations $V = (V_i)_{i=1}^n$ by solving the following convex optimization problem:

$$\begin{aligned}(T,V) \in \operatorname*{argmin}_{T,V_1,\dots,V_n} &\left\{\lambda\|T\|_* + \rho\|T\|_1\right.\\ &\left.+\textstyle\sum_{i=1}^n \ell(A_i, T+V_i) + \mu\|V_i\|_* + \nu\|V_i\|_1\right\}\,.\end{aligned}$$

As the noise $\varepsilon = (\varepsilon_i)_{i=1}^n$ in our model is sparse, we use the ℓ_1 loss as a natural candidate for ℓ. This choice has proven useful to denoise adjacency matrices [11]. Notice that taking $\mu = \nu = +\infty$ amounts enforces $V = 0$, grouping ε_i and V_i in a single sparse term.

4. ALGORITHM

We propose an algorithm to solve the above optimization problem. The objective function is non-differentiable, hence gradient descent methods can not be employed. The standard way to deal with this problem is to use proximal methods [19]. Given a non-differentiable, yet simple convex function $g(x)$ and a parameter $\tau > 0$, the proximal operator of g is given by:

$$\mathrm{prox}_{\tau g}(x) = \operatorname*{argmin}_z\{\|x-z\|_F^2/2 + \tau g(z)\}\,.$$

The objective function is the sum of simple functions whose proximal operators are known. The Douglas-Rachford (DR) algorithm [20] can be adapted to our purpose as follows. We introduce duplicate variables T_*, T_1, V_*, V_1 (for the regularizers) and Y (for the ℓ term) and add corresponding equality constraints. The loss becomes:

$$\begin{aligned}\mathcal{L}(T_*,T_1,V_*,V_1,Y) = {}&\lambda\|T_*\|_* + \rho\|T_1\|_1 + \chi_{\{T_*=T_1\}}\\ &+\textstyle\sum_{i=1}^n \|Y_i - A_i\|_1 + \mu\|V_{*,i}\|_* + \nu\|V_{i,1}\|_1\\ &+\textstyle\sum_{i=1}^n \chi_{\{V_{*,i}=V_{1,i}\}} + \chi_{\{V_{*,i}+T_*=Y_i\}}\end{aligned}$$

where χ_E denotes the characteristic function of the set E, which takes value 0 over E and $+\infty$ elsewhere. Let $X = (T_*, T_1, V_*, V_1, Y)$, the full loss thus writes

$$\mathcal{L}(X) = f(X) + \chi_{(T_1,V_1,Y)=P(T_*,V_*)}\,,$$

with P a linear operator summarizing the constraints and

$$f(X) = f_1(T_*) + f_2(T_1) + f_3(V_*) + f_4(V_1) + f_5(Y)\,,$$

where each f_i has known proximal operator.

We have $\mathrm{prox}_{\tau f}(X) = (\mathrm{prox}_{\tau f_1}(T_*), ..., \mathrm{prox}_{\tau f_5}(Y))$ from the definition of f. Furthermore, each $\mathrm{prox}_{\tau f_i}$ derives from the following [19]:

• ℓ_1 *norm.* Let $\mathrm{ST}_\tau(X) = \mathrm{sgn}(X)\odot(X-\tau)_+$ the soft thresholding operator. Then the proximal operator for the ℓ_1 norm is given by $\mathrm{prox}_{\tau\|\cdot\|_1}(X) = \mathrm{ST}_\tau(X)$. Consequently, $\mathrm{prox}_{\tau\|\cdot-A\|_1}(X) = \mathrm{ST}_\tau(X-A)+A$.

• *Nuclear norm.* Given a matrix X, let $X = U\mathrm{Diag}(\lambda)V^T$ be the singular value decomposition of X. Then we have $\mathrm{prox}_{\tau\|\cdot\|_*}(X) = U\mathrm{Diag}(\mathrm{ST}_\tau(\lambda))V^T$.

The proximal operator of the indicator function $\chi_{(T_1,V_1,Y)=P(T_*,V_*)}$ is the linear projection onto the linear subspace of constraints $\mathcal{C} = \{X \mid (T_1,V_1,Y) = P(T_*,V_*)\}$. This projection can be computed explicitly:

Proposition ([21], proposition 42)**.** *Let P be a linear operator. The projection of (x,y) onto the set $\{(x,y) \mid y = Px\}$ is given by $(\tilde{x},\tilde{y})$, with $\tilde{x} = (I+P^TP)^{-1}(P^Ty+x)$, $\tilde{y} = P\tilde{x}$.*

For $x \in \{V_*, V_1, Y\}$, let $\bar{x} = \frac{2}{2n+6}\sum_i x_i$. In our optimization problem, this projection writes:

$$\begin{cases}\tilde{T}_* = \frac{3}{2n+6}(T_*+T_1) + \bar{V}_1 - \bar{V}_* - \bar{Y}\\ \tilde{V}_{*,i} = -\frac{1}{2n+6}(T_*+T_1) - \frac{1}{3}\bar{V}_1 + \frac{1}{6}\bar{V}_* + \frac{1}{6}\bar{Y} + \frac{1}{3}(V_{*,i}+V_{1,i}+Y_i)\,,\end{cases}$$

And $\tilde{T}_1 = \tilde{T}_*, \tilde{V}_{1,i} = \tilde{V}_{*,i}, \tilde{Y}_i = \tilde{V}_{*,i} + \tilde{T}_*$. The complete scheme is detailed in Algorithm 1. It only requires choosing parameters θ for the total step size and τ for the proximal step size. Unless specified, the algorithm is run for 200 iterations for every experiment in this paper, with $\theta = 0.9$ and $\tau = 0.1$. Stability of the results was such that there was no need to adapt to each dataset. The algorithm described in this section as well as reproducible code for all the numerical experiments in this paper are available online.[1]

5. EXPERIMENTS ON SIMULATED DATA

5.1. A visual example

In order to get a first grasp on the decomposition's action on graph data sets, we first study a synthetic case of a population with 100 nodes split into five communities sparsely interacting with each other. We construct a binary template T describing five fully connected communities with random sizes, and no interactions between communities. In $n = 10$ occurrences, all nodes between two random indices interact with nodes between two other random indices, forming an additional block of interactions V_i. Finally, we randomly

[1] https://github.com/cmantoux/sparse-low-rank-decomposition

Algorithm 1: DR splitting for the SPLR decomposition

Initialize X_0 with $T_* = T_1 = \frac{1}{n}\sum_i A_i$,
$V_{*,i} = V_{1,i} = A_i - T_*$, $Y_i = A_i$
repeat
$Z^0_{t+1} = 2X_t - \mathrm{Proj}_{\mathcal{C}}(X_t)$
$Z^1_{t+1} = 2Z^0_{t+1} - \mathrm{prox}_{\tau f}(Z^0_{t+1})$
$X_{t+1} = (1-\theta)X_t + \theta Z^1_{t+1}$
until *convergence*
return $\mathrm{Proj}_{\mathcal{C}}(X_t)$

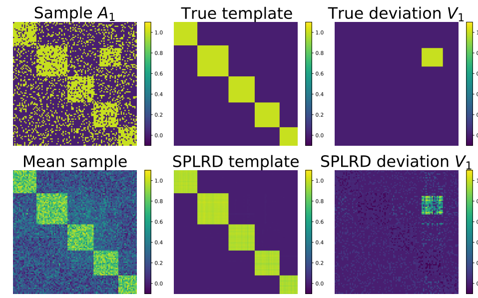

Fig. 1: Results for a simulated example with simple cluster structure.

flip 20% of the edges and get directed adjacency matrices A_i. The matrices (A_1, T, V_1) are shown in figure 1. We wish to identify the communities and the additional interactions V_i.

We apply our Sparse Low Rank Decomposition (SPLRD) algorithm on these 10 matrices (A_i), tuning the parameters $(\lambda, \rho, \mu, \nu)$ using the `forest_minimize` optimizer from the `scikit-optimize` library [22]. We select the tuple minimizing the RMSE for the deviations V_i, which naturally encourages a good estimation for T (as V_i is estimated from $A_i - T$). We use $\theta = 0.1$ and run the algorithm for 400 iterations because of the strong noise amplitude. The results are shown in Figure 1. The template is recovered accurately despite the small number of samples ($n = 10$). Most of the noise has been removed from the variation V_i, even though the estimation is less accurate than that of T. This is coherent since the template benefits from multiple samples, while the deviations are individual and the noise considered is strong. This result further highlights the interest of splitting the difference to the template into a low-rank term and noise: if both were regrouped in a single sparse term, the V_i's in this example could not be recovered. Up to a 0.1 threshold, 100% of the support is correctly recovered for T and 90% for V.

5.2. More complex simulated data

Experimental setup We now consider more general random samples. We generate non-symmetric sparse low-rank matrices T and V_i of the form UU^T, with U a sparse rectangular Gaussian matrix with full rank, and thus obtain a data set of matrices $A_i = T + V_i + \varepsilon_i$. The ranks of T and the V_i's, and the sparsity level in T, V and ε can be chosen freely. Here we use $n = 10$ samples, $m = 100$ nodes, $\mathrm{rk}(T) = \mathrm{rk}(V_i) = 10$, and 70% sparsity in T and the V_i's. The results we present stay stable when changing these settings. With the chosen values of rank and sparsity, non-zero coefficients in T and V have standard deviation close to 1. The noise ε_i also has sparsity 70%, and its non-zero coefficients follow a Gaussian distribution with standard deviation 1. In a second experiment, we simulate matrices that may represent actual weighted undirected networks. To that end we perform the same experiment, except now we impose that T and V have non-negative, symmetric coefficients. The noise ε is drawn from a symmetric sparse Gaussian distribution ε_0 thresholded to get non-negative final coefficients: $\varepsilon = \max(\varepsilon_0, -T - V)$.

We select $(\lambda, \rho, \mu, \nu)$ with the `forest_optimize` function of the `scikit-optimize` library with default arguments, and select the tuple with smallest RMSE for V. This metric is optimized on 5 random data sets and evaluated on 5 other random data sets drawn from the same distribution ($n = 10$ samples, $m = 100$ nodes).

Model evaluation For both data sets, we compute the relative RMSE for the estimation of T and V. In order to highlight the benefit of the joint estimation of T and V, we use the original method from [9] with the ℓ_1 loss to perform a sparse low-rank estimation on the empirical mean of the adjacency matrices $M = \frac{1}{n}\sum_i A_i$. We select the parameters on the training data set with `scikit-optimize` as for our model. This gives us an estimate T^M of T. We then apply the same method to $A_i - T^M$ with parameter selection on the training data set, and thus get an estimate V^M of V. Finally, we estimate the decomposition $A_i = T^S + \varepsilon_i^S$ in the sparse-only case $\mu = \nu = +\infty$. These results, as well as the naive decomposition $A_i = M + (A_i - M)$, are compared to our SPLRD solution (T^D, V^D).

In order to evaluate the network structure estimation capacity of the models, we also compare the relative RMSE of several standard graph features for the symmetric non-negative matrices. We compute the average weighted node degree d, the average shortest path length L and the clustering coefficient C. The average shortest path length L is computed using a decreasing function of the edge weights $\mathrm{cost}(w) = 1/(w + \alpha)$ ($\alpha = 0.1$), accounting for the cases where these weights represent a connection intensity rather than a cost. The weighted clustering coefficient C is defined following [23].

Results The numerical results are shown in Table 1. In both cases, our decomposition method greatly improves the estimation of the template compared to the denoised mean T^M. The recovery of T is achieved very accurately with a small number of samples ($n = 10$). Consequently, the estimator V^D also improves on V^M, which is only natural since the estimation of each V_i^M is based on a biased matrix $A_i - T^M$. It can be noticed that, while the estimation performance of the mean sample M and its denoised version T^M worsen when using non-negative coefficients, the performance of the SPLR decomposition stays good. This results points out the difference between the template T^D and the average sample M, which in this case is a biased template estimator. The results on graph features also show that the SPLR decomposition estimates the initial samples in a way that is consistent with a graph structure. This is not surprising for the average degree which depends linearly from the adjacency matrix, but the average shortest path length and the clustering coefficient estimations are also improved with respect to the noisy samples.

6. EXPERIMENTS ON REAL DATA

6.1. Airplane traffic network

We use a data set of networks from [24] listing US domestic airplane flights every hour for 10 days. We sum the total flight count per day to average day/night variability and get $n = 10$ directed weighted adjacency matrices. The obtained networks have $m = 299$ nodes corresponding to airports. The matrices have 95% null coefficients and rank at most 123, with very rapidly decaying singular values. We apply the SPLR decomposition with $(\lambda, \rho, \mu, \nu) = (10, 0.2, 5, 0.2)$, which allows to observe an interesting result while staying close to

Table 1: Mean and standard deviation of the relative RMSE. For matrices with unconstrained signs we show the error for T and V, and for symmetric non-negative coefficients (Sym+) the error for T, V and the graph features (GF) presented in the text.

		(T^D, V_i^D)	(T^M, V_i^M)	(T^S, ε_i^S)	Naïve
Any sign	T	$\mathbf{.04 \pm .01}$	$.42 \pm .08$	$.07 \pm .01$	$.55 \pm .10$
	V	$\mathbf{.33 \pm .02}$	$.49 \pm .02$	$.85 \pm .02$	$.91 \pm .05$
Sym+	T	$\mathbf{.07 \pm .02}$	$.55 \pm .22$	$.08 \pm .03$	$.64 \pm .27$
	V	$\mathbf{.30 \pm .01}$	$.55 \pm .02$	$.66 \pm .03$	$.77 \pm .03$
GF	d	$\mathbf{.01 \pm .01}$	$.09 \pm .04$	$.27 \pm .09$	$.27 \pm .09$
	L	$\mathbf{.15 \pm .07}$	$.29 \pm .08$	$.49 \pm .07$	$.49 \pm .07$
	C	$\mathbf{.03 \pm .01}$	$.35 \pm .03$	$.09 \pm .02$	$.09 \pm .02$

the data (10% relative RMSE between the $A_i's$ and the $T + V_i$'s accounting for the noise removed). There is no definite rule to select these parameters, as often with variational methods. We choose the parameters which allows to observe low-rank patterns while keeping the decomposition $T + V_i$ close to A_i.

The estimated template has rank 112 and sparsity 96%. The deviations have rank below 85 and average sparsity 96%. In comparison, the samples have rank at least 110, with the mean sample having rank 123. The template and the first three deviations are shown in Figure 2. The cities are ordered according to an algorithmically detected core/periphery structure [25], grouping the core nodes and the periphery nodes separately.

The recovered template T contains the regular traffic; it is high between the core airline hubs, weak between the core and the periphery and almost null between nodes from the periphery. The deviations V_i's account for the fluctuations in the flights planning, as well as events affecting airports. In Figure 2, V_1 is mainly constituted of fluctuations in the core and the impact of a few core cities onto the periphery nodes. This impact directly translates into a low-rank perturbation: the strongest pattern in V_1 is a reduced activity for New York's western airport (10th core node) maybe linked to a snow storm in New Jersey that day, resulting in a rank 2 perturbation. Two symmetric bright spots indicate an unusually high traffic between Denver and New York's eastern airport, due to the Denver football team taking part to the 2013 SuperBowl finals in New York that day. Similar observations can be made on the structure of V_2 and V_3.

6.2. Functional brain networks

We consider a data set of human brains analyzed with functional Magnetic Resonance Imaging in [26]. The nodes represent brain regions and the connections the temporal correlations between their activity signals. The study considers $n = 49$ subjects with $m = 312$ brain regions. Each subject is recorded during a resting period and while performing a listening task. For more details on the data processing, we refer the reader to [26]. The correlation matrices in the data set are not sparse but have very low rank, with only one or two leading singular values. Our decomposition model can be employed using $\rho = \nu = 0$ and turn into a low rank template + low rank deviation estimation method. Setting $\lambda = 50, \mu = 5$, we run the algorithm first separately on the 49 resting state and the 49 active listening matrices, then together on the 98 matrices. The three decompositions approximate the samples with relative RMSE at most 3.9%. The results are shown in Figure 3.

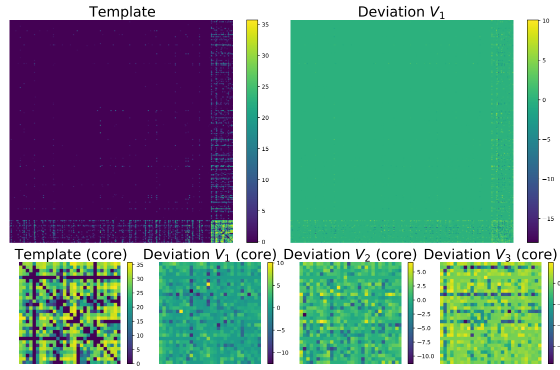

Fig. 2: Decomposition result for US flights data set. This figure shows the full template T and the full deviation V_1 (top), and a zoom on core cities for T, V_1, V_2 and V_3 (bottom). In the interests of readability, color scales differ across figures.

The templates T^{rest} and T^{task} have relatively low rank (resp. 70 and 166), whereas the template T^{full} has rank 284 out of 312, showing that our model is more relevant for each separate task than for the global data set. Over all three models, the deviation ranks range between 54 and 97. Moreover, the ℓ_2 distance between T^{task} and T^{full} is 37% larger than that between T^{rest} and T^{full}. It indicates that the best template for the whole data set is closer to the resting state brain. This observation can be linked with the results on simulated data with non-negative deviations V_i: the decomposition identifies a template that best fits the $T + V_i$ model rather than the average sample. The templates provide a sharper perspective than the means for which region's activity changes when performing a task.

7. CONCLUSION

We have presented a new method to decompose data sets of matrices into a sparse low-rank template, sparse low-rank deviations and noise. This model relies on observable properties of real-world adjacency matrices. We have proposed a procedure to fit this model by optimizing a convex loss function. This algorithm has proven very efficient to estimate the underlying template from a small number of samples, recovering the deviations with acceptable accuracy. We proved that this model is relevant to handle graph data sets and showed that the algorithm is able to recover graph features from noisy simulated data. This decomposition is suited to real networks like airplane flights and gives a meaningful interpretation for their variability.

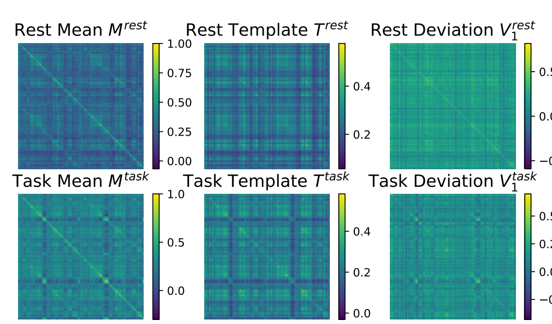

Fig. 3: Mean sample matrices M^{rest}, M^{task}, templates, and the first subject's deviations at rest (top) and performing a task (bottom).

8. REFERENCES

[1] Catalina Obando, Charlotte Rosso, Joshua Siegel, Maurizio Corbetta, and Fabrizio De Vico Fallani, "Temporal connection signatures of human brain networks after stroke," *arXiv:1907.10009 [q-bio, stat]*, July 2019, arXiv: 1907.10009.

[2] Steve Hanneke, Wenjie Fu, and Eric P. Xing, "Discrete temporal models of social networks," *Electronic Journal of Statistics*, vol. 4, pp. 585–605, 2010, Publisher: The Institute of Mathematical Statistics and the Bernoulli Society.

[3] Swarnendu Ghosh, Nibaran Das, Teresa Gonçalves, Paulo Quaresma, and Mahantapas Kundu, "The journey of graph kernels through two decades," *Computer Science Review*, vol. 27, pp. 88–111, Feb. 2018.

[4] Noga Alon and Assaf Naor, "Approximating the cut-norm via Grothendieck's inequality," in *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, Chicago, IL, USA, June 2004, STOC '04, pp. 72–80, Association for Computing Machinery.

[5] Jenine K Harris, *An introduction to exponential random graph modeling*, 2014, OCLC: 870698788.

[6] Swati Chandna and Pierre-Andre Maugis, "Nonparametric regression for multiple heterogeneous networks," *arXiv:2001.04938 [stat]*, Jan. 2020, arXiv: 2001.04938.

[7] Ivan Gutman and Bojana Borovićanin, "Nullity of Graphs: An Updated Survey," *Zbornik Radova*, , no. 22, pp. 137–154, 2011, Publisher: Matematički institut SANU.

[8] Samet Oymak and Babak Hassibi, "Finding Dense Clusters via "Low Rank + Sparse" Decomposition," *arXiv:1104.5186 [cs, math, stat]*, Apr. 2011, arXiv: 1104.5186 version: 1.

[9] Emile Richard, Pierre-Andre Savalle, and Nicolas Vayatis, "Estimation of Simultaneously Sparse and Low Rank Matrices," in *ICML 2012*, June 2012, arXiv: 1206.6474.

[10] Mattia Zorzi and Alessandro Chiuso, "Sparse plus low rank network identification: A nonparametric approach," *Automatica*, vol. 76, pp. 355–366, Feb. 2017.

[11] Taiju Kanada, Masaki Onuki, and Yuichi Tanaka, "Low-rank Sparse Decomposition of Graph Adjacency Matrices for Extracting Clean Clusters," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Nov. 2018, pp. 1153–1159, ISSN: 2640-0103.

[12] Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright, "Robust principal component analysis?," *Journal of the ACM*, vol. 58, no. 3, pp. 11:1–11:37, June 2011.

[13] V. Koltchinskii, K. Lounici, and A. B. Tsybakov, "Estimation of low-rank covariance function," *Stochastic Processes and their Applications*, vol. 126, no. 12, pp. 3952–3967, Dec. 2016.

[14] Emmanuel J. Candes and Yaniv Plan, "Matrix Completion With Noise," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 925–936, June 2010.

[15] Sheng-Long Zhou, Nai-Hua Xiu, Zi-Yan Luo, and Ling-Chen Kong, "Sparse and Low-Rank Covariance Matrix Estimation," *Journal of the Operations Research Society of China*, vol. 3, no. 2, pp. 231–250, June 2015.

[16] Tiago P. Peixoto, "Bayesian stochastic blockmodeling," in *Advances in Network Clustering and Blockmodeling*, Patrick Doreian, Vladimir Batagelj, and Anuska Ferligoj, Eds., Wiley Series in Computational and Quantitative Social Science, pp. 289–332. Wiley, Feb. 2020, arXiv: 1705.10225.

[17] Yuchao Dai and Hongdong Li, "Rank Minimization or Nuclear-Norm Minimization: Are We Solving the Right Problem?," in *2014 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, Nov. 2014, pp. 1–8.

[18] Emile Richard, Francis Bach, and Jean-Philippe Vert, "Intersecting singularities for multi-structured estimation," in *ICML 2013 - 30th International Conference on Machine Learning*, Atlanta, United States, June 2013, pp. –.

[19] Neal Parikh and Stephen Boyd, *Proximal Algorithms*, Now Publishers Inc., Jan. 2014.

[20] D. Gabay, "Chapter IX Applications of the Method of Multipliers to Variational Inequalities," in *Studies in Mathematics and Its Applications*, vol. 15, pp. 299–331. Elsevier, 1983.

[21] Gabriel Peyré, "Chapter 13 Convex Optimization," in *Mathematical foundations of data sciences*. 2018.

[22] Tim Head, MechCoder, Gilles Louppe, Iaroslav Shcherbatyi, fcharras, Zé Vinícius, cmmalone, Christopher Schröder, nel215, Nuno Campos, Todd Young, Stefano Cereda, Thomas Fan, rene rex, Kejia (KJ) Shi, Justus Schwabedal, carlosdanielcsantos, Hvass-Labs, Mikhail Pak, SoManyUsernamesTaken, Fred Callaway, Loïc Estève, Lilian Besson, Mehdi Cherti, Karlson Pfannschmidt, Fabian Linzberger, Christophe Cauet, Anna Gut, Andreas Mueller, and Alexander Fabisch, "scikit-optimize/scikit-optimize: v0.5.2," Mar. 2018.

[23] Tore Opsahl and Pietro Panzarasa, "Clustering in weighted networks," *Social Networks*, vol. 31, no. 2, pp. 155–163, May 2009.

[24] Matthew J. Williams and Mirco Musolesi, "Spatio-temporal networks: reachability, centrality and robustness," *Royal Society Open Science*, vol. 3, no. 6, pp. 160196, June 2016.

[25] Sadamori Kojaku and Naoki Masuda, "Core-periphery structure requires something else in the network," *New Journal of Physics*, vol. 20, no. 4, pp. 043012, Apr. 2018, Publisher: IOP Publishing.

[26] Leonhard Waschke, Mohsen Alavash, Jonas Obleser, and Julia Erb, "AUDADAPT," Nov. 2018, Publisher: OSF.